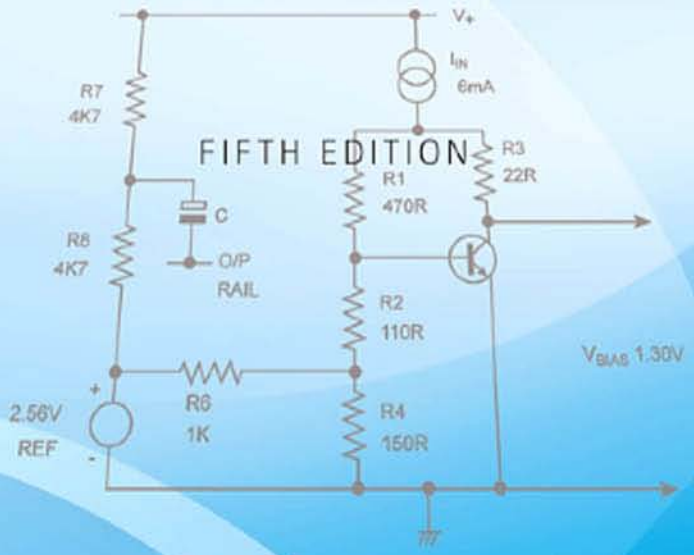
DOUGLAS SELF

AUDIO POWER AMPLIFIER DESIGN HANDBOOK

FIFTH EDITION



Focal Press

Audio Power Amplifier Design Handbook

This book is dedicated to Julie,
without whom it would not have happened.

Audio Power Amplifier Design Handbook

Fifth Edition

Douglas Self

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Focal Press is an imprint of Elsevier

Focal Press is an imprint of Elsevier
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
Linacre House, Jordan Hill, Oxford OX2 8DP, UK

First published 2009

British Library Cataloguing-in-Publication Data
Self, Douglas.
Audio power amplifier design handbook. – 5th ed.
1. Audio amplifiers—Design. 2. Power amplifiers—Design.
I. Title
621.3'81535—dc22

Library of Congress Control Number: 2009920721

ISBN: 978-0-240-52162-6

For information on all Focal Press publications visit our website at www.focalpress.com

Printed and bound in the United States of America

09 10 11 12 13 12 11 10 9 8 7 6 5 4 3 2 1

Contents

Acknowledgments

Heartfelt thanks to Gareth Connor of The Signal Transfer Company for practical help, never-failing encouragement, and for providing the facilities with which some of the experiments in this book were done.

I wish to thank Averil Donohoe for her help with some of the harder sums.

Preface to Fifth Edition

You will have noted from the increased weight of this book that it has been significantly expanded. The text has increased in size by more than 50%, and there are a hundred new illustrations.

There is a completely new chapter on the Class-XD system that I recently introduced at Cambridge Audio; an amplifier utilizing this system won an Innovation award at Chicago CES, January 2008.

There is also a big new chapter on balanced line inputs and balanced interconnections in general. These are becoming more and more common in the hi-fi field and have always been of prime importance in professional amplifier systems. This is a vital topic as without good interconnection technology the signal quality is irrevocably compromised before it gets anywhere near the actual power amplifier stage. This chapter also includes a lot of new material on ultra-low-noise design.

There is also a wholly new chapter on amplifier subsystems such as signal activation, 12 V trigger, level indication, and more. Amplifier input stages and voltage-amplifier stages now have separate chapters of their own.

I have added lots of new material on four-stage amplifier architectures, current-mirrors, power transistors with internal sensing diodes, amplifier bridging, distortion mechanisms, input stage common-mode distortion, double input stages, amplifier stability, output stages with gain, transformers and their hum fields, inrush current suppression, DC servo design, thermal protection, the subtleties of cooling fan control, line input stages, low-noise design, high- and low-pass filtering, testing and safety, infrared control, and much more. There is significantly more material on professional power amplifiers as used in sound reinforcement and PA applications.

I am aware there is still very little material on power MOSFETs in this book, as I still hold to the view that they are inevitably more nonlinear and harder to work with than bipolar transistors. I know that some people – including some I have much respect for – do not agree, but I find the evidence in both theory and practice to be convincing.

There has been some rearrangement to get a more logical layout of the subject matter. Your favorite topic has not been removed, but it might well have been moved.

As you will have gathered, I am still fascinated by the apparently simple but actually fiendishly complex business of making small signals bigger and applying them to a loudspeaker. An amplifier performs one of the simplest possible mathematical operations on a signal – multiplication by a constant. It is fascinating to see how much more complicated things get after that.

Part of the lure of electronics as a pursuit is the speed with which ideas can be turned into physical reality. In audio amplifier design, you very often need just a handful of components, a piece of prototype board, and a few minutes to see if the latest notion really is correct. If you come up with a brilliant new way of designing large concrete dams then it is going to take more than an afternoon to prove that it works.

You will also see, in Chapter 1, that in the last few years I have found no reason to alter my views on the pernicious irrationality of subjectivism. In that period I have repeatedly been involved in double-blind listening tests using experienced subjects and proper statistical analysis, which confirmed every time that if you cannot measure it you cannot hear it. Nevertheless the controversy rumbles on, although in a more logical world it would have been regarded as settled in the 1970s. I get a steady flow of emails supporting my position on this issue, but I fear I am still regarded in some quarters as the Gregor Eisenhorn of amplifier design.

There is in this book a certain emphasis on commercial manufacture, which I hope does not offend those purely interested in amateur construction or intellectual enquiry. In a commercial environment, if you want to sell something (for more than a very short time) it has to work – and keep working. This is still a valuable discipline if you are making a one-off design to test some new ideas; if the design is not reliable then it must be unsound in some way that may have more impact on what is going on than you think.

In a changing world, one of the many things that has changed is the nature of discussion on audio technologies. For many years *Wireless World* – later *Electronics World* – was a major forum for this, and I contributed many articles to it over 30 years; it has, however, now changed its emphasis. *Elektor* since its beginning has hosted serious audio articles and still does. The biggest change is of course the arrival of the Internet, which allows debate to proceed at a lightning pace compared with the old method of writing a letter and waiting for a month or two to see it published. Currently the only bulletin-board I frequent is DIYaudio.com; I personally think it is one of the best.

In producing this edition of the book it struck me frequently and forcibly how much has had to be omitted for reasons of space, despite the generous increase in its size. Audio power amplifier design, even if confined to solid-state amplifiers, and even if further confined to those with bipolar output stages, is already too big a field for one person to know everything. I certainly don't think I do.

The journey continues.

Douglas Self

Abbreviations

I have kept the number of abbreviations used to a minimum. However, those few are used extensively, so a list is given in case they are not all blindingly obvious:

BJT	Bipolar junction transistor
CFP	Complementary feedback pair
C/L	Closed loop
CM	Common mode
CMOS	Complementary metal oxide semiconductor
CMRR	Common-mode rejection ratio
CTF	Current timing factor
DF	Damping factor
DSP	Digital signal processing
EF	Emitter-follower
EFA	Emitter-follower added
EIN	Equivalent input noise
ESR	Equivalent series resistance
FEA	Finite element analysis
FET	Field-effect transistor
HF	Amplifier behavior above the dominant pole frequency, where the open-loop gain is usually falling at 6 dB/octave
IAE	Integrated absolute error
IC	Integrated circuit
IGBT	Insulated-gate bipolar transistor
I/P	Input
ISE	Integrated square error
LED	Light-emitting diode
LF	Relating to amplifier action below the dominant pole, where the open-loop gain is assumed to be essentially flat with frequency
LSN	Large-signal nonlinearity
MOSFET	Metal oxide semiconductor field-effect transistor
NF	Noise figure

NFB	Negative feedback
O/L	Open loop
O/P	Output
*P*1	The first O/L response pole, and its frequency in Hz (i.e. the −3 dB point of a 6 dB/octave roll-off)
*P*2	The second response pole, at a higher frequency
PA	Public address
PCB	Printed-circuit board
PDF	Probability density function
PPD	Power partition diagram
PSRR	Power-supply rejection ratio
PSU	Power-supply unit
PWM	Pulse width modulation
RF	Radio frequency
SID	Slew-induced distortion
SOA, SOAR	Safe operating area
SPL	Sound pressure level
Tempco	Temperature coefficient
THD	Total harmonic distortion
TID	Transient intermodulation distortion
TIM	Transient intermodulation
VAS	Voltage-amplifier stage
VCIS	Voltage-controlled current source
VCVS	Voltage-controlled voltage source
VI	Voltage/current

CHAPTER 1

Introduction and General Survey

The Economic Importance of Power Amplifiers

Audio power amplifiers are of considerable economic importance. They are built in their hundreds of thousands every year, and have a history extending back to the 1920s. It is therefore surprising there have been so few books dealing in any depth with solid-state power amplifier design.

The first aim of this text is to fill that need, by providing a detailed guide to the many design decisions that must be taken when a power amplifier is designed.

The second aim is to disseminate the results of the original work done on amplifier design in the last few years. The unexpected result of these investigations was to show that power amplifiers of extraordinarily low distortion could be designed as a matter of routine, without any unwelcome side-effects, so long as a relatively simple design methodology was followed. This methodology will be explained in detail.

Assumptions

To keep its length reasonable, a book such as this must assume a basic knowledge of audio electronics. I do not propose to plough through the definitions of frequency response, total harmonic distortion (THD) and signal-to-noise ratio; these can be found anywhere. Commonplace facts have been ruthlessly omitted where their absence makes room for something new or unusual, so this is not the place to start learning electronics from scratch. Mathematics has been confined to a few simple equations determining vital parameters such as open-loop gain; anything more complex is best left to a circuit simulator you trust. Your assumptions, and hence the output, may be wrong, but at least the calculations in between will be correct . . .

The principles of negative feedback as applied to power amplifiers are explained in detail, as there is still widespread confusion as to exactly how it works.

Origins and Aims

The core of this book is based on a series of eight articles originally published in *Electronics World* as 'Distortion in Power Amplifiers'. This series was primarily concerned with distortion as the most variable feature of power amplifier performance. You may have two units placed side by side,

one giving 2% THD and the other 0.0005% at full power, and both claiming to provide the ultimate audio experience. The ratio between the two figures is a staggering 4000:1, and this is clearly a remarkable state of affairs. One might be forgiven for concluding that distortion was not a very important parameter. What is even more surprising to those who have not followed the evolution of audio over the last two decades is that the more distortive amplifier will almost certainly be the more expensive. I shall deal in detail with the reasons for this astonishing range of variation.

The original series was inspired by the desire to invent a new output stage that would be as linear as Class-A, without the daunting heat problems. In the course of this work it emerged that output stage distortion was completely obscured by nonlinearities in the small-signal stages, and it was clear that these distortions would need to be eliminated before any progress could be made. The small-signal stages were therefore studied in isolation, using *model* amplifiers with low-power and very linear Class-A output stages, until the various overlapping distortion mechanisms had been separated out. It has to be said this was not an easy process. In each case there proved to be a simple, and sometimes well-known, cure and perhaps the most novel part of my approach is that all these mechanisms are dealt with, rather than one or two, and the final result is an amplifier with unusually low distortion, using only modest and safe amounts of global negative feedback.

Much of this book concentrates on the distortion performance of amplifiers. One reason is that this varies more than any other parameter – by up to a factor of 1000. Amplifier distortion was until recently an enigmatic field – it was clear that there were several overlapping distortion mechanisms in the typical amplifier, but it is the work reported here that shows how to disentangle them, so they may be separately studied and then, with the knowledge thus gained, minimized.

I assume here that distortion is a bad thing, and should be minimized; I make no apology for putting it as plainly as that. Alternative philosophies hold that as some forms of nonlinearity are considered harmless or even euphonic, they should be encouraged, or at any rate not positively discouraged. I state plainly that I have no sympathy with the latter view; to my mind the goal is to make the audio path as transparent as possible. If some sort of distortion is considered desirable, then surely the logical way to introduce it is by an outboard processor, working at line level. This is not only more cost-effective than generating distortion with directly heated triodes, but has the important attribute that *it can be switched off*. Those who have brought into being our current signal-delivery chain, i.e. mixing consoles, multitrack recorders, CDs, etc., have done us proud in the matter of low distortion, and to willfully throw away this achievement at the very last stage strikes me as curious at best.

In this book I hope to provide information that is useful to all those interested in power amplifiers. Britain has a long tradition of small and very small audio companies, whose technical and production resources may not differ very greatly from those available to the committed amateur. I hope this volume will be of service to both.

I have endeavored to address both the quest for technical perfection – which is certainly not over, as far as I am concerned – and also the commercial necessity of achieving good specifications at minimum cost.

The field of audio is full of statements that appear plausible but in fact have never been tested and often turn out to be quite untrue. For this reason, I have confined myself as closely as possible to facts that I have verified myself. This volume may therefore appear somewhat idiosyncratic in places. For example, field-effect transistor (FET) output stages receive much less coverage than bipolar ones because the conclusion appears to be inescapable that FETs are both more expensive and less linear; I have therefore not pursued the FET route very far. Similarly, most of my practical design experience has been on amplifiers of less than 300 W power output, and so heavy-duty designs for large-scale public address (PA) work are also under-represented. I think this is preferable to setting down untested speculation.

The Study of Amplifier Design

Although solid-state amplifiers have been around for some 40 years, it would be a great mistake to assume that everything possible is known about them. In the course of my investigations, I discovered several matters which, not appearing in the technical literature, appear to be novel, at least in their combined application:

- The need to precisely balance the input pair to prevent second-harmonic generation.
- The demonstration of how a beta-enhancement transistor increases the linearity and reduces the collector impedance of the voltage-amplifier stage (VAS).
- An explanation of why BJT output stages always distort more into 4 Ω than 8 Ω.
- In a conventional BJT output stage, quiescent current as such is of little importance. What is crucial is the voltage between the transistor emitters.
- Power FETs, though for many years touted as superior in linearity, are actually far less linear than bipolar output devices.
- In most amplifiers, the major source of distortion is not inherent in the amplifying stages, but results from avoidable problems such as induction of supply-rail currents and poor power-supply rejection.
- Any number of oscillograms of square waves with ringing have been published that claim to be the transient response of an amplifier into a capacitive load. In actual fact this ringing is due to the output inductor resonating with the load, and tells you precisely nothing about amplifier stability.

The above list is by no means complete.

As in any developing field, this book cannot claim to be the last word on the subject; rather it hopes to be a snapshot of the state of understanding at this time. Similarly, I certainly do not claim that this book is fully comprehensive; a work that covered every possible aspect of every conceivable power amplifier would run to thousands of pages. On many occasions I have found myself about to

write: '*It would take a whole book to deal properly with. . . .*' Within a limited compass I have tried to be innovative as well as comprehensive, but in many cases the best I can do is to give a good selection of references that will enable the interested to pursue matters further. The appearance of a reference means that I consider it worth reading, and not that I think it to be correct in every respect.

Sometimes it is said that discrete power amplifier design is rather unenterprising, given the enormous outpouring of ingenuity in the design of analog integrated circuits. Advances in op-amp design would appear to be particularly relevant. I have therefore spent some considerable time studying this massive body of material and I have had to regretfully conclude that it is actually a very sparse source of inspiration for new audio power amplifier techniques; there are several reasons for this, and it may spare the time of others if I quickly enumerate them here:

- A large part of the existing data refers only to small-signal MOSFETs, such as those used in (CMOS) op-amps, and is dominated by the ways in which they differ from BJTs, for example in their low transconductance. CMOS devices can have their characteristics customized to a certain extent by manipulating the width/length ratio of the channel.
- In general, only the earlier material refers to bipolar junction transistor (BJT) circuitry, and then it is often mainly concerned with the difficulties of making complementary circuitry when the only PNP transistors available are the slow lateral kind with limited beta and poor frequency response.
- Many of the CMOS op-amps studied are transconductance amplifiers, i.e. voltage difference in, current out. Compensation is usually based on putting a specified load capacitance across the high-impedance output. This does not appear to be a promising approach to making audio power amplifiers.
- Much of the op-amp material is concerned with the common-mode performance of the input stage. This is pretty much irrelevant to power amplifier design.
- Many circuit techniques rely heavily on the matching of device characteristics possible in IC fabrication, and there is also an emphasis on minimizing chip area to reduce cost.
- A good many IC techniques are only necessary because it is (or was) difficult to make precise and linear IC resistors. Circuit design is also influenced by the need to keep compensation capacitors as small as possible, as they take up a disproportionately large amount of chip area for their function.

The material here is aimed at all audio power amplifiers that are still primarily built from discrete components, which can include anything from 10 W mid-fi systems to the most rarefied reaches of what is sometimes called the 'high end', though the 'expensive end' might be a more accurate term. There are of course a large number of IC and hybrid amplifiers, but since their design details are fixed and inaccessible they are not dealt with here. Their use is (or at any rate should be) simply a matter of following the relevant application note. The quality and reliability of IC power amps has improved noticeably over the last decade, but low distortion and high power still remain the province of discrete circuitry, and this situation seems likely to persist for the foreseeable future.

Power amplifier design has often been treated as something of a black art, with the implication that the design process is extremely complex and its outcome not very predictable. I hope to show that this need no longer be the case, and that power amplifiers are now designable – in other words it is possible to predict reasonably accurately the practical performance of a purely theoretical design. I have done a considerable amount of research work on amplifier design, much of which appears to have been done for the first time, and it is now possible for me to put forward a design methodology that allows an amplifier to be designed for a specific negative-feedback factor at a given frequency, and to a large extent allows the distortion performance to be predicted. I shall show that this methodology allows amplifiers of extremely low distortion (sub-0.001% at 1 kHz) to be designed and built as a matter of routine, using only modest amounts of global negative feedback.

Misinformation in Audio

Few fields of technical endeavor are more plagued with errors, misstatements and confusion than audio. In the last 20 years, the rise of controversial and non-rational audio hypotheses, gathered under the title *Subjectivism* has deepened these difficulties. It is commonplace for hi-fi reviewers to claim that they have perceived subtle audio differences that cannot be related to electrical performance measurements. These claims include the alleged production of a 'three-dimensional sound stage and protests that the rhythm of the music has been altered'; these statements are typically produced in isolation, with no attempt made to correlate them to objective test results. The latter in particular appears to be a quite impossible claim.

This volume does not address the implementation of subjectivist notions, but confines itself to the measurable, the rational, and the repeatable. This is not as restrictive as it may appear; there is nothing to prevent you using the methodology presented here to design an amplifier that is technically excellent, and then gilding the lily by using whatever brands of expensive resistor or capacitor are currently fashionable, and doing the internal wiring with cable that costs more per meter than the rest of the unit put together. Such nods to subjectivist convention are unlikely to damage the real performance; this is, however, not the case with some of the more damaging hypotheses, such as the claim that negative feedback is inherently harmful. Reduce the feedback factor and you will degrade the real-life operation of almost any design.

Such problems arise because audio electronics is a more technically complex subject than it at first appears. It is easy to cobble together some sort of power amplifier that works, and this can give people an altogether exaggerated view of how deeply they understand what they have created. In contrast, no one is likely to take a 'subjective' approach to the design of an aeroplane wing or a rocket engine; the margins for error are rather smaller, and the consequences of malfunction somewhat more serious.

The subjectivist position is of no help to anyone hoping to design a good power amplifier. However, it promises to be with us for some further time yet, and it is appropriate to review it here and show why it need not be considered at the design stage. The marketing stage is of course another matter.

Science and Subjectivism

Audio engineering is in a singular position. There can be few branches of engineering science rent from top to bottom by such a basic division as the subjectivist/rationalist dichotomy. Subjectivism is still a significant issue in the hi-fi section of the industry, but mercifully has made little headway in professional audio, where an intimate acquaintance with the original sound, and the need to earn a living with reliable and affordable equipment, provides an effective barrier against most of the irrational influences. (Note that the opposite of subjectivist is not 'objectivist'. This term refers to the followers of the philosophy of Ayn Rand.)

Most fields of technology have defined and accepted measures of excellence; car makers compete to improve mph and mpg; computer manufacturers boast of MIPs (millions of instructions per second) and so on. Improvement in these real quantities is regarded as unequivocally a step forward. In the field of hi-fi, many people seem to have difficulty in deciding which direction forward is.

Working as a professional audio designer, I often encounter opinions which, while an integral part of the subjectivist offshoot of hi-fi, are treated with ridicule by practitioners of other branches of electrical engineering. The would-be designer is not likely to be encouraged by being told that audio is not far removed from witchcraft, and that no one truly knows what they are doing. I have been told by a subjectivist that the operation of the human ear is so complex that its interaction with measurable parameters lies forever beyond human comprehension. I hope this is an extreme position; it was, I may add, proffered as a flat statement rather than a basis for discussion.

I have studied audio design from the viewpoints of electronic design, psychoacoustics, and my own humble efforts at musical creativity. I have found complete skepticism towards subjectivism to be the only tenable position. Nonetheless, if hitherto unsuspected dimensions of audio quality are ever shown to exist, then I look forward keenly to exploiting them. At this point I should say that no doubt most of the esoteric opinions are held in complete sincerity.

The Subjectivist Position

A short definition of the subjectivist position on power amplifiers might read as follows:

- Objective measurements of an amplifier's performance are unimportant compared with the subjective impressions received in informal listening tests. Should the two contradict, the objective results may be dismissed.
- Degradation effects exist in amplifiers that are unknown to orthodox engineering science, and are not revealed by the usual objective tests.
- Considerable latitude may be employed in suggesting hypothetical mechanisms of audio impairment, such as mysterious capacitor shortcomings and subtle cable defects, without reference to the plausibility of the concept, or the gathering of objective evidence of any kind.

I hope that this is considered a reasonable statement of the situation; meanwhile the great majority of the paying public continue to buy conventional hi-fi systems, ignoring the expensive and esoteric high-end sector where the debate is fiercest.

It may appear unlikely that a sizeable part of an industry could have set off in a direction that is quite counter to the facts; it could be objected that such a loss of direction in a scientific subject would be unprecedented. This is not so.

Parallel events that suggest themselves include the destruction of the study of genetics under Lysenko in the USSR[1]. Another possibility is the study of parapsychology, now in deep trouble because after some 100 years of investigation it has not uncovered the ghost (sorry) of a repeatable phenomenon[2]. This sounds all too familiar. It could be argued that parapsychology is a poor analogy because most people would accept that there was nothing there to study in the first place, whereas nobody would assert that objective measurements and subjective sound quality have no correlation at all; one need only pick up the telephone to remind oneself what a 4 kHz bandwidth and 10% or so THD sounds like.

The most startling parallel I have found in the history of science is the almost forgotten affair of Blondlot and the N-rays[3]. In 1903, Rene Blondlot, a respected French physicist, claimed to have discovered a new form of radiation he called 'N-rays'. (This was shortly after the discovery of X-rays by Roentgen, so rays were in the air, as it were.) This invisible radiation was apparently mysteriously refracted by aluminum prisms; but the crucial factor was that its presence could only be shown by subjective assessment of the brightness of an electric arc allegedly affected by N-rays. No objective measurement appeared to be possible. To Blondlot, and at least 14 of his professional colleagues, the subtle changes in brightness were real, and the French Academy published more than 100 papers on the subject.

Unfortunately N-rays were completely imaginary, a product of the 'experimenter-expectancy' effect. This was demonstrated by American scientist Robert Wood, who quietly pocketed the aluminum prism during a demonstration, without affecting Bondlot's recital of the results. After this the N-ray industry collapsed very quickly, and while it was a major embarrassment at the time, it is now almost forgotten.

The conclusion is inescapable that it is quite possible for large numbers of sincere people to deceive themselves when dealing with subjective assessments of phenomena.

A Short History of Subjectivism

The early history of sound reproduction is notable for the number of times that observers reported that an acoustic gramophone gave results indistinguishable from reality. The mere existence of such statements throws light on how powerfully mindset affects subjective impressions. Interest in sound reproduction intensified in the postwar period, and technical standards such as DIN 45–500 were set, though they were soon criticized as too permissive. By the late 1960s it was widely accepted that the requirements for hi-fi would be satisfied by 'THD less than 0.1%, with no significant crossover distortion, frequency response 20 Hz–20 kHz and as little noise as possible, please'.

The early 1970s saw this expanded to include slew rates and properly behaved overload protection, but the approach was always scientific and it was normal to read amplifier reviews in which measurements were dissected but no mention made of listening tests.

Following the growth of subjectivism through the pages of one of the leading subjectivist magazines (*Hi-Fi News*), the first intimation of what was to come was the commencement of Paul Messenger's column 'Subjective Sounds' in September 1976, in which he said: '*The assessment will be (almost) purely subjective, which has both strengths and weaknesses, as the inclusion of laboratory data would involve too much time and space, and although the ear may be the most fallible, it is also the most sensitive evaluation instrument*.' This is subjectivism as expedient rather than policy. Significantly, none of the early installments contained references to amplifier sound. In March 1977, an article by Jean Hiraga was published vilifying high levels of negative feedback and praising the sound of an amplifier with 2% THD. In the same issue, Paul Messenger stated that a Radford valve amplifier sounded better than a transistor one, and by the end of the year the amplifier-sound bandwagon was rolling. Hiraga returned in August 1977 with a highly contentious set of claims about audible speaker cables, and after that no hypothesis was too unlikely to receive attention.

The Limits of Hearing

In evaluating the subjectivist position, it is essential to consider the known abilities of the human ear. Contrary to the impression given by some commentators, who call constantly for more psychoacoustical research, a vast amount of hard scientific information already exists on this subject, and some of it may be briefly summarized thus:

- The smallest step-change in amplitude that can be detected is about 0.3 dB for a pure tone. In more realistic situations it is 0.5–1.0 dB. This is about a 10% change[4].
- The smallest detectable change in frequency of a tone is about 0.2% in the band 500 Hz–2 kHz. In percentage terms, this is the parameter for which the ear is most sensitive[5].
- The least detectable amount of harmonic distortion is not an easy figure to determine, as there is a multitude of variables involved, and in particular the continuously varying level of program means that the level of THD introduced is also dynamically changing. With mostly low-order harmonics present the just-detectable amount is about 1%, though crossover effects can be picked up at 0.3%, and probably lower. There is certainly no evidence that an amplifier producing 0.001% THD sounds any cleaner than one producing 0.005%[6].

It is acknowledged that THD measurements, taken with the usual notch-type analyzer, are of limited use in predicting the subjective impairment produced by an imperfect audio path. With music, etc. intermodulation effects are demonstrably more important than harmonics. However, THD tests have the unique advantage that visual inspection of the distortion residual gives an experienced observer a great deal of information about the root cause of the nonlinearity. Many

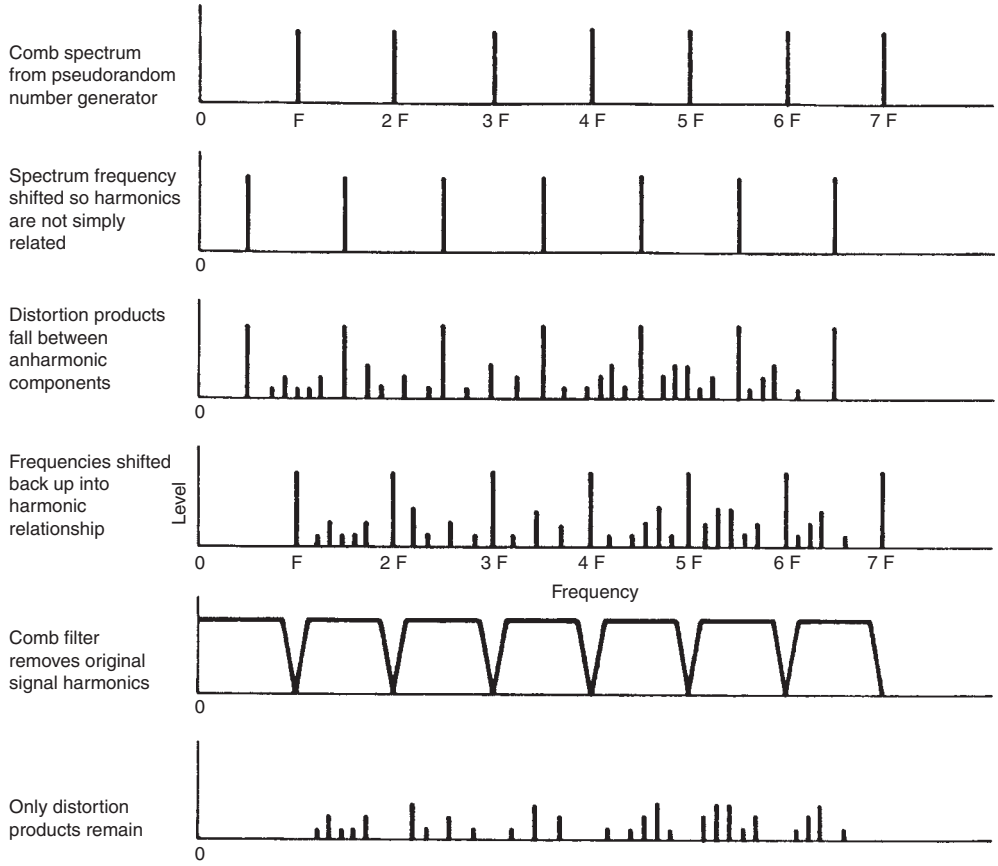


Figure 1.1: Basic principle of Belcher intermodulation test

other distortion tests exist which, while yielding very little information to the designer, exercise the whole audio bandwidth at once and correlate well with properly conducted tests for subjective impairment by distortion. The Belcher intermodulation test (the principle is shown in Figure 1.1) deserves more attention than it has received, and may become more popular now that DSP chips are cheaper.

One of the objections often made to THD tests is that their resolution does not allow verification that no nonlinearities exist at very low level – a sort of micro-crossover distortion. Hawksford, for example, has stated '*Low-level threshold phenomena . . . set bounds upon the ultimate transparency of an audio system*'[7], and several commentators have stated their belief that some metallic contacts consist of a net of so-called 'micro-diodes'. In fact, this kind of mischievous hypothesis can be disposed of using THD techniques.

I evolved a method of measuring THD down to 0.01% at 200 μV rms, and applied it to large electrolytics, connectors of varying provenance, and lengths of copper cable with and without alleged magic properties. The method required the design of an ultra-low noise (EIN = −150 dBu for a 10 Ω

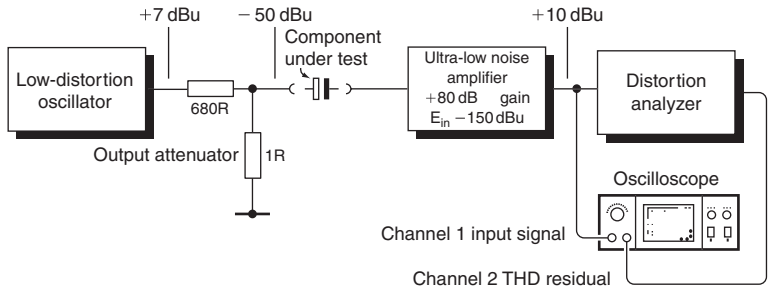


Figure 1.2: THD measurements at very low levels

source resistance) and very low THD[8]. The measurement method is shown in Figure 1.2; using an attenuator with a very low value of resistance to reduce the incoming signal keeps the Johnson noise to a minimum. In no case was any unusual distortion detected, and it would be nice to think that this red herring at least has been laid to rest.

- Interchannel crosstalk can obviously degrade stereo separation, but the effect is not detectable until it is worse than 20 dB, which would be a very bad amplifier indeed[9].
- Phase and group delay have been an area of dispute for a long time. As Stanley Lipshitz et al. have pointed out, these effects are obviously perceptible if they are gross enough; if an amplifier was so heroically misconceived as to produce the top half of the audio spectrum 3 hours after the bottom, there would be no room for argument. In more practical terms, concern about phase problems has centered on loudspeakers and their crossovers, as this would seem to be the only place where a phase shift might exist without an accompanying frequency-response change to make it obvious. Lipshitz appears to have demonstrated[10] that a second-order all-pass filter (an all-pass filter gives a frequency-dependent phase shift without level changes) is audible, whereas BBC findings reported by Harwood[11] indicate the opposite, and the truth of the matter is still not clear. This controversy is of limited importance to amplifier designers, as it would take spectacular incompetence to produce a circuit that included an accidental all-pass filter. Without such, the phase response of an amplifier is completely defined by its frequency response, and vice versa; in Control Theory this is Bode's Second Law[12], and it should be much more widely known in the hi-fi world than it is. A properly designed amplifier has its response roll-off points not too far outside the audio band, and these will have accompanying phase shifts; there is no evidence that these are perceptible[8].

The picture of the ear that emerges from psychoacoustics and related fields is not that of a precision instrument. Its ultimate sensitivity, directional capabilities and dynamic range are far more impressive than its ability to measure small level changes or detect correlated low-level signals like distortion harmonics. This is unsurprising; from an evolutionary viewpoint the functions of the ear are to warn of approaching danger (sensitivity and direction-finding being paramount) and for speech. In speech perception the identification of formants (the bands of harmonics from vocal-chord pulse excitation, selectively emphasized by vocal-tract resonances) and vowel/consonant

discriminations are infinitely more important than any hi-fi parameter. Presumably the whole existence of music as a source of pleasure is an accidental side-effect of our remarkable powers of speech perception: how it acts as a direct route to the emotions remains profoundly mysterious.

Articles of Faith: The Tenets of Subjectivism

All of the alleged effects listed below have received considerable affirmation in the audio press, to the point where some are treated as facts. The reality is that none of them has in the last 15 years proved susceptible to objective confirmation. This sad record is perhaps equalled only by students of parapsychology. I hope that the brief statements below are considered fair by their proponents. If not I have no doubt I shall soon hear about it:

- *Sine waves are steady-state signals that represent too easy a test for amplifiers, compared with the complexities of music.*

This is presumably meant to imply that sine waves are in some way particularly easy for an amplifier to deal with, the implication being that anyone using a THD analyzer must be hopelessly naive. Since sines and cosines have an unending series of non-zero differentials, steady hardly comes into it. I know of no evidence that sine waves of randomly varying amplitude (for example) would provide a more searching test of amplifier competence.

I hold this sort of view to be the result of anthropomorphic thinking about amplifiers, treating them as though they think about what they amplify. Twenty sine waves of different frequencies may be conceptually complex to us, and the output of a symphony orchestra even more so, but to an amplifier both composite signals resolve to a single instantaneous voltage that must be increased in amplitude and presented at low impedance. An amplifier has no perspective on the signal arriving at its input, but must literally take it as it comes.

- *Capacitors affect the signal passing through them in a way invisible to distortion measurements.*

Several writers have praised the technique of subtracting pulse signals passed through two different sorts of capacitor, claiming that the non-zero residue proves that capacitors can introduce audible errors. My view is that these tests expose only well-known capacitor shortcomings such as dielectric absorption and series resistance, plus perhaps the vulnerability of the dielectric film in electrolytics to reverse-biasing. No one has yet shown how these relate to capacitor audibility in properly designed equipment.

- *Passing an audio signal through cables, printed-circuit board (PCB) tracks or switch contacts causes a cumulative deterioration. Precious metal contact surfaces alleviate but do not eliminate the problem. This too is undetectable by tests for nonlinearity.*

Concern over cables is widespread, but it can be said with confidence that there is as yet not a shred of evidence to support it. Any piece of wire passes a sine wave with unmeasurable distortion, and so simple notions of inter-crystal rectification or ‘micro-diodes’ can be discounted, quite apart from

the fact that such behaviour is absolutely ruled out by established materials science. No plausible means of detecting, let alone measuring, cable degradation has ever been proposed.

The most significant parameter of a loudspeaker cable is probably its lumped inductance. This can cause minor variations in frequency response at the very top of the audio band, given a demanding load impedance. These deviations are unlikely to exceed 0.1 dB for reasonable cable constructions (say, inductance less than 4 μH). The resistance of a typical cable (say, 0.1 Ω) causes response variations across the band, following the speaker impedance curve, but these are usually even smaller at around 0.05 dB. This is not audible.

Corrosion is often blamed for subtle signal degradation at switch and connector contacts; this is unlikely. By far the most common form of contact degradation is the formation of an insulating sulfide layer on silver contacts, derived from hydrogen sulfide air pollution. This typically cuts the signal altogether, except when signal peaks temporarily punch through the sulfide layer. The effect is gross and seems inapplicable to theories of subtle degradation. Gold-plating is the only certain cure. It costs money.

- *Cables are directional, and pass audio better in one direction than the other.*

Audio signals are AC. Cables cannot be directional any more than 2+2 can equal 5. Anyone prepared to believe this nonsense will not be capable of designing amplifiers, so there seems no point in further comment.

- *The sound of valves is inherently superior to that of any kind of semiconductor.*

The 'valve sound' is one phenomenon that may have a real existence; it has been known for a long time that listeners sometimes prefer to have a certain amount of second-harmonic distortion added in[13], and most valve amplifiers provide just that, due to grave difficulties in providing good linearity with modest feedback factors. While this may well sound nice, hi-fi is supposedly about accuracy, and if the sound is to be thus modified it should be controllable from the front panel by a 'niceness' knob.

The use of valves leads to some intractable problems of linearity, reliability and the need for intimidatingly expensive (and, once more, nonlinear) iron-cored transformers. The current fashion is for exposed valves, and it is not at all clear to me that a fragile glass bottle, containing a red-hot anode with hundreds of volts DC on it, is wholly satisfactory for domestic safety.

A recent development in subjectivism is enthusiasm for single-ended directly heated triodes, usually in extremely expensive monoblock systems. Such an amplifier generates large amounts of second-harmonic distortion, due to the asymmetry of single-ended operation, and requires a very large output transformer as its primary carries the full DC anode current, and core saturation must be avoided. Power outputs are inevitably very limited at 10 W or less. In a recent review, the Cary CAD-300SEI triode amplifier yielded 3% THD at 9 W, at a cost of £3400[14]. And you still need to buy a pre-amp.

- *Negative feedback is inherently a bad thing; the less it is used, the better the amplifier sounds, without qualification.*

Negative feedback is not inherently a bad thing; it is an absolutely indispensable principle of electronic design, and if used properly has the remarkable ability to make just about every parameter better. It is usually global feedback that the critic has in mind. Local negative feedback is grudgingly regarded as acceptable, probably because making a circuit with no feedback of any kind is near impossible. It is often said that high levels of NFB enforce a low slew rate. This is quite untrue; and this thorny issue is dealt with in detail in Chapters 4 and 8. For more on slew rate, see also Ref. [15].

- *Tone controls cause an audible deterioration even when set to the flat position.*

This is usually blamed on 'phase shift'. At the time of writing, tone controls on a pre-amp badly damage its chances of street (or rather sitting-room) credibility, for no good reason. Tone controls set to 'flat' cannot possibly contribute any extra phase shift and must be inaudible. My view is that they are absolutely indispensable for correcting room acoustics, loudspeaker shortcomings, or tonal balance of the source material, and that a lot of people are suffering suboptimal sound as a result of this fashion. It is now commonplace for audio critics to suggest that frequency-response inadequacies should be corrected by changing loudspeakers. This is an extraordinarily expensive way of avoiding tone controls.

- *The design of the power supply has subtle effects on the sound, quite apart from ordinary dangers like ripple injection.*

All good amplifier stages ignore imperfections in their power supplies, op-amps in particular excelling at power-supply rejection ratio. More nonsense has been written on the subject of subtle PSU failings than on most audio topics; recommendations of hard-wiring the mains or using gold-plated 13 A plugs would seem to hold no residual shred of rationality, in view of the usual processes of rectification and smoothing that the raw AC undergoes. And where do you stop? At the local substation? Should we gold-plate the pylons?

- *Monobloc construction (i.e. two separate power amplifier boxes) is always audibly superior, due to the reduction in crosstalk.*

There is no need to go to the expense of monobloc power amplifiers in order to keep crosstalk under control, even when making it substantially better than the −20 dB that is actually necessary. The techniques are conventional; the last stereo power amplifier I designed managed an easy −90 dB at 10 kHz without anything other than the usual precautions. In this area dedicated followers of fashion pay dearly for the privilege, as the cost of the mechanical parts will be nearly doubled.

- *Microphony is an important factor in the sound of an amplifier, so any attempt at vibration damping is a good idea.*

Microphony is essentially something that happens in sensitive valve preamplifiers. If it happens in solid-state power amplifiers the level is so far below the noise it is effectively nonexistent.

Experiments on this sort of thing are rare (if not unheard of) and so I offer the only scrap of evidence I have. Take a microphone pre-amp operating at a gain of +70 dB, and tap the input

capacitors (assumed electrolytic) sharply with a screwdriver; the pre-amp output will be a dull thump, at low level. The physical impact on the electrolytics (the only components that show this effect) is hugely greater than that of any acoustic vibration; and I think the effect in power amps, if any, must be so vanishingly small that it could never be found under the inherent circuit noise.

Let us for a moment assume that some or all of the above hypotheses are true, and explore the implications. The effects are not detectable by conventional measurement, but are assumed to be audible. First, it can presumably be taken as axiomatic that for each audible defect some change occurs in the pattern of pressure fluctuations reaching the ears, and therefore a corresponding modification has occurred to the electrical signal passing through the amplifier. Any other starting point supposes that there is some other route conveying information apart from the electrical signals, and we are faced with magic or forces unknown to science. Mercifully no commentator has (so far) suggested this. Hence there must be defects in the audio signals, but they are not revealed by the usual test methods. How could this situation exist? There seem to be two possible explanations for this failure of detection: one is that the standard measurements are relevant but of insufficient resolution, and we should be measuring frequency response, etc., to thousandths of a decibel. There is no evidence whatsoever that such micro-deviations are audible under any circumstances.

An alternative (and more popular) explanation is that standard sine-wave THD measurements miss the point by failing to excite subtle distortion mechanisms that are triggered only by music, the spoken word, or whatever. This assumes that these music-only distortions are also left undisturbed by multi-tone intermodulation tests, and even the complex pseudorandom signals used in the Belcher distortion test[16]. The Belcher method effectively tests the audio path at all frequencies at once, and it is hard to conceive of a real defect that could escape it.

The most positive proof that subjectivism is fallacious is given by subtraction testing. This is the devastatingly simple technique of subtracting before and after amplifier signals and demonstrating that nothing audibly detectable remains.

It transpires that these alleged music-only mechanisms are not even revealed by music, or indeed anything else, and it appears the subtraction test has finally shown as nonexistent these elusive degradation mechanisms.

The subtraction technique was proposed by Baxandall in 1977[17]. The principle is shown in Figure 1.3; careful adjustment of the roll-off balance network prevents minor bandwidth variations from swamping the true distortion residual. In the intervening years the subjectivist camp has made no effective reply.

A simplified version of the test was introduced by Hafler[18]. This method is less sensitive, but has the advantage that there is less electronics in the signal path for anyone to argue about (see Figure 1.4). A prominent subjectivist reviewer, on trying this demonstration, was reduced to claiming that the passive switchbox used to implement the Hafler test was causing so much sonic degradation that all amplifier performance was swamped[19]. I do not feel that this is a tenable position. So far all experiments such as these have been ignored or brushed aside by the subjectivist camp; no attempt has been made to answer the extremely serious objections that this demonstration raises.

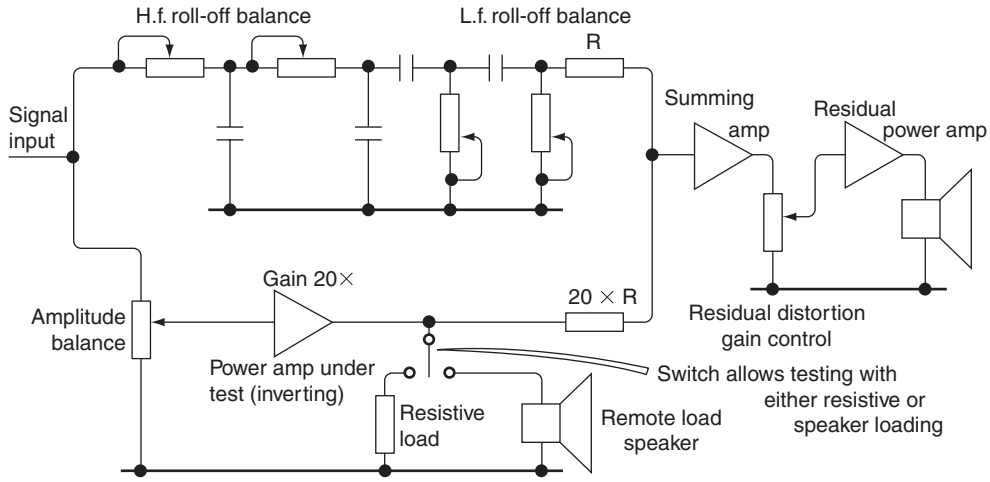


Figure 1.3: Baxandall cancelation technique

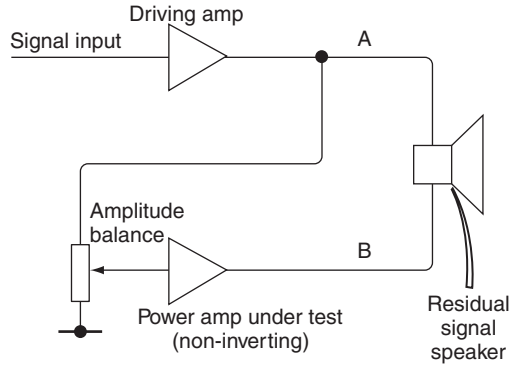


Figure 1.4: Hafler straight-wire differential test

In the 20 or so years that have elapsed since the emergence of the Subjectivist Tendency, no hitherto unsuspected parameters of audio quality have emerged.

The Length of the Audio Chain

An apparently insurmountable objection to the existence of non-measurable amplifier quirks is that recorded sound of almost any pedigree has passed through a complex mixing console at least once; prominent parts like vocals or lead guitar will almost certainly have passed through at least twice, once for recording and once at mix-down. More significantly, it must have passed through the potential quality bottleneck of an analog tape machine or more likely the A–D converters of digital equipment. In its long path from here to ear the audio passes through at least 100 op-amps, dozens of connectors, and several hundred meters of ordinary screened cable. If mystical degradations can occur, it defies reason to insist that those introduced by the last 1% of the path are the critical ones.

The Implications

This confused state of amplifier criticism has negative consequences. First, if equipment is reviewed with results that appear arbitrary, and which are in particular incapable of replication or confirmation, this can be grossly unfair to manufacturers who lose out in the lottery. Since subjective assessments cannot be replicated, the commercial success of a given make can depend entirely on the vagaries of fashion. While this is fine in the realm of clothing or soft furnishings, the hi-fi business is still claiming accuracy of reproduction as its raison d'être, and therefore you would expect the technical element to be dominant.

A second consequence of placing subjectivism above measurements is that it places designers in a most unenviable position. No degree of ingenuity or attention to technical detail can ensure a good review, and the pressure to adopt fashionable and expensive expedients (such as linear-crystal internal wiring) is great, even if the designer is certain that they have no audible effect for good or evil. Designers are faced with a choice between swallowing the subjectivist credo whole or keeping very quiet and leaving the talking to the marketing department.

If objective measurements are disregarded, it is inevitable that poor amplifiers will be produced, some so bad that their defects are unquestionably audible. In recent reviews[20] it was easy to find a £795 preamplifier (Counterpoint SA7) that boasted a feeble 12 dB disk overload margin (another pre-amp costing £2040 struggled up to 15 dB – Burmester 838/846) and another costing £1550 that could only manage a 1 kHz distortion performance of 1%, a lack of linearity that would have caused consternation 10 years ago (Quicksilver). However, by paying £5700 one could inch this down to 0.3% (Audio Research M100-2 monoblocs). This does not of course mean that it is impossible to buy an 'audiophile' amplifier that does measure well; another example would be the preamplifier/power amplifier combination that provides a very respectable disk overload margin of 31 dB and 1 kHz rated-power distortion below 0.003%, the total cost being £725 (Audiolab 8000C/8000P). I believe this to be a representative sample, and we appear to be in the paradoxical situation that the most expensive equipment provides the worst objective performance. Whatever the rights and wrongs of subjective assessment, I think that most people would agree that this is a strange state of affairs. Finally, it is surely a morally ambiguous position to persuade non-technical people that to get a really good sound they have to buy £2000 pre-amps and so on, when both technical orthodoxy and common sense indicate that this is quite unnecessary.

The Reasons Why

Some tentative conclusions are possible as to why hi-fi engineering has reached the pass that it has. I believe one basic reason is the difficulty of defining the quality of an audio experience; you cannot draw a diagram to communicate what something sounded like. In the same way, acoustical memory is more evanescent than visual memory. It is far easier to visualize what a London bus looks like than to recall the details of a musical performance. Similarly, it is difficult to 'look more closely': turning up the volume is more like turning up the brightness of a TV picture; once an optimal level is reached, any further increase becomes annoying, then painful.

It has been universally recognized for many years in experimental psychology, particularly in experiments about perception, that people tend to perceive what they want to perceive. This is often called the *experimenter-expectancy* effect; it is more subtle and insidious than it sounds, and the history of science is littered with the wrecked careers of those who failed to guard against it. Such self-deception has most often occurred in fields like biology, where although the raw data may be numerical, there is no real mathematical theory to check it against. When the only 'results' are vague subjective impressions, the danger is clearly much greater, no matter how absolute the integrity of the experimenter. Thus in psychological work great care is necessary in the use of impartial observers, double-blind techniques, and rigorous statistical tests for significance. The vast majority of subjectivist writings wholly ignore these precautions, with predictable results. In a few cases properly controlled listening tests have been done, and at the time of writing all have resulted in different amplifiers sounding indistinguishable. I believe the conclusion is inescapable that experimenter expectancy has played a dominant role in the growth of subjectivism.

It is notable that in subjectivist audio the 'correct' answer is always the more expensive or inconvenient one. Electronics is rarely as simple as that. A major improvement is more likely to be linked with a new circuit topology or new type of semiconductor, than with mindlessly specifying more expensive components of the same type; cars do not go faster with platinum pistons.

It might be difficult to produce a rigorous statistical analysis, but it is my view that the reported subjective quality of a piece of equipment correlates far more with the price than with anything else. There is perhaps here an echo of the Protestant work ethic: you must suffer now to enjoy yourself later. Another reason for the relatively effortless rise of subjectivism is the *me-too* effect; many people are reluctant to admit that they cannot detect acoustic subtleties as nobody wants to be labeled as insensitive, outmoded, or just plain deaf. It is also virtually impossible to absolutely disprove any claims, as the claimant can always retreat a fraction and say that there was something special about the combination of hardware in use during the disputed tests, or complain that the phenomena are too delicate for brutal logic to be used on them. In any case, most competent engineers with a taste for rationality probably have better things to do than dispute every controversial report.

Under these conditions, vague claims tend, by a kind of intellectual inflation, to gradually become regarded as facts. Manufacturers have some incentive to support the subjectivist camp as they can claim that only they understand a particular non-measurable effect, but this is no guarantee that the dice may not fall badly in a subjective review.

The Outlook

It seems unlikely that subjectivism will disappear for a long time, if ever, given the momentum that it has gained, the entrenched positions that some people have taken up, and the sadly uncritical way in which people accept an unsupported assertion as the truth simply because it is asserted with frequency and conviction. In an ideal world every such statement would be greeted by loud demands for evidence. However, the history of the world sometimes leads one to suppose pessimistically that people will believe anything. By analogy, one might suppose that subjectivism

would persist for the same reason that parapsychology has; there will always be people who will believe what they want to believe rather than what the hard facts indicate.

More than 10 years have passed since the above material on subjectivism was written, but there seems to be no reason to change a word of it. Amplifier reviews continue to make completely unsupportable assertions, of which the most obtrusive these days is the notion that an amplifier can in some way alter the 'timing' of music. This would be a remarkable feat to accomplish with a handful of transistors, were it not wholly imaginary.

During my sojourn at TAG-McLaren Audio, we conducted an extensive set of double-blind listening tests, using a lot of experienced people from various quarters of the hi-fi industry. An amplifier loosely based on the Otala four-stage architecture was compared with a Blameless three-stage architecture perpetrated by myself (these terms are fully explained in Chapter 2). The two amplifiers could not have been more different – the four-stage had complex lead-lag compensation and a buffered complementary feedback pair (CFP) output, while my three-stage had conventional Miller dominant-pole compensation. There were too many other detail differences to list here. After a rigorous statistical analysis the result – as you may have guessed – was that nobody could tell the two amplifiers apart.

Technical Errors

Misinformation also arises in the purely technical domain; I have also found some of the most enduring and widely held technical beliefs to be unfounded. For example, if you take a Class-B amplifier and increase its quiescent current so that it runs in Class-A at low levels, i.e. in Class-AB, most people will tell you that the distortion will be reduced as you have moved nearer to the full Class-A condition. This is untrue. A correctly configured amplifier gives more distortion in Class-AB, not less, because of the abrupt gain changes inherent in switching from A to B every cycle.

Discoveries like this can only be made because it is now straightforward to make testbed amplifiers with ultra-low distortion – lower than that which used to be thought possible. The reduction of distortion to the basic or inherent level that a circuit configuration is capable of is a fundamental requirement for serious design work in this field; in Class-B at least this gives a defined and repeatable standard of performance that in later chapters I name a Blameless amplifier, so called because it avoids error rather than claiming new virtues.

It has proved possible to take the standard Class-B power amplifier configuration, and by minor modifications reduce the distortion to below the noise floor at low frequencies. This represents approximately 0.0005–0.0008% THD, depending on the exact design of the circuitry, and the actual distortion can be shown to be substantially below this if spectrum-analysis techniques are used to separate the harmonics from the noise.

The Performance Requirements for Amplifiers

This section is not a recapitulation of international standards, which are intended to provide a minimum level of quality rather than extend the art. It is rather my own view of what you should

be worrying about at the start of the design process, and the first items to consider are the brutally pragmatic ones related to keeping you in business and out of prison.

Safety

In the drive to produce the finest amplifier ever made, do not forget that the Prime Directive of audio design is – Thou Shalt Not Kill. Every other consideration comes a poor second, not only for ethical reasons, but also because one serious lawsuit will close down most audio companies forever.

Reliability

If you are in the business of manufacturing, you had better make sure that your equipment keeps working, so that you too can keep working. It has to be admitted that power amplifiers – especially the more powerful ones – have a reputation for reliability that is poor compared with most branches of electronics. The 'high end' in particular has gathered to itself a bad reputation for dependability[21].

Power Output

In commercial practice, this is decided for you by the marketing department. Even if you can please yourself, the power output capability needs careful thought as it has a powerful and nonlinear effect on the cost.

The last statement requires explanation. As the output power increases, a point is reached when single output devices are incapable of sustaining the thermal dissipation; parallel pairs are required, and the price jumps up. Similarly, transformer laminations come in standard sizes, so the transformer size and cost will also increase in discrete steps.

Domestic hi-fi amplifiers usually range from 20 to 150 W into 8 Ω, though with a scattering of much higher powers. PA units will range from 50 W, for foldback purposes (i.e. the sound the musician actually hears, to monitor his/her playing, as opposed to that thrown out forwards by the main PA stacks, also called stage monitoring) to 1 kW or more. Amplifiers of extreme high power are not popular, partly because the economies of scale are small, but mainly because it means putting all your eggs in one basket, and a failure becomes disastrous. This is accentuated by the statistically unproven but almost universally held opinion that high-power solid-state amplifiers are inherently less reliable than others.

If an amplifier gives a certain output into 8 Ω, it will not give exactly twice as much into 4 Ω loads; in fact it will probably be much less than this, due to the increased resistive losses in 4 Ω operation, and the way that power alters as the square of voltage. Typically, an amplifier giving 180 W into 8 Ω might be expected to yield 260 W into 4 Ω and 350 W into 2 Ω, if it can drive so low a load at all. These figures are approximate, depending very much on power supply design.

Nominally 8 Ω loudspeakers are the most common in hi-fi applications. The 'nominal' title accommodates the fact that all loudspeakers, especially multi-element types, have marked changes

in input impedance with frequency, and are only resistive at a few spot frequencies. Nominal 8 Ω loudspeakers may be expected to drop to at least 6 Ω in some part of the audio spectrum. To allow for this, almost all amplifiers are rated as capable of 4 Ω as well as 8 Ω loads. This takes care of almost any nominal 8 Ω speaker, but leaves no safety margin for nominal 4 Ω designs, which are likely to dip to 3 Ω or less. Extending amplifier capability to deal with lower load impedances for anything other than very short periods has serious cost implications for the power-supply transformer and heat-sinking; these already represent the bulk of the cost.

The most important thing to remember in specifying output power is that you have to increase it by an awful lot to make the amplifier significantly louder. We do not perceive acoustic power as such – there is no way we could possibly integrate the energy liberated in a room, and it would be a singularly useless thing to perceive if we could. It is much nearer the truth to say that we perceive pressure. It is well known that power in watts must be quadrupled to double sound pressure level (SPL), but this is not the same as doubling subjective loudness; this is measured in Sones rather than dB above threshold, and some psychoacousticians have reported that doubling subjective loudness requires a 10 dB rather than 6 dB rise in SPL, implying that amplifier power must be increased tenfold, rather than merely quadrupled[22]. It is at any rate clear that changing from a 25 W to a 30 W amplifier will not give an audible increase in level.

This does not mean that fractions of a watt are never of interest. They can matter either in pursuit of maximum efficiency for its own sake, or because a design is only just capable of meeting its output specification.

Some hi-fi reviewers set great value on very high peak current capability for short periods. While it is possible to think up special test waveforms that demand unusually large peak currents, any evidence that this effect is important in use is so far lacking.

Frequency Response

This can be dealt with crisply; the minimum is 20 Hz–20 kHz, ±0.5 dB, though there should never be any *plus* about it when solid-state amplifiers are concerned. Any hint of a peak before the roll-off should be looked at with extreme suspicion, as it probably means doubtful HF stability. This is less true of valve amplifiers, where the bandwidth limits of the output transformer mean that even modest NFB factors tend to cause peaking at both high and low ends of the spectrum.

Having dealt with the issue crisply, there is no hope that everyone will agree that this is adequate. CDs do not have the built-in LF limitations of vinyl and could presumably encode the barometric pressure in the recording studio if this was felt to be desirable, and so an extension to −0.5 dB at 5 or 10 Hz is perfectly feasible. However, if infrabass information does exist down at these frequencies, no domestic loudspeaker will reproduce them.

Noise

There should be as little as possible without compromising other parameters. The noise performance of a power amplifier is not an irrelevance[23], especially in a domestic setting.

Distortion

Once more, a sensible target might be: *as little as possible without messing up something else*. This ignores the views of those who feel a power amplifier is an appropriate device for adding distortion to a musical performance. Such views are not considered in the body of this book; it is, after all, not a treatise on fuzz-boxes or other guitar effects.

I hope that the techniques explained in this book have a relevance beyond power amplifiers. Applications obviously include discrete op-amp-based preamplifiers[24], and extend to any amplifier aiming at static or dynamic precision.

My philosophy is the simple one that distortion is bad and high-order distortion is worse. The first part of this statement is, I suggest, beyond argument, and the second part has a good deal of evidence to back it. The distortion of the nth harmonic should be weighted by $n^2/4$ worse, according to many authorities[25]. This leaves the second harmonic unchanged, but scales up the third by 9/4, i.e. 2.25 times, the fourth by 16/4, i.e. 4 times, and so on. It is clear that even small amounts of high-order harmonics could be unpleasant, and this is one reason why even modest crossover distortion is of such concern.

Digital audio now routinely delivers the signal with less than 0.002% THD, and I can earnestly vouch for the fact that analog console designers work furiously to keep the distortion in long complex signal paths down to similar levels. I think it an insult to allow the very last piece of electronics in the chain to make nonsense of these efforts.

I would like to make it clear that I do not believe that an amplifier yielding 0.001% THD is going to sound much better than its fellow giving 0.002%. However, if there is ever a scintilla of doubt as to what level of distortion is perceptible, then using the techniques I have presented it should be possible to routinely reduce the THD below the level at which there can be any rational argument.

I am painfully aware that there is a school of thought that regards low THD as inherently immoral, but this is to confuse electronics with religion. The implication is that very low THD can only be obtained by huge global NFB factors that require heavy dominant-pole compensation that severely degrades slew rate; the obvious flaw in this argument is that once the compensation is applied the amplifier no longer has a large global NFB factor, and so its distortion performance presumably reverts to mediocrity, further burdened with a slew rate of 4 V per fortnight.

To me low distortion has its own aesthetic and philosophical appeal; it is satisfying to know that the amplifier you have just designed and built is so linear that there simply is no realistic possibility of it distorting your favorite material. Most of the linearity-enhancing strategies examined in this book are of minimal cost (the notable exception being resort to Class-A) compared with the essential heat-sinks, transformer, etc., and so why not have ultra-low distortion? Why put up with more than you must?

Damping Factor

Audio amplifiers, with a few very special exceptions[26], approximate to perfect voltage sources, i.e. they aspire to a zero output impedance across the audio band. The result is that amplifier output is

unaffected by loading, so that the frequency-variable impedance of loudspeakers does not give an equally variable frequency response, and there is some control of speaker cone resonances.

While an actual zero impedance is impossible, a very close approximation is possible if large negative-feedback factors are used. (Actually, a judicious mixture of voltage and current feedback will make the output impedance zero, or even negative – i.e. increasing the loading makes the output voltage increase. This is clever, but usually pointless, as will be seen.) Solid-state amplifiers are quite happy with lots of feedback, but it is usually impractical in valve designs.

Damping factor (DF) is defined as the ratio of the load impedance R_{load} to the amplifier output resistance R_{out}:

$$\text{Damping factor} = \frac{R_{load}}{R_{out}} \qquad \text{Equation 1.1}$$

A solid-state amplifier typically has output resistance of the order of 0.05 Ω, so if it drives an 8 Ω speaker we get a damping factor of 160 times. This simple definition ignores the fact that amplifier output impedance usually varies considerably across the audio band, increasing with frequency as the negative feedback factor falls; this indicates that the output *resistance* is actually more like an inductive reactance. The presence of an output inductor to give stability with capacitive loads further complicates the issue.

Mercifully, damping factor as such has very little effect on loudspeaker performance. A damping factor of 160 times, as derived above, seems to imply a truly radical effect on cone response – it implies that resonances and such have been reduced by 160 times as the amplifier output takes an iron grip on the cone movement. Nothing could be further from the truth.

The resonance of a loudspeaker unit depends on the total resistance in the circuit. Ignoring the complexities of crossover circuitry in multi-element speakers, the total series resistance is the sum of the speaker coil resistance, the speaker cabling and, last of all, the amplifier output impedance. The values will be typically 7, 0.5, and 0.05 Ω respectively, so the amplifier only contributes 0.67% to the total, and its contribution to speaker dynamics must be negligible.

The highest output impedances are usually found in valve equipment, where global feedback including the output transformer is low or nonexistent; values around 0.5 Ω are usual. However, idiosyncratic semiconductor designs sometimes also have high output resistances; see Olsher[27] for a design with $R_{out} = 0.6\,\Omega$, which I feel is far too high.

This view of the matter was practically investigated and fully confirmed by James Moir as far back as 1950[28], though this has not prevented periodic resurgences of controversy.

The only reason to strive for a high damping factor – which can, after all, do no harm – is the usual numbers game of impressing potential customers with specification figures. It is as certain as anything can be that the subjective difference between two amplifiers, one with a DF of 100 and the other boasting 2000, is undetectable by human perception. Nonetheless, the specifications look

very different in the brochure, so means of maximizing the DF may be of some interest. This is examined further in Chapter 8.

Absolute Phase

Concern for absolute phase has for a long time hovered ambiguously between real audio concerns like noise and distortion, and the subjective realm where solid copper is allegedly audible. Absolute phase means the preservation of signal phase all the way from microphone to loudspeaker, so that a drum impact that sends an initial wave of positive pressure towards the live audience is reproduced as a similar positive pressure wave from the loudspeaker. Since it is known that the neural impulses from the ear retain the periodicity of the waveform at low frequencies, and distinguish between compression and rarefaction, there is a prima facie case for the audibility of absolute phase.

It is unclear how this applies to instruments less physical than a kickdrum. For the drum the situation is simple – you kick it, the diaphragm moves outwards and the start of the transient must be a wave of compression in the air (followed almost at once by a wave of rarefaction). But what about an electric guitar? A similar line of reasoning – plucking the string moves it in a given direction, which gives such and such a signal polarity, which leads to whatever movement of the cone in the guitar amp speaker cabinet – breaks down at every point in the chain. There is no way to know how the pickups are wound, and indeed the guitar will almost certainly have a switch for reversing the phase of one of them. I also suggest that the preservation of absolute phase is not the prime concern of those who design and build guitar amplifiers.

The situation is even less clear if more than one instrument is concerned, which is of course almost all the time. It is very difficult to see how two electric guitars played together could have a 'correct' phase in which to listen to them.

Recent work on the audibility of absolute phase[29,30] shows it is sometimes detectable. A single tone flipped back and forth in phase, providing it has a spiky asymmetrical waveform and an associated harsh sound, will show a change in perceived timbre and, according to some experimenters, a perceived change in pitch. A monaural presentation has to be used to yield a clear effect. A complex sound, however, such as that produced by a musical ensemble, does not in general show a detectable difference.

Proposed standards for the maintenance of absolute phase have just begun to appear[31], and the implication for amplifier designers is clear; whether absolute phase really matters or not, it is simple to maintain phase in a power amplifier and so it should be done (compare a complex mixing console, where correct phase is absolutely vital, and there are hundreds of inputs and outputs, all of which must be in phase in every possible configuration of every control). In fact, it probably already has been done, even if the designer has not given absolute phase a thought, because almost all power amplifiers use series negative feedback, and this is inherently non-inverting. Care is, however, required if there are stages such as balanced line input amplifiers before the power amplifier itself; if the hot and cold inputs get swapped by mistake then the amplifier output will be phase inverted.

Amplifier Formats

When the first edition of this book appeared in 1996, the vast majority of domestic amplifiers were two-channel stereo units. Since then there has been a great increase in other formats, particularly in multichannel units having seven or more channels for audio-visual use, and in single-channel amplifiers built into subwoofer loudspeakers.

Multichannel amplifiers come in two kinds. The most cost-effective way to build a multichannel amplifier is to put as many power amplifier channels as convenient on each PCB, and group them around a large toroidal transformer that provides a common power supply for all of them. While this keeps the costs down there are inevitable compromises on interchannel crosstalk and rejection of the transformer's stray magnetic fields. The other method is to make each channel (or, in some cases, each pair of channels) into a separate amplifier module with its own transformer, power supply, heat-sinks, and separate input and output connections – a sort of multiple-monobloc format. The modules usually share a microcontroller housekeeping system but nothing else. This form of construction gives much superior interchannel crosstalk, as the various audio circuits need have no connection with each other, and much less trouble with transformer hum as the modules are relatively long and thin so that a row of them can be fitted into a chassis, and thus the mains transformer can be put right at one end and the sensitive input circuitry right at the other. Inevitably this is a more expensive form of construction.

Subwoofer amplifiers are single channel and of high power. There seems to be a general consensus that the quality of subwoofer amplifiers is less critical than that of other amplifiers, and this has meant that both Class-G and Class-D designs have found homes in subwoofer enclosures. Subwoofer amplifiers differ from others in that they often incorporate their own specialized filtering (typically at 200 Hz) and equalization circuitry.

References

[1] M. Gardner, Fads & Fallacies in the Name of Science, Dover (Chapter 12), pp. 140–151.

[2] F.M. David, Investigating the paranormal, Nature 320 (13 March 1986).

[3] J. Randi, Flim-Flam! Psychics, ESP Unicorns and Other Delusions, Prometheus Books, 1982, pp. 196–198.

[4] J.D. Harris, Loudness discrimination, J. Speech Hear. Dis. Monogr. (Suppl. 11) pp. 1–63.

[5] B.C.J. Moore, Relation between the critical bandwidth k the frequency-difference limen, J. Acoust. Soc. Am. 55 p. 359.

[6] J. Moir, Just detectable distortion levels, Wireless World (February 1981) pp. 32–34.

[7] M. Hawksford, The Essex echo, Hi-fi News & RR (May 1986) p. 53.

[8] D. Self, Ultra-low-noise amplifiers & granularity distortion, JAES (November 1987) pp. 907–915.

[9] H. Harwood, Shorter, Stereophony and the effect of crosstalk between left and right channels, BBC Engineering Monograph No. 52.

[10] S. Lipshitz et al., On the audibility of midrange phase distortion in audio systems, JAES (September 1982) pp. 580–595.

[11] H. Harwood, Audibility of phase effects in loudspeakers, Wireless World (January 1976) pp. 30–32.

[12] S. Shinners, Modern Control System Theory and Application, Addison-Wesley, p. 310.

[13] G. King, Hi-fi reviewing, Hi-fi News & RR (May 1978) p. 77.

[14] R. Harley, Review of Cary CAD-300SEI single-ended triode amplifier, Stereophile (September 1995) p. 141.

[15] P. Baxandall, Audio power amplifier design, Wireless World (January 1978) p. 56.

[16] R.A. Belcher, A new distortion measurement, Wireless World (May 1978) pp. 36–41.

[17] P. Baxandall, Audible amplifier distortion is not a mystery, Wireless World (November 1977) pp. 63–66.

[18] D. Hafler, A listening test for amplifier distortion, Hi-fi News & RR (November 1986) pp. 25–29.

[19] M. Colloms, Hafler XL-280 test, Hi-fi News & RR (June 1987) pp. 65–67.

[20] Hi-fi Choice, The Selection, Sportscene (1986).

[21] R.H. Lawry, High end difficulties, Stereophile (May 1995) p. 23.

[22] B.J. Moore, An Introduction to the Psychology of Hearing, Academic Press, 1982, pp. 48–50.

[23] L. Fielder, Dynamic range issues in the Modern Digital Audio Environment, JAES p. 43.

[24] D. Self, Advanced preamplifier design, Wireless World (November 1976) p. 41.

[25] J. Moir, Just detectable distortion levels, Wireless World (February 1981) p. 34.

[26] P.G.L. Mills, M.O.J. Hawksford, Transconductance power amplifier systems for current-driven loudspeakers, JAES p. 37.

[27] D. Olsher, Times One RFS400 power amplifier review, Stereophile (August 1995) p. 187.

[28] J. Moir, Transients and loudspeaker damping, Wireless World (May 1950) p. 166.

[29] R.A. Greiner, D.E. Melton, A quest for the audibility of polarity, Audio (December 1993) p. 40.

[30] R.A. Greiner, D.E. Melton, Observations on the audibility of acoustic polarity, JAES p. 42.

[31] AES, Draft AES recommended practice standard for professional audio – conservation of the polarity of audio signals, inserted in: JAES p. 42.

CHAPTER 2

Power Amplifier Architecture and Negative Feedback

Amplifier Architectures

This grandiose title simply refers to the large-scale structure of the amplifier; that is, the block diagram of the circuit one level below that representing it as a single white block labeled Power Amplifier. Almost all solid-state amplifiers have a three-stage architecture as described below, though they vary in the detail of each stage. Two-stage architectures have occasionally been used, but their distortion performance is not very satisfactory. Four-stage architectures have been used in significant numbers, but they are still much rarer than three-stage designs, and usually involve relatively complex compensation schemes to deal with the fact that there is an extra stage to add phase shift and potentially imperil high-frequency stability.

The Three-Stage Amplifier Architecture

The vast majority of audio amplifiers use the conventional architecture, shown in Figure 2.1, and so it is dealt with first. There are three stages, the first being a transconductance stage (differential voltage in, current out), the second a transimpedance stage (current in, voltage out), and the third a unity-voltage-gain output stage. The second stage clearly has to provide all the voltage gain and I have therefore called it the voltage-amplifier stage or VAS. Other authors have called it the

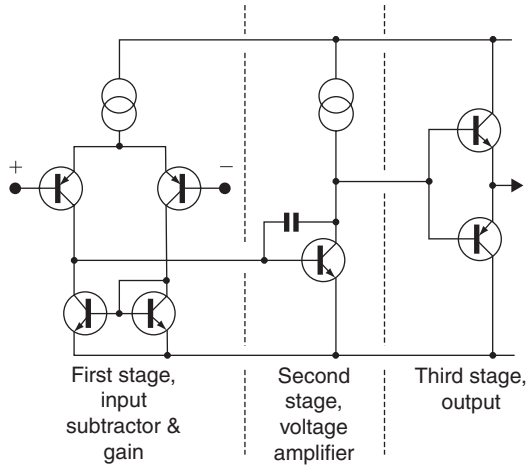


Figure 2.1: The three-stage amplifier structure. There is a transconductance stage, a transadmittance stage (the VAS), and a unity-gain buffer output stage

pre-driver stage but I prefer to reserve this term for the first transistors in output triples. This three-stage architecture has several advantages, not least being that it is easy to arrange things so that interaction between stages is negligible. For example, there is very little signal voltage at the input to the second stage, due to its current-input (virtual-earth) nature, and therefore very little on the first stage output; this minimizes Miller phase shift and possible Early effect in the input devices.

Similarly, the compensation capacitor reduces the second stage output impedance, so that the nonlinear loading on it due to the input impedance of the third stage generates less distortion than might be expected. The conventional three-stage structure, familiar though it may be, holds several elegant mechanisms such as this. They will be fully revealed in later chapters. Since the amount of linearizing global negative feedback (NFB) available depends upon amplifier open-loop gain, how the stages contribute to this is of great interest. The three-stage architecture always has a unity-gain output stage – unless you really want to make life difficult for yourself – and so the total forward gain is simply the product of the transconductance of the input stage and the transimpedance of the VAS, the latter being determined solely by the Miller capacitor C_{dom}, except at very low frequencies. Typically, the closed-loop gain will be between + 20 and + 30 dB. The NFB factor at 20 kHz will be 25–40 dB, increasing at 6 dB/octave with falling frequency until it reaches the dominant pole frequency $P1$, when it flattens out. What matters for the control of distortion is the amount of NFB available, rather than the open-loop bandwidth, to which it has no direct relationship. In my *Electronics World* Class-B design, the input stage g_m is about 9 mA/V, and C_{dom} is 100 pF, giving an NFB factor of 31 dB at 20 kHz. In other designs I have used as little as 26 dB (at 20 kHz) with good results.

Compensating a three-stage amplifier is relatively simple; since the pole at the VAS is already dominant, it can be easily increased to lower the HF negative-feedback factor to a safe level. The local NFB working on the VAS through C_{dom} has an extremely valuable linearizing effect.

The conventional three-stage structure represents at least 99% of the solid-state amplifiers built, and I make no apology for devoting much of this book to its behavior. I am quite sure I have not exhausted its subtleties.

The Two-Stage Amplifier Architecture

In contrast with the three-stage approach, the architecture in Figure 2.2 is a two-stage amplifier, the first stage being once more a transconductance stage, though now without a guaranteed low impedance to accept its output current. The second stage combines VAS and output stage in one block; it is inherent in this scheme that the VAS must double as a phase splitter as well as a generator of raw gain. There are then two quite dissimilar signal paths to the output, and it is not at all clear that trying to break this block down further will assist a linearity analysis. The use of a phase-splitting stage harks back to valve amplifiers, where it was inescapable, as a complementary valve technology has so far eluded us.

Paradoxically, a two-stage amplifier is likely to be more complex in its gain structure than a three-stage. The forward gain depends on the input stage g_m, the input stage collector load (because

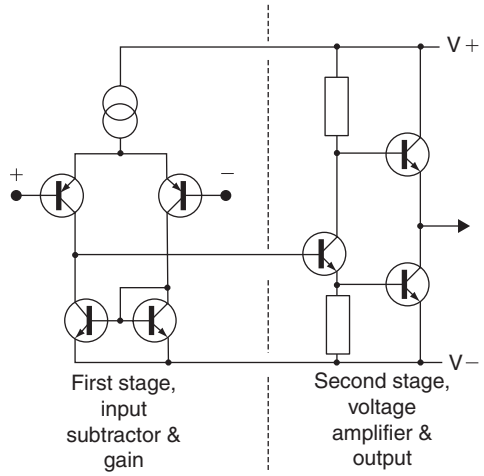


Figure 2.2: The two-stage amplifier structure. A voltage-amplifier output follows the same transconductance input stage

the input stage can no longer be assumed to be feeding a virtual earth) and the gain of the output stage, which will be found to vary in a most unsettling manner with bias and loading. Choosing the compensation is also more complex for a two-stage amplifier, as the VAS/phase splitter has a significant signal voltage on its input and so the usual pole-splitting mechanism that enhances Nyquist stability by increasing the pole frequency associated with the input stage collector will no longer work so well. (I have used the term Nyquist stability, or Nyquist oscillation, throughout this book to denote oscillation due to the accumulation of phase shift in a global NFB loop, as opposed to local parasitics, etc.)

The LF feedback factor is likely to be about 6 dB less with a 4 Ω load, due to lower gain in the output stage. However, this variation is much reduced above the dominant pole frequency, as there is then increasing local NFB acting in the output stage.

Here are two examples of two-stage amplifiers: Linsley-Hood[1] and Olsson[2]. The two-stage amplifier offers little or no reduction in parts cost, is harder to design, and in my experience invariably gives a poor distortion performance.

The Four-Stage Amplifier Architecture

The best-known example of a four-stage architecture is probably that published by Lohstroh and Otala in their influential paper, which was confidently entitled 'An audio power amplifier for ultimate quality requirements' and appeared in December 1973[3]. A simplified circuit diagram of their design is shown in Figure 2.3. One of their design objectives was the use of a low value of overall feedback, made possible by heavy local feedback in the first three amplifier stages, in the form of emitter degeneration; the closed-loop gain was 32 dB (40 times) and the feedback factor 20 dB, allegedly flat across the audio band. Another objective was the elimination of so-called transient intermodulation distortion, which after many years of argument and futile debate has

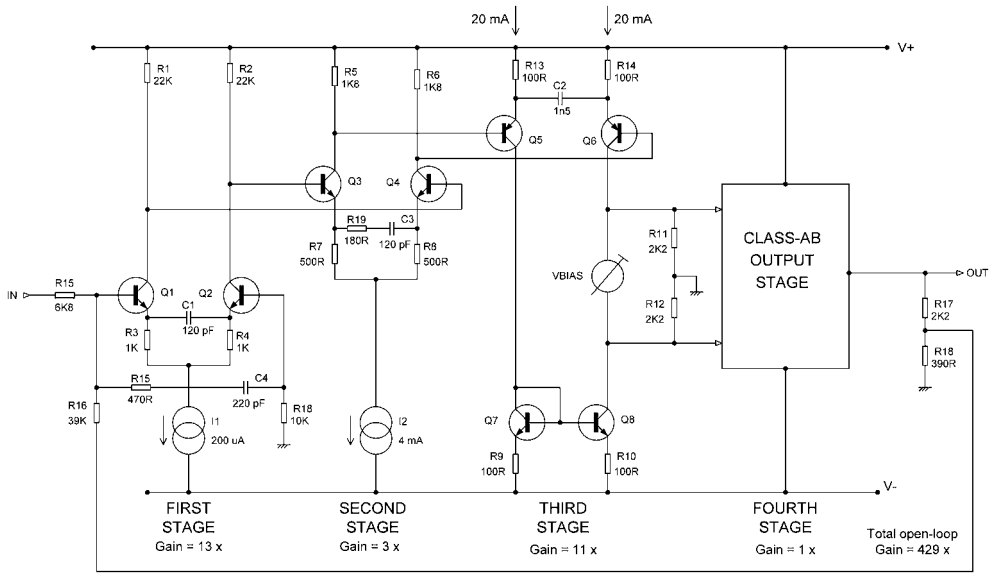


Figure 2.3: A simplified circuit diagram of the Lohstroh and Otala four-stage power amplifier. The gain figures for each stage are as quoted in the original paper

at last been accepted to mean nothing more than old-fashioned slew-rate limiting. To this end dominant-pole compensation was avoided in this design. The compensation scheme that was used was complex, but basically the lead capacitors C1, C2 and the lead-lag network R19, C3 were intended to cancel out the internal poles of the amplifier. According to Lohstroh and Otala, these lay between 200 kHz and 1 MHz, but after compensation the open-loop frequency response had its first pole at 1 MHz. A final lag compensation network R15, C4 was located outside the feedback loop. An important point is that the third stage was heavily loaded by the two resistors R11, R12. The emitter-follower (EF)-type output stage was biased far into Class-AB by a conventional V_{be}-multiplier, drawing 600 mA of quiescent current. As explained later in Chapter 6, this gives poor linearity when you run out of the Class-A region.

You will note that the amplifier uses shunt feedback; this certainly prevents any possibility of common-mode distortion in the input stage, as there is no common-mode voltage, but it does have the frightening drawback of going berserk if the source equipment is disconnected, as there is then a greatly increased feedback factor, and high-frequency instability is pretty much inevitable. Input common-mode nonlinearity is dealt with in Chapter 4, where it is shown that in normal amplifier designs it is of negligible proportions, and certainly not a good reason to adopt overall shunt feedback.

Many years ago I was asked to put a version of this amplifier circuit into production for one of the major hi-fi companies of the time. It was not a very happy experience. High-frequency stability was very doubtful and the distortion performance was distinctly unimpressive, being in line with that quoted in the original paper as 0.09% at 50 W, 1 kHz[3]. After a few weeks of struggle the

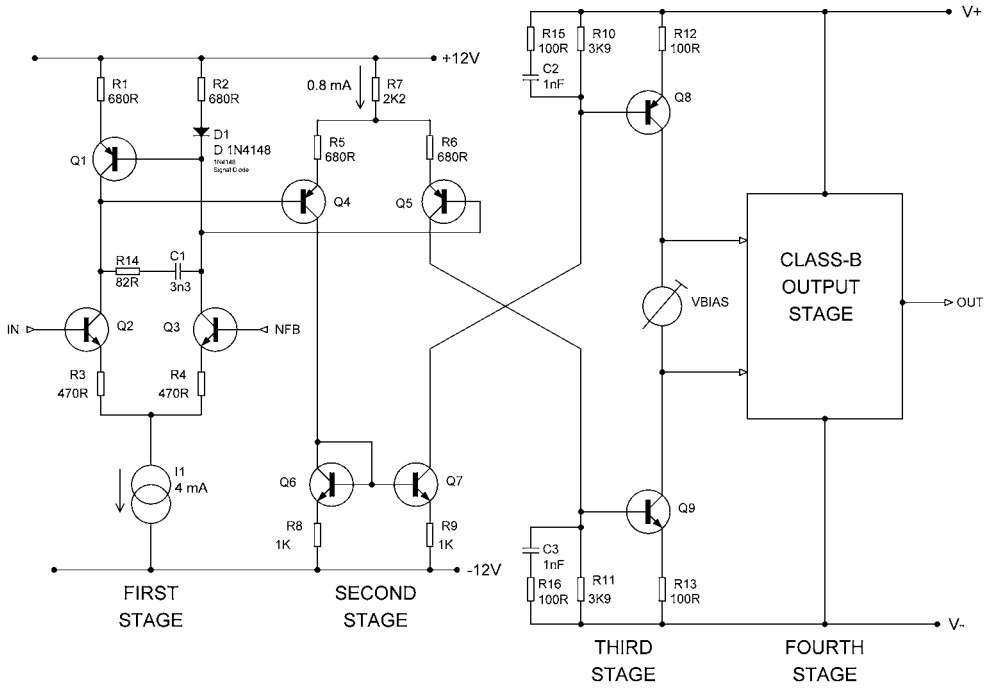


Figure 2.4: Four-stage amplifier architecture of a commercial amplifier

four-stage architecture was abandoned and a more conventional (and much more tractable) three-stage architecture was adopted instead.

Another version of the four-stage architecture is shown in Figure 2.4; it is a simplified version of a circuit used for many years by another of the major hi-fi companies. There are two differential stages, the second one driving a push–pull VAS Q8, Q9. Once again the differential stages have been given a large amount of local negative feedback in the form of emitter degeneration. Compensation is by the lead-lag network R14, C1 between the two input stage collectors and the two lead-lag networks R15, C2 and R16, C3 that shunt the collectors of Q5, Q7 in the second differential stage. Unlike the Lohstroh and Otala design, series overall feedback was used, supplemented with an op-amp DC servo to control the DC offset at the output.

Having had some experience with this design (no, it's not one of mine) I have to report that while in general the amplifier worked soundly and reliably, it was unduly fussy about transistor types and the distortion performance was not of the best.

The question now obtrudes itself: what is gained by using the greater complexity of a four-stage architecture? So far as I can see at the moment, little or nothing. The three-stage architecture appears to provide as much open-loop gain as can be safely used with a conventional output stage; if more is required then the Miller compensation capacitor can be reduced, which will also improve the maximum slew rates. A four-stage architecture does, however, present some interesting possibilities for using nested Miller compensation, a concept which has been extensively used in op-amps.

Power Amplifier Classes

For a long time the only amplifier classes relevant to high-quality audio were Class-A and Class-AB. This is because valves were the only active devices, and Class-B valve amplifiers generated so much distortion that they were barely acceptable even for public address purposes. All amplifiers with pretensions to high fidelity operated in push–pull Class-A.

Solid-state gives much more freedom of design; all of the amplifier classes below have been commercially exploited. This book deals in detail with Classes A, AB, B, D and G, and this certainly covers the vast majority of solid-state amplifiers. For the other classes plentiful references are given so that the intrigued can pursue matters further. In particular, my book *Self On Audio*[4] contains a thorough treatment of all known audio amplifier classes, and indeed suggests some new ones.

Class-A

In a Class-A amplifier current flows continuously in all the output devices, which enables the nonlinearities of turning them on and off to be avoided. They come in two rather different kinds, although this is rarely explicitly stated, which work in very different ways. The first kind is simply a Class-B stage (i.e. two emitter-followers working back to back) with the bias voltage increased so that sufficient current flows for neither device to cut off under normal loading. The great advantage of this approach is that it cannot abruptly run out of output current; if the load impedance becomes lower than specified then the amplifier simply takes brief excursions into Class-AB, hopefully with a modest increase in distortion and no seriously audible distress.

The other kind could be called the controlled-current-source (VCIS) type, which is in essence a single emitter-follower with an active emitter load for adequate current-sinking. If this latter element runs out of current capability it makes the output stage clip much as if it had run out of output voltage. This kind of output stage demands a very clear idea of how low an impedance it will be asked to drive before design begins.

Valve textbooks will be found to contain enigmatic references to classes of operation called AB1 and AB2; in the former grid current did not flow for any part of the cycle, but in the latter it did. This distinction was important because the flow of output-valve grid current in AB2 made the design of the previous stage much more difficult.

AB1 or AB2 has no relevance to semiconductors, for in BJTs base current always flows when a device is conducting, while in power FETs gate current never does, apart from charging and discharging internal capacitances.

Class-AB

This is not really a separate class of its own, but a combination of A and B. If an amplifier is biased into Class-B, and then the bias further increased, it will enter AB. For outputs below a certain level both output devices conduct, and operation is Class-A. At higher levels, one device will be turned completely off as the other provides more current, and the distortion jumps upward at this point as

AB action begins. Each device will conduct between 50% and 100% of the time, depending on the degree of excess bias and the output level.

Class-AB is less linear than either A or B, and in my view its only legitimate use is as a fallback mode to allow Class-A amplifiers to continue working reasonably when faced with a low-load impedance.

Class-B

Class-B is by far the most popular mode of operation, and probably more than 99% of the amplifiers currently made are of this type. Most of this book is devoted to it. My definition of Class-B is that unique amount of bias voltage which causes the conduction of the two output devices to overlap with the greatest smoothness and so generate the minimum possible amount of crossover distortion.

Class-C

Class-C implies device conduction for significantly less than 50% of the time, and is normally only usable in radio work, where an LC circuit can smooth out the current pulses and filter harmonics. Current-dumping amplifiers can be regarded as combining Class-A (the correcting amplifier) with Class-C (the current-dumping devices); however, it is hard to visualize how an audio amplifier using devices in Class-C only could be built. I regard a Class-B stage with no bias voltage as working in Class-C.

Class-D

These amplifiers continuously switch the output from one rail to the other at a supersonic frequency, controlling the mark/space ratio to give an average representing the instantaneous level of the audio signal; this is alternatively called pulse width modulation (PWM). Great effort and ingenuity has been devoted to this approach, for the efficiency is in theory very high, but the practical difficulties are severe, especially so in a world of tightening EMC legislation, where it is not at all clear that a 200 kHz high-power square wave is a good place to start. Distortion is not inherently low[5], and the amount of global negative feedback that can be applied is severely limited by the pole due to the effective sampling frequency in the forward path. A sharp cut-off low-pass filter is needed between amplifier and speaker, to remove most of the RF; this will require at least four inductors (for stereo) and will cost money, but its worst feature is that it will only give a flat frequency response into one specific load impedance.

Chapter 13 in this book is devoted to Class-D. Important references to consult for further information are Goldberg and Sandler[6] and Hancock[7].

Class-E

This is an extremely ingenious way of operating a transistor so that it has either a small voltage across it or a small current through it almost all the time, so that the power dissipation is kept very low[8]. Regrettably this is an RF technique that seems to have no sane application to audio.

Class-F

There is no Class-F, as far as I know. This seems like a gap that needs filling . . .

Class-G

This concept was introduced by Hitachi in 1976 with the aim of reducing amplifier power dissipation. Musical signals have a high peak/mean ratio, spending most of the time at low levels, so internal dissipation is much reduced by running from low-voltage rails for small outputs, switching to higher rails current for larger excursions[9,10].

The basic series Class-G with two rail voltages (i.e. four supply rails, as both voltages are ±) is shown in Figure 2.5. Current is drawn from the lower ±V1 supply rails whenever possible; should the signal exceed ±V1, TR6 conducts and D3 turns off, so the output current is now drawn entirely from the higher ±V2 rails, with power dissipation shared between TR3 and TR6. The inner stage TR3, TR4 is usually operated in Class-B, although AB or A are equally feasible if the output stage bias is suitably increased. The outer devices are effectively in Class-C as they conduct for significantly less than 50% of the time.

In principle movements of the collector voltage on the inner device collectors should not significantly affect the output voltage, but in practice Class-G is often considered to have poorer linearity than Class-B because of glitching due to charge storage in commutation diodes D3, D4.

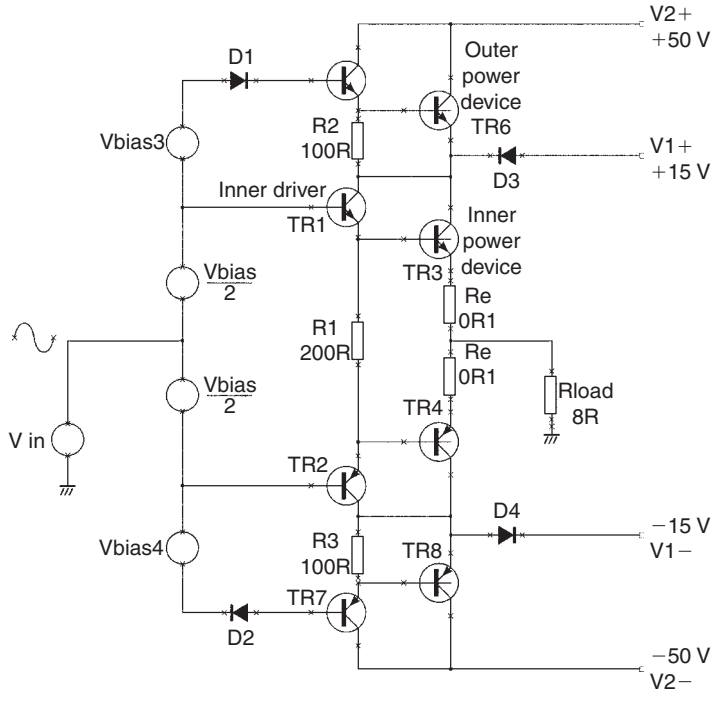


Figure 2.5: Class-G series output stage. When the output voltage exceeds the transition level, D3 or D4 turn off and power is drawn from the higher rails through the outer power devices

However, if glitches occur they do so at moderate power, well displaced from the crossover region, and so appear relatively infrequently with real signals.

An obvious extension of the Class-G principle is to increase the number of supply voltages. Typically the limit is three. Power dissipation is further reduced and efficiency increased as the average voltage from which the output current is drawn is kept closer to the minimum. The inner devices operate in Class-B/AB as before, and the middle devices are in Class-C. The outer devices are also in Class-C, but conduct for even less of the time.

To the best of my knowledge three-level Class-G amplifiers have only been made in Shunt mode, as described below, probably because in Series mode the cumulative voltage drops become too great and compromise the efficiency gains. The extra complexity is significant, as there are now six supply rails and at least six power devices, all of which must carry the full output current. It seems most unlikely that this further reduction in power consumption could ever be worthwhile for domestic hi-fi.

A closely related type of amplifier is Class-G Shunt[11]. Figure 2.6 shows the principle; at low outputs only Q3, Q4 conduct, delivering power from the low-voltage rails. Above a threshold set by Vbias3 and Vbias4, D1 or D2 conduct and Q6, Q8 turn on, drawing current from the high-voltage rails, with D3, D4 protecting Q3, Q4 against reverse bias. The conduction periods of the Q6, Q8 Class-C devices are variable, but inherently less than 50%. Normally the low-voltage section runs in Class-B to minimize dissipation. Such shunt Class-G arrangements are often called 'commutating amplifiers'.

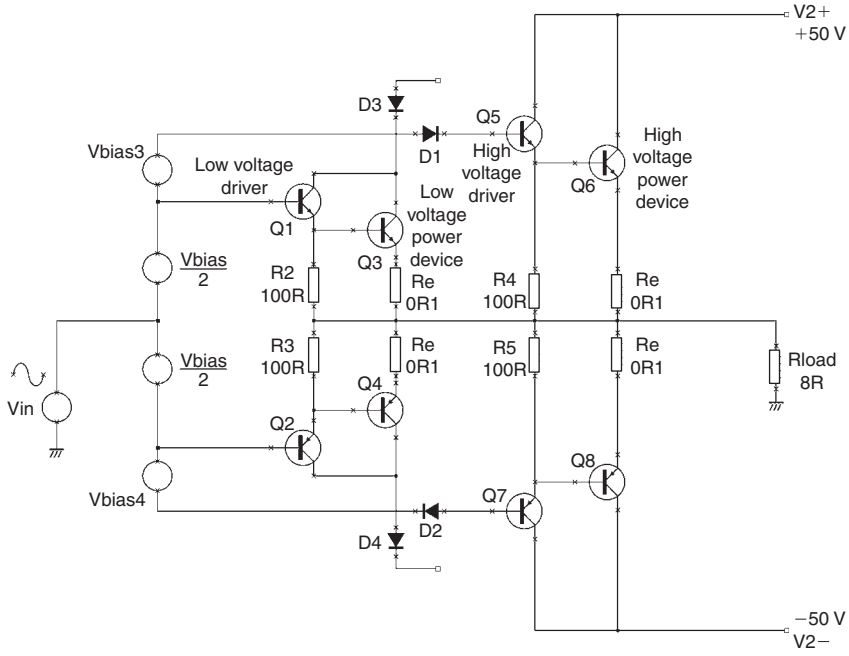


Figure 2.6: A Class-G Shunt output stage, composed of two EF output stages with the usual drivers. Vbias3, Vbias4 set the output level at which power is drawn from the higher rails

Some of the more powerful Class-G Shunt PA amplifiers have three sets of supply rails to further reduce the average voltage drop between rail and output. This is very useful in large PA amplifiers.

Chapter 12 in this book is devoted to Class-G.

Class-H

Class-H is once more basically Class-B, but with a method of dynamically boosting the single supply rail (as opposed to switching to another one) in order to increase efficiency[12]. The usual mechanism is a form of bootstrapping. Class-H is occasionally used to describe Class-G as above; this sort of confusion we can do without.

Class-S

Class-S, so named by Dr Sandman[13], uses a Class-A stage with very limited current capability, backed up by a Class-B stage connected so as to make the load appear as a higher resistance that is within the first amplifier's capability. The method used by the Technics SE-A100 amplifier is extremely similar[14]. I hope that this necessarily brief catalog is comprehensive; if anyone knows of other bona fide classes I would be glad to add them to the collection. This classification does not allow a completely consistent nomenclature; for example, Quad-style current-dumping can only be specified as a mixture of Classes A and C, which says nothing about the basic principle of operation, which is error correction.

Variations on Class-B

The solid-state Class-B three-stage amplifier has proved both successful and flexible, so many attempts have been made to improve it further, usually by trying to combine the efficiency of Class-B with the linearity of Class-A. It would be impossible to give a comprehensive list of the changes and improvements attempted, so I give only those that have been either commercially successful or particularly thought-provoking to the amplifier-design community.

Error-Correcting Amplifiers

This refers to error-cancelation strategies rather than the conventional use of negative feedback. This is a complex field, for there are at least three different forms of error correction, of which the best known is error feedforward as exemplified by the groundbreaking Quad 405[15]. Other versions include error feedback and other even more confusingly named techniques, some at least of which turn out on analysis to be conventional NFB in disguise. For a highly ingenious treatment of the feedforward method see a design by Giovanni Stochino[16]. A most interesting recent design using the Hawksford correction topology has recently been published by Jan Didden[17].

Non-Switching Amplifiers

Most of the distortion in Class-B is crossover distortion, and results from gain changes in the output stage as the power devices turn on and off. Several researchers have attempted to avoid this by ensuring that each device is clamped to pass a certain minimum current at all times[18]. This approach has certainly been exploited commercially, but few technical details have been published. It is not intuitively obvious (to me, anyway) that stopping the diminishing device current in its tracks will give less crossover distortion (see also Chapter 10).

Current-Drive Amplifiers

Almost all power amplifiers aspire to be voltage sources of zero output impedance. This minimizes frequency-response variations caused by the peaks and dips of the impedance curve, and gives a universal amplifier that can drive any loudspeaker directly.

The opposite approach is an amplifier with a sufficiently high output impedance to act as a constant-current source. This eliminates some problems – such as rising voice-coil resistance with heat dissipation – but introduces others such as control of the cone resonance. Current amplifiers therefore appear to be only of use with active crossovers and velocity feedback from the cone[19].

It is relatively simple to design an amplifier with any desired output impedance (even a negative one), and so any compromise between voltage and current drive is attainable. The snag is that loudspeakers are universally designed to be driven by voltage sources, and higher amplifier impedances demand tailoring to specific speaker types[20].

The Blomley Principle

The goal of preventing output transistors from turning off completely was introduced by Peter Blomley in 1971[21]; here the positive/negative splitting is done by circuitry ahead of the output stage, which can then be designed so that a minimum idling current can be separately set up in each output device. However, to the best of my knowledge this approach has not yet achieved commercial exploitation.

I have built Blomley amplifiers twice (way back in 1975) and on both occasions I found that there were still unwanted artefacts at the crossover point, and that transferring the crossover function from one part of the circuit to another did not seem to have achieved much. Possibly this was because the discontinuity was narrower than the usual crossover region and was therefore linearized even less effectively by negative feedback that reduces as frequency increases. I did not have the opportunity to investigate very deeply and this is not to be taken as a definitive judgment on the Blomley concept.

Geometric Mean Class-AB

The classical explanations of Class-B operation assume that there is a fairly sharp transfer of control of the output voltage between the two output devices, stemming from an equally abrupt switch in conduction from one to the other. In practical audio amplifier stages this is indeed the

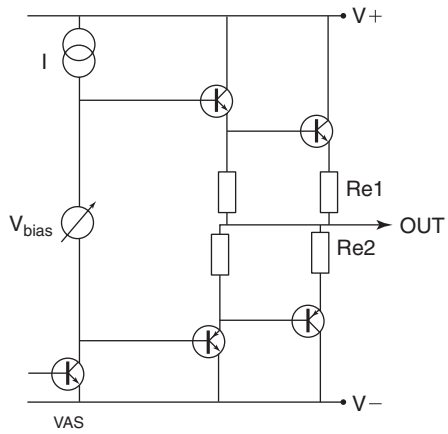


Figure 2.7: A conventional double emitter-follower output stage with emitter resistors Re shown

case, but it is not an inescapable result of the basic principle. Figure 2.7 shows a conventional output stage, with emitter resistors Re1, Re2 included to increase quiescent-current stability and allow current sensing for overload protection; it is these emitter resistances that to a large extent make classical Class-B what it is.

However, if the emitter resistors are omitted, and the stage biased with two matched diode junctions, then the diode and transistor junctions form a *translinear loop*[22], around which the junction voltages sum to zero. This links the two output transistor currents I_p, I_n in the relationship $I_n \cdot I_p$ = constant, which in op-amp practice is known as Geometric-Mean Class-AB operation. This gives smoother changes in device current at the crossover point, but this does not necessarily mean lower THD. Such techniques are not very practical for discrete power amplifiers; first, in the absence of the very tight thermal coupling between the four junctions that exists in an IC, the quiescent-current stability will be atrocious, with thermal runaway and spontaneous combustion a near certainty. Second, the output device bulk emitter resistance will probably give enough voltage drop to turn the other device off anyway, when current flows. The need for drivers, with their extra junction-drops, also complicates things.

A new extension of this technique is to redesign the translinear loop so that $1/I_n + 1/I_p$ = constant, this being known as Harmonic-Mean Class-AB operation[23]. It is too early to say whether this technique (assuming it can be made to work outside an IC) will be of use in reducing crossover distortion and thus improving amplifier performance.

Nested Differentiating Feedback Loops

This is a most ingenious but conceptually complex technique for significantly increasing the amount of NFB that can be applied to an amplifier. I wish I could tell you how well it works but I have never found the time to investigate it practically. For the original paper see Cherry[24], but it's tough going mathematically. A more readable account was published in *Electronics Today International* in 1983, and included a practical design for a 60 W NDFL amplifier[25].

Amplifier Bridging

When two power amplifiers are driven with anti-phase signals and the load connected between their outputs, with no connection to ground, this is called bridging. It is a convenient and inexpensive way to turn a stereo amplifier into a more powerful mono amplifier. It is called bridging because if you draw the four output transistors with the load connected between them, it looks something like the four arms of a Wheatstone bridge (see Figure 2.8). Doubling the voltage across a load of the same resistance naturally quadruples the output power – in theory. In harsh reality the available power will be considerably less, due to the power supply sagging and extra voltage losses in the two output stages. In most cases you will get something like three times the power rather than four, the ratio depending on how seriously the bridge mode was regarded when the initial design was done. It has to be said that in many designs the bridging mode looks like something of an afterthought.

In Figure 2.8 an 8 Ω load has been divided into two 4 Ω halves, to underline the point that the voltage at their center is zero, and so both amplifiers are effectively driving 4 Ω loads to ground, with all that that implies for increased distortion and increased losses in the output stages. A unity-gain inverting stage is required to generate the anti-phase signal; nothing fancy is required and the simple shunt-feedback stage shown does the job nicely. I have used it in several products. The resistors in the inverter circuit need to be kept as low in value as possible to reduce their Johnson noise contribution, but not of course so low that the op-amp distortion is increased by driving them; this is not too hard to arrange as the op-amp will only be working over a small fraction of its voltage output capability, because the power amplifier it is driving will clip a long time before the op-amp does. The capacitor assures stability – it causes a roll-off of 3 dB down at 5 MHz, so it does not in any way imbalance the audio frequency response of the two amplifiers.

You sometimes see the statement that bridging reduces the distortion seen across the load because the push–pull action causes cancelation of the distortion products. In brief, it is not true. Push–pull systems can only cancel even-order distortion products, and in a well-found amplifier these are in short supply. In such an amplifier the input stage and the output stage will both be symmetrical (it is hard to see why anyone would choose them to be anything else) and produce only odd-order

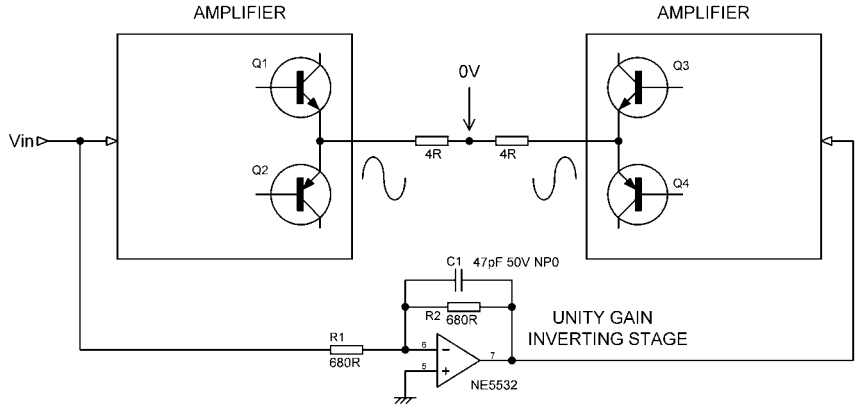


Figure 2.8: Bridging two power amplifiers to create a single, more powerful amplifier

harmonics, which will not be canceled. The only asymmetrical stage is the VAS, and the distortion contribution from that is, or at any rate should be, very low. In reality, switching to bridging mode will almost certainly increase distortion, because as noted above, the output stages are now in effect driving 4 Ω loads to ground instead of 8 Ω.

Fractional Bridging

I will now tell you how I came to invent the strange practice of 'fractional bridging'. I was tasked with designing a two-channel amplifier module for a multichannel unit. Five of these modules fitted into the chassis, and if each one was made independently bridgeable, you got a very flexible system that could be configured for anywhere between five and ten channels of amplification. The normal output of each amplifier was 85 W into 8 Ω, and the bridged output was about 270 W as opposed to the theoretical 340 W. And now the problem. The next unit up in the product line had modules that gave 250 W into 8 Ω unbridged, and the marketing department felt that having the small modules giving more power than the large ones was really not on; I'm not saying they were wrong. The problem was therefore to create an amplifier that only doubled its power when bridged. Hmm!

One way might have been to develop a power supply with deliberately poor regulation, but this implies a mains transformer with high-resistance windings that would probably have overheating problems. Another possibility was to make the bridged mode switch in a circuit that clipped the input signal before the power amplifiers clipped. The problem is that building a clipping circuit that does not exhibit poor distortion performance below the actual clipping level is actually surprisingly difficult – think about the nonlinear capacitance of signal diodes. I worked out a way to do it, but it took up an amount of PCB area that simply wasn't available. So the ultimate solution was to let one of the power amplifiers do the clipping, which it does cleanly because of the high level of negative feedback, and the fractional bridging concept was born.

Figure 2.9 shows how it works. An inverter is still used to drive the anti-phase amplifier, but now it is configured with a gain *G* that is less than unity. This means that the in-phase amplifier will

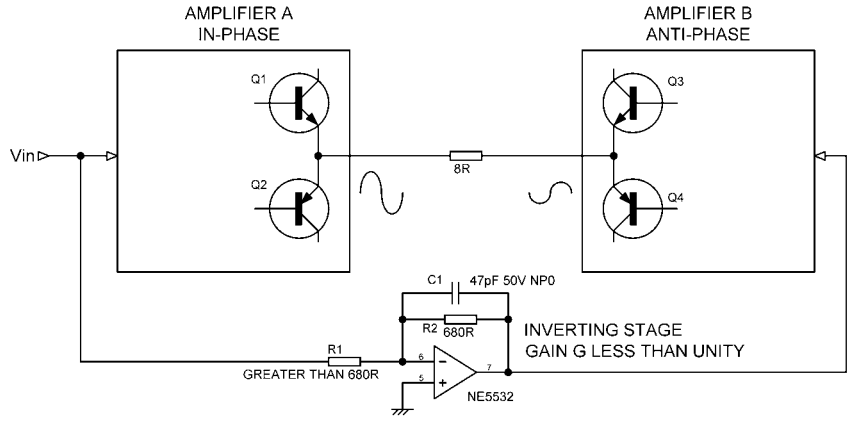


Figure 2.9: Fractional bridging of two power amplifiers to give doubled rather than quadrupled power output

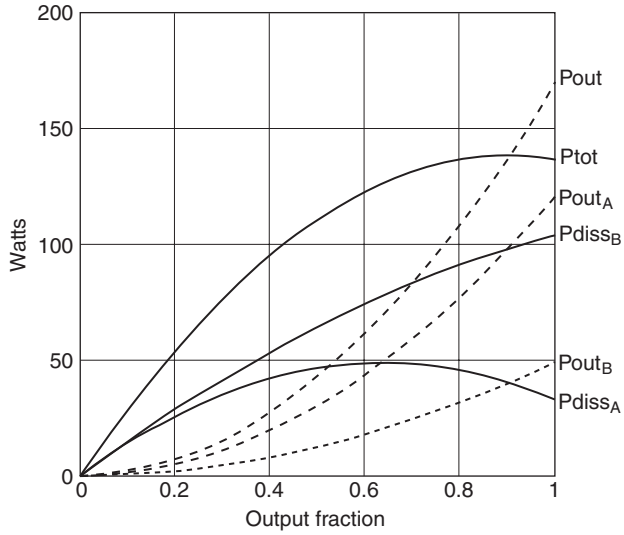


Figure 2.10: The variation of power output and power dissipation of two fractionally bridged power amplifiers, with a bridging fraction of 0.41 to give doubled rather than quadrupled power output

clip when the anti-phase amplifier is still well below maximum output, and the bridged output is therefore restricted. Double output power means an output voltage increased by root-2 or 1.41 times, and so the anti-phase amplifier is driven with a signal attenuated by a factor of 0.41, which I call the bridging fraction, giving a total voltage swing across the load of 1.41 times. It worked very well, the product was a considerable success, and no salesmen were plagued with awkward questions about power output ratings.

There are two possible objections to this cunning plan, the first being that it is obviously inefficient compared with a normal Class-B amplifier. Figure 2.10 shows how the power is dissipated in the pair of amplifiers; this is derived from basic calculations and ignores output stage losses. P_{dissA} is the power dissipated in the in-phase amplifier A, and varies in the usual way for a Class-B amplifier with a maximum at 63% of the maximum voltage output. P_{dissB} is the dissipation in anti-phase amplifier B that receives a smaller drive signal and so never reaches its dissipation maximum; it dissipates more power because it is handling the same current but has more voltage left across the output devices, and this is what makes the overall efficiency low. P_{tot} is the sum of the two amplifier dissipations. The dotted lines show the output power contribution from each amplifier, and the total output power in the load.

The bridging fraction can of course be set to other values to get other maximum outputs. The lower it is, the lower the overall efficiency of the amplifier pair, reaching the limiting value when the bridging fraction is zero. In this (quite pointless) situation the anti-phase amplifier is simply being used as an expensive alternative to connecting one end of the load to ground, and so it dissipates a lot of heat. Figure 2.11 shows how the maximum efficiency (which always occurs at maximum output) varies with the bridging fraction. When it is unity, we get normal Class-B operation and the maximum efficiency is the familiar figure of 78.6%; when it is zero the overall efficiency is halved to 39.3%, with a linear variation between these two extremes.

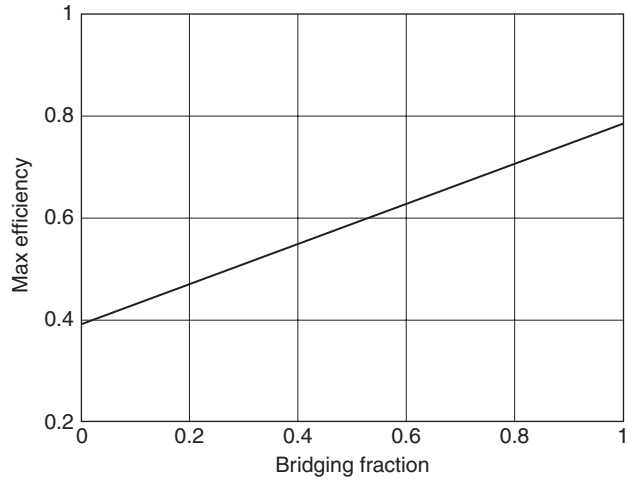


Figure 2.11: The variation of maximum efficiency of two fractionally bridged power amplifiers with bridging fraction

The second possible objection is that you might think it is a grievous offence against engineering ethics to deliberately restrict the output of an amplifier for marketing reasons, and you might be right, but it kept people employed, including me. Nevertheless, given the current concerns about energy, perhaps this sort of thing should not be encouraged. Chapter 9 gives another example of devious engineering, where I describe how an input clipping circuit (the one I thought up in an attempt to solve this problem, in fact) can be used to emulate the performance of a massive low-impedance power supply or a complicated regulated power supply. I have given semi-serious thought to writing a book called *How to Cheat with Amplifiers*.

AC- and DC-Coupled Amplifiers

All power amplifiers are either AC-coupled or DC-coupled. The first kind have a single supply rail, with the output biased to be halfway between this rail and ground to give the maximum symmetrical voltage swing; a large DC-blocking capacitor is therefore used in series with the output. The second kind have positive and negative supply rails, and the output is biased to be at 0 V, so no output DC-blocking is required in normal operation.

The Advantages of AC-Coupling

1. The output DC offset is always zero (unless the output capacitor is leaky).
2. It is very simple to prevent turn-on thump by purely electronic means; there is no need for an expensive output relay. The amplifier output must rise up to half the supply voltage at turn-on, but providing this occurs slowly there is no audible transient. Note that in many designs this is not simply a matter of making the input bias voltage rise slowly, as it also takes time for the DC feedback to establish itself, and it tends to do this with a snap action when a threshold is reached.

The last AC-coupled power amplifier I designed (which was in 1980, I think) had a simple RC time-constant and diode arrangement that absolutely constrained the VAS collector voltage to rise slowly at turn-on, no matter what the rest of the circuitry was doing – cheap but very effective.

3. No protection against DC faults is required, providing the output capacitor is voltage-rated to withstand the full supply rail. A DC-coupled amplifier requires an expensive and possibly unreliable output relay for dependable speaker protection.

4. The amplifier should be more easy to make short-circuit proof, as the output capacitor limits the amount of electric charge that can be transferred each cycle, no matter how low the load impedance. This is speculative; I have no data as to how much it really helps in practice.

5. AC-coupled amplifiers do not in general appear to require output inductors for stability. Large electrolytics have significant equivalent series resistance (ESR) and a little series inductance. For typical amplifier output sizes the ESR will be of the order of 100 mΩ; this resistance is probably the reason why AC-coupled amplifiers rarely had output inductors, as it is often enough resistance to provide isolation from capacitive loading and so gives stability. Capacitor series inductance is very low and probably irrelevant, being quoted by one manufacturer as 'a few tens of nanohenrys'. The output capacitor was often condemned in the past for reducing the low-frequency damping factor (DF), for its ESR alone is usually enough to limit the DF to 80 or so. As explained above, this is not a technical problem because 'damping factor' means virtually nothing.

The Advantages of DC-Coupling

1. No large and expensive DC-blocking capacitor is required. On the other hand, the dual supply will need at least one more equally expensive reservoir capacitor, and a few extra components such as fuses.

2. In principle there should be no turn-on thump, as the symmetrical supply rails mean the output voltage does not have to move through half the supply voltage to reach its bias point – it can just stay where it is. In practice the various filtering time-constants used to keep the bias voltages free from ripple are likely to make various sections of the amplifier turn on at different times, and the resulting thump can be substantial. This can be dealt with almost for free, when a protection relay is fitted, by delaying the relay pull-in until any transients are over. The delay required is usually less than a second.

3. Audio is a field where almost any technical eccentricity is permissible, so it is remarkable that AC-coupling appears to be the one technique that is widely regarded as unfashionable and unacceptable. DC-coupling avoids any marketing difficulties.

4. Some potential customers will be convinced that DC-coupled amplifiers give better speaker damping due to the absence of the output capacitor impedance. They will be wrong, as explained in Chapter 1, but this misconception has lasted at least 40 years and shows no sign of fading away.

5. Distortion generated by an output capacitor is avoided. This is a serious problem, as it is not confined to low frequencies, as is the case in small-signal circuitry (see page 212).

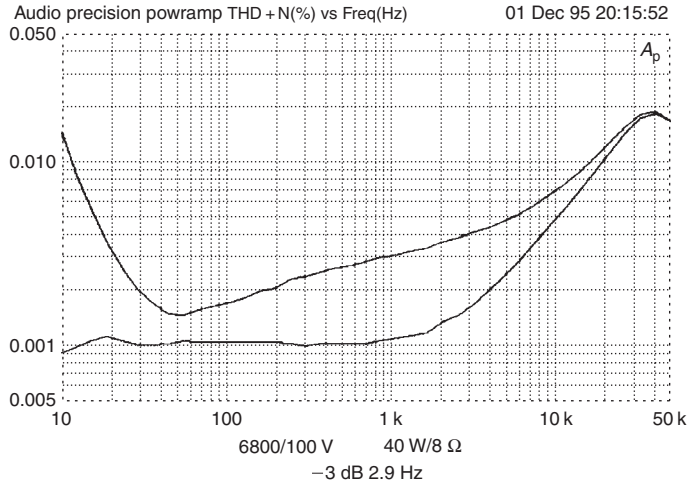


Figure 2.12: The extra distortion generated by a 6800 μF electrolytic delivering 40 W into 8 Ω. Distortion rises as frequency falls, as for the small-signal case, but at this current level there is also added distortion in the mid-band

For a 6800 μF output capacitor driving 40 W into an 8 Ω load, there is significant mid-band third harmonic distortion at 0.0025%, as shown in Figure 2.12. This is at least five times more than the amplifier generates in this part of the frequency range. In addition, the THD rise at the LF end is much steeper than in the small-signal case, for reasons that are not yet clear. There are two cures for output capacitor distortion. The straightforward approach uses a huge output capacitor, far larger in value than required for a good low-frequency response. A 100,000 μF/40 V Aerovox from BHC eliminated all distortion, as shown in Figure 2.13. An allegedly 'audiophile' capacitor gives some interesting results; a Cerafine Supercap of only moderate size (4700 μF/63 V) gave the result shown in Figure 2.14, where the mid-band distortion is gone but the LF distortion rise remains. What special audio properties this component is supposed to have are unknown; as far as I know electrolytics are never advertised as 'low mid-band THD', but that seems to be the case here. The volume of the capacitor case is about twice as great as conventional electrolytics of the same value, so it is possible the crucial difference may be a thicker dielectric film than is usual for this voltage rating.

Either of these special capacitors costs more than the rest of the amplifier electronics put together. Their physical size is large. A DC-coupled amplifier with protective output relay will be a more economical option.

A little-known complication with output capacitors is that their series reactance increases the power dissipation in the output stage at low frequencies. This is counter-intuitive as it would seem that any impedance added in series must reduce the current drawn and hence the power dissipation. In fact it is the load phase shift that increases the amplifier dissipation.

6. The supply currents can be kept out of the ground system. A single-rail AC amplifier has half-wave Class-B currents flowing in the 0 V rail, and these can have a serious effect on distortion and crosstalk performance.

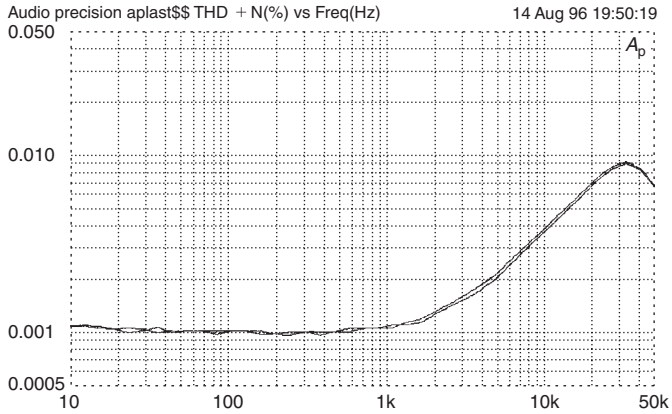


Figure 2.13: Distortion with and without a very large output capacitor, the BHC Aerovox 100,000 μF/40 V (40 W/8 W). Capacitor distortion is eliminated

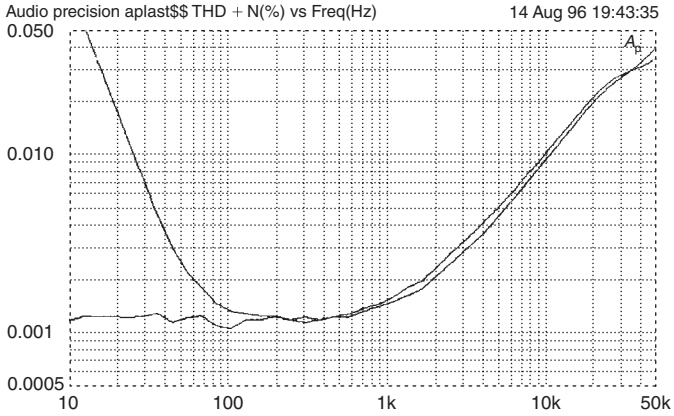


Figure 2.14: Distortion with and without an 'audiophile' Cerafine 4700 μF/63 V capacitor. Mid-band distortion is eliminated but LF rise is much the same as the standard electrolytic

Negative Feedback in Power Amplifiers

It is not the role of this book to step through elementary theory that can be easily found in any number of textbooks. However, correspondence in audio and technical journals shows that considerable confusion exists on negative feedback as applied to power amplifiers; perhaps there is something inherently mysterious in a process that improves almost all performance parameters simply by feeding part of the output back to the input, but inflicts dire instability problems if used to excess. I therefore deal with a few of the less obvious points here; more information is provided in Chapter 8.

The main use of NFB in power amplifiers is the reduction of harmonic distortion, the reduction of output impedance, and the enhancement of supply-rail rejection. There are also analogous improvements in frequency response and gain stability, and reductions in DC drift.

The basic feedback equation is dealt with in a myriad of textbooks, but it is so fundamental to power amplifier design that it is worth a look here. In Figure 2.15, the open-loop amplifier is

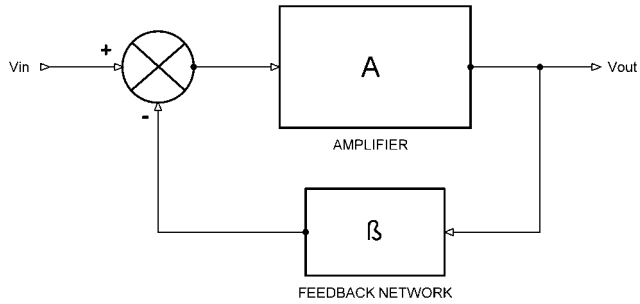


Figure 2.15: A simple negative-feedback system with an amplifier with open-loop gain *A* and a feedback network with a 'gain', which is less than 1, of β

the big block with open-loop gain *A*. The negative-feedback network is the block marked β; this could contain anything, but for our purposes it simply scales down its input, multiplying it by β, and is usually in the form of a potential divider. The funny round thing with the cross on is the conventional control theory symbol for a block that adds or subtracts and does nothing else.

Firstly, it is pretty clear that one input to the subtractor is simply V_{in}, and the other is $V_{out} \cdot \beta$, so subtract these two, multiply by *A*, and you get the output signal V_{out}:

$$V_{in} = A(V_{in} - \beta \cdot V_{out})$$

Collect the V_{out} values together and you get:

$$V_{out}(1 + A\beta) = A \cdot V_{in}$$

So:

$$\frac{V_{out}}{V_{in}} = \frac{A}{1 + A\beta} \qquad \text{Equation 2.1}$$

This is the feedback equation, and it could not be more important. The first thing it shows is that negative feedback stabilizes the gain. In real-life circuitry *A* is a high but uncertain and variable quantity, while β is firmly fixed by resistor values. Looking at the equation, you can see that the higher *A* is, the less significant the 1 on the bottom is; the *A* values cancel out, and so with high *A* the equation can be regarded as simply:

$$\frac{V_{out}}{V_{in}} = \frac{1}{\beta}$$

This is demonstrated in Table 2.1, where β is set at 0.04 with the intention of getting a closed-loop gain of 25 times. With a low open-loop gain of 100, the closed-loop gain is only 20, a long way short of 25. But as the open-loop gain increases, the closed-loop gain gets closer to the target. If you look at the bottom two rows, you will see that an increase in open-loop gain of more than a factor of 2 only alters the closed-loop gain by a trivial second decimal place.

Table 2.1: How the closed-loop gain gets closer to the target as the open-loop gain increases

1 Desired C/L gain	2 β NFB fraction	3 *A* O/L gain	4 NFB factor	5 C/L gain	6 O/L error	7 C/L error
25	0.04	100	5	20.00	1	0.2
25	0.04	1000	41	24.39	1	0.0244
25	0.04	10,000	401	24.94	1	0.0025
25	0.04	40,000	1601	24.98	1	0.0006
25	0.04	100,000	4001	24.99	1	0.0002

Negative feedback is, however, capable of doing much more than stabilizing gain. Anything untoward happening in the amplifier block *A*, be it distortion or DC drift, or any of the other ills that electronics is prone to, is also reduced by the negative-feedback factor (NFB factor for short). This is equal to:

$$\text{NFB factor} = \frac{1}{1 + A\beta} \qquad \text{Equation 2.2}$$

and it is tabulated in the fourth column in Table 2.1. To show why this factor is vitally important, Figure 2.16 shows the same scenario as Figure 2.11, with the addition of a voltage V_d to the output of *A*; this represents noise, DC drift, or anything that can cause a voltage error, but what is usually most interesting to the practitioners of amplifier design is its use to represent distortion.

Repeating the simple algebra we did before, and adding in V_d, we get:

$$V_{out} = A(V_{in} - \beta \cdot V_{out}) + V_d$$

$$V_{out}(1 + A\beta) = A \cdot V_{in} + V_d$$

$$\frac{V_{out}}{V_{in}} = \frac{A}{1 + A\beta} + \frac{V_d}{1 + A\beta}$$

So the effect of V_d has been decreased by the feedback factor:

$$\frac{1}{1 + A\beta}$$

In other words, the higher the open-loop gain *A* compared with the gain demanded by β, the lower the distortion. Since we are usually dealing with high values of *A*, the 1 on the bottom of the fraction has very little effect and doubling the open-loop gain halves the distortion. This effect is illustrated in the fifth and sixth columns of Table 2.1, which adds an error of magnitude 1 to the output of the amplifier; the closed-loop error is then simply the reciprocal of the NFB factor for each value of open-loop gain.

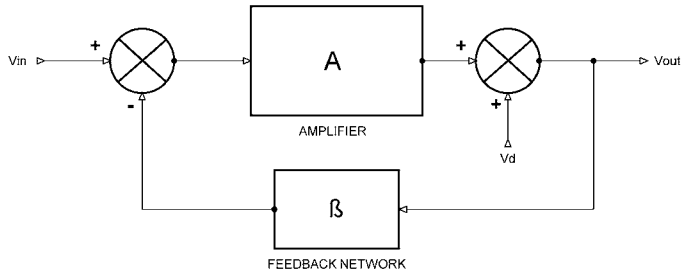


Figure 2.16 The negative-feedback system with an error signal V_d added to the output of the amplifier

In simple circuits with low open-loop gain you just apply negative feedback and that is the end of the matter. In a typical power amplifier, which cannot be operated without NFB, if only because it would be saturated by its own DC offset voltages, there are several stages that may accumulate phase shift, and simply closing the loop usually brings on severe Nyquist oscillation at HF. This is a serious matter, as it will not only burn out any tweeters that are unlucky enough to be connected, but can also destroy the output devices by overheating, as they may be unable to turn off fast enough at ultrasonic frequencies.

The standard cure for this instability is compensation. A capacitor is added, usually in Miller-integrator format, to roll off the open-loop gain at 6 dB/octave, so it reaches unity loop-gain before enough phase shift can build up to allow oscillation. This means the NFB factor varies strongly with frequency, an inconvenient fact that many audio commentators seem to forget.

It is crucial to remember that a distortion harmonic, subjected to a frequency-dependent NFB factor as above, will be reduced by the NFB factor corresponding to its own frequency, not that of its fundamental. If you have a choice, generate low-order rather than high-order distortion harmonics, as the NFB deals with them much more effectively.

Negative feedback can be applied either locally (i.e. to each stage, or each active device) or globally, in other words right around the whole amplifier. Global NFB is more efficient at distortion reduction than the same amount distributed as local NFB, but places much stricter limits on the amount of phase shift that may be allowed to accumulate in the forward path (more on this later in this chapter).

Above the dominant-pole frequency, the VAS acts as a Miller integrator, and introduces a constant 90° phase lag into the forward path. In other words, the output from the input stage must be in quadrature if the final amplifier output is to be in phase with the input, which to a close approximation it is. This raises the question of how the 90° phase shift is accommodated by the negative-feedback loop; the answer is that the input and feedback signals applied to the input stage are there subtracted, and the small difference between two relatively large signals with a small phase shift between them has a much larger phase shift. This is the signal that drives the VAS input of the amplifier.

Solid-state power amplifiers, unlike many valve designs, are almost invariably designed to work at a fixed closed-loop gain. If the circuit is compensated by the usual dominant-pole method, the HF open-loop gain is also fixed, and therefore so is the important negative-feedback factor. This is in contrast to valve amplifiers, where the amount of negative feedback applied was regarded

as a variable, and often user-selectable, parameter; it was presumably accepted that varying the negative-feedback factor caused significant changes in input sensitivity. A further complication was serious peaking of the closed-loop frequency response at both LF and HF ends of the spectrum as negative feedback was increased, due to the inevitable bandwidth limitations in a transformer-coupled forward path. Solid-state amplifier designers go cold at the thought of the customer tampering with something as vital as the NFB factor, and such an approach is only acceptable in cases like valve amplification where global NFB plays a minor role.

Some Common Misconceptions about Negative Feedback

All of the comments quoted below have appeared many times in the hi-fi literature. All are wrong.

Negative feedback is a bad thing. Some audio commentators hold that, without qualification, negative feedback is a bad thing. This is of course completely untrue and based on no objective reality. Negative feedback is one of the fundamental concepts of electronics, and to avoid its use altogether is virtually impossible; apart from anything else, a small amount of local NFB exists in every common-emitter transistor because of the internal emitter resistance. I detect here distrust of good fortune; the uneasy feeling that if something apparently works brilliantly then there must be something wrong with it.

A low negative-feedback factor is desirable. Untrue – global NFB makes just about everything better, and the sole effect of too much is HF oscillation, or poor transient behavior on the brink of instability. These effects are painfully obvious on testing and not hard to avoid unless there is something badly wrong with the basic design.

In any case, just what does *low* mean? One indicator of imperfect knowledge of negative feedback is that the amount enjoyed by an amplifier is almost always badly specified as *so many decibels* on the very few occasions it is specified at all – despite the fact that most amplifiers have a feedback factor that varies considerably with frequency. A decibel figure quoted alone is meaningless, as it cannot be assumed that this is the figure at 1 kHz or any other standard frequency.

My practice is to quote the NFB factor at 20 kHz, as this can normally be assumed to be above the dominant pole frequency, and so in the region where open-loop gain is set by only two or three components. Normally the open-loop gain is falling at a constant 6 dB/octave at this frequency on its way down to intersect the unity-loop-gain line and so its magnitude allows some judgment as to Nyquist stability. Open-loop gain at LF depends on many more variables such as transistor beta, and consequently has wide tolerances and is a much less useful quantity to know. This is dealt with in more detail in the chapter on voltage-amplifier stages.

Negative feedback is a powerful technique, and therefore dangerous when misused. This bland truism usually implies an audio Rake's Progress that goes something like this: an amplifier has too much distortion, and so the open-loop gain is increased to augment the NFB factor. This causes HF instability, which has to be cured by increasing the compensation capacitance. This is turn reduces the slew-rate capability, and results in a sluggish, indolent, and generally bad amplifier.

The obvious flaw in this argument is that the amplifier so condemned no longer has a high NFB factor, because the increased compensation capacitor has reduced the open-loop gain at HF;

therefore feedback itself can hardly be blamed. The real problem in this situation is probably unduly low standing current in the input stage; this is the other parameter determining slew rate.

NFB may reduce low-order harmonics but increases the energy in the discordant higher harmonics. A less common but recurring complaint is that the application of global NFB is a shady business because it transfers energy from low-order distortion harmonics – considered musically consonant – to higher-order ones that are anything but. This objection contains a grain of truth, but appears to be based on a misunderstanding of one article in an important series by Peter Baxandall[26] in which he showed that if you took an amplifier with only second-harmonic distortion, and then introduced NFB around it, higher-order harmonics were indeed generated as the second harmonic was fed back round the loop. For example, the fundamental and the second harmonic intermodulate to give a component at third-harmonic frequency. Likewise, the second and third intermodulate to give the fifth harmonic. If we accept that high-order harmonics should be numerically weighted to reflect their greater unpleasantness, there could conceivably be a rise rather than a fall in the weighted THD when negative feedback is applied.

All active devices, in Class A or B (including FETs, which are often erroneously thought to be purely square law), generate small amounts of high-order harmonics. Feedback could and would generate these from nothing, but in practice they are already there.

The vital point is that if enough NFB is applied, all the harmonics can be reduced to a lower level than without it. The extra harmonics generated, effectively by the distortion of a distortion, are at an extremely low level providing a reasonable NFB factor is used. This is a powerful argument against low feedback factors like 6 dB, which are most likely to increase the weighted THD. For a full understanding of this topic, a careful reading of the Baxandall series is absolutely indispensable.

A low open-loop bandwidth means a sluggish amplifier with a low slew rate. Great confusion exists in some quarters between open-loop bandwidth and slew rate. In truth open-loop bandwidth and slew rate are nothing to do with each other, and may be altered independently. Open-loop bandwidth is determined by compensation C_{dom}, VAS β, and the resistance at the VAS collector, while slew rate is set by the input stage standing current and C_{dom}. C_{dom} affects both, but all the other parameters are independent (see Chapter 3 for more details).

In an amplifier, there is a maximum amount of NFB you can safely apply at 20 kHz; this does not mean that you are restricted to applying the same amount at 1 kHz, or indeed 10 Hz. The obvious thing to do is to allow the NFB to continue increasing at 6 dB/octave – or faster if possible – as frequency falls, so that the amount of NFB applied doubles with each octave as we move down in frequency, and we derive as much benefit as we can. This obviously cannot continue indefinitely, for eventually open-loop gain runs out, being limited by transistor beta and other factors. Hence the NFB factor levels out at a relatively low and ill-defined frequency; this frequency is the open-loop bandwidth, and for an amplifier that can never be used open-loop, has very little importance.

It is difficult to convince people that this frequency is of no relevance whatever to the speed of amplifiers, and that it does not affect the slew rate. Nonetheless, it is so, and any first-year electronics textbook will confirm this. High-gain op-amps with sub-1 Hz bandwidths and

blindingly fast slewing are as common as the grass (if somewhat less cheap) and if that does not demonstrate the point beyond doubt then I really do not know what will.

Limited open-loop bandwidth prevents the feedback signal from immediately following the system input, so the utility of this delayed feedback is limited. No linear circuit can introduce a pure time delay; the output must begin to respond at once, even if it takes a long time to complete its response. In the typical amplifier the dominant-pole capacitor introduces a 90° phase shift between input pair and output at all but the lowest audio frequencies, but this is not a true time delay. The phrase delayed feedback is often used to describe this situation, and it is a wretchedly inaccurate term; if you really delay the feedback to a power amplifier (which can only be done by adding a time-constant to the feedback network rather than the forward path) it will quickly turn into the proverbial power oscillator as sure as night follows day.

Amplifier Stability and NFB

In controlling amplifier distortion, there are two main weapons. The first is to make the linearity of the circuitry as good as possible before closing the feedback loop. This is unquestionably important, but it could be argued it can only be taken so far before the complexity of the various amplifier stages involved becomes awkward. The second is to apply as much negative feedback as possible while maintaining amplifier stability. It is well known that an amplifier with a single time-constant is always stable, no matter how high the feedback factor. The linearization of the VAS by local Miller feedback is a good example. However, more complex circuitry, such as the generic three-stage power amplifier, has more than one time-constant, and these extra poles will cause poor transient response or instability if a high feedback factor is maintained up to the higher frequencies where they start to take effect. It is therefore clear that if these higher poles can be eliminated or moved upward in frequency, more feedback can be applied and distortion will be less for the same stability margins. Before they can be altered – if indeed this is practical at all – they must be found and their impact assessed.

The dominant-pole frequency of an amplifier is, in principle, easy to calculate; the mathematics is very simple (see Chapter 3). In practice, two of the most important factors, the effective beta of the VAS and the VAS collector impedance, are only known approximately, so the dominant pole frequency is a rather uncertain thing. Fortunately this parameter in itself has no effect on amplifier stability. What matters is the amount of feedback at high frequencies.

Things are different with the higher poles. To begin with, where are they? They are caused by internal transistor capacitances and so on, so there is no physical component to show where the roll-off is. It is generally regarded as fact that the next poles occur in the output stage, which will use power devices that are slow compared with small-signal transistors. Taking the Class-B design in Chapter 7, the TO92 MPSA06 devices have an F_t of 100 MHz, the MJE340 drivers about 15 MHz (for some reason this parameter is missing from the data sheet) and the MJ802 output devices an F_t of 2.0 MHz. Clearly the output stage is the prime suspect. The next question is at what frequencies these poles exist. There is no reason to suspect that each transistor can be modeled by one simple pole.

There is a huge body of knowledge devoted to the art of keeping feedback loops stable while optimizing their accuracy; this is called Control Theory, and any technical bookshop will yield some intimidatingly fat volumes called things like 'Control System Design'. Inside, system stability is tackled by Laplace-domain analysis, eigenmatrix methods, and joys like the Lyapunov stability criterion. I think that makes it clear that you need to be pretty good at mathematics to appreciate this kind of approach.

Even so, it is puzzling that there seems to have been so little application of Control Theory to audio amplifier design. The reason may be that so much Control Theory assumes that you know fairly accurately the characteristics of what you are trying to control, especially in terms of poles and zeros.

One approach to appreciating negative feedback and its stability problems is SPICE simulation. Some SPICE simulators have the ability to work in the Laplace or s-domain, but my own experiences with this have been deeply unhappy. Otherwise respectable simulator packages output complete rubbish in this mode. Quite what the issues are here I do not know, but it does seem that s-domain methods are best avoided. The approach suggested here instead models poles directly as poles, using RC networks to generate the time-constants. This requires minimal mathematics and is far more robust. Almost any SPICE simulator – evaluation versions included – should be able to handle the simple circuit used here.

Figure 2.17 shows the basic model, with SPICE node numbers. The scheme is to idealize the situation enough to highlight the basic issues and exclude distractions like nonlinearities or clipping. The forward gain is simply the transconductance of the input stage multiplied by the transadmittance of the VAS integrator. An important point is that with correct parameter values, the current from the input stage is realistic, and so are all the voltages.

The input differential amplifier is represented by G. This is a standard SPICE element – the VCIS, or voltage-controlled current source. It is inherently differential, as the output current from Node 4 is the scaled difference between the voltages at Nodes 3 and 7. The scaling factor of 0.009 sets the input stage transconductance (g_m) to 9 mA/V, a typical figure for a bipolar input with some local feedback. Stability in an amplifier depends on the amount of negative feedback available at 20 kHz.

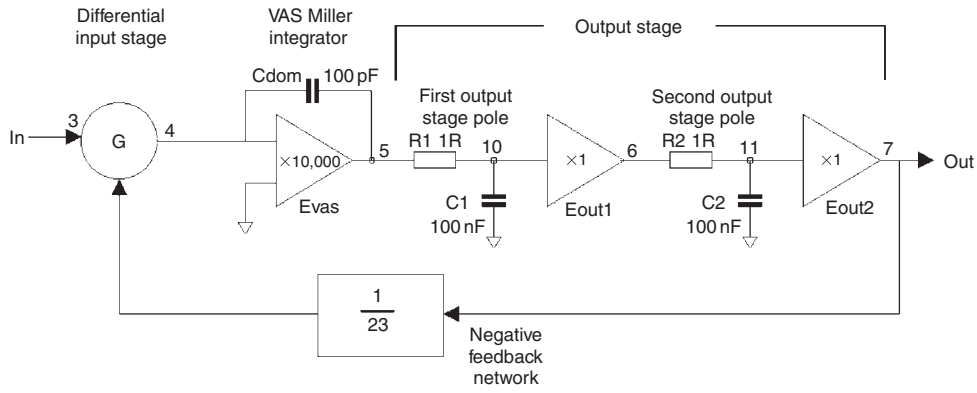


Figure 2.17: Block diagram of system for SPICE stability testing

This is set at the design stage by choosing the input g_m and C_{dom}, which are the only two factors affecting the open-loop gain. In simulation it would be equally valid to change g_m instead; however, in real life it is easier to alter C_{dom} as the only other parameter this affects is slew rate. Changing input stage transconductance is likely to mean altering the standing current and the amount of local feedback, which will in turn impact input stage linearity.

The VAS with its dominant pole is modeled by the integrator E_{vas}, which is given a high but finite open-loop gain, so there really is a dominant pole $P1$ created when the gain demanded becomes equal to that available. With $C_{dom} = 100\,pF$ this is below 1 Hz. With infinite (or as near infinite as SPICE allows) open-loop gain the stage would be a perfect integrator. A explained elsewhere, the amount of open-loop gain available in real versions of this stage is not a well-controlled quantity, and $P1$ is liable to wander about in the 1–100 Hz region; fortunately this has no effect at all on HF stability. C_{dom} is the Miller capacitor that defines the transadmittance, and since the input stage has a realistic transconductance C_{dom} can be set to 100 pF, its usual real-life value. Even with this simple model we have a nested feedback loop. This apparent complication here has little effect, so long as the open-loop gain of the VAS is kept high.

The output stage is modeled as a unity-gain buffer, to which we add extra poles modeled by R1, C1 and R2, C2. Eout1 is a unity-gain buffer internal to the output stage model, added so the second pole does not load the first. The second buffer Eout2 is not strictly necessary as no real loads are being driven, but it is convenient if extra complications are introduced later. Both are shown here as a part of the output stage but the first pole could equally well be due to input stage limitations instead; the order in which the poles are connected makes no difference to the final output. Strictly speaking, it would be more accurate to give the output stage a gain of 0.95, but this is so small a factor that it can be ignored.

The component values used to make the poles are of course completely unrealistic, and chosen purely to make the maths simple. It is easy to remember that 1 Ω and 1 μF make up a 1 μs time-constant. This is a pole at 159 kHz. Remember that the voltages in the latter half of the circuit are realistic, but the currents most certainly are not.

The feedback network is represented simply by scaling the output as it is fed back to the input stage. The closed-loop gain is set to 23 times, which is representative of many power amplifiers.

Note that this is strictly a linear model, so the slew-rate limiting that is associated with Miller compensation is not modeled here. It would be done by placing limits on the amount of current that can flow in and out of the input stage.

Figure 2.18 shows the response to a 1 V step input, with the dominant pole the only time element in the circuit. (The other poles are disabled by making C1, C2 0.00001 pF, because this is quicker than changing the actual circuit.) The output is an exponential rise to an asymptote of 23 V, which is exactly what elementary theory predicts. The exponential shape comes from the way that the error signal that drives the integrator becomes less as the output approaches the desired level. The error, in the shape of the output current from G, is the smaller signal shown; it has been multiplied by 1000 to get mA onto the same scale as volts. The speed of response is inversely proportional to

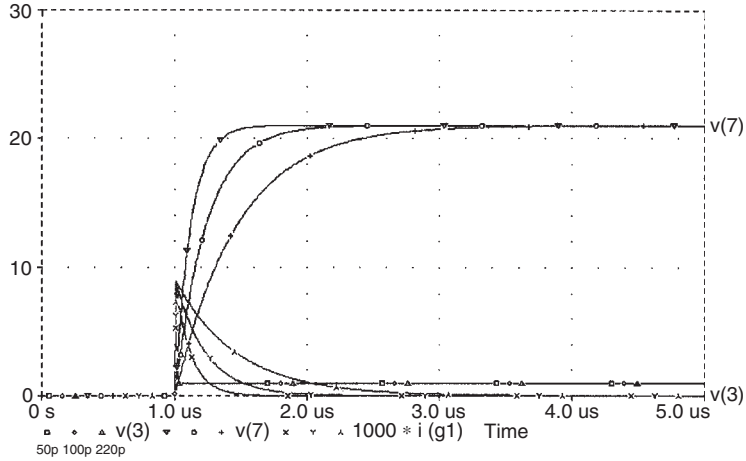


Figure 2.18: SPICE results in the time domain. As C_{dom} increases, the response V(7) becomes slower, and the error i(g1) declines more slowly. The input is the step-function V(3) at the bottom

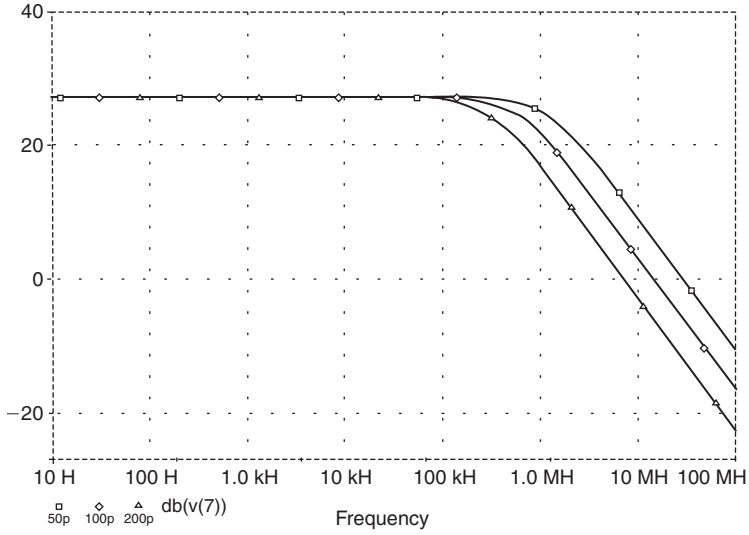


Figure 2.19: SPICE simulation in the frequency domain. As the compensation capacitor is increased, the closed-loop bandwidth decreases proportionally

the size of C_{dom}, and is shown here for values of 50 and 220 pF as well as the standard 100 pF. This simulation technique works well in the frequency domain, as well as the time domain. Simply tell SPICE to run an AC simulation instead of a TRANS (transient) simulation. The frequency response in Figure 2.19 exploits this to show how the closed-loop gain in an NFB amplifier depends on the open-loop gain available. Once more elementary feedback theory is brought to life. The value of C_{dom} controls the bandwidth, and it can be seen that the values used in the simulation do not give a very extended response compared with a 20 kHz audio bandwidth.

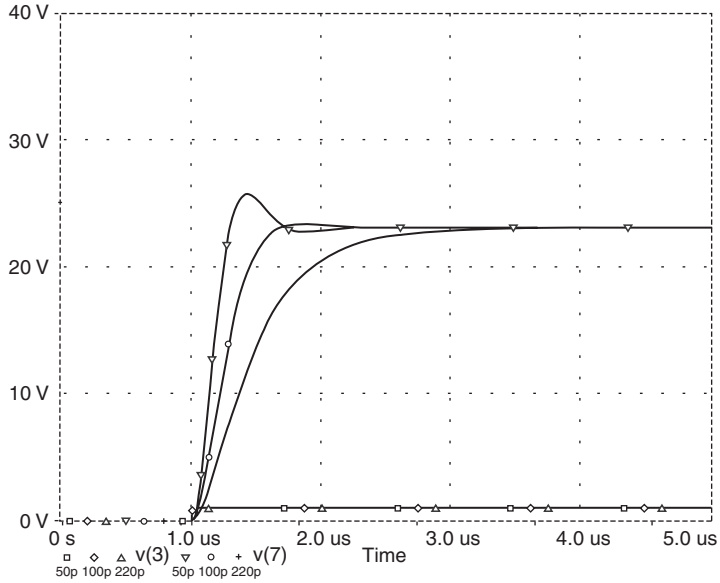


Figure 2.20: Adding a second pole *P*2 causes overshoot with smaller values C_{dom}, but cannot bring about sustained oscillation

In Figure 2.20, one extra pole *P*2 at 1.59 MHz (a time-constant of only 100 ns) is added to the output stage, and C_{dom} stepped through 50, 100 and 200 pF as before: 100 pF shows a slight overshoot that was not there before; with 50 pF there is a serious overshoot that does not bode well for the frequency response. Actually, it's not that bad; Figure 2.21 returns to the frequency-response domain to show that an apparently vicious overshoot is actually associated with a very mild peaking in the frequency domain.

From here on C_{dom} is left set to 100 pF, its real value in most cases. In Figure 2.22 *P*2 is stepped instead, increasing from 100 ns to 5 μs, and while the response gets slower and shows more overshoot, the system does not become unstable. The reason is simple: sustained oscillation (as opposed to transient ringing) in a feedback loop requires positive feedback, which means that a total phase shift of 180° must have accumulated in the forward path, and reversed the phase of the feedback connection. With only two poles in a system the phase shift cannot reach 180°. The VAS integrator gives a dependable 90° phase shift above *P*1, being an integrator, but *P*2 is instead a simple lag and can only give 90° phase lag at infinite frequency. So, even this very simple model gives some insight. Real amplifiers do oscillate if C_{dom} is too small, so we know that the frequency response of the output stage cannot be meaningfully modeled with one simple lag.

As President Nixon is alleged to have said: 'Two wrongs don't make a right – so let's see if three will do it!' Adding in a third pole *P*3 in the shape of another simple lag gives the possibility of sustained oscillation. This is case A in Table 2.2.

Stepping the value of *P*2 from 0.1 to 5 μs with $P3 = 500$ ns in Figure 2.23 shows that damped oscillation is present from the start. Figure 2.23 also shows over 50 μs what happens when the amplifier is made very unstable (there are degrees of this) by setting $P2 = 5$ μs and $P3 = 500$ ns.

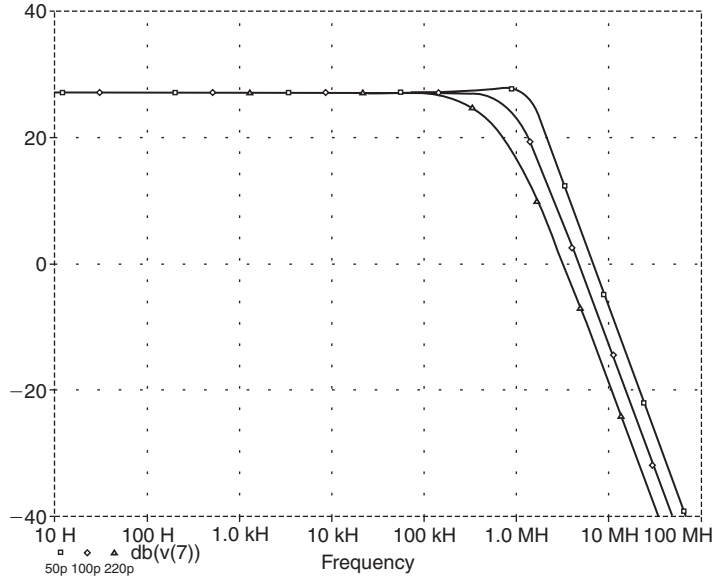


Figure 2.21: The frequency responses that go with the transient plots of Figure 2.20. The response peaking for C_{dom} = 50 pF is very small compared with the transient overshoot

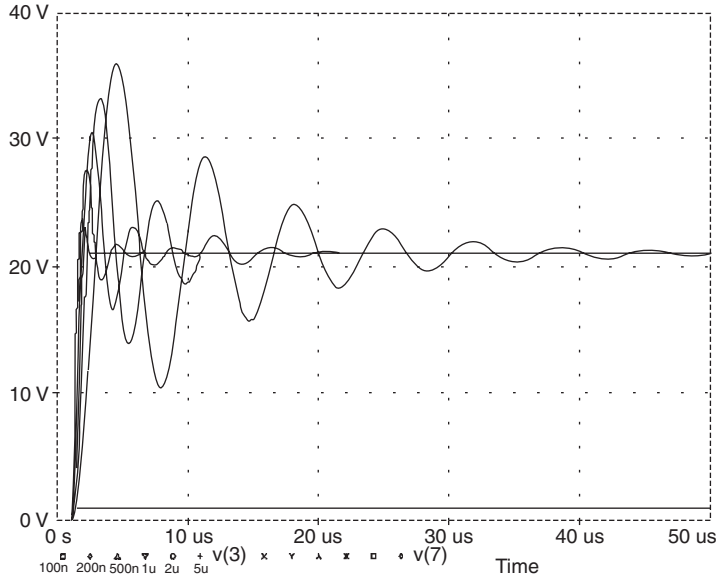


Figure 2.22: Manipulating the *P*2 frequency can make ringing more prolonged but it is still not possible to provoke sustained oscillation

It still takes time for the oscillation to develop, but exponentially diverging oscillation like this is a sure sign of disaster. Even in the short time examined here the amplitude has exceeded a rather theoretical half a kilovolt. In reality oscillation cannot increase indefinitely, if only because the supply rail voltages would limit the amplitude. In practice slew-rate limiting is probably the major controlling factor in the amplitude of high-frequency oscillation.

Table 2.2: Instability onset: *P*2 is increased until sustained oscillation occurs

Case	C_{dom}	*P*2	*P*3	*P*4	*P*5	*P*6	
A	100p	0.45	0.5	–	–		200 kHz
B	100p	0.5	0.2	0.2	–		345 kHz
C	100p	0.2	0.2	0.2	0.01		500 kHz
D	100p	0.3	0.2	0.1	0.05		400 kHz
E	100p	0.4	0.2	0.1	0.01		370 kHz
F	100p	0.2	0.2	0.1	0.05	0.02	475 kHz

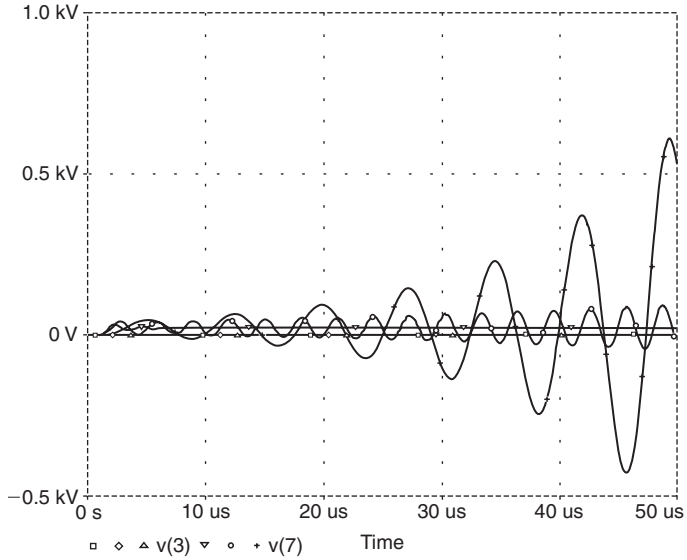


Figure 2.23: Adding a third pole makes possible true instability with exponentially increasing amplitude of oscillation. Note the unrealistic voltage scale on this plot

We have now modeled a system that will show instability. But does it do it right? Sadly, no. The oscillation is about 200 kHz, which is a rather lower frequency than is usually seen when an amplifier misbehaves. This low frequency stems from the low *P*2 frequency we have to use to provoke oscillation; apart from anything else this seems out of line with the known fT of power transistors. Practical amplifiers are likely to take off at around 500 kHz to 1 MHz when C_{dom} is reduced, and this seems to suggest that phase shift is accumulating quickly at this sort of frequency. One possible explanation is that there are a large number of poles close together at a relatively high frequency.

A fourth pole can be simply added to Figure 2.17 by inserting another RC-buffer combination into the system. With $P2 = 0.5\,\mu s$ and $P3 = P4 = 0.2\,\mu s$, instability occurs at 345 kHz, which is a step towards a realistic frequency of oscillation. This is case B in Table 2.2.

When a fifth output stage pole is grafted on, so that $P3 = P4 = P5 = 0.2\,\mu s$ the system just oscillates at 500 kHz with *P*2 set to 0.01 μs. This takes us close to a realistic frequency of oscillation. Rearranging the order of poles so $P2 = P3 = P4 = 0.2\,\mu s$, while $P5 = 0.01\,\mu s$, is tidier, and the stability results are of course the same; this is a linear system so the order does not matter. This is case C in Table 2.2.

Having *P*2, *P*3, and *P*4 all at the same frequency does not seem very plausible in physical terms, so case D shows what happens when the five poles are staggered in frequency. *P*2 needs to be increased to 0.3 μs to start the oscillation, which is now at 400 kHz. Case E is another version with five poles, showing that if *P*5 is reduced *P*2 needs to be doubled to 0.4 μs for instability to begin.

In the final case F, a sixth pole is added to see if this permitted sustained oscillation is above 500 kHz. This seems not to be the case; the highest frequency that could be obtained after a lot of pole twiddling was 475 kHz. This makes it clear that this model is of limited accuracy (as indeed are all models – it is a matter of degree) at high frequencies, and that further refinement is required to gain further insight.

Maximizing the NFB

Having hopefully freed ourselves from fear of feedback, and appreciating the dangers of using only a little of it, the next step is to see how much can be used. It is my view that the amount of negative feedback applied should be maximized at all audio frequencies to maximize linearity, and the only limit is the requirement for reliable HF stability. In fact, global or Nyquist oscillation is not normally a difficult design problem in power amplifiers; the HF feedback factor can be calculated simply and accurately, and set to whatever figure is considered safe. (Local oscillations and parasitics are beyond the reach of design calculations and simulations, and cause much more trouble in practice.)

In classical Control Theory, the stability of a servomechanism is specified by its *phase margin*, the amount of extra phase shift that would be required to induce sustained oscillation, and its *gain margin*, the amount by which the open-loop gain would need to be increased for the same result. These concepts are not very useful in audio power amplifier work, where many of the significant time-constants are only vaguely known. However, it is worth remembering that the phase margin will never be better than 90°, because of the phase lag caused by the VAS Miller capacitor; fortunately this is more than adequate.

In practice designers must use their judgment and experience to determine an NFB factor that will give reliable stability in production. My own experience leads me to believe that when the conventional three-stage architecture is used, 30 dB of global feedback at 20 kHz is safe, providing an output inductor is used to prevent capacitive loads from eroding the stability margins. I would say that 40 dB was distinctly risky, and I would not care to pin it down any more closely than that.

The 30 dB figure assumes simple dominant-pole compensation with a 6 dB/octave roll-off for the open-loop gain. The phase and gain margins are determined by the angle at which this slope cuts the horizontal unity-loop-gain line. (I am deliberately terse here; almost all textbooks give a very full treatment of this stability criterion.) An intersection of 12 dB/octave is definitely unstable. Working within this, there are two basic ways in which to maximize the NFB factor:

1. While a 12 dB/octave gain slope is unstable, intermediate slopes greater than 6 dB/octave can be made to work. The maximum usable is normally considered to be 10 dB/octave, which gives a phase margin of 30°. This may be acceptable in some cases, but I think it cuts it a little fine. The steeper fall in gain means that more NFB is applied at lower frequencies, and so less distortion is produced. Electronic circuitry only provides slopes in multiples of 6 dB/octave, so 10 dB/octave

requires multiple overlapping time-constants to approximate a straight line at an intermediate slope. This gets complicated, and this method of maximizing NFB is not popular.

2. The gain slope varies with frequency, so that maximum open-loop gain and hence NFB factor is sustained as long as possible as frequency increases; the gain then drops quickly, at 12 dB/octave or more, but flattens out to 6 dB/octave before it reaches the critical unity loop–gain intersection. In this case the stability margins should be relatively unchanged compared with the conventional situation. This approach is dealt with in Chapter 8.

Overall Feedback versus Local Feedback

It is one of the fundamental principles of negative feedback that if you have more than one stage in an amplifier, each with a fixed amount of open-loop gain, it is more effective to close the feedback loop around all the stages, in what is called an overall or global feedback configuration, rather than applying the feedback locally by giving each stage its own feedback loop. I hasten to add that this does not mean you cannot or should not use local feedback *as well* as overall feedback – indeed, one of the main themes of this book is that it is a very good idea, and indeed probably the only practical route to very low distortion levels. This is dealt with in more detail in the chapters on input stages and voltage-amplifier stages.

It is worth underlining the effectiveness of overall feedback because some of the less informed audio commentators have been known to imply that overall feedback is in some way decadent or unhealthy, as opposed to the upright moral rigor of local feedback. The underlying thought, insofar as there is one, appears to be that overall feedback encloses more stages each with their own phase shift, and therefore requires compensation which will reduce the maximum slew rate. The truth, as is usual with this sort of moan, is that this could happen if you get the compensation all wrong; so get it right – it isn't hard.

It has been proposed on many occasions that if there is an overall feedback loop, the output stage should be left outside it. I have tried this, and believe me, it is not a good idea. The distortion produced by an output stage so operated is jagged and nasty, and I think no one could convince themselves it was remotely acceptable if they had seen the distortion residuals.

Figure 2.24 shows a negative-feedback system based on that in Figure 2.12, but with two stages. Each has its own open-loop gain A, its own NFB factor β, and its own open-loop error V_d added to the output of the amplifier. We want to achieve the same closed-loop gain of 25 as in Table 2.1 and we will make the wild assumption that the open-loop error of 1 in that table is now distributed equally between the two amplifiers A1 and A2. There are many ways the open- and closed-loop gains could be distributed between the two sections, but for simplicity we will give each section a closed-loop gain of 5; this means the conditions on the two sections are identical. The open-loop gains are also equally distributed between the two amplifiers so that their product is equal to column 3 in Table 2.1. The results are shown in Table 2.3: columns 1–7 show what's happening in each loop, and columns 8 and 9 give the results for the output of the two loops together, assuming for simplicity that the errors from each section can be simply added together; in other words there is no partial cancelation due to differing phases and so on.

This final result is compared with the overall feedback case of Table 2.1 in Table 2.4, where column 1 gives total open-loop gain, and column 2 is a copy of column 7 in Table 2.1 and gives the closed-loop error for the overall feedback case. Column 3 gives the closed-loop error for the two-stage feedback case, and it is brutally obvious that splitting the overall feedback situation into two local feedback stages has been a pretty bad move. With a modest total open-loop gain of 100, the local feedback system is almost twice as bad. Moving up to total open-loop gains that are more realistic for real power amplifiers, the factor of deterioration is between six and 40 times – an amount that cannot be ignored. With higher open-loop gains the ratio gets even worse. Overall feedback is totally and unarguably superior at dealing with all kinds of amplifier errors, though in this book distortion is often the one at the front of our minds.

While there is space here to give only one illustration in detail, you may be wondering what happens if the errors are not equally distributed between the two stages; the signal level at the

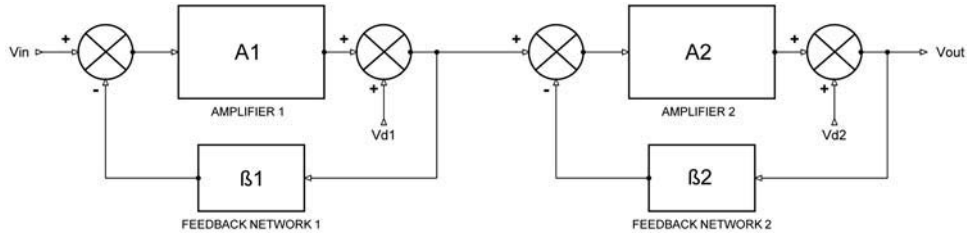


Figure 2.24: A negative-feedback system with two stages, each with its own feedback loop. There is no overall negative-feedback path

Table 2.3: Open-loop gain and closed-loop errors in the two loops

1 Desired C/L gain	2 β1 NFB fraction	3 *A*1 O/L gain	4 NFB factor	5 C/L gain	6 O/L error	7 C/L error	8 Total C/L gain	9 Total C/L error
5	0.2	10.00	3.00	3.333	0.5	0.1667	11.11	0.3333
5	0.2	31.62	7.32	4.317	0.5	0.0683	18.64	0.1365
5	0.2	100	21.00	4.762	0.5	0.0238	22.68	0.0476
5	0.2	200	41.00	4.878	0.5	0.0122	23.80	0.0244
5	0.2	316.23	64.25	4.922	0.5	0.0078	24.23	0.0156

Table 2.4: Overall NFB gives a lower closed-loop error for the same total open-loop gain. The error ratio increases as the open-loop gain increases

1 *A* Total O/L gain	2 Overall NFB C/L error	3 Two-stage NFB C/L error	4 Error ratio
100	0.2000	0.3333	1.67
1000	0.0244	0.1365	5.60
10,000	0.0025	0.0476	19.10
40,000	0.0006	0.0244	39.05
100,000	0.0002	0.0156	62.28

output of the second stage will be greater than that at the output of the first stage, so it is plausible (but by no means automatically true in the real world) that the second stage will generate more distortion than the first. If this is so, and we stick with the assumption that open-loop gain is equally distributed between the two stages, then the best way to distribute the closed-loop gain is to put most of it in the first stage so we can get as high a feedback factor as possible in the second stage. As an example, take the case where the total open-loop gain is 40,000.

Assume that all the distortion is in the second stage, so its open-loop error is 1 while that of the first stage is zero. Now redistribute the total closed-loop gain of 25 so the first stage has a closed-loop gain of 10 and the second stage has a closed-loop gain of 2.5. This gives a closed-loop error of 0.0123, which is about half of 0.0244, the result we got with the closed-loop gain equally distributed. Clearly things have been improved by applying the greater part of the local negative feedback where it is most needed. But our improved figure is still about 20 times worse than if we had used overall feedback.

In a real power amplifier, the situation is of course much more complex than this. To start with, there are usually three rather than two stages, the distortion produced by each one is level-dependent, and in the case of the voltage-amplifier stage the amount of local feedback (and hence also the amount of overall feedback) varies with frequency. Nonetheless, it will be found that overall feedback always gives better results.

Maximizing Linearity before Feedback

Make your amplifier as linear as possible before applying NFB has long been a cliché. It blithely ignores the difficulty of running a typical solid-state amplifier without any feedback, to determine its basic linearity.

Virtually no dependable advice on how to perform this desirable linearization has been published. The two factors are the basic linearity of the forward path, and the amount of negative feedback applied to further straighten it out. The latter cannot be increased beyond certain limits or high-frequency stability is put in peril, whereas there seems no reason why open-loop linearity could not be improved without limit, leading us to what in some senses must be the ultimate goal – a distortionless amplifier. This book therefore takes as one of its main aims the understanding and improvement of open-loop linearity; as it proceeds we will develop circuit blocks culminating in some practical amplifier designs that exploit the techniques presented here.

References

[1] J. Linsley-Hood, Simple Class-A amplifier, Wireless World (April 1969) p. 148.

[2] B. Olsson, Better audio from non-complements? Electronics World (December 1994) p. 988.

[3] J. Lohstroh, M. Otala, An audio power amplifier for ultimate quality requirements, IEEE Trans. Audio Electroacoustics AU-21 (6) (December 1973).

[4] D. Self, Self On Audio, second ed., Newnes, 2006, Chapter 32.

[5] B. Attwood, Design parameters important for the optimisation of PWM (Class-D) amplifiers, JAES 31 (November 1983) p. 842.

[6] J.M. Goldberg, M.B. Sandler, Noise shaping and pulse-width modulation for all-digital audio power amplifier, JAES 39 (February 1991) p. 449.

[7] J.A. Hancock, Class-D amplifier using MOSFETS with reduced minority carrier lifetime, JAES 39 (September 1991) p. 650.

[8] A. Peters, Class-E RF amplifiers, IEEE J. Solid-State Circuits (June 1975) p. 168.

[9] L. Feldman, Class-G high-efficiency hi-fi amplifier, Radio-Electronics (August 1976) p. 47.

[10] F. Raab, Average efficiency of Class-G power amplifiers, IEEE Trans. Consumer Electronics CE-22 (May 1986) p. 145.

[11] T. Sampei et al., Highest efficiency & super quality audio amplifier using MOS-power FETs in Class-G, IEEE Trans. Consumer Electronics CE-24 (August 1978) p. 300.

[12] P. Buitendijk, A 40 W integrated car radio audio amplifier, IEEE Conf. Consumer Electronics, 1991 session, THAM 12.4, p. 174 (Class-H).

[13] A. Sandman, Class S: a novel approach to amplifier distortion, Wireless World (September 1982) p. 38.

[14] R. Sinclair (Ed.), Audio and Hi-fi Handbook, Newnes, 1993, p. 541

[15] P.J. Walker, Current dumping audio amplifier, Wireless World (December 1975) p. 560.

[16] G. Stochino, Audio design leaps forward? Electronics World (October 1994) p. 818.

[17] J. Didden, paX – a power amplifier with error correction, Elektor (April/May 2008).

[18] S. Tanaka, A new biasing circuit for Class-B operation, JAES (January/February 1981) p. 27.

[19] P.G.L. Mills, M.O.J. Hawksford, Transconductance power amplifier systems for current-driven loudspeakers, JAES 37 (March 1989) p. 809.

[20] R. Evenson, Audio amplifiers with tailored output impedances, Preprint for November 1988 AES Convention, Los Angeles.

[21] P. Blomley, A new approach to Class-B, Wireless World (February 1971) p. 57.

[22] B. Gilbert, in: C. Toumazou, F.G. Lidgey and D.G. Haigh (Eds.), Current Mode Circuits from a Translinear Viewpoint, Chapter 2: Analogue IC Design: The Current-Mode Approach, IEEE, 1990.

[23] F. Thus, Compact bipolar Class AB output stage, IEEE J. Solid-State Circuits (December 1992) p. 1718.

[24] E. Cherry, Nested differentiating feedback loops in simple audio power amplifiers, JAES 30 (5) (May 1982) p. 295.

[25] E. Cherry, Designing NDFL amps, Electronics Today International (April/May 1983).

[26] P. Baxandall, Audio power amplifier design: Part 5, Wireless World (December 1978) 53. (This superb series of articles had six parts and ran on roughly alternate months, starting in Jan 1978.)

CHAPTER 3

The General Principles of Power Amplifiers

How a Generic Amplifier Works

Figure 3.1 shows a very conventional power amplifier circuit; it is as standard as possible. A great deal has been written about this configuration, though the subtlety and quiet effectiveness of the topology are usually overlooked, and the explanation below therefore touches on several aspects that seem to be almost unknown. The circuit has the merit of being docile enough to be made into a functioning amplifier by someone who has only the sketchiest of notions as to how it works.

The input differential pair implements one of the few forms of distortion cancelation that can be relied upon to work reliably without adjustment – this is because the transconductance of the input pair is determined by the physics of transistor action rather than matching of ill-defined parameters such as beta; the logarithmic relation between I_c and V_{be} is proverbially accurate over some eight or nine decades of current variation.

The voltage signal at the voltage-amplifier stage (hereafter VAS) transistor base is typically a couple of millivolts, looking rather like a distorted triangle wave. Fortunately the voltage here is of little more than academic interest, as the circuit topology essentially consists of a transconductance amp (voltage-difference input to current output) driving into a transresistance (current-to-voltage converter) stage. In the first case the exponential V_{be}/I_c law is straightened out by the differential-pair action, and in the second the global (overall) feedback factor at LF is sufficient to linearize the VAS, while at HF shunt negative feedback (hereafter NFB) through C_{dom} conveniently takes over VAS linearization while the overall feedback factor is falling.

The behavior of Miller dominant-pole compensation in this stage is actually exceedingly elegant, and not at all a case of finding the most vulnerable transistor and slugging it. As frequency rises and C_{dom} begins to take effect, negative feedback is no longer applied globally around the whole amplifier, which would include the higher poles, but instead is seamlessly transferred to a purely local role in linearizing the VAS. Since this stage effectively contains a single gain transistor, any amount of NFB can be applied to it without stability problems.

The amplifier operates in two regions; the LF, where open-loop (O/L) gain is substantially constant, and HF, above the dominant-pole breakpoint, where the gain is decreasing steadily at 6 dB/octave. Assuming the output stage is unity gain, three simple relationships define the gain in these two regions:

$$LFgain = g_m \cdot \beta \cdot R_c \qquad \text{Equation 3.1}$$

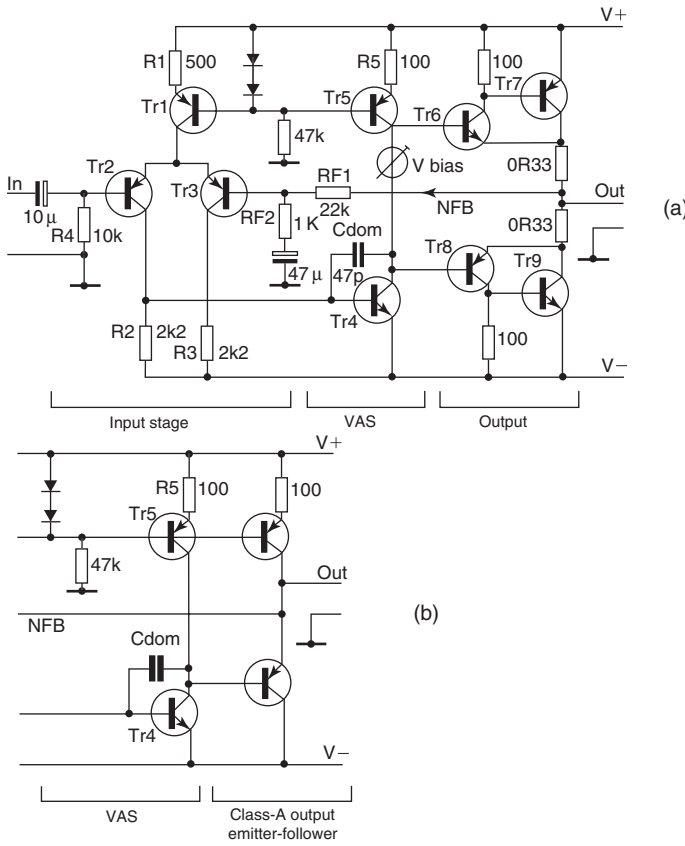


Figure 3.1: (a) A conventional Class-B power-amp circuit. (b) With small-signal Class-A output emitter-follower replacing Class-B output to make a model amplifier

At least one of the factors that set this (beta) is not well controlled and so the LF gain of the amplifier is to a certain extent a matter of pot luck; fortunately this does not matter, so long as it is high enough to give a suitable level of NFB to eliminate LF distortion. The use of the word *eliminate* is deliberate, as will be seen later. Usually the LF gain, or HF local feedback factor, is made high by increasing the effective value of the VAS collector impedance R_c, either by the use of current-source collector load, or by some form of bootstrapping.

The other important relations are:

$$HFgain = \frac{g_m}{\omega \cdot C_{dom}} \qquad \text{Equation 3.2}$$

Dominant-pole frequency

$$P1 = \frac{1}{C_{dom} \cdot \beta \cdot R_c} \qquad \text{Equation 3.3}$$

Where:

$$\omega = 2 \cdot \pi \cdot frequency$$

In the HF region, things are distinctly more difficult as regards distortion, for while the VAS is locally linearized, the global feedback factor available to linearize the input and output stages is falling steadily at 6 dB/octave. For the time being we will assume that it is possible to define an HF gain (say, *N* dB at 20 kHz), which will assure stability with practical loads and component variations. Note that the HF gain, and therefore both HF distortion and stability margin, are set by the simple combination of the input stage transconductance and one capacitor, and most components have no effect on it at all.

It is often said that the use of a high VAS collector impedance provides a current drive to the output devices, often with the implication that this somehow allows the stage to skip quickly and lightly over the dreaded crossover region. This is a misconception – the collector impedance falls to a few kilohms at HF, due to increasing local feedback through C_{dom}, and in any case it is very doubtful if true current drive would be a good thing: calculation shows that a low-impedance voltage drive minimizes distortion due to beta-unmatched output halves[1], and it certainly eliminates the effect of Distortion 4, described below.

The Advantages of the Conventional

It is probably not an accident that the generic configuration is by a long way the most popular, though in the uncertain world of audio technology it is unwise to be too dogmatic about this sort of thing. The generic configuration has several advantages over other approaches:

- The input pair not only provides the simplest way of making a DC-coupled amplifier with a dependably small output offset voltage, but can also (given half a chance) completely cancel the second-harmonic distortion that would be generated by a single-transistor input stage. One vital condition for this must be met; the pair must be accurately balanced by choosing the associated components so that the two collector currents are equal. (The *typical* component values shown in Figure 3.1 do *not* bring about this most desirable state of affairs.)
- The input devices work at a constant and near-equal V_{ce}, giving good thermal balance.
- The input pair has virtually no voltage gain so no low-frequency pole can be generated by Miller effect in the TR2 collector-base capacitance. All the voltage gain is provided by the VAS stage, which makes for easy compensation. Feedback through C_{dom} lowers VAS input and output impedances, minimizing the effect of input-stage capacitance, and the output-stage capacitance. This is often known as pole-splitting[2]; the pole of the VAS is moved downwards in frequency to become the dominant pole, while the input-stage pole is pushed up in frequency.
- The VAS Miller compensation capacitance smoothly transfers NFB from a global loop that may be unstable, to the VAS local loop that cannot be. It is quite wrong to state that *all* the benefits of feedback are lost as the frequency increases above the dominant pole, as the VAS is still being linearized. This position of C_{dom} also swamps the rather variable C_{cb} of the VAS transistor.

The Distortion Mechanisms

My original series of articles on amplifier distortion listed seven important distortion mechanisms, all of which are applicable to any Class-B amplifier, and do not depend on particular circuit arrangements. As a result of further experimentation and further thought, I have now increased this to ten.

In the typical amplifier THD is often thought to be simply due to the Class-B nature of the output stage, which is linearized less effectively as the feedback factor falls with increasing frequency. This is, however, only true when all the removable sources of distortion have been eliminated. In the vast majority of amplifiers in production, the true situation is more complex, as the small-signal stages can generate significant distortion of their own, in at least two different ways; this distortion can easily exceed output stage distortion at high frequencies. It is particularly inelegant to allow this to occur given the freedom of design possible in the small-signal section.

If the ills that a Class-B stage is prone to are included then there are eight major distortion mechanisms. Note that this assumes that the amplifier is not overloaded in any way, and therefore is not suffering from:

- activation of any overload protection circuitry;
- overloading not affecting protection circuitry (for example, insufficient current to drive the output stage due to a VAS current source running set to too low a value);
- slew-rate limiting;
- defective or out-of-tolerance components.

It also assumes the amplifier has proper global or Nyquist stability and does not suffer from any parasitic oscillations; the latter, if of high enough frequency, cannot be seen on the average oscilloscope and tend to manifest themselves only as unexpected increases in distortion, sometimes at very specific power outputs and frequencies.

In Figure 3.2 an attempt has been made to show the distortion situation diagrammatically, indicating the location of each mechanism within the amplifier. Distortion 8 is not shown as there is no output capacitor.

The first four distortion mechanisms are inherent to any three-stage amplifier.

Distortion 1: Input Stage Distortion

This concerns nonlinearity in the input stage. If this is a carefully balanced differential pair then the distortion is typically only measurable at HF, rises at 18 dB/octave, and is almost pure third harmonic. If the input pair is unbalanced (which from published circuitry it usually is) then the HF distortion emerges from the noise floor earlier, as frequency increases, and rises at 12 dB/octave as it is mostly second harmonic.

This mechanism is dealt with in Chapter 4.

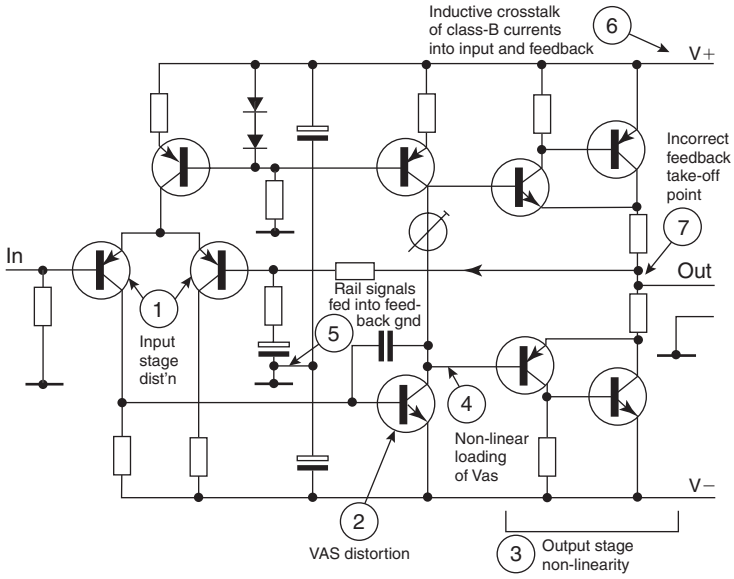


Figure 3.2: The location of the first seven major distortion mechanisms. The eighth (capacitor distortion) is omitted for clarity

Distortion 2: VAS Distortion

Nonlinearity in the voltage-amplifier stage (which I call the VAS for brevity) surprisingly does not always figure in the total distortion. If it does, it remains constant until the dominant-pole frequency $P1$ is reached, and then rises at 6 dB/octave. With the configurations discussed here it is always second harmonic.

Usually the level is very low due to linearizing negative feedback through the dominant-pole capacitor. Hence if you crank up the local VAS open-loop gain, for example by cascoding or putting more current-gain in the local VAS–C_{dom} loop, and attend to Distortion 4 below, you can usually ignore VAS distortion.

This mechanism is dealt with in Chapter 5.

Distortion 3: Output Stage Distortion

Nonlinearity in the output stage, which is naturally the obvious source. This in a Class-B amplifier will be a complex mix of large-signal distortion and crossover effects, the latter generating a spray of high-order harmonics, and in general rising at 6 dB/octave as the amount of negative feedback decreases. Large-signal THD worsens with 4 Ω loads and worsens again at 2 Ω. The picture is complicated by dilatory switch-off in the relatively slow output devices, ominously signaled by supply current increasing in the top audio octaves.

These mechanisms are dealt with in Chapters 6 and 7.

Distortion 4: VAS-Loading Distortion

This is loading of the VAS by the nonlinear input impedance of the output stage. When all other distortion sources have been attended to, this is the limiting distortion factor at LF (say, below 2 kHz); it is simply cured by buffering the VAS from the output stage. Magnitude is essentially constant with frequency, though the overall effect in a complete amplifier becomes less as frequency rises and feedback through C_{dom} starts to linearize the VAS.

This mechanism is dealt with in Chapter 7.

The next three distortion mechanisms are in no way inherent; they may be reduced to unmeasurable levels by simple precautions. They are what might be called topological distortions, in that they depend wholly on the arrangement of wiring and connections, and on the physical layout of the amplifier.

Distortion 5: Rail-Decoupling Distortion

Nonlinearity caused by large rail-decoupling capacitors feeding the distorted signals on the supply lines into the signal ground. This seems to be the reason that many amplifiers have rising THD at low frequencies. Examining one commercial amplifier kit, I found that rerouting the decoupler ground return reduced the THD at 20 Hz by a factor of 3.

This mechanism is dealt with in Chapter 7.

Distortion 6: Induction Distortion

This is nonlinearity caused by induction of Class-B supply currents into the output, ground, or negative-feedback lines. This was highlighted by Cherry[3] but seems to remain largely unknown; it is an insidious distortion that is hard to remove, though when you know what to look for on the THD residual it is fairly easy to identify. I suspect that a large number of commercial amplifiers suffer from this to some extent.

This mechanism is dealt with in Chapter 7.

Distortion 7: NFB Take-Off Distortion

This is nonlinearity resulting from taking the NFB feed from slightly the wrong place near where the power-transistor Class-B currents sum to form the output. This may well be another very prevalent defect.

This mechanism is dealt with in Chapter 7.

The next two distortion mechanisms relate to circuit components that are non-ideal or poorly chosen.

Distortion 8: Capacitor Distortion

In its most common manifestation this is caused by the non-ideal nature of electrolytic capacitors. It rises as frequency falls, being strongly dependent on the signal voltage across the capacitor.

The most common sources of nonlinearity are the input DC-blocking capacitor or the feedback network capacitor; the latter is more likely as it is much easier to make an input capacitor large enough to avoid the problem. It causes serious difficulties if a power amplifier is AC-coupled, i.e. has a series capacitor at the output, but this is rare these days.

It can also occur in ceramic capacitors that are nominally of the NP0/C0G type but actually have a significant voltage coefficient, when they are used to implement Miller dominant-pole compensation.

This mechanism is dealt with in detail in Chapter 7.

Distortion 9: Magnetic Distortion

This arises when a signal at amplifier output level is passed through a ferromagnetic conductor. Ferromagnetic materials have a nonlinear relationship between the current passing through them and the magnetic flux it creates, and this induces voltages that add distortion to the signal. The effect has been found in output relays and also speaker terminals. The terminals appeared to be made of brass but were actually plated steel.

This mechanism is also dealt with in detail in Chapter 7.

Distortion 10: Input Current Distortion

This distortion is caused when an amplifier input is driven from a significant source impedance. The input current taken by the amplifier is nonlinear, even if the output of the amplifier is distortion free, and the resulting voltage drop in the source impedance introduces distortion.

This mechanism is purely a product of circuit design, rather than layout or component integrity, but it has been put in a category of its own because, unlike the inherent Distortions 1–4, it is a product of the interfacing between the amplifier and the circuitry upstream of it.

This mechanism is dealt with in Chapter 4.

Distortion 11: Premature Overload Protection

The overload protection of a power amplifier can be implemented in many ways, but without doubt the most popular method is the use of VI limiters that shunt signal current away from the inputs to the output stage. In their simplest and most common form, these come into operation relatively gradually as their set threshold is exceeded, and introduce distortion into the signal long before they close it down entirely. It is therefore essential to plan a sufficient safety margin into the output stage so that the VI limiters are never near activation in normal use. This issue is examined more closely in Chapter 17.

Other methods of overload protection that trigger and then latch the amplifier into a standby state cannot generate this distortion, but if this leads to repeated unnecessary shutdowns it will be a good deal more annoying than occasional distortion.

Nonexistent or Negligible Distortions

Having set down what might be called the Eleven Great Distortions, we must pause to put to flight a few paper tigers …

The first is common-mode distortion in the input stage, a specter that haunts the correspondence columns. Since it is fairly easy to make an amplifier with less than <0.00065% THD (1 kHz) without paying any attention at all to this issue it cannot be too serious a problem. It is perhaps a slight exaggeration to call it nonexistent, as under special circumstances it can be seen, but it is certainly unmeasurable under normal circumstances.

If the common-mode voltage on the input pair is greatly increased, then a previously negligible distortion mechanism is indeed provoked. This increase is achieved by reducing the C/L gain to between 1 and 2×; the input signal is now much larger for the same output, and the feedback signal must match it, so the input stage experiences a proportional increase in common-mode voltage.

The distortion produced by this mechanism increases as the square of the common-mode voltage, and therefore falls rapidly as the closed-loop gain is increased back to normal values. It therefore appears that the only precautions required against common-mode distortion are to ensure that the closed-loop gain is at least five times (which is no hardship, as it almost certainly is anyway) and to use a tail-current source for the input pair, which again is standard practice. This issue is dealt with in more detail in the chapter on power amplifier input stages.

The second distortion conspicuous by its absence in the list is the injection of distorted supply-rail signals directly into the amplifier circuitry. Although this putative mechanism has received a lot of attention[4], dealing with Distortion 5 above by proper grounding seems to be all that is required; once more, if triple-zero THD can be attained using simple unregulated supplies and without paying any attention to the power-supply rejection ratio (PSRR) beyond keeping the amplifier free from hum (which it reliably can be) then there seems to be no problem. There is certainly no need for regulated supply rails to get a good performance. PSRR does need careful attention if the hum/noise performance is to be of the first order, but a little RC filtering is usually all that is needed. This topic is dealt with in Chapter 9.

A third mechanism of very doubtful validity is thermal distortion, allegedly induced by parameter changes in semiconductor devices whose instantaneous power dissipation varies over a cycle. This would surely manifest itself as a distortion rise at very low frequencies, but it simply does not happen. There are several distortion mechanisms that can give a THD rise at LF, but when these are eliminated the typical distortion trace remains flat down to at least 10 Hz. The worst thermal effects would be expected in Class-B output stages where dissipation varies wildly over a cycle; however, drivers and output devices have relatively large junctions with high thermal inertia. Low frequencies are of course also where the NFB factor is at its maximum. This contentious issue is dealt with at greater length in Chapter 6.

To return to our list of the unmagnificent eleven, note that only Distortion 3 is directly due to output stage nonlinearity, though Distortions 4–7 all result from the Class-B nature of the typical output stage. Distortions 8–11 can happen in any amplifier, whatever its operating class.

The Performance of a Standard Amplifier

The THD curve for the standard amplifier is shown in Figure 3.3. As usual, distortion increases with frequency, and as we shall see later, would give grounds for suspicion if it did not. The flat part of the curve below 500 Hz represents non-frequency-sensitive distortion rather than the noise floor, which for this case is at the 0.0005% level. Above 500 Hz the distortion rises at an increasing rate, rather than a constant number of dB/octave, due to the combination of Distortions 1–4. (In this case, Distortions 5–7 have been carefully eliminated to keep things simple; this is why the distortion performance looks good already, and the significance of this should not be overlooked.) It is often written that having distortion constant across the audio band is a good thing – a most unhappy conclusion, as the only practical way to achieve this with a normal Class-B amplifier is to *increase* the distortion at LF, for example by allowing the VAS to distort significantly.

It should now be clear why it is hard to wring linearity out of such a snake-pit of contending distortions. A circuit-value change is likely to alter at least two of the distortion mechanisms, and probably change the O/L gain as well; in the coming chapters I shall demonstrate how each distortion mechanism can be measured and manipulated in isolation.

Open-Loop Linearity and How to Determine It

Improving something demands measuring it, and thus it is essential to examine the open-loop linearity of power-amp circuitry. This cannot be done directly, so it is necessary to measure the NFB factor and calculate open-loop distortion from closed-loop measurements. The closed-loop gain is normally set by input sensitivity requirements.

Measuring the feedback factor is at first sight difficult, as it means determining the open-loop gain. Standard methods for measuring op-amp open-loop gain involve breaking feedback loops and

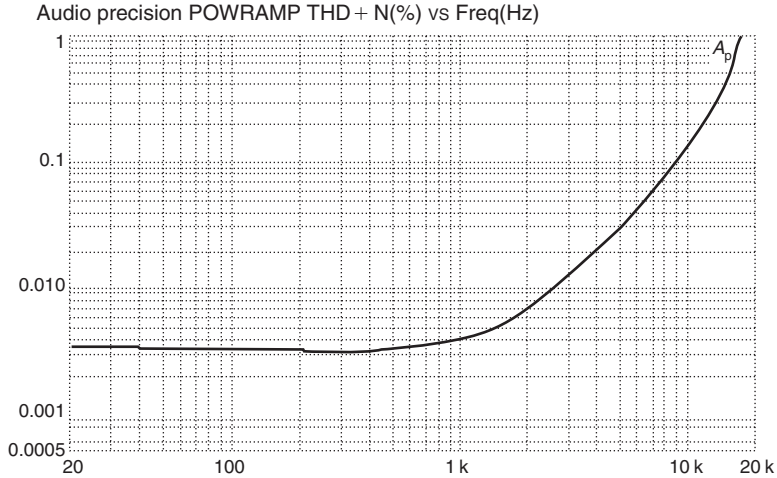


Figure 3.3: The distortion performance of the Class-B amplifier in Figure 3.1

manipulating C/L gains, procedures that are likely to send the average power amplifier into fits. Nonetheless the need to measure this parameter is inescapable as a typical circuit modification – e.g. changing the value of R2 changes the open-loop gain as well as the linearity, and to prevent total confusion it is essential to keep a very clear idea of whether an observed change is due to an improvement in O/L linearity or merely because the O/L gain has risen. It is wise to keep a running check on this as work proceeds, so the direct method of open-loop gain measurement shown in Figure 3.4 was evolved.

Direct Open-Loop Gain Measurement

The amplifier shown in Figure 3.1 is a differential amplifier, so its open-loop gain is simply the output divided by the voltage difference between the inputs. If output voltage is kept constant by providing a constant swept-frequency voltage at the positive input, then a plot of open-loop gain versus frequency is obtained by measuring the error-voltage between the inputs, and referring this to the output level. This gives an upside-down plot that rises at HF rather than falling, as the differential amplifier requires more input for the same output as frequency increases, but the method is so quick and convenient that this can be lived with. Gain is plotted in dB with respect to the chosen output level (+16 dBu in this case) and the actual gain at any frequency can be read off simply by dropping the minus sign. Figure 3.5 shows the plot for the amplifier in Figure 3.1.

The HF-region gain slope is always 6 dB/octave unless you are using something special in the way of compensation, and by the Nyquist rules must continue at this slope until it intersects the horizontal line representing the feedback factor, if the amplifier is stable. In other words, the slope is not being accelerated by other poles until the loop gain has fallen to unity, and this provides a simple way of putting a lower bound on the next pole *P*2; the important *P*2 frequency (which is usually somewhat mysterious) must be above the intersection frequency if the amplifier is seen to be stable.

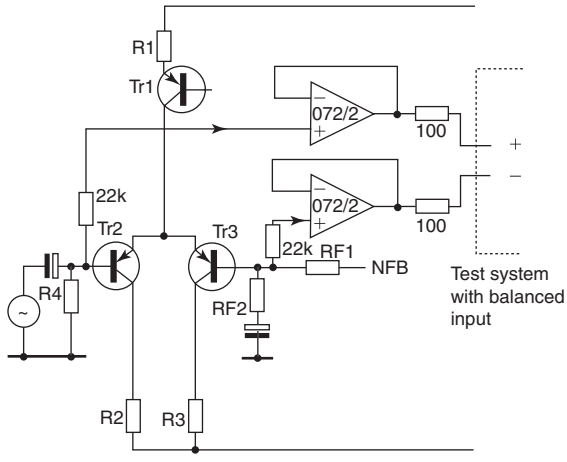


Figure 3.4: Test circuit for measuring open-loop gain directly. The accuracy with which high O/L gains can be measured depends on the test-gear CMRR

Given test gear with a sufficiently high common-mode rejection ratio (CMRR) balanced input, the method of Figure 3.4 is simple; just buffer the differential inputs from the cable capacitance with TL072 buffers, which place negligible loading on the circuit if normal component values are used. In particular be wary of adding stray capacitance to ground to the negative input, as this directly imperils amplifier stability by adding an extra feedback pole. Short wires from power amplifier to buffer IC can usually be unscreened as they are driven from low impedances.

The test-gear input CMRR defines the maximum open-loop gain measurable; I used an Audio Precision System-1 without any special alignment of CMRR. A calibration plot can be produced by feeding the two buffer inputs from the same signal; this will probably be found to rise at 6dB/octave, being set by the inevitable input asymmetries. This must be low enough for amplifier error signals to be above it by at least 10dB for reasonable accuracy. The calibration plot will flatten out at low frequencies, and may even show an LF rise due to imbalance of the test-gear input-blocking capacitors; this can make determination of the lowest pole *P*1 difficult, but this is not usually a vital parameter in itself.

Using Model Amplifiers

Distortions 1 and 2 can dominate amplifier performance and need to be studied without the manifold complications introduced by a Class-B output stage. This can be done by reducing the circuit to a *model* amplifier that consists of the small-signal stages alone, with a very linear Class-A emitter-follower attached to the output to allow driving the feedback network; here *small signal* refers to current rather than voltage, as the model amplifier should be capable of giving a full power-amp voltage swing, given sufficiently high rail voltages. From Figure 3.2 it is clear that this will allow study of Distortions 1 and 2 in isolation, and using this approach it will prove relatively easy to design a small-signal amplifier with negligible distortion across the audio band, and this is the only sure foundation on which to build a good power amplifier.

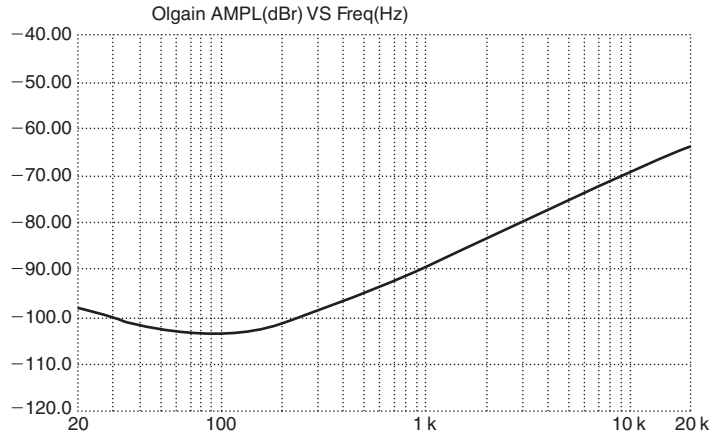


Figure 3.5: Open-loop gain versus frequency plot for Figure 3.1. Note that the curve rises as gain falls, because the amplifier error is the actual quantity measured

A typical plot combining Distortions 1 and 2 from a model amp is shown in Figure 3.6, where it can be seen that the distortion rises with an accelerating slope, as the initial rise at 6 dB/octave from the VAS is contributed to and then dominated by the 12 dB/octave rise in distortion from an unbalanced input stage.

The model can be powered from a regulated current-limited PSU to cut down the number of variables, and a standard output level chosen for comparison of different amplifier configurations; the rails and output level used for the results in this work were ±15 V and +16 dBu. The rail voltages can be made comfortably lower than the average amplifier HT rail, so that radical bits of circuitry can be tried out without the creation of a silicon cemetery around your feet. It must be remembered that some phenomena such as input-pair distortion depend on absolute output level, rather than the proportion of the rail voltage used in the output swing, and will be increased by a mathematically predictable amount when the real voltage swings are used.

The use of such model amplifiers requires some caution, and gives no insight into BJT output stages, whose behavior is heavily influenced by the sloth and low current gain of the power devices. As a general rule, it should be possible to replace the small-signal output with a real output stage and get a stable and workable power amplifier; if not, then the model is probably dangerously unrealistic.

The Concept of the Blameless Amplifier

Here I introduce the concept of what I have chosen to call a *Blameless* audio power amplifier. This is an amplifier designed so that all the easily defeated distortion mechanisms have been rendered negligible. (Note that the word *Blameless* has been carefully chosen *not* to imply perfection, but merely the avoidance of known errors.) Such an amplifier gives about 0.0005% THD at 1 kHz and approximately 0.003% at 10 kHz when driving 8 Ω. This is much less THD than a Class-B

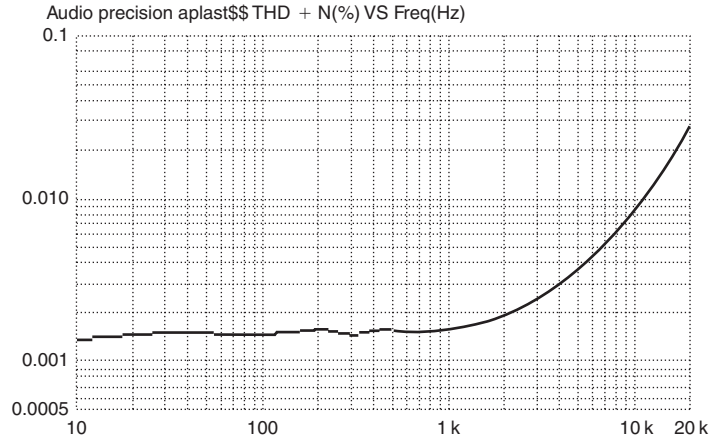


Figure 3.6: The distortion from a model amplifier, produced by the input pair and the voltage-amplifier stage. Note increasing slope as input pair distortion begins to add to VAS distortion

amplifier is normally expected to produce, but the performance is repeatable, predictable, and definitely does not require large global feedback factors.

Distortion 1 cannot be totally eradicated, but its onset can be pushed well above 20 kHz by the use of local feedback. Distortion 2 (VAS distortion) can be similarly suppressed by cascoding or beta-enhancement, and Distortions 4–7 can be made negligible by simple topological methods. All these measures will be detailed later. This leaves Distortion 3, which includes the intractable Class-B problems, i.e. crossover distortion (Distortion 3b) and HF switch-off difficulties (Distortion 3c). Minimizing 3b requires a Blameless amplifier to use a BJT output rather than FETs.

A Blameless Class-B amplifier essentially shows crossover distortion only, so long as the load is no heavier than 8 Ω; this distortion increases with frequency as the amount of global NFB falls. At 4 Ω loading an extra distortion mechanism (3a) generates significant third harmonic.

The importance of the Blameless concept is that it represents the best distortion performance obtainable from straightforward Class-B. This performance is stable and repeatable, and varies little with transistor type as it is not sensitive to variable quantities such as beta.

Blamelessness is a condition that can be defined with precision, and is therefore a standard other amplifiers can be judged against. A Blameless design represents a stable point of departure for more radical designs, such as the Trimodal concept in Chapter 10. This may be the most important use of the idea.

References

[1] B. Oliver, Distortion in complementary-pair Class-B amplifiers, Hewlett-Packard Journal (February 1971) p. 11.

[2] D. Feucht, Handbook of Analog Circuit Design, Academic Press, 1990, p. 256 (pole-splitting).

[3] E. Cherry, A new distortion mechanism in Class-B amplifiers, JAES (May 1981) p. 327.

[4] G. Ball, Distorting power supplies, Electronics & Wireless World (December 1990) p. 1084.

CHAPTER 4

The Input Stage

'A beginning is the time for taking the most delicate care that the balances are correct.'

Frank Herbert, Dune

The Role of the Input Stage

The input stage of an amplifier performs the critical duty of subtracting the feedback signal from the input, to generate the error signal that drives the output. It is almost invariably a differential transconductance stage; a voltage-difference input results in a current output that is essentially insensitive to the voltage at the output port. Its design is also frequently neglected, as it is assumed that the signals involved must be small, and that its linearity can therefore be taken lightly compared with that of the VAS or the output stage. This is quite wrong, for a misconceived or even mildly wayward input stage can easily dominate the HF distortion performance.

The input transconductance is one of the two parameters setting HF open-loop (O/L) gain, and therefore has a powerful influence on stability and transient behavior as well as distortion. Ideally the designer should set out with some notion of how much O/L gain at 20 kHz will be safe when driving worst-case reactive loads (this information should be easier to gather now there is a way to measure O/L gain directly), and from this a suitable combination of input transconductance and dominant-pole Miller capacitance can be chosen.

Many of the performance graphs shown here are taken from a *model* (small-signal stages only) amplifier with a Class-A emitter-follower output, at +16 dBu on ±15 V rails; however, since the output from the input pair is in current form, the rail voltage in itself has no significant effect on the linearity of the input stage; it is the current swing at its output that is the crucial factor.

Distortion from the Input Stage

The motivation for using a differential pair as the input stage of an amplifier is usually its low DC offset. Apart from its inherently lower offset due to the cancelation of the V_{be} voltages, it has the important added advantage that its standing current does not have to flow through the feedback network. However, a second powerful reason, which seems less well known, is that linearity is far superior to single-transistor input stages. Figure 4.1 shows three versions, in increasing order of sophistication. The resistor-tail version at 1a has poor CMRR and PSRR and is generally a false economy of the shabbiest kind; it will not be further considered here. The mirrored version at 1c has the best balance, as well as twice the transconductance of 1b.

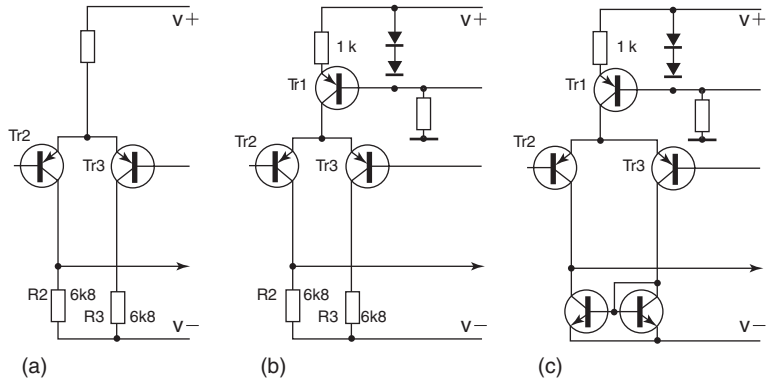


Figure 4.1: Three versions of an input pair. (a) Simple tail resistor. (b) Tail-current source. (c) With collector current-mirror to give inherently good I_c balance

At first sight, the input stage should generate a minimal proportion of the overall distortion because the voltage signals it handles are very small, appearing as they do upstream of the VAS that provides almost all the voltage gain. However, above the first pole frequency $P1$, the current required to drive C_{dom} dominates the proceedings, and this remorselessly doubles with each octave, thus:

$$i_{pk} = \omega \cdot C_{dom} \cdot V_{pk} \qquad \text{Equation 4.1}$$

where $\omega = 2 \cdot \pi \cdot frequency$.

For example, the current required at 100 W (8 Ω) and 20 kHz, with a 100 pF C_{dom}, is 0.5 mA peak, which may be a large proportion of the input standing current, and so the linearity of transconductance for large current excursions will be of primary importance if we want low distortion at high frequencies.

Curve A in Figure 4.2 shows the distortion plot for a model amplifier (at +16 dBu output), designed so the distortion from all other sources is negligible compared with that from the carefully balanced input stage; with a small-signal Class-A stage this reduces to making sure that the VAS is properly linearized. Plots are shown for both 80 and 500 kHz measurement bandwidths, in an attempt to show both HF behavior and the vanishingly low LF distortion. It can be seen that the distortion is below the noise floor until 10 kHz, when it emerges and heaves upwards at a precipitous 18 dB/octave. This rapid increase is due to the input stage signal current doubling with every octave, to feed C_{dom}; this means that the associated third-harmonic distortion will quadruple with every octave increase. Simultaneously the overall NFB available to linearize this distortion is falling at 6 dB/octave since we are almost certainly above the dominant-pole frequency $P1$, and so the combined effect is an octuple or 18 dB/octave rise. If the VAS or the output stage were generating distortion this would be rising at only 6 dB/octave, and so would look quite different on the plot.

This nonlinearity, which depends on the rate of change of the output voltage, is the nearest thing that exists to the late unlamented transient intermodulation distortion (TID), an acronym that has now fallen out of fashion. It was sometimes known by the alias transient intermodulation (TIM).

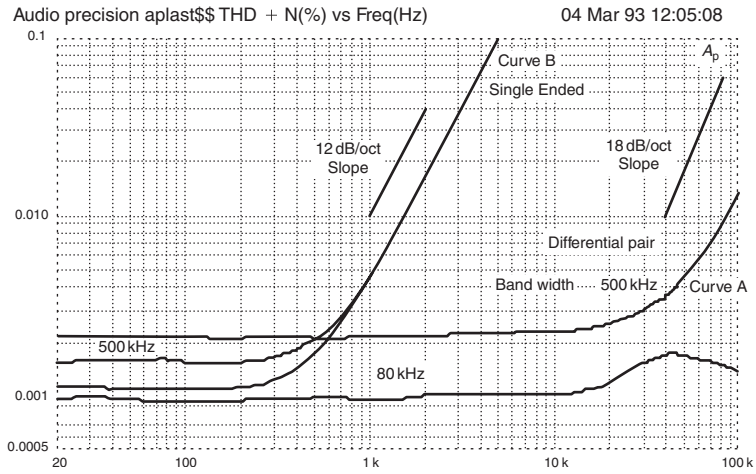


Figure 4.2: Distortion performance of model amplifier-differential pair at A compared with singleton input at B. The singleton generates copious second-harmonic distortion

Slew-induced distortion (SID) is a better description of the effect, but implies that slew-limiting is responsible, which is not the case.

If the input pair is *not* accurately balanced, then the situation is more complex. Second- as well as third-harmonic distortion is now generated, and by the same reasoning this has a slope nearer to 12 dB/octave; this vital point is examined more closely below.

All the input stages in this book are of the PNP format shown in Figure 4.1. One reason for this is that PNP bipolar transistors are claimed to have lower recombination noise than their NPN complements, though how much difference this makes in practice is doubtful. Another reason is that this puts the VAS transistor at the bottom of the circuit diagram and its current source at the top, which somehow seems the visually accessible arrangement.

BJTs versus FETs for the Input Stage

At every stage in the design of an amplifier, it is perhaps wise to consider whether BJTs or FETs are the best devices for the job. I may as well say at once that the predictable V_{be}/I_c relationship and much higher transconductance of the bipolar transistor make it, in my opinion, the best choice for all three stages of a generic power amplifier. The position is briefly summarized below.

Advantages of the FET Input Stage

There is no base current with FETs, so this is eliminated as a source of DC offset errors. However, it is wise to bear in mind that FET gate leakage currents increase very rapidly with temperature, and under some circumstances may need to be allowed for.

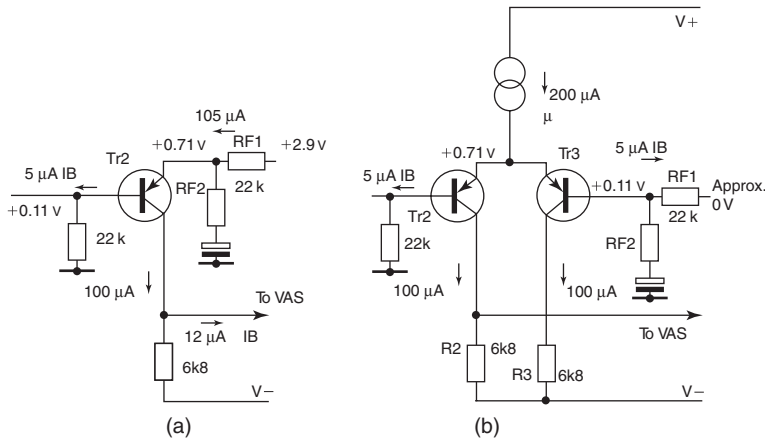


Figure 4.3: Singleton and differential pair input stages, showing typical DC conditions. The large DC offset of the singleton is mainly due to all the stage current flowing through the feedback resistor RF1

Disadvantages of FET Input Stage

1. The undegenerated transconductance is low compared with BJTs. There is much less scope for linearizing the input stage by adding degeneration in the form of source resistors, and so an FET input stage will be very nonlinear compared with a BJT version degenerated to give the same low transconductance.

2. The V_{gs} offset spreads will be high. Having examined many different amplifier designs, it seems that in practice it is essential to use dual FETs, which are relatively very expensive and not always easy to obtain. Even then, the V_{gs} mismatch will probably be greater than V_{be} mismatch in a pair of cheap discrete BJTs; for example, the 2N5912 N-channel dual FET has a specified maximum V_{gs} mismatch of 15 mV. In contrast the V_{be} mismatches of BJTs, especially those taken from the same batch (which is the norm in production) will be much lower, at about 2–3 mV, and usually negligible compared with DC offset caused by unbalanced base currents.

3. The noise performance will be inferior if the amplifier is being driven from a low-impedance source, say 5 kΩ or less. This is almost always the case.

Singleton Input Stage versus Differential Pair

Using a single input transistor (Figure 4.3a) may seem attractive, where the amplifier is capacitor-coupled or has a separate DC servo; it at least promises strict economy. However, any cost saving would be trivial, and the snag is that this singleton configuration has no way to cancel the second harmonics generated in copious quantities by its strongly curved exponential V_{in}/I_{out} characteristic[1]. The result is shown in Figure 4.2, curve B, where the distortion is much higher, though rising at the slower rate of 12 dB/octave.

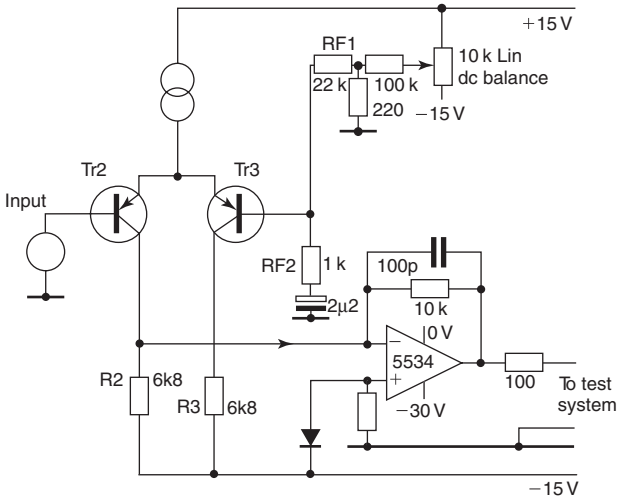


Figure 4.4: Test circuit for examining input stage distortion in isolation. The shunt-feedback op-amp is biased to provide the right DC conditions for TR2

The Input Stage Distortion in Isolation

Examining the slope of the distortion plot for the whole amplifier is instructive, but for serious research we need to measure input stage nonlinearity in isolation. This can be done with the test circuit of Figure 4.4. The op-amp uses shunt feedback to generate an appropriate AC virtual earth at the input-pair output. Note that this current-to-voltage conversion op-amp requires a third −30 V rail to allow the i/p pair collectors to work at a realistic DC voltage, i.e. about one diode-worth above the −15 V rail. The op-amp feedback resistor can be scaled as convenient, to stop op-amp clipping, without the input stage knowing anything has changed. The DC balance of the pair can be manipulated by the potentiometer, and it is instructive to see the THD residual diminish as balance is approached, until at its minimum amplitude it is almost pure third harmonic.

The differential pair has the great advantage that its transfer characteristic is mathematically highly predictable[2]. The output current is related to the differential input voltage V_{in} by:

$$I_{out} = I_e \cdot \tanh(-V_{in}/2V_t) \qquad \text{Equation 4.2}$$

(where V_t is the usual thermal voltage of about 26 mV at 25°C, and I_e is the tail current).

Two vital facts derived from this equation are that the transconductance (g_m) is maximal at $V_{in} = 0$, when the two collector currents are equal, and that the value of this maximum is proportional to the tail current I_e. Device beta does not figure in the equation, and the performance of the input pair is not significantly affected by transistor type.

Figure 4.5a shows the linearizing effect of local feedback or degeneration on the voltage-in/current-out law; Figure 4.5b plots transconductance against input voltage and shows clearly how the peak transconductance value is reduced, but the curve made flatter and linear over a wider

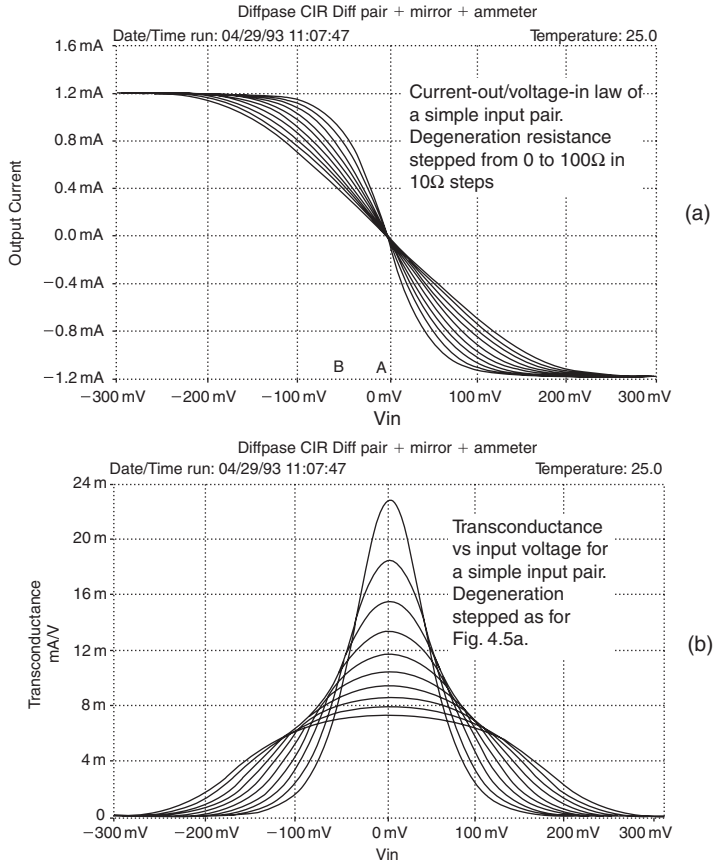


Figure 4.5: Effect of degeneration on input pair *V/I* law, showing how transconductance is sacrificed in favor of linearity (SPICE simulation)

operating range. Simply adding emitter degeneration markedly improves the linearity of the input stage, but the noise performance is slightly worsened, and of course the overall amplifier feedback factor has been reduced for, as previously shown, the vitally important HF closed-loop gain is determined solely by the input transconductance and the value of the dominant-pole capacitor.

Input Stage Balance

Exact DC balance of the input differential pair is absolutely essential in power amplifiers. It still seems almost unknown that minor deviations from equal I_c in the pair seriously upset the second-harmonic cancelation, by moving the operating point from A to B in Figure 4.5a. The average slope of the characteristic is greatest at A, so imbalance also reduces the open-loop gain if serious enough. The effect of small amounts of imbalance is shown in Figure 4.6 and Table 4.1; for an input of −45 dBu a collector-current imbalance of only 2% gives a startling worsening of linearity, with THD increasing from 0.10% to 0.16%; for 10% imbalance this deteriorates badly to 0.55%. Unsurprisingly, imbalance in the other direction ($I_{c1} > I_{c2}$) gives similar results.

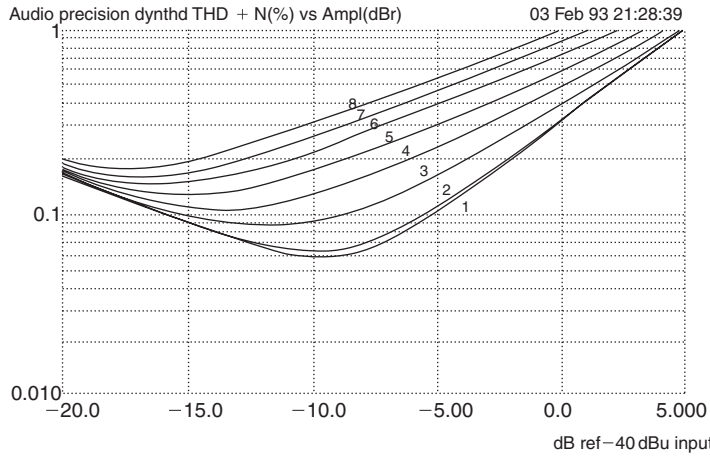


Figure 4.6: Effect of collector-current imbalance on an isolated input pair; the second harmonic rises well above the level of the third if the pair moves away from balance by as little as 2%

Table 4.1: Key to Figure 4.6

Curve no.	I_c imbalance (%)
1	0
2	0.5
3	2.2
4	3.6
5	5.4
6	6.9
7	8.5
8	10

Imbalance is defined as deviation of I_c (per device) from that value which gives equal currents in the pair.

This explains the complex distortion changes that accompany the apparently simple experiment of altering the value of R2[3]. We might design an input stage like in Figure 4.7a, where R1 has been selected as 1 k by uninspired guesswork and R2 made highish at 10 k in a plausible but wholly misguided attempt to maximize O/L gain by minimizing loading on the Q1 collector. R3 is also 10 k to give the stage a notional 'balance', though unhappily this is a visual rather than an electrical balance. The asymmetry is shown in the resulting collector currents; the design generates a lot of avoidable second harmonic distortion, displayed in the 10 k curve of Figure 4.8.

Recognizing the crucial importance of DC balance, the circuit can be rethought as in Figure 4.7b. If the collector currents are to be roughly equal, then R2 must be about 2 × R1, as both have about 0.6 V across them. The dramatic effect of this simple change is shown in the 2 k2 curve of Figure 4.8; the improvement is accentuated as the O/L gain has also increased by some 7 dB, though this has only a minor effect on the closed-loop linearity compared with the improved balance of the input pair. R3 has been excised as it contributes very little to input stage balance.

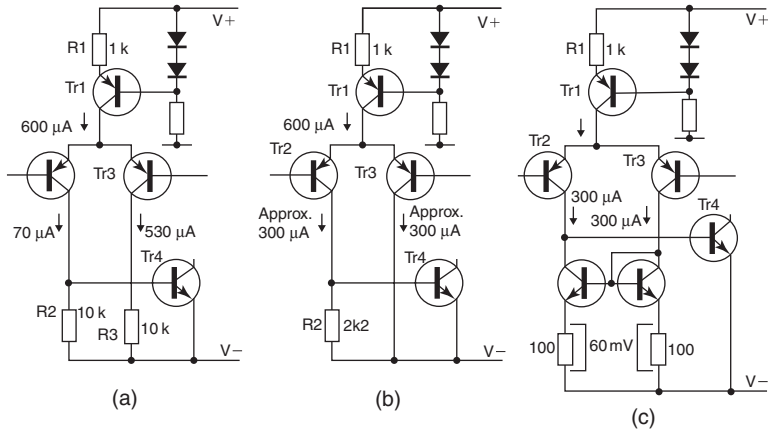


Figure 4.7: Improvements to the input pair. (a) Poorly designed version. (b) Better; partial balance by correct choice of R2. (c) Best; near-perfect I_c balance enforced by mirror

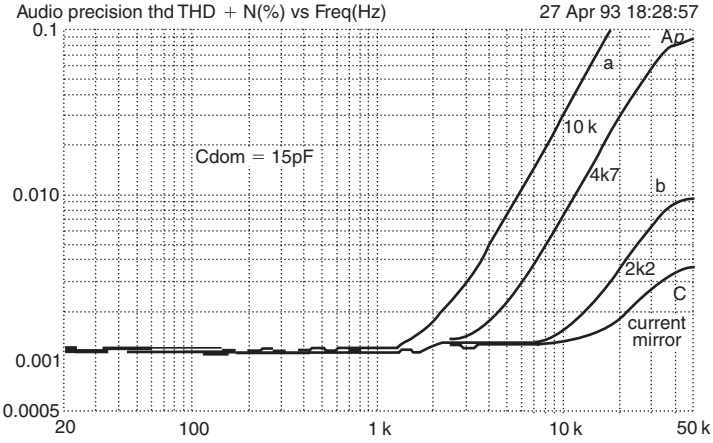


Figure 4.8: Distortion of model amplifier. (a) Unbalanced with R2 = 10 k. (b) Partially balanced with R = 2 k2. (c) Accurately balanced by current-mirror

There are very few references in the literature on the importance of collector-current balance in differential pairs; one worth looking up is an article in *Wireless World* by Eric Taylor that appeared in August 1977[4].

The Joy of Current-Mirrors

Although the input pair can be approximately balanced by the correct values for R1 and R2, we remain at the mercy of several circuit tolerances. Figure 4.6 shows that balance is critical, needing an accuracy of 1% or better for optimal linearity and hence low distortion at HF, where the input pair works hardest. The standard current-mirror configuration in Figure 4.7c forces the two collector currents very close to equality, giving correct cancelation of the second harmonic; the great improvement that results is seen in the current-mirror curve in Figure 4.8. There is also less

DC offset due to unequal base currents flowing through input and feedback resistances; I often find that a power-amplifier improvement gives at least two separate benefits.

It will be noticed that both the current-mirror transistors have very low collector-emitter voltages; the diode-connect one has just its own V_{be}, while the other sustains the V_{be} of the VAS transistor, or two V_{be} values if the VAS has been enhanced with an emitter-follower. This means that they can be low-voltage types with a high beta, which improves the mirror action.

The hyperbolic-tangent law also holds for the mirrored pair[5], though the output current swing is twice as great for the same input voltage as the resistor-loaded version. This doubled output is given at the same distortion as for the unmirrored version, as input-pair linearity depends on the input voltage, which has not changed. Alternatively, we can halve the input and get the same output, which with a properly balanced pair generating third harmonic only will give one-quarter the distortion – a most pleasing result.

The input mirror is made from discrete transistors, regretfully foregoing the V_{be}-matching available to IC designers, and so it needs its own emitter-degeneration resistors to ensure good current-matching. A voltage drop across the current-mirror emitter resistors in the range 30–60 mV will be enough to make the effect of V_{be} tolerances on distortion negligible; if degeneration is omitted then there is significant variation in HF distortion performance with different specimens of the same transistor type. Current mirrors can be made using a signal diode such as the 1N4148 instead of the diode-connected transistor, but this gives poor matching, saves little if any money, and is generally to be deprecated.

Putting a current-mirror in a well-balanced input stage increases the total O/L gain by at least 6 dB, and by up to 15 dB if the stage was previously poorly balanced; this needs to be taken into account in setting the compensation. Another happy consequence is that the slew rate is roughly doubled, as the input stage can now source and sink current into C_{dom} without wasting it in a collector load. If C_{dom} is 100 pF, the slew rate of Figure 4.7b is about 2.8 V/μs up and down, while Figure 4.7c gives 5.6 V/μs. The unbalanced pair in Figure 4.7a displays further vices by giving 0.7 V/μs positive-going and 5 V/μs negative-going.

In the world of op-amp design, the utilization of both outputs from the input differential stage is called 'phase summing'. Herpy[6] gives some interesting information on alternative ways to couple the input stage to the VAS, though some of them look unpromising for power amplifier use.

Better Current-Mirrors

The simple mirror has well-known residual base-current errors, as demonstrated in Figure 4.9 (emitter degeneration resistors are omitted for clarity, and all transistors are assumed to be identical to keep things simple). In Figure 4.9a, Q1 turns on as much as necessary to absorb the current I_{c1} into its collector, and Q2, which perforce has the same V_{be}, turns on exactly the same current. But I_{c1} is not the same as I_{in}, because two helpings of base current I_{b1} and I_{b2} have been siphoned off it. (It is helpful at this point to keep a firm grip on the idea that a bipolar transistor is a voltage-operated device *not* a current-operated device, and the base currents are not 'what turns the transistors on' but the unwanted effect of finite beta. It does not help that beta is sometimes called 'current gain' when

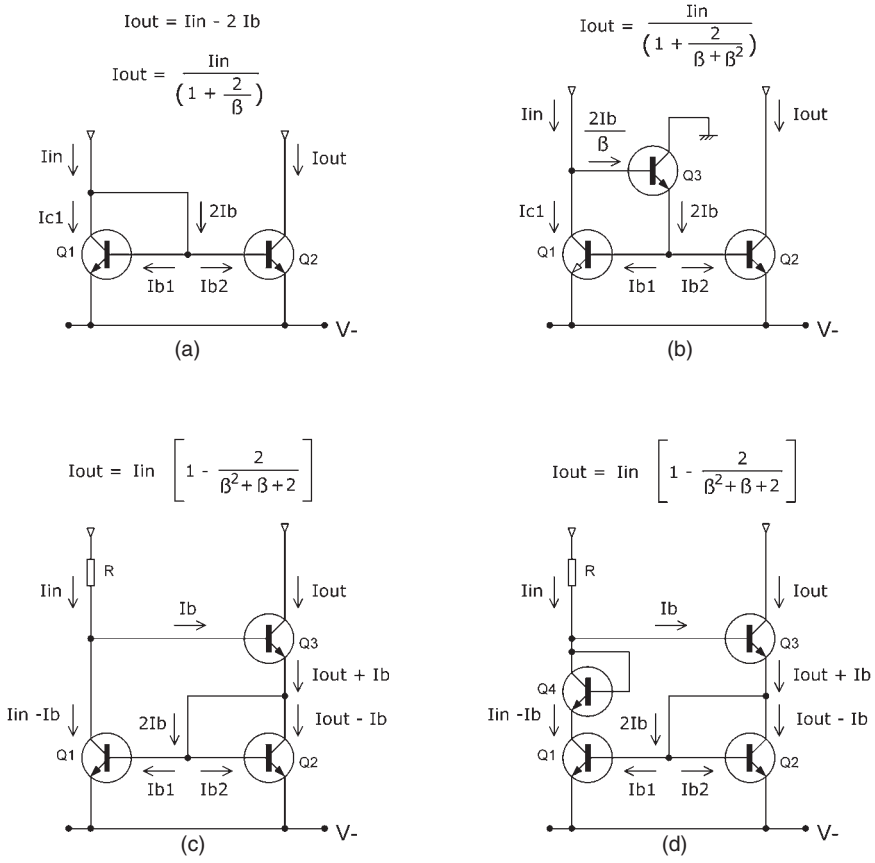


Figure 4.9: Current-mirrors and their discontents. (a) The basic mirror has base-current errors. (b) The EFA circuit reduces these. (c) The Wilson mirror greatly reduces these. (d) A further improvement to Wilson by equalizing the Vce's of Q1 and Q2

it is nothing of the sort.) Therefore, I_{out} is going to be less than I_{in} by twice I_b, and I_b will not be a linear function of I_{in}, because beta varies with collector current. Note that this problem occurs even though our transistors have been assumed to be perfectly matched for both beta and V_{be}.

A great deal of effort has been put into making good current-mirrors by op-amp designers, and we can cheerfully exploit the results. A good source is Ref. [7]. One way to reduce the base-current problem is to add a third transistor, as in Figure 4.9b. This reduces the base current bled away from the input current by a factor of beta – the beta of Q3. If this configuration has an official name I don't know it, and I have always called it the emitter-follower added (EFA) circuit. Another way is shown in Figure 4.9c. This is the famous Wilson current-mirror, which unlike the previous versions uses negative feedback. Q3 is a voltage-follower, and Q1, Q2 a basic current-mirror, and Q1. If the current through Q3 should tend to increase, the current-mirror pulls current away from Q3 base and turns it off a bit. We are assuming that R exists in some form (for it would make little sense to have a low impedance feeding a current-mirror); in our case it is the collector impedance of one of the input pair transistors. An important feature of the Wilson is the way the base-current

Table 4.2: Current-in/-out ratios for the smiple current-mirror and more sophisticated versions, as transistor beta varies

Beta	Simple mirror I_{out}/I_{in}	EFA I_{out}/I_{in}	Wilson I_{out}/I_{in}
1	0.33333	0.50000	0.60000
2	0.50000	0.75000	0.80000
5	0.71429	0.93750	0.94595
10	0.83333	0.98214	0.98361
25	0.92593	0.99693	0.99705
50	0.96154	0.99922	0.99923
100	0.98039	0.99980	0.99980
150	0.98684	0.99991	0.99991
200	0.99010	0.99995	0.99995
250	0.99206	0.99997	0.99997
500	0.99602	0.99999	0.99999

errors cancel, as shown in the diagram. It really is a beautiful sight. The input/output equations are given for each version, and it is equally clear that the EFA and the Wilson have beta-squared terms that make the denominators of the fractions much closer to unity. The calculated results are shown in Table 4.2, and it is clear that both the EFA and the Wilson are far superior to the simple mirror, but this superiority lessens as beta increases. The Wilson comes out slightly better than the EFA at very low betas, but at betas of 25 or more (and hopefully the beta won't be lower than that in small-signal transistors, even if they are high-voltage types) there is really very little between the two of them.

So far we have not looked at the influence of Early effect on mirror accuracy; for our purposes it is probably very small, but it is worth noting that in Figure 4.9c Q1 has a V_{ce} of two V_{be} drops while Q2 has a V_{ce} of only one V_{be}. If you are feeling perfectionist, the mirror in Figure 4.9d has an added diode-connected transistor Q4 that reduces the V_{ce} of Q1 to a single V_{be} drop.

So how much benefit can be gained by using more sophisticated current-mirrors? In some studies I have made of advanced input stages with very good linearity (not ready for publication yet, I'm afraid) I found that a simple mirror could introduce more nonlinearity than the input stage itself, that the three-transistor Wilson improved things greatly, and the four-transistor version even more.

In practical measurements, when I tried replacing the standard mirror with a Wilson in a Blameless amplifier the improvement in the distortion performance was marginal at best, for as usual most of the distortion was coming from the output stage. That does not mean we should never look at ways of improving the small-signal stages; when the bulletproof distortionless output stage finally appears, we want to be ready.

Improving Input Stage Linearity

Even if the input pair has a current-mirror, we may still feel that the HF distortion needs further reduction; after all, once it emerges from the noise floor it octuples with each doubling of

frequency, and so it is well worth postponing the evil day until as far as possible up the frequency range. The input pair shown has a conventional value of tail current. We have seen that the stage transconductance increases with I_c, and so it is possible to increase the g_m by increasing the tail current, and then return it to its previous value (otherwise C_{dom} would have to be increased proportionately to maintain stability margins) by applying local NFB in the form of emitter-degeneration resistors. This ruse powerfully improves input linearity, despite its rather unsettling flavour of something for nothing. The transistor nonlinearity can here be regarded as an internal nonlinear emitter resistance r_e, and what we have done is to reduce the value of this (by increasing I_c) and replace the missing part of it with a linear external resistor Re.

For a single device, the value of r_e can be approximated by:

$$r_e = 25/I_c \ \Omega \text{ (for } I_c \text{ in mA)} \qquad \text{Equation 4.3}$$

Our original stage at Figure 4.10a has a per-device I_c of 600 µA, giving a differential (i.e. mirrored) g_m of 23 mA/V and $r_e = 41.6\,\Omega$. The improved version in Figure 4.10b has $I_c = 1.35$ mA and so $r_e = 18.6\,\Omega$; therefore emitter degeneration resistors of $22\,\Omega$ are required to reduce the g_m back to its original value, as $18.6 + 22 = 40.6\,\Omega$, which is near enough. The distortion measured by the circuit of Figure 4.4 for a −40 dBu input voltage is reduced from 0.32% to 0.032%, which is an extremely valuable linearization, and will translate into a distortion reduction at HF of about five times for a complete amplifier; for reasons that will emerge later the full advantage is rarely gained. The distortion remains a visually pure third harmonic, so long as the input pair remains balanced. Clearly this sort of thing can only be pushed so far, as the reciprocal-law reduction of r_e is limited by practical values of tail current. A name for this technique seems to be lacking; *constant-*g_m *degeneration* is descriptive but rather a mouthful.

The standing current is roughly doubled so we have also gained a higher slew rate; it has theoretically increased from 10 to 20 V/µs, and once again we get two benefits for the price of one inexpensive modification.

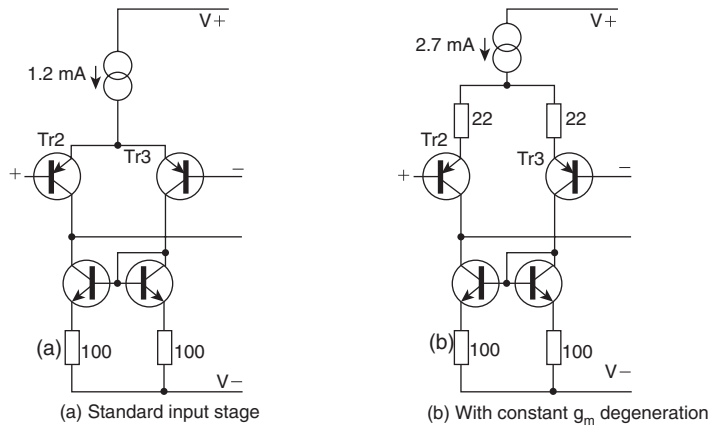


Figure 4.10: Input pairs before and after constant-g_m degeneration, showing how to double stage current while keeping transconductance constant; distortion is reduced by about 10 times

It is, however, not all benefit when we add emitter-degeneration resistors. The extra resistances will generate Johnson noise, increasing the total noise from the input stage. Differing values for the two resistors due to the usual tolerances will increase the input offset voltage. If the resistor matching is $\alpha\%$, the tail current is I_{tail}, and the degeneration resistors have the value R_e, the extra offset voltage V_{off} is given by:

$$V_{off} = \left(\frac{\alpha}{100}\right) \times \left(\frac{I_{tail} \times R_e}{2}\right) \qquad \text{Equation 4.4}$$

Thus for 100 Ω 1% resistors and a tail current of 6 mA, the extra offset voltage is 3 mV, which is small compared with the offsets due to the base currents flowing in the input and feedback resistances. This looks like one issue you need not worry about.

When a mirrored input stage is degenerated in this way, it is important to realize that its transconductance can only be very roughly estimated from the value of the emitter resistors. An input pair with a tail current of 4 mA and 22 Ω emitter resistors has a g_m of 25.6 mA/V, which represents an effective *V*/*I* conversion resistance of 39.0 Ω, the extra resistance being the internal r_e values of the transistors (remember that the input voltage is shared between two emitter resistors, apparently halving the current swing, but it is doubled again by the presence of the current-mirror). In this case the value of the emitter resistors gives a very poor estimate of the g_m. When 100 Ω emitter resistors are used with a tail current of 4 mA the g_m is 8.18 mA/V, representing an effective *V*/*I* conversion resistance of 122 Ω, which makes the estimate somewhat better but still more than 20% out. Increasing the tail current to 6 mA, which is the value used in the designs in this book, changes those values to 34.2 and 118 Ω, because the internal r_e values are reduced, but the estimates are still some way off. If more accurate figures are wanted at the design stage then SPICE simulation will usually be faster and better than manual calculation.

Further Improving Input Linearity

If we are seeking still better linearity, various techniques exist, but before deploying them we need to get a handle on the signal levels the input stage will be handling; the critical factor is the input voltage, by which I mean the differential voltage across the input stage – what might be called the error voltage created by the global feedback. Using Equation 3.2 (in Chapter 3) it is straightforward to work out the input voltage to the input stage for a given input stage g_m, C_{dom} value, and frequency; these parameters give us the open-loop gain, and we can then work back from the output voltage to the input voltage. The closed-loop gain and resulting global feedback factor are not involved except insofar as the feedback factor determines how much the input stage distortion is reduced when the loop is closed. However much the reduction by global feedback, an input stage that is twice as linear remains twice as linear.

Let us take an example typical of the designs in this book, with 100 Ω input degeneration resistors and a resulting g_m of 8.18 mA/V, and a C_{dom} of 100 pF. The worst-case frequency is 20 kHz, and we will assume a 50 W/8 Ω output level. This gives an input voltage of −28.3 dBu, which for the

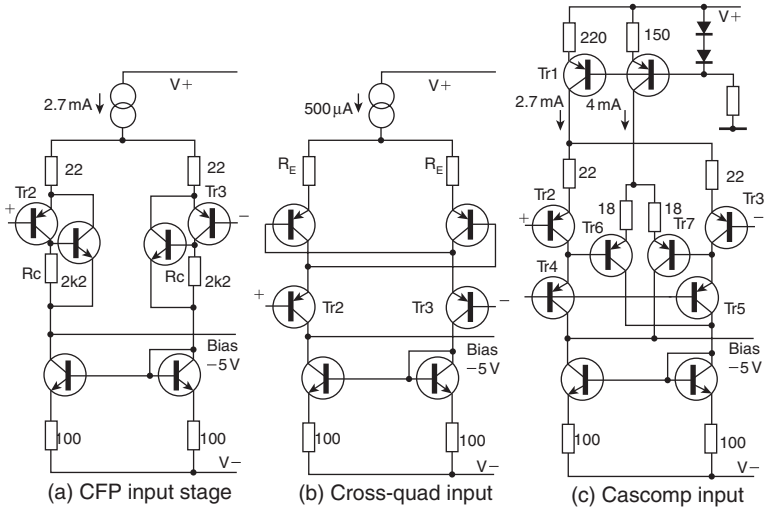


Figure 4.11: Some enhanced differential pairs. (a) The complementary feedback pair. (b) The cross-quad. (c) The cascomp

discussion below is rounded to −30 dBu. For a 100 W/8 Ω output level the input voltage would be −25.3 dBu.

Whenever it is necessary to increase the linearity of a circuit, it is often a good approach to increase the *local* feedback factor, because if this operates in a tight local NFB loop there is often little effect on the overall global-loop stability. A reliable method is to replace the input transistors with complementary feedback (CFP or Sziklai) pairs, as shown in the stage of Figure 4.11a. If an isolated input stage is measured using the test circuit of Figure 4.4, the constant-g_m degenerated version shown in Figure 4.10b yields 0.35% third-harmonic distortion for a −30 dBu input voltage, while the CFP version of Figure 4.11a gives 0.045%, a very valuable improvement of almost eight times. (Note that the input level here is 10 dB up on the −40 dBu input level used for the example in the previous section, which is both more realistic and gets the distortion well clear of the noise floor.) When this stage is put to work in a model amplifier, the third-harmonic distortion at a given frequency is roughly halved, assuming all other distortion sources have been appropriately minimized; the reason for the discrepancy is not currently known. However, given the high slope of input stage distortion, this only extends the low-distortion regime up in frequency by less than an octave (see Figure 4.12).

A compromise is required in the CFP circuit on the value of R_c, which sets the proportion of the standing current that goes through the NPN and PNP devices on each side of the stage. A higher value of R_c gives better linearity (see Table 4.3 for more details on this) but potentially more noise, due to the lower collector current in the NPN devices that are the inputs of the input stage, as it were, causing them to perform less well with the relatively low source resistances. 2k2 seems to be a good compromise value for R_c; it gives a collector current of 320 µA.

Several other elaborations of the basic input pair are possible, although almost unknown in the audio community. We are lucky in power-amp design as we can tolerate a restricted input

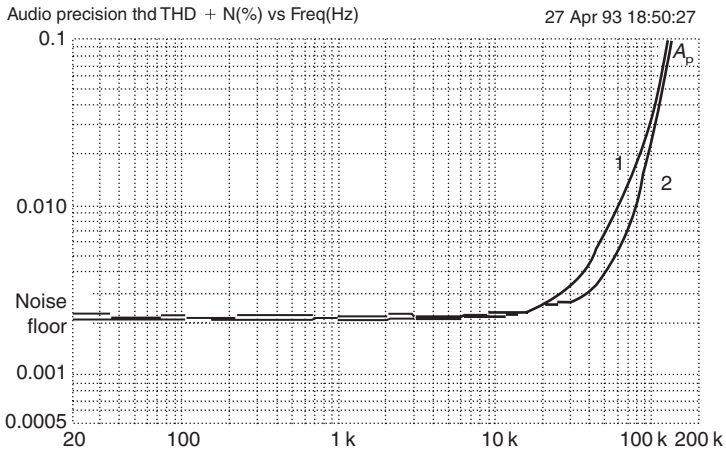


Figure 4.12: Whole-amplifier THD with normal and CFP input stages; input stage distortion only shows above noise floor at 20 kHz, so improvement occurs above this frequency. The noise floor appears high as the measurement bandwidth is 500 kHz

Table 4.3: Summary of measured input stage linearity

Type	Input level (dBu)	R_{degen} (Ω)	THD (%)	Notes	Figure
Simple	−40	0	0.32		4.10a
Simple	−40	22	0.032		4.10b
Simple	−30	22	0.35		4.10b
CFP	−30	22	0.045	R_c = 2k2	4.11a
CFP	−30	39	0.058	R_c = 1k5	4.11a
CFP	−30	39	0.039	R_c = 2k2	4.11a
CFP	−30	39	0.026	R_c = 4k7	4.11a
CFP	−30	39	0.022	R_c = 10k	4.11a
Cascomp	−30	50	0.016		4.11c

common-mode range that would be unusable in an op-amp, giving the designer great scope. Complexity in itself is not a serious disadvantage as the small-signal stages of the typical amplifier are of almost negligible cost compared with mains transformers, heat-sinks, etc.

Two established methods to produce a linear input transconductance stage (often referred to in op-amp literature simply as a transconductor) are the cross-quad[8] and the cascomp[9] configurations. The cross-quad input stage (Figure 4.11b) works by imposing the input voltage to each half across two base–emitter junctions in series, one in each arm of the circuit. In theory the errors due to nonlinear r_e of the transistors is divided by beta, but in practice the reduction in distortion is modest. The cross-quad nonetheless gives a useful reduction in input distortion when operated in isolation, but is hard to incorporate in a practical amplifier because it relies on very low source resistances to tame the negative conductances inherent in its operation. If you just drop it into a normal power amplifier circuit with the usual source resistances in the input and feedback arms it will promptly latch up, with one side or the other turning hard on. This does not seem like

a good start to an amplifier design, despite the seductive simplicity of the circuit, and with some lingering regret it will not be considered further here.

The cascomp (Figure 4.11c) does not have problems with negative impedances, but it is significantly more complicated and significantly more complex to design. TR2, TR3 are the main input pair as before, delivering current through cascode transistors TR4, TR5 (this cascoding does not in itself affect linearity), which, since they carry almost the same current as TR2, TR3, duplicate the input V_{be} errors at their emitters. These error voltages are sensed by error diff-amp TR6, TR7, whose output currents are summed with the main output in the correct phase for error correction. By careful optimization of the (many) circuit variables, distortion at −30 dBu input can be reduced to about 0.016% with the circuit values shown, which handily beats the intractable cross-quad. Sadly, this effort provides very little further improvement in whole-amplifier HF distortion over the simpler CFP input, as other distortion mechanisms are coming into play, one of which is the finite ability of the VAS to source current into the other end of C_{dom}.

Table 4.3 summarizes the performance of the various types of input stage.

Increasing the Output Capability

The standing current in the input pair tail is one of the parameters that defines the maximum slew rate, the other being the size of the dominant-pole Miller capacitor. The value of this capacitor is usually fixed by the requirements of stability, but increasing the tail current can increase slew rate without directly affecting stability so long as the degeneration resistors are adjusted to keep the input stage transconductance at the desired value.

Unfortunately there are limits to how much this current can be increased; the input bias currents increase, as do the voltage drops across the degeneration resistors, and both these factors increase the spread of DC offset voltage. The ultimate limit is of course the power dissipation in the input stage; if you take a 6 mA tail current, which is the value I commonly use, and ±50 V supply rails, the dissipation in each input transistor is 150 mW to a close approximation, and there is clearly not a vast amount of scope for increasing this. There is also the point that hot input devices are more susceptible to stray air currents and therefore we can expect more drift.

Op-amp designers face the same problems, exacerbated by the need to keep currents and dissipations to a much lower level than those permissible in a power amplifier. Much ingenuity has therefore been expended in devising input stages that do not work in Class-A, like the standard differential pair, but operate in what might be called Class-AB; they have a linear region for normal input levels, but can turn on much more than the standing current when faced with large inputs. Typically there is an abrupt change in transconductance and linearity is much degraded as the input stage enters the high-current mode. The first input stage of this type was designed by W.E. Hearn in 1970, and it appeared in the Signetics NE531 op-amp[10]. Another such stage was put forward by Van de Plassche[11]. Both types have been used successfully in the standard three-stage architecture by Giovanni Stochino[12].

The rest of this chapter deals only with the standard input differential amplifier.

Input Stage Cascode Configurations

Cascoding is the addition of a common-base stage to the collector of a common-emitter amplifier, to prevent the stage output from affecting the common-emitter stage, or to define its operating collector voltage. The word is a contraction of 'cascade to cathode', which tells you at once that, like so many circuit techniques, it dates back to the valve era[13]. It can often be usefully applied to the standard input differential amplifier. The basic principle of a cascode amplifier stage is shown in Figure 4.13a. There is a common-emitter amplifier Q1, directly coupled to a common-base stage Q2. The common-base stage gives no increase in the transconductance of the overall stage, as it simply passes the collector current of Q1 onto the current-source collector load I1, less a small amount that is the base current of Q2. The important job that it does do is to hold the collector of Q1 at a substantially constant voltage; the voltage of biasing voltage source V1, minus the V_{be} voltage of Q2. This constant collector voltage for Q1 gives two benefits; the frequency response of the stage is improved because there is no longer local negative feedback through the collector-base capacitance of Q1, and the stage gain is potentially both greater and more linear because the Early effect (the modulation of collector current I_c by V_{ce}) can no longer occur in Q1. The V_{ce} of Q1 is now both lower and constant – a V_{ce} of 5 V is usually quite enough – and the consequent reduction of heating in Q1 can have indirect benefits in reducing thermal drift. This configuration will be met with again in the chapter on voltage-amplifier stages.

When the cascoding principle is applied to the input stage of a power amplifier, we get the configuration shown in Figure 4.13b, with the DC conditions indicated. The circuit is inverted compared with the single-transistor example so it corresponds with the other input stages in this book. If the bases of the input devices Q1, Q2 are at 0 V, which is usually the case, their collectors need to be held at something like −5 V for correct operation.

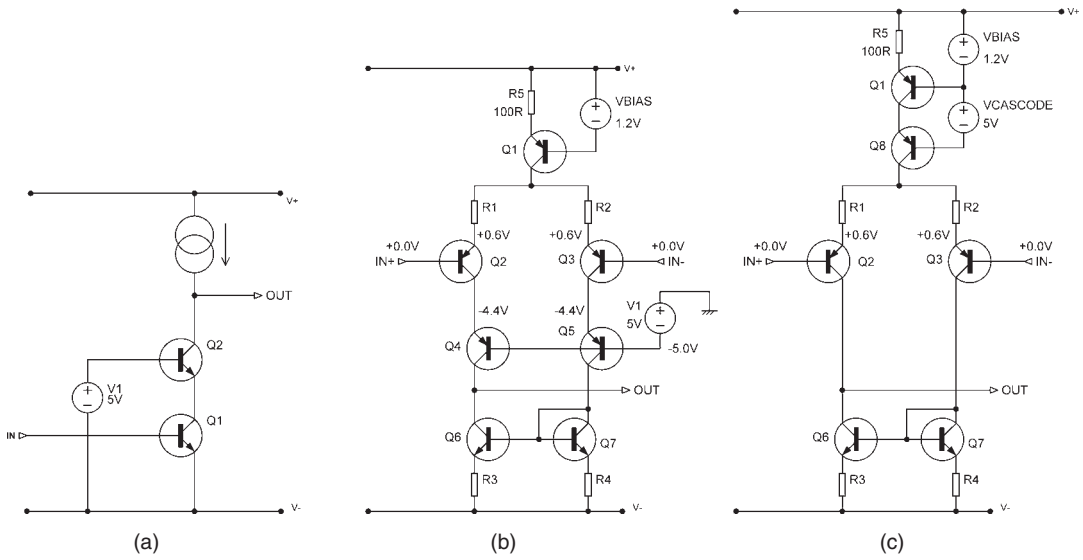


Figure 4.13: Cascode configurations. (a) The basic cascode concept. (b) Cascoding applied to the input devices of a differential input stage, with DC conditions shown. (c) Cascoding applied to the tail-current source

Cascoding an input stage does nothing to improve the linearity of the stage itself, as there is no appreciable voltage swing on the input device collectors due to the low-impedance current input of a typical VAS stage; it can, however, in some circumstances reduce input current distortion as it allows high-beta, low-V_{ce} input devices to be used. See later in this chapter, in the section on input current distortion, where it is shown that sometimes there are real benefits in hum rejection to be obtained by cascoding the input pair tail current source, as shown in Figure 4.13c. In specialized circumstances, for example where the closed-loop gain of the amplifier is lower than usual, cascoding the input stage can actually make linearity worse (see the section on input stage common-mode distortion below).

Isolating the input device collector capacitance from the VAS input sometimes allows C_{dom} to be slightly reduced for the same stability margins, but the improvement is marginal. A more significant advantage is the reduction of the high V_{ce} that the input devices work at. This allows them to run cooler, and so be less susceptible to drift caused by air currents. This is dealt with in more detail in the chapter on DC servos.

Double Input Stages

Two input stages, one the complement of the other, are quite often used to drive both the top and bottom of a push–pull VAS (see Figure 5.11 and following in Chapter 5). Their operation is just the same as for a single input stage, and both emitter degeneration and the use of current-mirrors are recommended as before. If the input bases are connected directly together, as is usual, there may be some cancelation of input currents (see the section later in this chapter on why this may be important) but this cannot be relied upon much because of the poor beta-matching of discrete transistors, and those of differing polarity at that. Chapter 5 includes an interesting example of a series input stage using one NPN and one PNP transistor (see Figure 5.14).

The use of double input stages should give 3 dB less noise due to arithmetical summing of the signals but rms-summing of the input stage noise, but I haven't had the opportunity to test this myself.

Input Stage Common-Mode Distortion

This does not appear to exist at detectable levels in normal amplifier circuitry, and by this I mean I am assuming that the input stage has emitter-degeneration resistors and a current-mirror, as previously described. A much higher common-mode (CM) voltage on the input stage than normally exists is required to produce a measurable amount of distortion. If an amplifier is operated at a low closed-loop gain such as one or two times, so that both input and feedback signals are much larger than usual, this puts a large CM voltage on the input stage, and distortion at HF is unexpectedly high, despite the much increased NFB factor. This distortion is mainly second harmonic. The immediate cause is clearly the increased CM voltage on the input devices, but the exact mechanism is at present unclear.

Table 4.4: How amplifier distortion varies as the common-mode voltage is altered

Closed-loop gain (×)	CM voltage (V rms)	15 kHz THD measured (%)	15 kHz THD calculated (%)
1.00	10.00	0.0112	0.00871
1.22	8.20	0.00602	0.00585
1.47	**6.81**	**0.00404**	**0.00404**
2.00	5.00	0.00220	0.00218
23	0.43	–	0.000017

Table 4.4 shows distortion increasing as closed-loop gain is reduced, with input increased to keep the output level constant at 10 V rms. A model amplifier (i.e. one with the output stage replaced by a small-signal Class-A stage, as described in Chapter 3) was used because the extra phase shift of a normal output stage would have made stability impossible to obtain at such low closed-loop gains; the basic circuit without any input stage modifications is shown in Figure 4.14. This is an excellent illustration of the use of a model amplifier to investigate input stage distortions without the extra complications of a Class-B output stage driving a load. For some reason long forgotten, NPN input devices were used, so the diagrams appear upside-down relative to most of the input stages in this book.

This version of a model amplifier has a couple of points of interest; you will note that the input degeneration resistors R2, R3 have been increased from the usual value of 100R to 220R to help achieve stability by reducing the open-loop gain. The output stage is a push–pull Class-A configuration, which has twice the drive capability than the usual constant-current version, chosen so it could drive the relatively low-value feedback resistance R5 (needed to keep the noise down as the lower feedback arm is higher in value than usual because of the low gain) without an increase in distortion; there was no other load on the output stage apart from the distortion analyzer. I have used this configuration extensively in the past in discrete-component preamplifiers, for example in Ref. [14]. It is very linear, because the push–pull action halves the current swing in the emitter-follower, and thoroughly stable and dependable, but does require regulated supply rails to work properly.

Tests were done at 10 V rms output and data taken at 15 kHz, so the falling global NFB factor with frequency allowed the distortion to be far enough above the noise floor for accurate measurement. The closed-loop gain was altered by changing the lower feedback arm Rfb2. The THD plots can be seen in Figure 4.15, which provided the data for Table 4.4. It can be seen that the distortion goes up at 6 dB/octave.

It appears THD is proportional to CM voltage squared. Taking the measured THD at a closed-loop gain of 1.47× as a reference, scaling it by the square of the gain gives the figures in the rightmost column, which correspond quite nicely with the measured THD figures. Some extra higher-order distortion was coming in at a closed-loop gain of 1.00, so the square law is less accurate there.

Thus, assuming the square law, the THD at 1.47 times gain (0.00404%), when scaled down for a more realistic closed-loop gain of 23, is reduced by a factor of $(23/1.47)^2 = 245$, giving

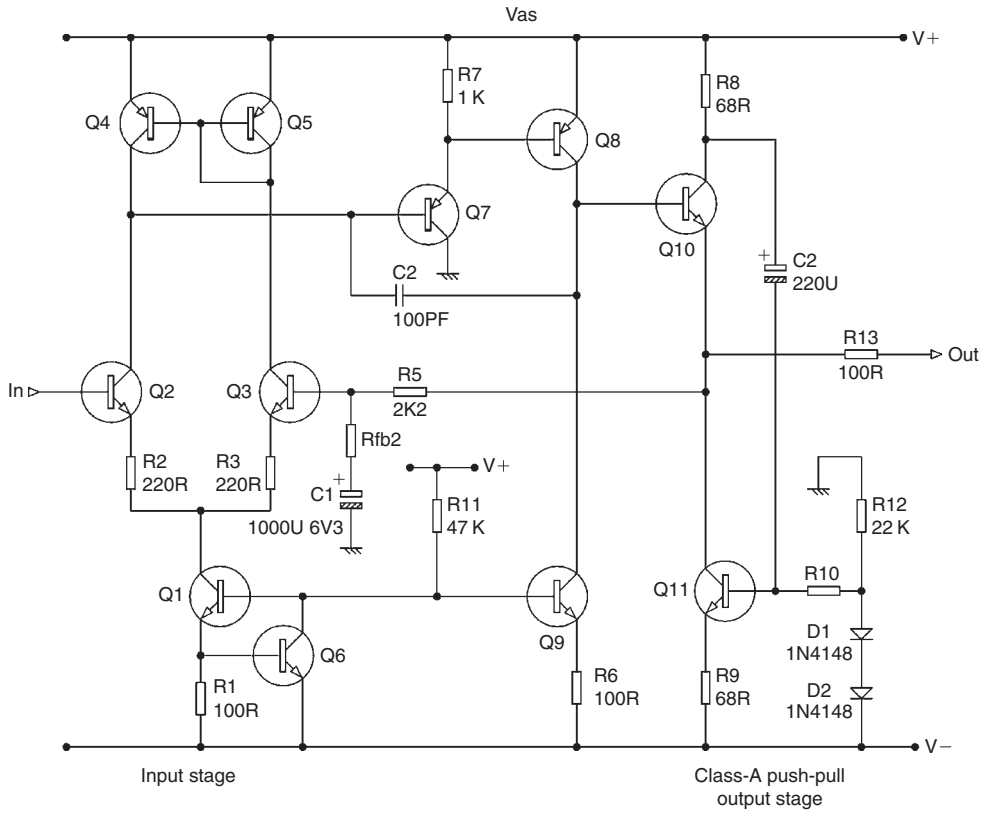


Figure 4.14: A model amplifier (output stage replaced by a small-signal Class-A stage) with low closed-loop gain

a negligible 0.000017% at 15 kHz. In terms of practical amplifier design, there are other things to worry about.

And yet … I was curious as to the actual distortion mechanism, and decided to probe deeper, without any expectation that the answer would be directly useful in amplifier design until we had made a lot more progress in other areas of nonlinearity; however, there was always the possibility that the knowledge gained would be applicable to other problems, and it would certainly come in handy if it was necessary to design a low-gain power amplifier for some reason. Giovanni Stochino and I therefore investigated this issue back in 1996, and at the end of a lot of thought and international faxing, I felt I could put down the following statements:

- Reducing the V_{ce} of the input devices by inserting capacitively decoupled resistors into the collector circuits, as shown in Figure 4.16a, makes the CM distortion worse. Altering the V+ supply rail (assuming NPN input devices) has a similar effect; less V_{ce} means more distortion. V_{ce} has a powerful effect on the HF THD. This seems to indicate that the nonlinearity is due to either Early effect (an increase in the effective beta of a BJT as the V_{ce} increases, due to narrowing of the effective base width) or the modulation of V_{bc},

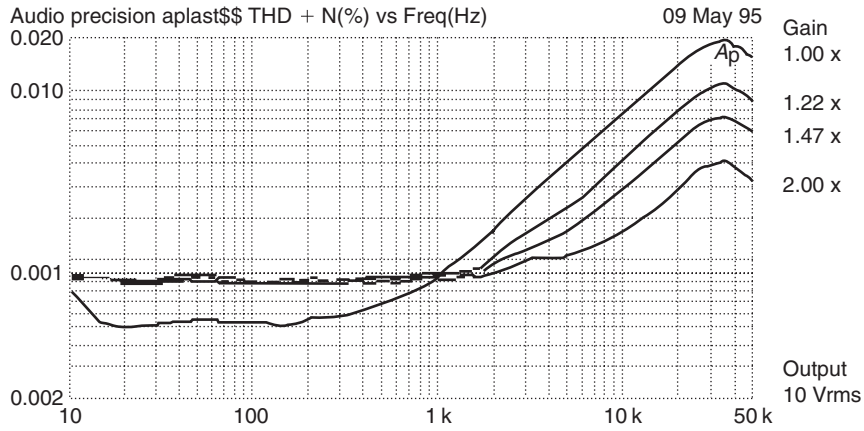


Figure 4.15: THD plots from model amplifier with various low closed-loop gains. Output 10 V rms

the base-collector capacitance, or very possibly a mixture of the two. Very little seems to have been published on these sort of nonlinearities; but Taylor[15] is well worth reading.

- The effect cannot be found in SPICE simulation, which is a bit disconcerting. It must therefore originate either in imbalances in transistor parameters, which do not exist in SPICE unless you put them in, *or* in second-order effects that are not modeled by SPICE. A particular suspect here is the fact that SPICE models the Early effect as linear with V_{ce}.
  I was told by Edward Cherry[16] that SPICE would need to include the second-order term in the Early voltage model, rather than use a linear law, before the effect could be simulated.
- The HF distortion does not alter *at all* when the input devices are changed, so the THD mechanism cannot depend on beta or Early voltage, as these would vary between device samples. I know this finding makes little sense, but I checked it several times, always with the same result.

Giovanni and I therefore concluded that if the problem is due to the Early effect, it should be possible to eliminate it by cascoding the input device collectors and driving the cascode bases with a suitable CM voltage so that the input device V_{ce} remains constant. I tried this, and found that if the CM voltage was derived from the amplifier output, via a variable attenuator, it allowed only partial nulling of the distortion. With a CM voltage of 6.81 V rms, the drive to the cascode for best second-harmonic nulling was only 131 mV rms, which made little sense.

A more effective means of reducing the common-mode nonlinearity was suggested to me by Giovanni Stochino[17]. Driving the input cascode bases directly from the input tail, rather than an output-derived signal, completely eliminates the HF distortion effect. It is not completely established as to why this works so much better than driving a bootstrap signal from the output, but Giovanni feels that it is because the output signal is phase-shifted compared with the input, and I suspect he is right. At any rate it is now possible to make a low-gain amplifier with very, very low HF distortion, which is rather pleasing. The bootstrapped cascode input configuration is shown in Figure 4.16b, and the impressive THD results are plotted in Figure 4.17.

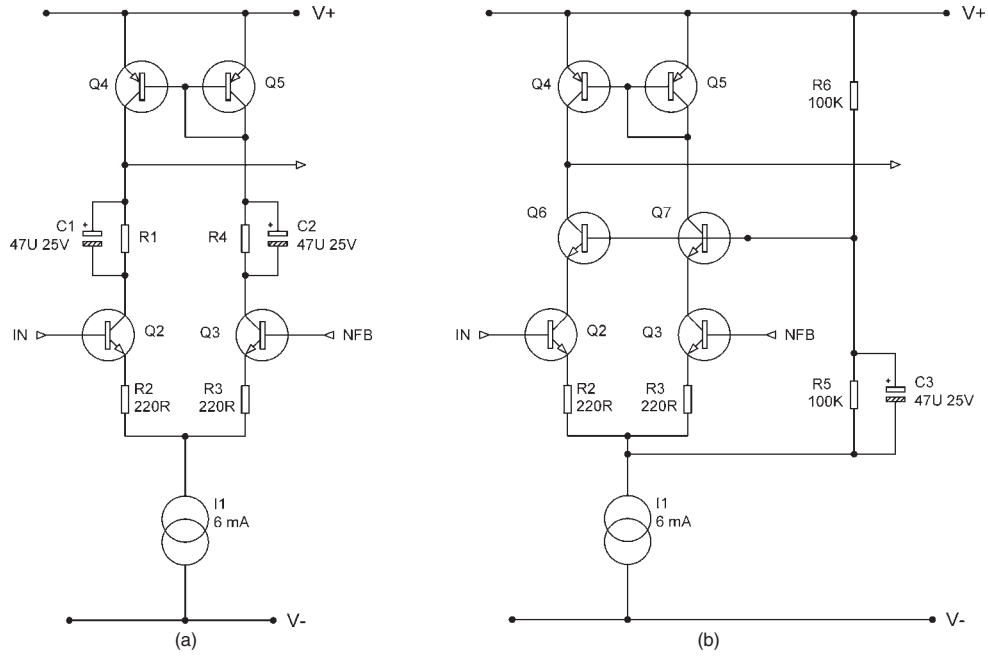


Figure 4.16: (a) Method of reducing input device V_{ce}. (b) A method of driving the bases of an input cascode structure directly from the input tail

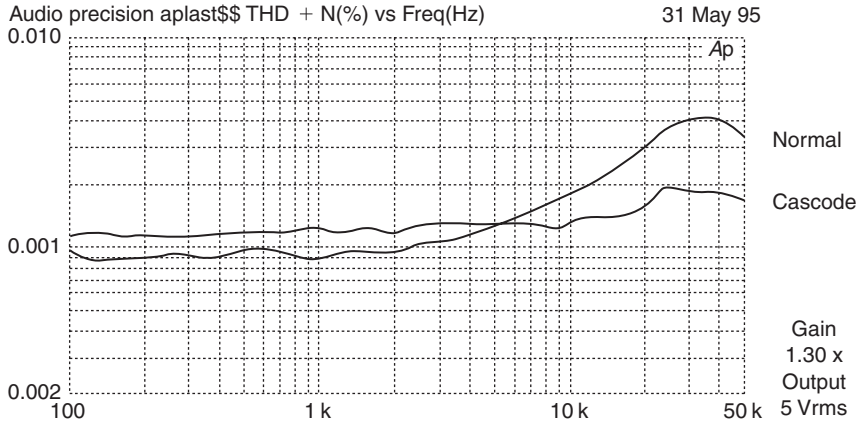


Figure 4.17: Showing how the input cascode completely eliminates the HF distortion effect. Output 5 V rms

And there, for the moment, the matter rests. If the closed-loop gain of your amplifier is low, you need to worry about common-mode distortion, but there is a fix. However, in most cases it is too low to worry about.

Input Current Distortion

When power amplifiers are measured, the input is normally driven from a low-impedance signal generator. Some test gear, such as the much-loved but now obsolete Audio Precision System-1, has

selectable output impedance options of 50, 150, and 600 Ω. The lowest value available is almost invariably used because:

1. it minimizes the Johnson noise from the source resistance;
2. it minimizes level changes due to loading by the amplifier input impedance;
3. it minimizes the possibility of hum, etc. being picked by the input.

This is all very sensible, and exactly the way I do it myself – 99% of the time. There are, however, two subtle effects that can be missed if the amplifier is always tested this way. These are:

1. distortion caused by the nonlinear input currents drawn by the typical amplifier;
2. hum caused by ripple modulation of the same input currents.

Note that (1) is not the same effect as the excess distortion produced by FET-input op-amps when driven from significant source impedances; this is due to their nonlinear input capacitances to the IC substrate, and has no equivalent in power amplifiers made of discrete transistors.

Figure 4.18 shows both the effects. The amplifier under test was a conventional Blameless design with an EF output stage comprising a single pair of sustained-beta bipolar power transistors; the circuit can be seen in Figure 4.19. The output power was 50 W into 8 Ω. The bottom trace is the distortion + noise with the usual source impedance of 50 Ω, and the top one shows how much worse the THD is with a source impedance of 3.9 k. The intermediate traces are for 2.2 k and 1.1 k source resistances. The THD residual shows both second-harmonic distortion and 100 Hz ripple components, the ripple dominating at low frequencies, while at higher frequencies the distortion dominates. The presence of ripple is signaled by the dip in the top trace at 100 Hz, where distortion products and ripple have partially canceled, and the distortion analyzer has settled on the

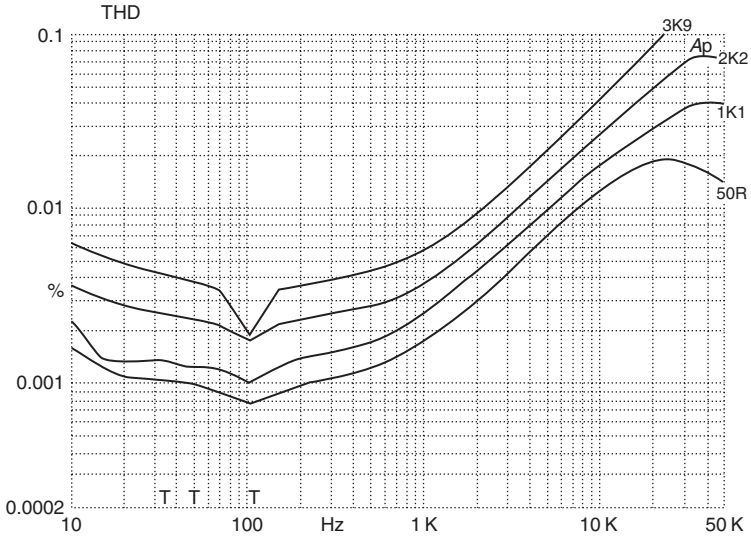


Figure 4.18: Second-harmonic distortion and 100 Hz ripple get worse as the source impedance rises from 50 Ω to 3.9 k; 50 W into 8 Ω

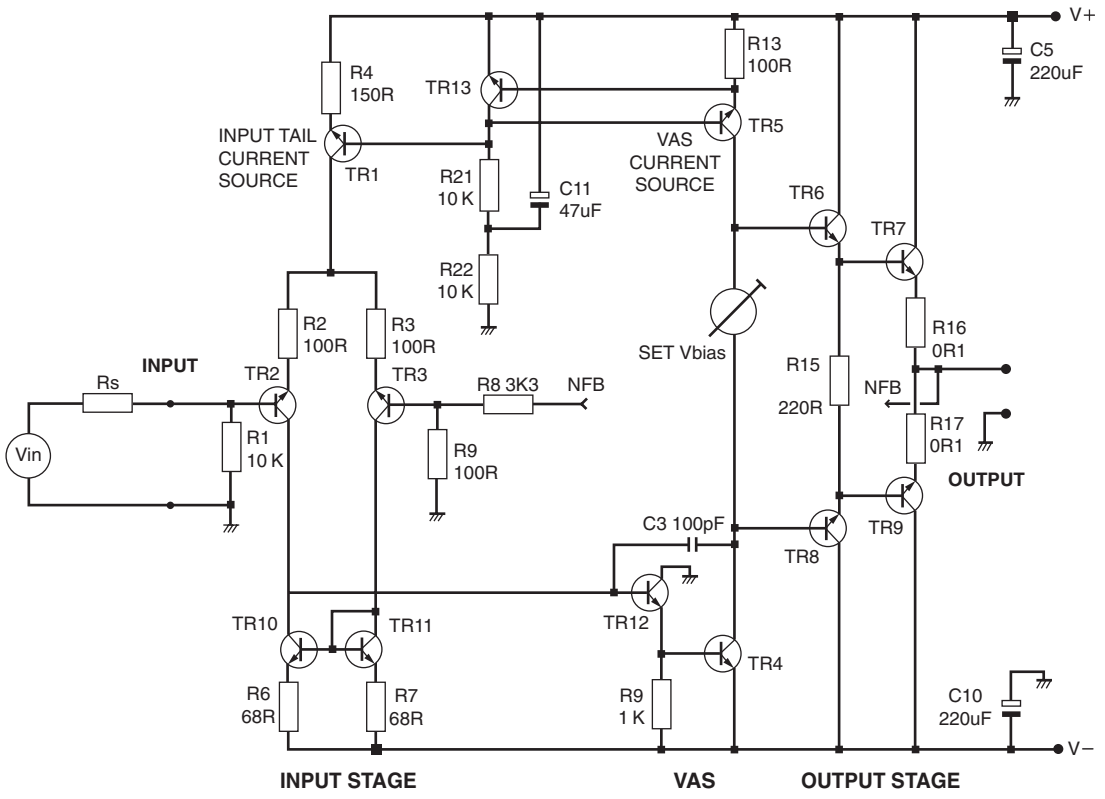


Figure 4.19: Simplified circuit of a typical Blameless power amplifier, with negative-feedback control of VAS current source TR5 by TR13. The bias voltage generated is also used by the input tail source TR1

minimum reading. The amount of degradation from both ripple and distortion is proportional to the source impedance.

The input currents are not a problem in many cases, where the preamplifier is driven by an active preamplifier, or by a buffer internal to the power amplifier. Competent active preamplifiers have a low output impedance, often around 50–100 Ω and sometimes less – there are no great technical difficulties involved in reducing it to a few ohms. This is to minimize high-frequency losses in cable capacitance. (I have just been hearing of a system with 10 meters of cable between pre-amp and power amp.) However, some active designs seem to take this issue less seriously than they should and active pre-amp output impedances of up to 1 k are not unknown. To the best of my knowledge pre-amp output impedances have never been made deliberately low to minimize power amplifier input current distortion, but it would certainly be no bad thing.

There are two scenarios where the input source resistance is considerably higher than the desirable 50–100 Ω. If a so-called 'passive pre-amp' is used then the output impedance is both much higher and volume-setting dependent. A 10 k volume potentiometer, which is the lowest value likely to be practical if the loading on source equipment is to be kept low, has a maximum output impedance of one-quarter the track resistance, i.e. 2.5 k, at its mid-point setting.

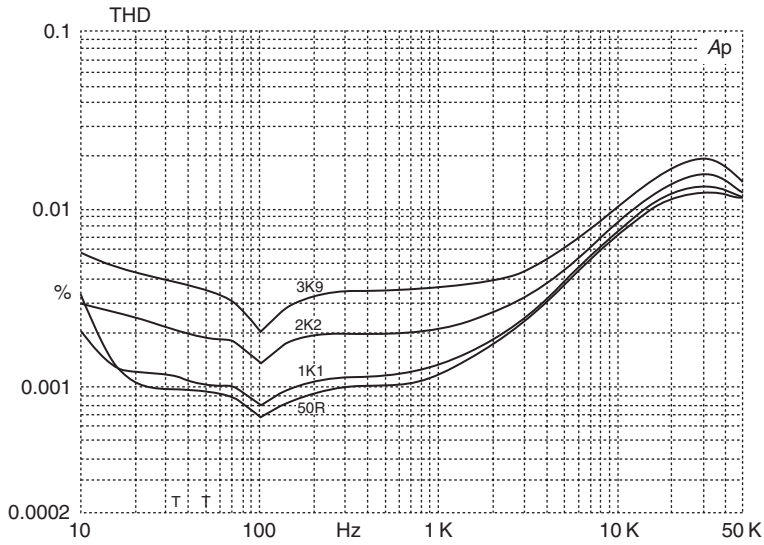


Figure 4.20: There is less introduction of ripple and distortion with high-beta input transistors and the same set of source resistances as in Figure 4.17

It is also possible for significant source resistance to exist inside the power amplifier unit – for example, there might be an balanced input amplifier that, while it has a very low output impedance itself, may have a resistive gain control network between it and the power amp. The value of this potentiometer is not likely to be less than 5 k, and is more likely to be 10 k, so once again we are faced with a maximum 2.5 k source resistance at the mid-point setting. (Assuming the input amplifier is a 5532 or equally capable op-amp, there would be no difficulty in driving a 2 k or even a 1 k pot without its loading introducing measurable extra distortion; this would reduce the source resistance and also the Johnson noise generated.) However, I digress (more on amplifier input circuitry in Chapter 20).

So, we have a problem, or rather two of them, in the form of extra ripple and extra distortion, and the first step to curing it is to understand the mechanisms involved. Since the problems get worse in proportion to the source impedance, it seems very likely that the input transistor base currents are directly to blame for both, so an obvious option is to minimize these currents by using transistors with the highest available beta in the input pair. In this amplifier the input pair were originally ZTX753, with a beta range of 70–200. Replacing these with BC556B input devices (beta range 180–460) gives the result in Figure 4.20, which shows a useful improvement in THD above 1 kHz; distortion at 10 kHz drops from 0.04% to 0.01%. Our theory that the base currents are to blame is clearly correct. The bottom trace is the reference 50 Ω source plot with the original ZTX753s, and the gap between this and our new result demonstrates that the problem has been reduced but certainly not eliminated.

The power amplifier used for the experiments here is very linear when fed from a low source impedance, and it might well be questioned as to why the input currents drawn are distorted if the output is beautifully distortion-free. The reason is of course that global negative feedback constrains the output to be linear – because this is where the NFB is taken from – although the internal signals of the amplifier are not necessarily linear, but whatever is required to keep

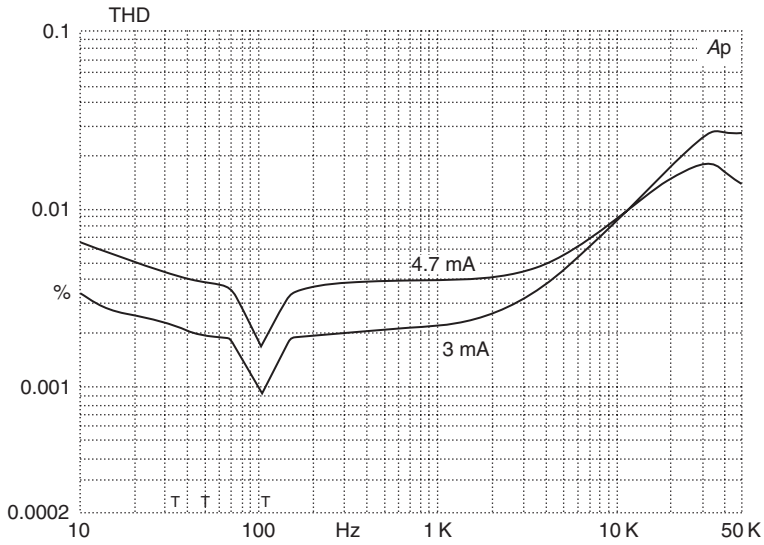


Figure 4.21: Reducing the tail current improves things at low frequencies but increases HF distortion above 10 kHz. The notches at 100 Hz indicate that the ripple content is still substantial

the output linear. The VAS is known to be nonlinear, so if the amplifier output is sinusoidal the collector currents of the input pair clearly are not. Even if they were, the beta of the input transistors is not constant so the base currents drawn by them would still be nonlinear.

It is also possible to get a reduction in hum and distortion by reducing the input pair tail current, but this very important parameter also affects input stage linearity and the slew rate of the whole amplifier. Figure 4.21 shows the result. The problem is reduced – though far from eliminated – but the high-frequency THD has actually got worse because of poorer linearity in the input stage. This is not a promising route to follow: no matter how much the tail current is reduced the problem will not be eliminated.

Both the ripple and THD effects consequent on the base currents drawn could be eliminated by using FETs instead of bipolars in the input stage. The drawbacks are:

1. Poor V_{gs} matching, which means that a DC servo becomes essential to control the amplifier output DC offset. Dual FETs do exist but they are discouragingly expensive.
2. Low transconductance, which means the stage cannot be linearized by local feedback as the raw gain is just not available.
3. Although there will be negligible DC gate currents, there might well be problems with nonlinear input capacitance, as there are with FET-input op-amps.

Once again, this is not a promising route; the use of FETs will create more problems than it solves.

The distortion problem looks rather intractable; one possible total cure is to put a unity-gain buffer between input and amplifier. The snag (for those seeking the highest possible performance) is that

any op-amp will compromise the noise and distortion of a Blameless amplifier. It is quite correct to argue that this does not matter, as any pre-amp hooked up to the power amp will have op-amps in it anyway, but the pre-amp is a different box, a different project, and possibly has a different designer, so philosophically this does not appeal to everyone. If a balanced input is required then an op-amp stage is mandatory (unless you prefer transformers, which of course have their own problems). The best choice for the op-amp is either the commonplace but extremely capable 5532 (which is pretty much distortion-free, but not alas noise-free, though it is very quiet) or the very expensive but very quiet AD797. A relatively new alternative is the LM4562, which has lower noise than a 5532, but at present they are a good deal more expensive.

The ripple problem, however, has a more elegant solution. If there is ripple in the input base current, then clearly there is some ripple in the tail current. This is not normally detectable because the balanced nature of the input stage cancels it out. A significant input source impedance upsets this balance, and the ripple appears. The tail is fed from a simple constant-current source TR1, and this is clearly not a mathematically perfect circuit element. Investigation showed that the cause of the tail-current ripple contamination is Early effect in this transistor, which is effectively fed with a constant bias voltage A tapped off from the VAS negative-feedback current source; the problem is *not* due to ripple in the bias voltage. (Early effect is the modulation of transistor collector current caused by changing the V_{ce}; as a relatively minor aspect of bipolar transistor behavior it is modeled by SPICE simulators in a rather simplistic way, by assuming a linear V_{ce}/I_c relationship.) Note that this kind of negative-feedback current source could control the tail current instead of the VAS current, which might well reduce the ripple problem, but the biasing system is arranged this way as it gives faster positive slewing. Another option is two separate negative-feedback current sources.

The root cause of our hum problem is therefore the modulation of the V_{ce} of TR1 by ripple on the positive rail, and this variation is easily eliminated by cascoding, as shown in Figure 4.22. This forces TR1 emitter and collector to move up and down together, preventing V_{ce} variations. It completely eradicates the ripple components, but leaves the input-current distortion unaltered, giving the results in Figure 4.23, where the upper trace is now degraded only by the extra distortion introduced by a 2 k source impedance; note that the 100 Hz cancelation notch has disappeared. The reference 50 Ω source plot is below it.

The voltage at A that determines the V_{ce} of TR1 is not critical. It must be sufficiently below the positive supply rail for TR1 to have enough V_{ce} to conduct properly, and it must be sufficiently above ground to give the input pair enough common-mode range. I usually split the biasing chain R21, R22 in half, as shown, so C11 is working with the maximum resistance to filter out rail noise and ripple, and biasing the cascode transistor from the mid-point works very well. Note that this is preferable to biasing the cascode transistor with a fixed voltage (e.g. from a Zener diode) for a non-obvious reason. It means that an untried amplifier will start up earlier when you are cautiously increasing the supply rail voltages by nervous manipulation of a variable transformer, and the earlier it starts the less damage will be done if there is something wrong.

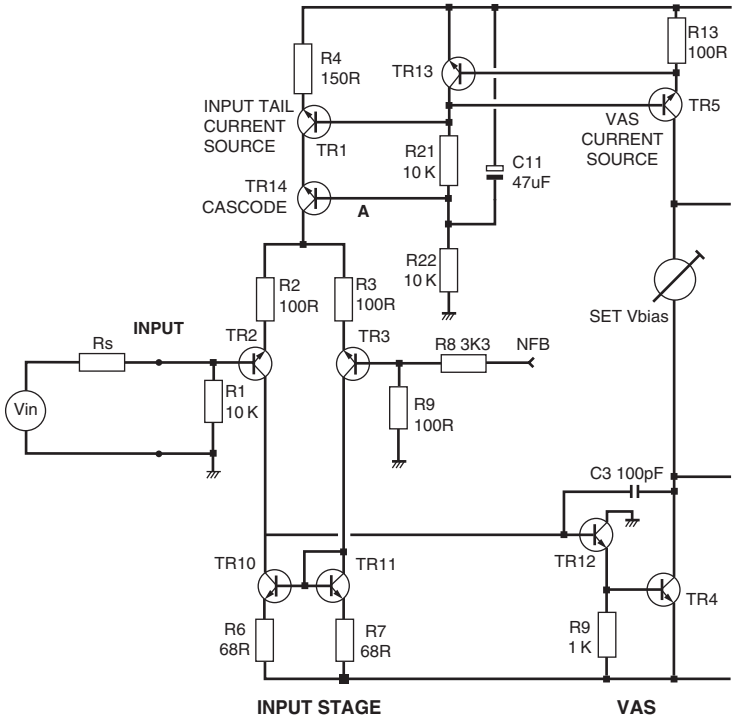


Figure 4.22: Cascoding the input tail – one method of biasing the cascode

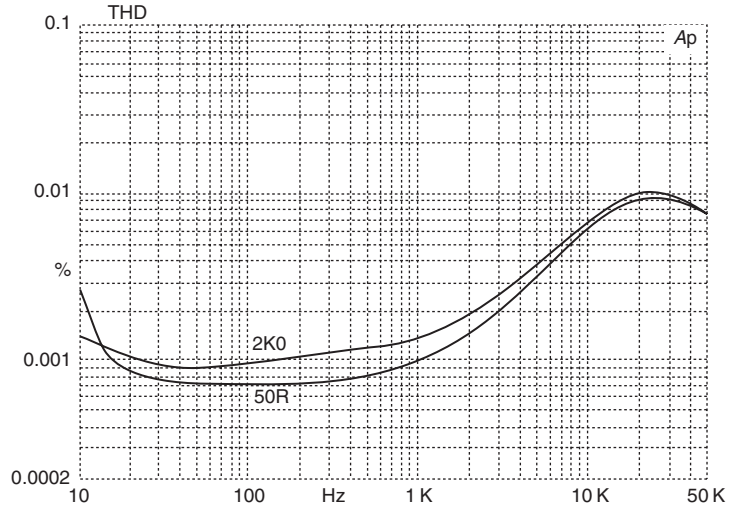


Figure 4.23: Cascoding the input tail removes the ripple problem, but not the extra distortion

An alternative, though rather less elegant, approach to preventing ripple injection is simply to smooth the positive rail with an RC filter before applying it to the tail-current source. The resulting voltage drop in the R part means that a separate tail-current source-biasing voltage must now be generated, and the C will have to be a high-voltage component as it has to withstand almost all the positive rail voltage.

At the end of the day the cascode approach will probably be cheaper as well as more elegant. And you can always put 'cascoded input stage!' in your publicity material.

It may have occurred to the reader that simply balancing the impedances seen by the two inputs will cancel out the unwanted noise and distortion. This is not very practical as with discrete transistors there is no guarantee that the two input devices will have the same beta. (I know there are such things as dual bipolars, but once more the cost is depressing.) This also implies that the feedback network will have to have its impedance raised to equal that at the input, which would give unnecessarily high levels of Johnson noise. It is of course impossible where the source resistance is variable, as when the amplifier is being fed from a volume-control potentiometer.

Another line of enquiry is canceling out the input current by applying an equal and opposite current, generated elsewhere in the input stage, to the input. This kind of stratagem is popular in some BJT-input op-amps, where it is called 'input bias-current cancelation'; it is hard to see how to apply it to an input stage made with discrete transistors because creating the cancelation currents relies on having closely matched betas in all the devices. Even if it were possible, there would almost certainly be a penalty in the shape of increased noise. Op-amps such as the OP27, which has input bias cancelation, have gained a certain notoriety for giving disappointing noise results. At first sight it appears that the OP27 is quieter than the 5534/5532; its e_n is 3.2 nV/rtHz compared with 4 nV/rtHz for the 5534. However, on practical measurement, the OP27 is often slightly noisier, and this is believed to be because the OP27 input bias-current circuitry generates common-mode noise. When the impedances on the two inputs are equal all is well, but when they are different the common-mode noise does not cancel, and this effect seems to be enough to degrade the noise performance significantly. If you want to pursue the matter of input bias cancelation further (and it has to be said that some of the circuitry is most ingenious and well worth studying), a good reference is Dostal[18].

Since neither of these approaches look very promising, what else can be done? It seems likely that the CFP input stage described earlier in this chapter would give lower values of input current distortion, as the base currents of the NPN transistors that can potentially flow in external source resistances (or, indeed, the feedback network source impedance) are much lower. A simple differential pair draws an input current from 0 to 49 μA over the input voltage range (from SPICE using MPSA42/MPSA92, tail current 6 mA) while the CFP draws 0–5.3 μA. I have not yet assessed the comparative linearity of the two currents but it looks as though there might be an order of magnitude improvement here.

The discussion above has focused on the effects of a significant source impedance at the input to the power amplifier. But a power amplifier, like an op-amp, has two inputs, and that not used for the signal input is used for the feedback connection. The current that this input draws from the feedback network will also lead to extra distortion, by exactly the same mechanism. If the feedback network consisted of, say, a 47 k upper arm and a 2 k2 lower arm, giving a closed-loop gain of 22.4 times, the source impedance seen by the input will be 2 k1, and we can expect to see some serious extra distortion, as shown in Figure 4.11a above. This is an important point; if this problem exists in an amplifier design, then no amount of work that attempts to improve the linearity of input stage

transconductance or the VAS will improve matters in the slightest, and I suspect that in many cases this has been a source of intractable grief for amplifier designers. In the next part of this chapter, I emphasize that the impedance of the feedback network should be kept as low as practicable to minimize the Johnson noise it generates and to minimize offset voltages. If this philosophy is followed, the feedback network source impedance as seen by the amplifier input will be too low (around 100 Ω) for the input current distortion from this part of the circuit to be measurable above the noise floor.

To summarize, if the system design requires or permits an op-amp at the input, then both the hum and distortion problems that the input currents create are removed with no further effort. If a significant source resistance is inescapable, for whatever reason, then cascoding the input pair tail cures the ripple problem but not the extra distortion. Using high-beta input transistors reduces both problems but does not eliminate them. When considering input current distortion, do not forget the feedback network has its own source impedance.

Input Stage Noise and How to Reduce It

The noise performance of a power amplifier is defined by its input stage, and so the issue is examined here. Power-amp noise is not an irrelevance; a powerful amplifier will have a high voltage gain, and this can easily result in a faint but irritating hiss from efficient loudspeakers even when all volume controls in the system are fully retarded. In the design considered here the equivalent input noise (EIN) has been measured at −120 dBu, which is only 7 or 8 dB worse than a first-class microphone preamplifier; the inferiority is largely due to the source resistances seen by the input devices being higher than the usual 150 Ω microphone impedance. By way of demonstration, halving the impedance of the usual feedback network (22 k and 1 k) reduces the EIN further by about 2 dB.

Amplifier noise is defined by a combination of the active devices at the input and the surrounding resistances. The operating conditions of the input transistors themselves are set by the demands of linearity and slew rate, so there is little freedom of design here; however, the collector currents are already high enough to give near-optimal noise figures with the low source impedances (a few hundred ohms) that we have here, so this is not too great a problem. Noise figure is a weak function of I_c, so minor tweakings of the tail current make no detectable difference. We certainly have the choice of input device type; there are many more possibles if we have relatively low rail voltages. Noise performance is, however, closely bound up with source impedance, and we need to define this before device selection.

Looking therefore to the passives, there are several resistances generating Johnson noise in the input, as in Figure 4.24, and the only way to reduce this noise is to reduce them in value. The obvious candidates are R2, R3, the input stage degeneration resistors, and R9, which determines the output impedance of the negative-feedback network. There is also another unseen component: the source resistance of the preamplifier or whatever upstream. Even if this equipment were miraculously noise-free, its output resistance would still generate Johnson noise. If the preamplifier had, say, a 20 k

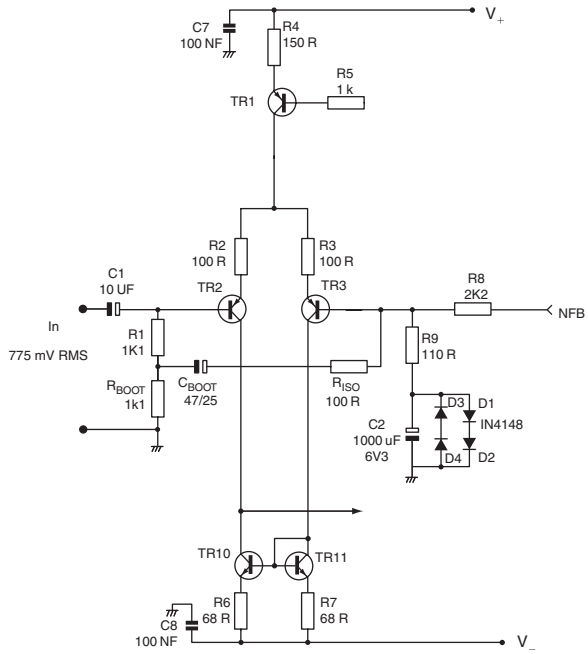


Figure 4.24: Stable input bootstrapping from the feedback point. R_{iso} is essential for HF stability; with 100 Ω, as shown, the input impedance is 13 kΩ

volume pot at its output (not a good idea, as this gives a poor gain structure and cable-dependent HF losses, but that is another story) then the source resistance could be a maximum of 5 k, which would almost certainly generate enough Johnson noise to dominate the power amplifier's noise behavior. However, there is nothing that power-amp designers can do about this, so we must content ourselves with minimizing the noise-generating resistances we do have control over.

Noise from the input degeneration resistors R2, R3 is the price we pay for linearizing the input stage by running it at a high current, and then bringing its transconductance down to a usable value by adding linearizing local negative feedback. These resistors cannot be reduced if the HF NFB factor is then to remain constant, for C_{dom} will have to be proportionally increased, reducing slew rate. With the original 22 k–1 k NFB network, these resistors degrade the noise performance by 1.7 dB. (This figure, like all other noise measurements given here, assumes a 50 Ω external source resistance.)

If we cannot alter the input degeneration resistors, then the only course left is the reduction of the NFB network impedance, and this sets off a whole train of consequences. If R8 is reduced to 2 k2, then R9 becomes 100 Ω, and this reduces noise output from −93.5 to −95.4 dBu. If degeneration resistors R2, R3 were not present, the respective figures would be −95.2 and −98.2 dBu; these and other noise measurements are summarized in Table 4.5. However, input resistor R1 must also be reduced to 2 k2 to maintain DC balance, and this is too low an input impedance for direct connection to the outside world. If we accept that the basic amplifier will have a low input impedance, there are two ways to deal with it. The simplest is to decide that a balanced line input

Table 4.5: How output noise varies with different input devices and NFB impedances

Input devices	NFB network	Input pair degen. R2, R3	Measured output noise (dBu)
MPSA56	22k – 1k	100 Ω	−93.5
MPSA56	2k2 – 100 Ω	100 Ω	−95.4
MPSA56	22k – 1k	0	−95.2
MPSA56	2k2 – 100 Ω	0	−98.2
2SA970	2k2 – 100 Ω	100 Ω	−97.2
2SB737	2k2 – 100 Ω	100 Ω	−97.3

is essential; this puts an op-amp stage before the amplifier proper, buffers the low input impedance, and can provide a fixed source impedance to allow the HF and LF bandwidths to be properly defined by an RC network using non-electrolytic capacitors. The usual practice of slapping an RC network on an unbuffered amplifier input must be roundly condemned as the source impedance is unknown, and so therefore is the roll-off point – a major stumbling block for subjectivist reviewing, one would have thought. The disadvantage is that adding even a quiet op-amp upstream will create more noise than the whole power amplifier; this is dealt with in more detail in Chapter 20 on power amplifier input systems.

Another approach is to have a low-resistance DC path at the input but a high AC impedance; in other words to use the fine old practice of input bootstrapping. Now this requires a low-impedance unity-gain-with-respect-to-input point to drive the bootstrap capacitor, and the only one available is at the amplifier inverting input, i.e. the base of TR3. While this node has historically been used for the purpose of input bootstrapping, it has only been done with simple circuitry employing very low feedback factors. There is very real reason to fear that any monkey business with the feedback point (TR3 base) will add shunt capacitance, creating a feedback pole that will degrade HF stability. There is also the awkward question of what will happen if the input is left open-circuit.

The input can in fact be safely bootstrapped; Figure 4.23 shows how. The total DC resistance of R1 and R_{boot} is equal to that of R8, giving input balance, and their central point is driven by C_{boot}. The value of R9 has been increased from 100 to 110 Ω to allow for the loading of R_{iso} and R_{boot} on the feedback point, and so the closed-loop gain is kept unchanged. Connecting C_{boot} directly to the feedback point did not produce gross instability, but it did seem to increase susceptibility to odd bits of parasitic oscillation. R_{iso} was then added to isolate the feedback point from stray capacitance, and this seemed to effect a complete cure. The input could be left open-circuit without any apparent ill-effects, though this is not good practice if loudspeakers are connected. A value for R_{iso} of 220 Ω increases the input impedance to 7.5 k, and 100 Ω raises it to 13.3 k, safely above the 10 k standard value for a bridging impedance. Despite successful tests, I must admit to a few lingering doubts about the HF stability of this approach, and it might be as well to consider it as experimental until more experience is gained.

One more consequence of a low-impedance NFB network is the need for feedback capacitor C2 to be proportionally increased to maintain LF response, and prevent capacitor distortion from

causing a rise in THD at low frequencies; it is the latter requirement that determines the value. (This is a separate distortion mechanism from the seven originally identified, and is given the title Distortion 8.) This demands a value of 1000 μF, necessitating a low-rated voltage such as 6V3 if the component is to be of reasonable size. This means that C2 needs protective shunt diodes in both directions, because if the amplifier fails it may saturate in either direction. Examination of the distortion residual shows that the onset of conduction of back-to-back diodes will cause a minor increase in THD at 10 Hz, from less than 0.001% to 0.002%, even at the low power of 20 W/8 Ω. It is not my practice to tolerate such gross nonlinearity; therefore four diodes are used in the final circuit, and this eliminates the distortion effect. It could be argued that a possible reverse bias of 1.2 V does not protect C2 very well, but at least there will be no explosion.

We can now consider alternative input devices to the MPSA56, which was never intended as a low-noise device. Several high-beta low-noise types such as 2SA970 give an improvement of about 1.8 dB with the low-impedance NFB network. Specialized low-R_b devices like 2SB737 give little further advantage (possibly 0.1 dB) and it is probably best to go for one of the high-beta types, as this will minimize both input-current distortion (as described above) and DC offsets (explored in the next section).

It could be argued that the above complications are a high price to pay for a noise reduction of some 2 dB; however, with the problems comes a definite advantage, for the above NFB network modification also significantly improves the output DC offset performance.

Noise Sources in Power Amplifiers

It is instructive to go a little deeper into the sources of noise inside a power amplifier, to see what determines it and how (and if) it can be improved. The measurements quoted in this section were made on a different power amplifier, though its circuitry was essentially the same as that we have already examined. The closed-loop gain in this case was 30.6 dB rather than 27.2 dB, so at first sight it appears a little noisier. The measured output noise (with the input terminated in 50 Ω) was −92.0 dBu. It is relatively easy to calculate what proportion of this comes from Johnson noise in the circuit resistances and what proportion from the input transistors. The only resistances that contribute significantly are the feedback network and the emitter degeneration resistors; the feedback network had an effective source resistance of 92 Ω, which gives a Johnson noise voltage of −132.6 dBu for a bandwidth of 22 kHz at 25°C. The emitter degeneration resistors were 100 Ω, which generate −132.2 dBu of Johnson noise each. They are effectively in series from the point of view of noise and so their noise output sums in the usual rms manner, and gives a total noise for the degeneration resistors of −129.2 dBu, an increase of 3 dB. Adding the feedback network noise to this gives a total of −127.6 dBu. This looks very low, but of course it is referred to the amplifier input. We add the closed-loop gain of 30.6 dB and we get a predicted output noise – from these resistors alone – of −97.0 dBu.

However, the measured output noise was −92.0 dBu, and so the 5.0 dB difference must be due to the input transistor pair. If the resistances and transistors were making equal contributions to the noise output that figure would have been 3 dB, so it looks as though the transistors are generating

the greater part of the noise, and in fact dominating the noise situation. If we want to improve the noise performance of the amplifier, that is the area we need to attack. Working from the measured output noise of −92.0 dBu, subtracting the closed-loop gain of 30.6 dB gives us a measured EIN of −122.6 dBu. We have calculated just above that the Johnson noise of the feedback and emitter degeneration resistances was −127.6 dBu, so if we subtract that from −122.6 dBu we should get the portion of the EIN that is due to the transistors. Performing the rms subtraction yields an EIN figure of −124.2 dBu, and we can now see if that figure agrees with the well-established theory of bipolar transistor noise, and if the theory gives any guidance on reducing the transistor noise contribution.

In this particular amplifier design the input transistors were MPSA92, chosen for their high voltage capability rather than their noise performances; in fact no noise data appears to have been published for these devices so it is rather difficult to say exactly how much the noise performance was compromised by this choice. The input transistors were running at the relatively high collector current of 3 mA each, the choice of this value being driven by the need to make the input stage linear and obtain a satisfactory maximum slew rate. The question suggests itself: would a further increase in collector current reduce the transistor component of the amplifier noise?

Noise in Bipolar Transistors

To understand the noise behavior of discrete bipolar transistors, it is necessary to delve a little deeper into their internal operation than is usually required, and take account of imperfections that do not appear in the simplest transistor models. I give here a quick summary rather than a thorough analysis; the latter can be found in many textbooks. Two important transistor parameters for understanding noise are R_{bb}, the base spreading resistance, and R_e, the intrinsic emitter resistance. R_{bb} is a real physical resistance – what is called an *extrinsic* resistance. The second parameter R_e is an expression of the V_{be}/I_c slope and not a physical resistance at all, and it is therefore called an *intrinsic* resistance.

Noise in bipolar transistors is best dealt with by assuming we have a noiseless transistor with a theoretical noise voltage source in series with the base and a theoretical noise current source connected from base to ground. These sources are usually just described as the 'voltage noise' and the 'current noise' of a transistor.

(1) The voltage noise V_n has two components, one of which is the Johnson noise generated in the base spreading resistance R_{bb}; the other is the collector current (I_c) shot noise creating a noise voltage across R_e, the intrinsic emitter resistance. Shot noise occurs simply because an electric current is a stream of discrete electric charges and not a continuous fluid, and it increases as the square root of the current. The two components can be represented thus:

$$\text{Voltage noise density } V_n = \sqrt{4kTR_{bb} + 2(kT)^2/(qI_c)} \text{ in V/rtHz (usually nV/rtHz)} \qquad \text{Equation 4.5}$$

The first part of this equation is the usual expression for Johnson noise, and is fixed for a given transistor type by the physical value of R_{bb}; the lower this is, the better. The absolute temperature

is obviously a factor but there is not usually much you can do about this. The second (shot noise) part of the equation decreases as collector current I_c increases; this is because as I_c increases, R_e decreases proportionally while the shot noise only increases as the square root of I_c. These factors are all built into the second part of the equation. The overall result is that the total v_n falls – though relatively slowly – as collector current increases, approaching asymptotically the level of noise set by the first part of the equation. There is no way you can reduce this except by changing to another type of transistor with a lower R_{bb}.

There is an extra voltage noise source resulting from flicker noise produced by the base current flowing through R_{bb}; this is only significant at high collector currents and low frequencies due to its 1/*f* nature, and is usually not included in design calculations unless low-frequency quietness is a special requirement.

(2) The current noise I_n is mainly produced by the shot noise of the steady current I_b flowing through the transistor base. This means it increases as the square root of I_b increases. Naturally I_b increases with I_c. Current noise is given by:

$$\text{Current noise density } i_n = \sqrt{2qI_b} \text{ in A/rtHz (usual values are in pA)} \qquad \text{Equation 4.6}$$

So, for a fixed collector current, you get less current noise with high-beta transistors because there is less base current. Such transistors usually have a $V_{ce(max)}$ that is too low for use in most power amplifiers; one solution to this would be a cascode input stage, as described earlier, which would take most of the voltage strain off the input devices. However, as we shall see, at the kind of source resistances we are dealing with, the current noise makes only a minor contribution to the total, and cascoding is probably not worthwhile for this reason alone.

The existence of current noise as well as voltage noise means that in general it is not possible to minimize transistor noise just by increasing the collector current to the maximum value the device can take. Increasing I_c certainly reduces voltage noise, but it increases current noise. Hence there is an optimum collector current for each value of source resistance, where the contributions are equal. Because both voltage and current mechanisms are proportional to the square root of I_c, they change relatively slowly as it is altered, and the noise curve is rather flat at the bottom (see Figure 4.25). There is no need to control collector current with great accuracy to obtain the optimum noise performance.

I want to emphasize here that this is a simplified noise model, not least because in practice both voltage and current noise densities vary with frequency. I have also ignored 1/*f* noise. However, it gives the essential insight into what is happening and leads to the right design decisions so we will put our heads down and press on.

A quick example shows how this works. In an audio power amplifier we want the source impedances seen by the input transistors to be as low as possible, to minimize Johnson noise, and to minimize the effects of input current distortion, as described elsewhere in this chapter. The output impedance of the source equipment will, if we are lucky, simply be the value of the output resistor required to give stability when driving cable capacitance, i.e. about 100 Ω. It is also usually

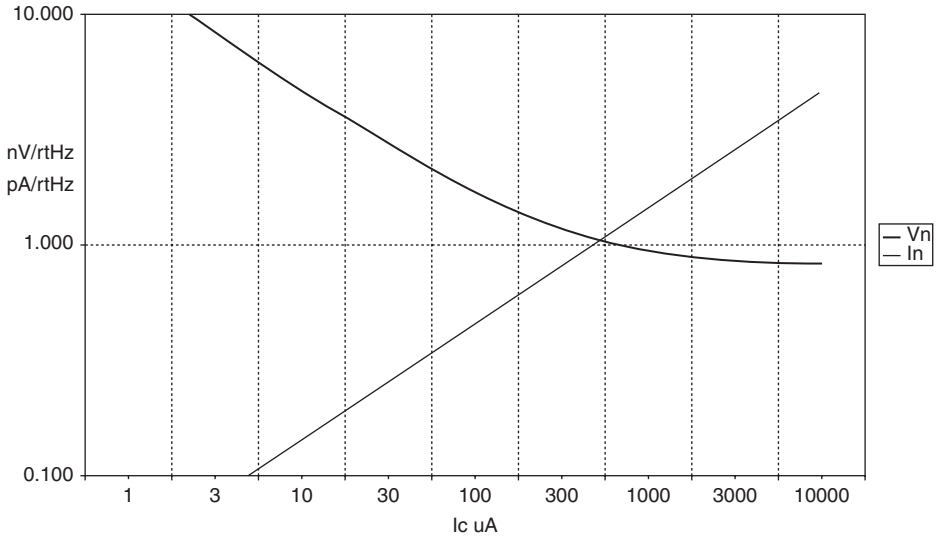


Figure 4.25: How voltage noise density V_n and current noise density I_n vary with collector current I_c in a generic transistor. As I_c increases the voltage noise falls to a lower limit while the current noise continuously increases

Table 4.6: The summation of Johnson noise from the source resistance with transistor noise

1 R_{source} (Ω)	2 R_{source} Johnson (nV/rtHz)	3 R_{source} Johnson BW (nV)	4 R_{source} Johnson BW (dBu)	5 Transistor noise incl. I_n in R_s (nV/rtHz)	6 Transistor noise plus R_s Johnson (nV/rtHz)	7 Noise in BW (nV)	8 Noise in BW (dBu)	9 Noise figure (dB)
1	0.128	19.0	−152.2	0.93	0.94	139.7	−134.9	17.3
10	0.406	60.2	−142.2	0.93	1.02	150.9	−134.2	8.0
100	1.283	190.3	−132.2	0.94	1.59	236.3	−130.3	1.9
1000	4.057	601.8	−122.2	1.73	4.41	654.4	−121.5	0.7
10,000	12.830	1903.0	−112.2	14.64	19.46	2886.9	−108.6	3.6
100,000	40.573	6017.9	−102.2	146.06	151.59	22,484.8	−90.7	11.4

possible to design the negative-feedback network so it has a similar source impedance (see Figure 4.23 for an example). So let us look at optimizing the noise from a single transistor faced with a 100 Ω source resistance.

A few assumptions need to be made. The temperature is 25°C, the bandwidth is 22 kHz, and the R_{bb} of our transistor is 40 Ω, which seems like an average value. (Why don't they put this on spec sheets any more?) The hfe (beta) is 150. Set the I_c to 1 mA, which is plausible for an amplifier input stage, step the source resistance from 1 to 100,000 Ω and the calculations come out as in Table 4.6.

The first column shows the source resistance, and the second column the Johnson noise density it generates by itself. Factor in the bandwidth and you get the third and fourth columns, which show the actual noise voltage in two different ways. The fifth column is the aggregate noise density from

the transistor, obtained by taking the rms sum of the voltage noise and the voltage generated by the current noise flowing in the source resistance. The sixth column gives total noise density when we sum the source resistance noise density with the transistor noise density. Factor in the bandwidth again, and the resultant noise voltage is given in columns 7 and 8. The last column gives the noise figure (NF), which is the amount by which the combination of transistor and source resistance is noisier than the source resistance alone. In other words, it tells how close we have got to theoretical perfection, which would be an NF of 0 dB.

The results are, I hope, instructive. The results for 100 Ω show that the transistor noise is less than the source resistance noise, and we know at once that the amount by which we can improve things by twiddling the transistor operating conditions is pretty limited. The results for the other source resistances are worth looking at. The lowest noise output (−134.9 dBu) is achieved by the lowest source resistance of 1 Ω, as you would expect, but the NF is very poor at 17.3 dB; this gives you some idea why it is hard to design quiet moving-coil head amplifiers. The best NF, and the closest approach to theoretical perfection, is with 1000 Ω, but this is attained with a *greater* noise output than 100 Ω. As the source resistance increases further, the NF begins to get worse again; a transistor with an I_c of 1 mA has relatively high current noise and does not perform well with high source resistances.

You will note that we started off with what in most areas of electronics would be a high collector current – 1 mA. In fact this is too low for amplifier input stages designed to my philosophy, and most of the examples in this book have a 6 mA tail current, which splits into 3 mA in each device; this value is chosen to allow linearization of the input pair and give a good slew rate, rather than from noise considerations. So we dial an I_c of 3 mA into our spreadsheet, and we find there is a slight improvement for our 100 Ω source resistance case, but only a marginal 0.2 dB (see Table 4.7, which this time skips the intermediate calculations and just gives the results).

For 1 Ω things are 0.7 dB better, due to slightly lower voltage noise, and for 100,000 Ω they are worse by no less than 9.8 dB as the current noise is much increased. So let's get radical and increase I_c to 10 mA. Unfortunately this makes the 100 Ω noise worse, and we have lost our slender 0.2 dB improvement. This theoretical result is backed up by practical experience, where it is found

Table 4.7: How input device collector current affects noise figure

	I_c = 3 mA		I_c = 10 mA		I_c = 10 mA, 2SB737		I_c = 100 μA	
R_{source} (Ω)	Noise (dBu)	Noise figure (dB)	Noise (dBu)	Noise figure (dB)	Noise (dBu)	Noise figure (dB)	Noise (dBu)	Noise figure (dB)
1	−135.6	16.6	−135.9	16.3	−145.9	6.3	−129.9	22.3
10	−134.8	7.4	−135.1	7.1	−140.9	1.3	−129.7	12.5
100	−130.5	1.7	−130.3	1.9	−131.5	0.7	−127.9	4.3
1000	−120.6	1.6	−118.5	3.7	−118.6	3.6	−121.5	0.7
10,000	−105.3	6.9	−100.7	11.4	−100.7	11.4	−111.6	0.6
10,0000	−86.2	16.0	−81.0	21.2	−81.0	21.2	−98.6	3.6

that increasing the tail current from 6 mA (3 mA per device) to 20 mA (10 mA per device) gives no significant reduction in the noise output.

For 1 Ω the noise is 0.3 dB better – hardly a triumph – and for the higher source resistances things get rapidly worse, the 100,000 Ω noise increasing by another 5.2 dB. It therefore appears that a collector current of 3 mA is actually pretty much optimal for noise with our 100 Ω source resistance, even though it was originally chosen for other reasons.

Let us now pluck out the 'ordinary' transistor and replace it with a specialized low-R_{bb} part like the much-lamented 2SB737 (now regrettably obsolete), which has a superbly low R_{bb} of 2 Ω. The noise output at 1 Ω plummets by 10 dB, showing just how important low R_{bb} is under these conditions; for a more practical 100 Ω source resistance noise drops by a useful 1.0 dB, as you might expect.

As an aside, let's go back to the ordinary transistor and cut its I_c right down to 100 μA, giving the last two columns in Table 4.7. Compared with I_c = 3 mA, noise with the 1 Ω source degrades by 5.7 dB, and with the 100 Ω source by 2.6 dB, but with the 100,000 Ω source there is a hefty 12.4 dB improvement, showing why BJT inputs for high impedances use low collector currents.

This shows you why transistor amplifiers with high source resistances are run with low collector currents. If you're stuck with such a situation, JFETs can give better noise performance than BJTs; JFETs are not dealt with here for reasons already explained, their low transconductance and poor V_{gs} matching.

We therefore conclude that our theoretical noise output with I_c = 3 mA and R_s = 100 Ω will be −130.5 dBu, with an NF of 1.7 dB. However, these calculations are dealing with a single transistor and a single source resistance; in a differential input stage there are two transistors, and if we assume equal source resistances of 100 Ω for each one, as explained above, the noise output has to be increased by 3 dB as we are adding two non-correlated noise voltages. This gives us a theoretical noise output of −127.5 dBu, which, it has to be said, does not match up particularly well with the practical figure of −124.2 dBu that we deduced in the previous section. There are several reasons for this: to make the explanation manageable in the space available we have had to ignore some minor sources of extra noise, the frequency dependence of the voltage and current noise sources, and we have used a generic transistor. There is not much choice about the latter as manufacturers tend not to publish R_{bb} or noise data for the high-voltage transistors that are used in audio power amplifiers.

It therefore seems pretty clear that we are not going to get any significant improvement in power amplifier noise by altering the input device conditions. It could of course be argued that there is no point in making it any quieter, because a pair of discrete transistors with a low source impedance are about as quiet as it gets, and pretty much anything you put in front of it is going to dominate the noise situation. This issue is developed further in the chapter on power amplifier input systems, which deals with balanced input amplifiers and so on. On the other hand . . .

Reducing Input Transistor Noise

However, let's assume that for some reason it is really important that the power amplifier itself be as quiet as humanly possible. As we saw earlier, in a practical amplifier the transistor noise is

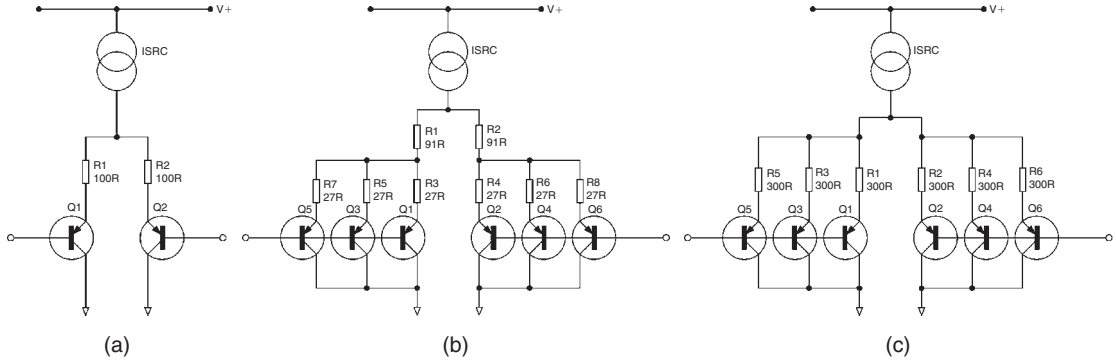


Figure 4.26: (a) Normal input stage with 100 Ω degeneration resistors. (b) Multiple input devices with small current-sharing resistors. (c) Multiple input devices with split emitter-degeneration resistors

the dominant source, so this needs to be addressed first. A reliable method of doing this, often used in moving-coil preamplifiers, is the use of multiple transistors in parallel. The gain will sum arithmetically but the noise from each transistor will be uncorrelated and therefore subject to rms-summing. In other words, two transistors will be 3 dB quieter than one, three transistors 4.8 dB quieter, and four transistors 6 dB quieter. There are obvious practical limits to a process where every 3 dB improvement means doubling the number of devices, and you soon start thinking about grains of corn on chessboards, but putting four transistors into each side of an input stage is quite feasible; the cost of the small-signal part of the amplifier will still be a very small fraction of the cost of power devices, heat-sinks, mains transformer and so on. The main thing that needs to be taken into account is current-sharing between the devices. Figure 4.26 shows two different ways that this could be implemented, assuming that it is desired to keep the emitter-degeneration resistors at their usual value of 100 Ω. In Figure 4.26b three 27 Ω resistors effectively in parallel give 9 Ω, which with a 91 Ω resistor very handily keeps the total degeneration resistance at exactly 100 Ω. In Figure 4.26c the values work out equally neatly. The tail-current source may have to be increased in value, but not necessarily trebled, for as we have seen the noise performance varies quite slowly as collector current changes.

If even now you want to make the amplifier quieter, you must turn your attention back to the circuit resistances and their Johnson noise. Start by getting the input and feedback source impedances very low. In the case of the feedback network, its source resistance is determined by the lower resistor in the feedback arm (R9 in Figure 4.23) and reducing this means the upper resistor (R8) has to be proportionally decreased to keep the closed-loop gain the same; the limit on this process will be the power dissipation in the upper resistor. In amplifiers I have designed this is commonly a 1 W part, as anything more capable tends to be inconveniently big.

The emitter degeneration resistors also produce noise, and the values of these may need to be reduced, with an eye to the fact that this will decrease the linearization of the input stage, and it will also be necessary to alter the compensation to maintain the same stability margins. This is not very satisfactory, and you will have to think hard if you really want to impair the distortion performance in the pursuit of the lowest possible noise.

Offset and Match: The DC Precision Issue

The same components that dominate amplifier noise performance also determine the output DC offset; looking at Figure 4.27, if R9 is reduced to minimize the source resistance seen by TR3, then the value of R8 must be scaled to preserve the same closed-loop gain, and this reduces the voltage drops caused by input transistor base currents.

Most of my amplifier designs have assumed that a ±50 mV output DC offset is acceptable. This allows DC trimpots, offset servos, etc. to be gratefully dispensed with. However, it is not in my nature to leave well enough alone, and it could be argued that ±50 mV is on the high side for a top-flight amplifier. I have therefore reduced this range as much as possible without resorting to a servo; the required changes have already been made when the NFB network was reduced in impedance to minimize Johnson noise (see page 104).

With the usual range of component values, the DC offset is determined not so much by input transistor V_{be} mismatch, which tends to be only 5 mV or so, but more by a second mechanism – imbalance in beta. This causes imbalance of the base currents (I_b) drawn thorough input bias resistor R1 and feedback resistor R8, and the cancellation of the voltage drops across these components is therefore compromised.

A third source of DC offset is non-ideal matching of input degeneration resistors R2, R3. Here they are 100 Ω, with 300 mV dropped across each, so two 1% components at opposite ends of their tolerance bands could give a maximum offset of 6 mV. In practice this is most unlikely, and the error from this source will probably not exceed 2 mV.

There are several ways to reduce DC offset. First, low-power amplifiers with a single output pair must be run from modest HT rails and so the requirement for high-V_{ce} input transistors can be relaxed. This allows higher beta devices to be used, directly reducing I_b. The 2SA970 devices used in this design have a beta range of 350–700, compared with 100 or less for MPSA06/56. Note the pinout is not the same.

On page 104, we reduced the impedance of the feedback network by a factor of 4.5, and the offset component due to I_b imbalance is reduced by the same ratio. We might therefore hope to keep the DC output offset for the improved amplifier to within ±15 mV without trimming or servos. Using high-beta input devices, the I_b errors did not exceed ±15 mV for 10 sample pairs (*not* all from the same batch) and only three pairs exceeded ±10mV. The I_b errors are now reduced to the same order of magnitude as V_{be} mismatches, and so no great improvement can be expected from further reduction of circuit resistances. Drift over time was measured at less than 1 mV, and this seems to be entirely a function of temperature equality in the input pair.

Figure 4.27 shows the ideal DC conditions in a perfectly balanced input stage, assuming $\beta = 400$, compared with a set of real voltages and currents from the prototype amplifier. In the latter case, there is a typical partial cancelation of offsets from the three different mechanisms, resulting in a creditable output offset of −2.6 mV.

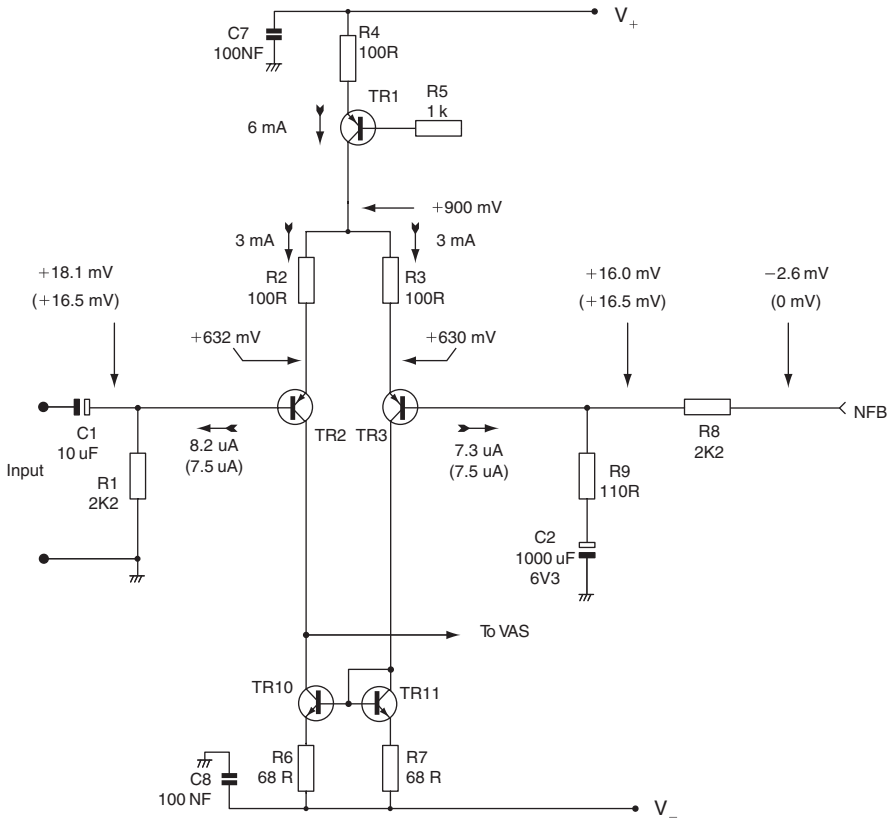


Figure 4.27: The measured DC conditions in a real input stage. Ideal voltages and currents for perfectly matched components are shown in parentheses

The Input Stage and the Slew Rate

This is another parameter that is usually assumed to be set by the input stage, and has a close association with HF distortion. A brief summary is therefore given here, but the subject is dealt with in much greater depth in Chapter 8.

An amplifier's slew rate is proportional to the input stage's maximum-current capability, most circuit configurations being limited to switching the whole of the tail current to one side or the other. The usual differential pair can only manage half of this, as with the output slewing negatively half the tail current is wasted in the input collector load R2. The addition of an input current-mirror, as advocated above, will double the slew rate in both directions as this inefficiency is abolished. With a tail current of 1.2 mA a mirror improves the slew rate from about 5 to 10 V/μs (for $C_{dom} = 100\,pF$). The constant-g_m degeneration method of linearity enhancement in Figure 4.9 further increases it to 20 V/μs.

In practice slew rates are not the same for positive- and negative-going directions, especially in the conventional amplifier architecture that is the main focus of this book; this issue is examined in Chapter 8.

Input Stage Conclusions

Hopefully this chapter has shown that input stage design is not something to be taken lightly if low noise, low distortion, and low offset are desired. A good design choice even for very high quality requirements is a constant-g_m degenerated input pair with a degenerated current-mirror; the extra cost of the mirror will be trivial.

References

[1] P.P. Gray, R.G. Meyer, Analysis and Design of Analog Integrated Circuits, Wiley, 1984, p. 172 (exponential law of singleton).

[2] P.P. Gray, R.G. Meyer, Analysis and Design of Analog Integrated Circuits, Wiley, 1984, p. 194 (tanh law of simple pair).

[3] D. Self, Sound Mosfet design, Electronics & Wireless World (September 1990) p. 760 (varying input balance with R2).

[4] E. Taylor, Distortion in low noise amplifiers, Wireless World (August 1977) p. 32.

[5] P.P. Gray, R.G. Meyer, Analysis and Design of Analog Integrated Circuits, Wiley, 1984, p. 256 (tanh law of current-mirror pair).

[6] M. Herpy, Analog Integrated Circuits, Wiley-Interscience, p. 118.

[7] R.L. Geiger, P.E. Allen, N.R. Strader, VLSI Design Techniques for Analog and Digital Circuits, McGraw-Hill, 1990.

[8] D. Feucht, Handbook of Analog Circuit Design, Academic Press, 1990, p. 432 (cross-quad).

[9] P. Quinn, IEEE International Solid-State Circuits Conference, THPM 14.5, p. 188 (cascomp).

[10] W.E. Hearn, Fast slewing monolithic operational amplifier, IEEE J. Solid State Circuits SC6 (February 1971) 20–24 (AB input stage).

[11] R.J. Van de Plassche, A wide-band monolithic instrumentation amplifier, IEEE J. Solid State Circuits SC10 (December 1975) pp. 424–431.

[12] G. Stochino, Ultra-fast amplifier, Electronics & Wireless World (December 1996) p. 835.

[13] R.W. Hickman, F.V. Hunt, On electronic voltage stabilizers, Review of Scientific Instruments 10 (January 1939) pp. 6–21.

[14] D. Self, A high performance preamplifier, Wireless World (February 1979) p. 40.

[15] E. Taylor, Distortion in low noise amplifiers, Wireless World (1977) p. 29.

[16] E. Cherry, Private communication, June 1996.

[17] G. Stochino, Private communication, May 1996.

[18] J. Dostal, Operational Amplifiers, Butterworth-Heinemann, 1993, p. 65.

CHAPTER 5

The Voltage-Amplifier Stage

The voltage-amplifier stage (or VAS) has often been regarded as the most critical part of a power amplifier, since it not only provides all the voltage gain but also must give the full output voltage swing. (The input stage may give substantial transconductance gain, but the output is in the form of a current.) However, as is not uncommon in audio, all is not quite as it appears. A well-designed VAS will contribute relatively little to the overall distortion total of an amplifier, and if even the simplest steps are taken to linearize it further, its contribution sinks out of sight.

As a starting point, Figure 5.1 shows the distortion plot of a model amplifier with a Class-A output (±15 V rails, +16 dBu out) as per Chapter 3, where no special precautions have been taken to linearize the input stage or the VAS; output stage distortion is negligible. It can be seen that the distortion is below the noise floor at LF; however, the distortion slowly rising from about 1 kHz is coming from the VAS. At higher frequencies, where the VAS 6 dB/octave rise becomes combined with the 12 or 18 dB/octave rise of input stage distortion, we can see the distortion slope of accelerating steepness that is typical of many amplifier designs.

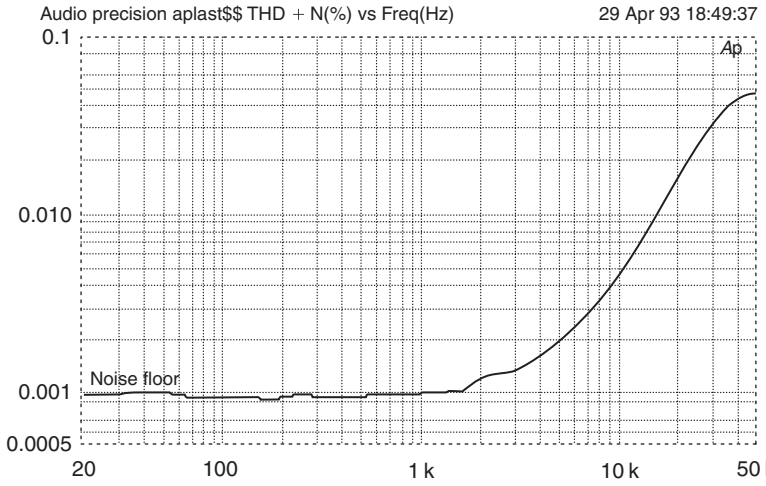


Figure 5.1: THD plot for model amp showing very low distortion (below noise floor) at LF, and increasing slope from 2 to 20 kHz. The ultimate flattening is due to the 80 kHz measurement bandwidth

As previously explained, the main reason why the VAS generates relatively little distortion is because at LF global feedback linearizes the whole amplifier, while at HF the VAS is linearized by local NFB through C_{dom}.

Measuring VAS Distortion in Isolation

Isolating the VAS distortion for study requires the input pair to be specially linearized, or else its steeply rising distortion characteristic will swamp the VAS contribution. This is most easily done by degenerating the input stage; this also reduces the open-loop gain, and the reduced feedback factor mercilessly exposes the nonlinearity of the VAS. This is shown in Figure 5.2, where the 6 dB/octave slope suggests that this must originate in the VAS, and increases with frequency solely because the compensation is rolling off the global feedback factor. To confirm that this distortion is due solely to the VAS, it is necessary to find a method for experimentally varying VAS linearity while leaving all other circuit parameters unchanged. Figure 5.3 shows my arrangement for doing this by varying the VAS V-voltage; this varies the proportion of its characteristic over which the VAS swings, and thus only alters the effective VAS linearity, as the important input stage conditions remain unchanged. The current-mirror must go up and down with the VAS emitter for correct operation, and so the V_{ce} of the input devices also varies, but this has no significant effect, as can be proved by the unchanged behavior on inserting cascode stages in the input transistor collectors.

VAS Operation

The typical VAS topology as shown in Figure 5.4a is a classical common-emitter voltage-amplifier stage, with a current-drive input into the base. The small-signal characteristics, which set open-loop gain and so on, can be usefully simulated by the SPICE model shown in Figure 5.5, of a VAS reduced to its conceptual essentials. G is a current-source whose value is controlled by the voltage difference between R_{in} and RF2, and represents the differential transconductance input stage.

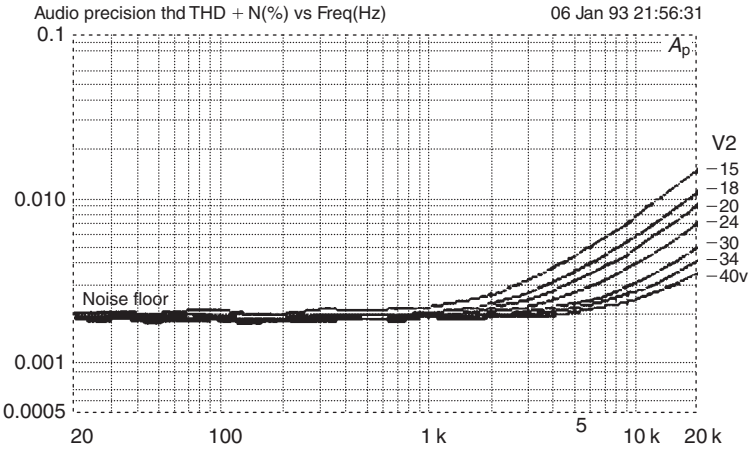


Figure 5.2: The change in HF distortion resulting from varying V− in the VAS test circuit. The VAS distortion is only revealed by degenerating the input stage with 100 Ω resistors

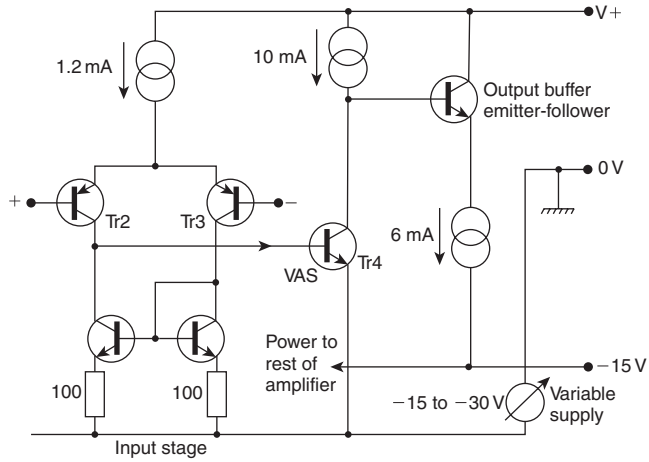


Figure 5.3: VAS distortion test circuit. Although the input pair mirror moves up and down with the VAS emitter, the only significant parameter being varied is the available voltage swing at the VAS collector

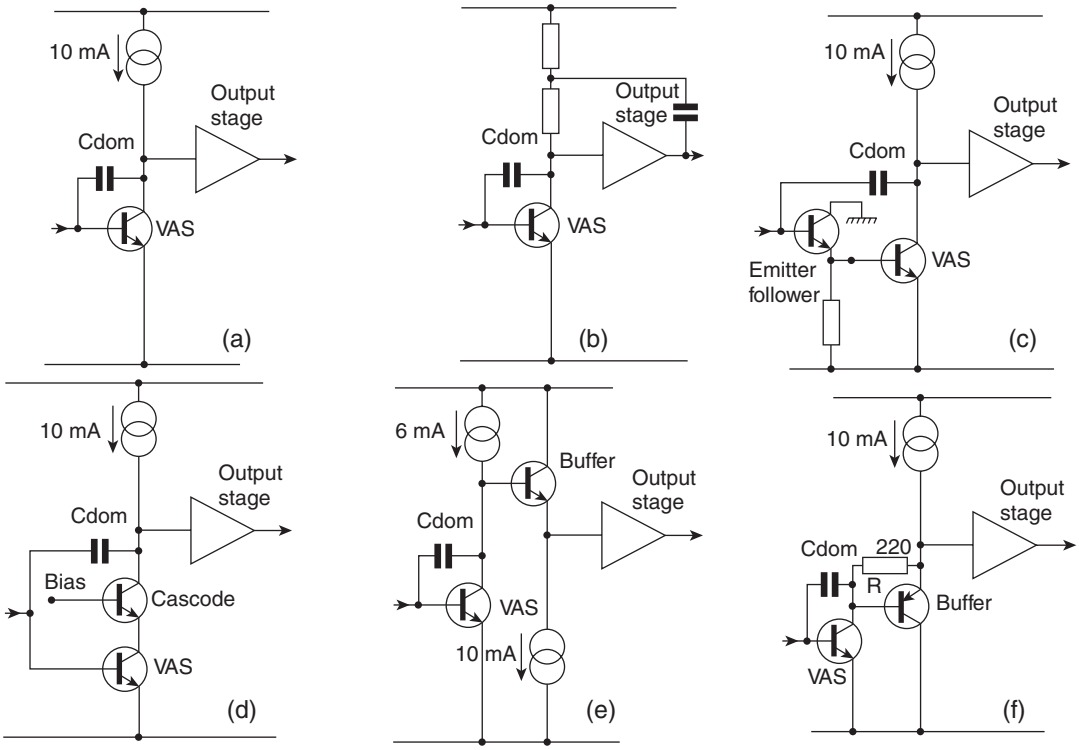


Figure 5.4: Six variations on a VAS. (a) Conventional VAS with current-source load. (b) Conventional VAS with bootstrapped load. (c) Increase in local NFB by adding beta-enhancing emitter-follower. (d) Increase in local NFB by cascoding VAS. (e) Buffering the VAS collector from the output stage. (f) Alternative buffering, bootstrapping VAS load R

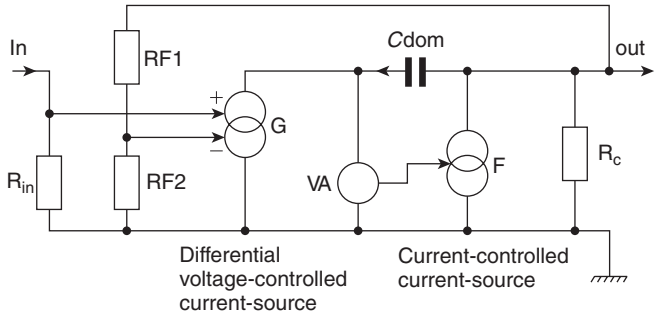


Figure 5.5: Conceptual SPICE model of differential input stage (G) and VAS (F). The current in F is beta times the current in VA

F represents the VAS transistor, and is a current source yielding a current of beta times that sensed flowing through 'ammeter' VA, which by SPICE convention is a voltage source set to 0 V; the value of beta, representing current gain as usual, models the relationship between VAS collector current and base current. R_c represents the total VAS collector impedance, a typical real value being 22 k. With suitable parameter values, this simple model provides a good demonstration of the relationships between gain, dominant-pole frequency, and input stage current that were introduced in Chapter 3. Injecting a small-signal current into the output node from an extra current source also allows the fall of impedance with frequency to be examined.

The overall voltage gain clearly depends linearly on beta, which in real transistors may vary widely. Working on the trusty engineering principle that what cannot be controlled must be made irrelevant, local shunt NFB through C_{dom} sets the crucial HF gain that controls Nyquist stability. The LF gain below the dominant-pole frequency $P1$ remains variable (and therefore so does $P1$) but is ultimately of little importance; if there is an adequate NFB factor for overall linearization at HF then there are unlikely to be problems at LF, where the gain is highest. As for the input stage, the linearity of the VAS is not greatly affected by transistor type, given a reasonably high beta.

VAS Distortion

VAS distortion arises from the fact that the V_{be}/I_c transfer characteristic of a common-emitter amplifier is curved, being a portion of an exponential[1]. This characteristic generates predominantly second-harmonic distortion, which in a closed-loop amplifier will increase at 6 dB/octave with frequency.

VAS distortion does not get worse for more powerful amplifiers as the stage traverses a constant proportion of its characteristic as the supply rails are increased. This is not true of the input stage; increasing output swing increases the demands on the transconductance amp as the current to drive C_{dom} increases. The increased V_{ce} of the input devices does not measurably affect their linearity.

It is ironic that VAS distortion only becomes clearly visible when the input pair is excessively degenerated – a pious intention to 'linearize before applying feedback' can in fact make the

closed-loop distortion worse by reducing the open-loop gain and hence the NFB factor available to linearize the VAS. In a real (non-model) amplifier with a distortive output stage the deterioration will be worse.

Linearizing the VAS: Active-Load Techniques

As described in Chapter 3, it is important that the local open-loop gain of the VAS (that existing inside the local feedback loop closed by C_{dom}) be high, so that the VAS can be linearized, and therefore a simple resistive load is unusable.

Increasing the value of R_c will decrease the collector current of the VAS transistor, reducing its transconductance and getting you back where you started in terms of voltage gain.

One way to ensure enough local loop gain is to use an active load to increase the effective collector impedance at TR4 and thus increase the raw voltage gain; either bootstrapping or a current source will do this effectively, though the current source is perhaps more dependable, and is the usual choice for hi-fi or professional amplifiers. The bootstrap promises more O/P swing, as the collector of TR4 can in theory soar like a lark above the V+ rail; under some circumstances this can be the overriding concern, and bootstrapping is alive and well in applications such as automotive power amps that must make the best possible use of a restricted supply voltage[2].

Both active-load techniques have another important role: ensuring that the VAS stage can source enough current to properly drive the upper half of the output stage in a positive direction, right up to the rail. If the VAS collector load was a simple resistor to +V, then this capability would certainly be lacking.

It may not be immediately obvious how to check that impedance-enhancing measures are working properly, but it is actually fairly simple. The VAS collector impedance can be determined by the simple expedient of shunting the VAS collector to ground with decreasing resistance until the open-loop gain reading falls by 6 dB, indicating that the collector impedance is equal to the current value of the test resistor.

The popular current-source version is shown in Figure 5.4a. This works well, though the collector impedance is limited by the effective output resistance R_o of the VAS and the current-source transistors[3], which is another way of saying that the improvement is limited by the Early effect.

It is often stated that this topology provides current drive to the output stage; this is only partly true. It is important to realize that once the local NFB loop has been closed by adding C_{dom} the impedance at the VAS output falls at 6 dB/octave for frequencies above $P1$. With typical values the impedance is only a few kΩ at 10 kHz, and this hardly qualifies as current drive at all.

Collector-load bootstrapping (Figure 5.4b) works in most respects as well as a current-source load, for all its old-fashioned look. Conventional capacitor bootstrapping has been criticized for prolonging recovery from clipping; I have no evidence to offer on this myself, but a more subtle drawback definitely does exist – with bootstrapping the LF open-loop gain is dependent on amplifier

output loading. The effectiveness of bootstrapping depends crucially on the output stage gain being unity or very close to it; however, the presence of the output-transistor emitter resistors means that there will be a load-dependent gain loss in the output stage, which in turn significantly alters the amount by which the VAS collector impedance is increased; hence the LF feedback factor is dynamically altered by the impedance characteristics of the loudspeaker load and the spectral distribution of the source material. This has a special significance if the load is an 'audiophile' speaker that may have impedance dips down to 2 Ω, in which case the gain loss is serious. If anyone needs a new audio-impairment mechanism to fret about, then I humbly offer this one in the confident belief that its effects, while measurable, are not of audible significance. Possibly this is a more convincing reason for avoiding bootstrapping than alleged difficulties with recovery from clipping.

Another drawback of bootstrapping is that the standing DC current through the VAS, and hence the bias generator, varies with rail voltage. Setting and maintaining the quiescent conditions is quite difficult enough already, so an extra source of possible variation is decidedly unwelcome.

A less well-known but more dependable form of bootstrapping is available if the amplifier incorporates a unity-gain buffer between the VAS collector and the output stage; this is shown in Figure 5.4f, where R_c is the collector load, defining the VAS collector current by establishing the V_{be} of the buffer transistor across itself. This is constant, and R_c is therefore bootstrapped and appears to the VAS collector as a constant-current source. In this sort of topology a VAS current of 3 mA is quite sufficient, compared with the 10 mA standing current in the buffer stage. The VAS would in fact work well with lower collector currents down to 1 mA, but this tends to compromise linearity at the high-frequency, high-voltage corner of the operating envelope, as the VAS collector current is the only source for driving current into C_{dom}.

VAS Enhancements

Figure 5.2 shows VAS distortion only, clearly indicating the need for further improvement over that given inherently by C_{dom} if our amplifier is to be as good as possible. The virtuous approach might be to try to straighten out the curved VAS characteristic, but in practice the simplest method is to increase the amount of *local* negative feedback through C_{dom}. Equation 3.1 in Chapter 3 shows that the LF gain (i.e. the gain before C_{dom} is connected) is the product of input stage transconductance, TR4 beta and the collector impedance R_c. The last two factors represent the VAS gain and therefore the amount of local NFB can be augmented by increasing either. Note that so long as the value of C_{dom} remains the same, the global feedback factor at HF is unchanged and so stability is not affected.

The effective beta of the VAS can be substantially increased by replacing the VAS transistor with a Darlington, or in other words putting an emitter-follower before it (Figure 5.4c). Adding an extra stage to a feedback amplifier always requires thought, because if significant additional phase shift is introduced, the global loop stability can suffer. In this case the new stage is inside the C_{dom} Miller loop and so there is little likelihood of trouble from this. The function of such an emitter-follower is sometimes described as 'buffering the input stage from the VAS' but that is totally misleading, as its true function is linearization by enhancement of local NFB through C_{dom}.

Alternatively the VAS collector impedance can be increased to get more local gain. This is straightforwardly done with a cascode configuration (see Figure 5.4d), but it should be said at once that the technique is only really useful when the VAS is not directly driving a markedly nonlinear impedance, such as that at the input of a Class-B output stage. Otherwise this nonlinear loading renders it largely a cosmetic feature. Assuming for the moment that this problem is dealt with either by use of a Class-A output or by VAS buffering, the drop in distortion is dramatic, as for the beta-enhancement method. The gain increase is ultimately limited by the Early effect in the cascode and current-source transistors, and more seriously by the loading effect of the next stage, but it is of the order of 10 times and gives a useful effect. This is shown by curves A and B in Figure 5.6, where once more the input stage of a model amplifier has been over-degenerated with 100 Ω emitter resistors to bring out the VAS distortion more clearly. Note that in both cases the slope of the distortion increase is 6 dB/octave. Curve C shows the result when a standard undegenerated input pair is combined with the cascoded VAS; the distortion is submerged in the noise floor for most of the audio band, being well below 0.001%. I think this justifies my contention that input stage and VAS distortions need not be problems; we have all but eliminated Distortions 1 and 2 from the list of eight in Chapter 3.

Using a cascode transistor also allows the use of a high-beta transistor for the VAS; these typically have a limited V_{ceo} that cannot withstand the high rail voltages of a high-power amplifier. There is a small loss of available voltage swing, but only about 300 mV, which is usually tolerable. Experiment shows that there is nothing to be gained by cascoding the current-source collector load.

A cascode topology is often used to improve frequency response, by isolating the upper collector from the C_{bc} of the lower transistor. In this case the frequency response is deliberately defined by C_{dom}, so this appears irrelevant, but in fact it is advantageous that C_{bc} – which carries the double demerit of being unpredictable and signal-dependent – is rendered harmless. Thus compensation is determined only by a well-defined passive component.

It is hard to say which technique is preferable; the beta-enhancing emitter-follower circuit is slightly simpler than the cascode version, which requires extra bias components, but the cost

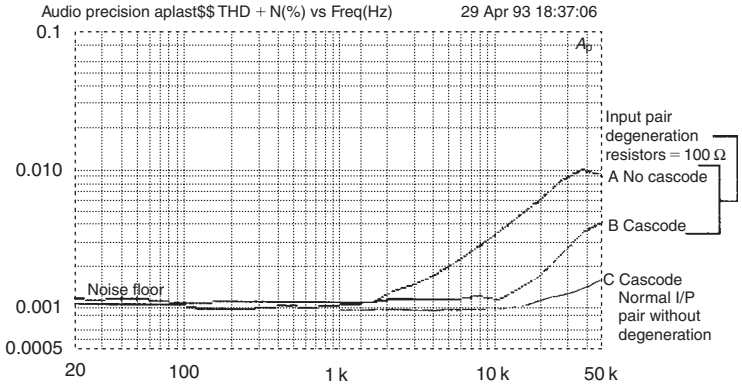


Figure 5.6: The reduction of VAS distortion possible by cascoding. The results from adding an emitter-follower to the VAS, as an alternative method of increasing local VAS feedback, are very similar

difference is tiny. When wrestling with these kinds of financial decisions it is as well to remember that the cost of a small-signal transistor is often less than a fiftieth of that of an output device, and the entire small-signal section of an amplifier usually represents less than 1% of the total cost, when heavy metal such as the mains transformer and heat-sinks are included.

Note that although the two VAS-linearizing approaches look very different, the basic strategy of increased *local* feedback is the same. Either method, properly applied, will linearize a VAS into invisibility.

Some More VAS Variations

The VAS configurations shown in Figure 5.4 by no means exhaust the possibilities. Figure 5.7a shows a different version of Figure 5.4c, where an emitter-follower has been added inside the local NFB loop. In Figure 5.7a the emitter-follower Q1 is now a PNP device. Its operating V_{ce} is limited to the V_{be} voltage of Q2, i.e. about 0.6V, but this should be enough. The collector current of Q1 is set by the value of R1, which could equally well be connected to ground as the signal voltage on its emitter is very small. (It is worth pointing out again that the presence of R2 does *not* mean that local negative feedback is being applied to the VAS, as it is a transadmittance stage with a current, not voltage, input.) This configuration has been used by Yamaha in recent designs, where the current through R1 has been pressed into further use for biasing a pair of diodes that define the VAS current-source reference voltage. Making double use of internal currents for biasing like this is ingenious but not always a good idea – it can lead to unexpectedly exotic behavior on clipping. Presumably it works all right in this case.

Figure 5.7b shows a variation on the cascoded VAS suggested by Hawksford[4]. The intention is apparently to reduce the V_{ce} variation on the VAS transistor Q1 by bootstrapping the cascode transistor Q2 from the emitter of Q1. Note that the emitter resistor R2 is not present to introduce

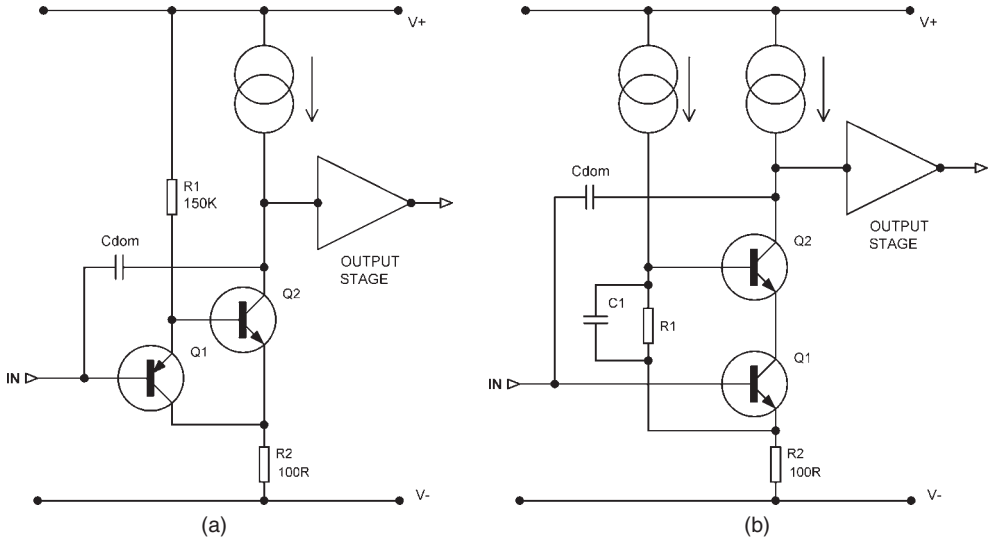


Figure 5.7: Some more VAS variations

local negative feedback; it is normally put there to allow current-sensing and also over-current protection of the VAS transistor when output stage overload circuitry is operating.

I remember trying this scheme out many years ago, but found no improvement in the overall distortion performance of the amplifier. This was almost certainly because, as previously stated, the distortion produced by a linearized VAS *without* using this enhancement is already well below the much more intractable distortion produced by a Class-B output stage.

VAS Operating Conditions

It is important to operate the VAS stage at a sufficiently high quiescent current. If a non-balanced VAS configuration is used then this current is fixed by the current-source load; it must be high enough to allow enough current drive for the top half of the output stage when the lowest load impedance contemplated is being driven to full output. The value of the current required obviously depends somewhat upon the design of the output stage. A high VAS quiescent current also has the potential to improve the maximum slew rate, but as described in Chapter 8, there are several important provisos to this. Typical quiescent current values are 5–20 mA.

If a VAS configuration without an emitter-follower is being used, then note must be taken of the base current it will draw from the input stage; it should not be allowed to unbalance the input transistor collector currents significantly.

The primary limitation on the VAS quiescent current is the dissipation of the transistors that make up this part of the circuit. There is a strong motivation to use TO92 package transistors as they have higher beta than medium-power devices in TO5 or TO-225 format; hence there is more gain available for local negative feedback and so greater linearity. However, the need to withstand high collector voltages in powerful amplifiers works against this as high-V_{ce} devices always have lower beta. As an example, the MPSA42 is often used in the VAS position; it can sustain 300 V but the minimum beta is only a humble 25. Its maximum dissipation is 625 mW.

Table 5.1 gives a quick comparison of the important parameters for a high-beta, low-V_{ce} transistor such as the BC337, the MPSA42, and the medium-power MJE340, which comes in a TO-225 format. (The beta figures in the table need a word of explanation. Firstly, the BC337 comes in three beta classifications: a -16 suffix means beta in the range 100–250, -25 means 160–400, and -40 means 250–630. The -25 variant seems to be by far the most common. Secondly, the minimum beta spec for MPSA42 is 25, which is actually less than the minimum of 30 for the much more power-capable MJE340. However, in real life the beta of MPSA42 is reliably higher than for MJE340, and it gives noticeably more linear results.)

Table 5.1 Parameters of possible VAS transistors

Type	Beta	$V_{ce(max)}$	$P_{diss(max)}$	Package
BC337-25	160–400	50 V	625 mW	TO92
MPSA42	25 min	300 V	625 mW	TO92
MJE340	30–240	300 V	20 W	TO-225

Let us examine these options, assuming the VAS quiescent current is chosen to be 10 mA.

If you use the BC337-25 then the supply rails are limited to ±25 V by V_{ce} max, and this restricts the theoretical maximum output power to 39 W into 8 Ω; in practice you'd be lucky to get 30 W. A BC337-25 could of course be cascoded with a higher-power device, to shield it from the high voltage; the extra complication is not great, and the cascoding itself may improve linearity, but the need to give the lower high-beta transistor a couple of volts of V_{ce} to work in will lead to an asymmetrical voltage-swing capability. So far as I can recall I haven't tried this approach but it looks highly plausible.

If you opt for the MPSA42 with its 300 V V_{ce} max, then supply rail voltages as such are not going to be a problem. However, if the VAS quiescent current is 10 mA, the amplifier supply rails will be limited to about ±50 V if the maximum package dissipation is taken as 500 mW, to provide some margin of safety. As it is, a TO92 package dissipating 500 mW is disconcertingly (and painfully) hot, but this sort of operation does not in my experience lead to reliability problems; I have used it many times in commercial designs and it works and keeps working. Small bent-metal heat-sinks that solder into the PCB are available for the TO92, and these are well worth using if you are pushing the dissipation envelope; I have a packet of brass ones in front of me that are labeled with a thermal resistance of 36°C/W. It is also good practice to use substantial PCB pads with thick tracks attached, so as much heat as possible can be lost down the legs of the transistor.

Another possibility, suggested to me by Glen Kleinschmidt, is the use of two TO92 VAS transistors in parallel, using the small resistors (circa 56 Ω) that are usually placed in the VAS emitter circuit for current limiting (*not* for local negative feedback) to ensure proper current sharing. The same approach could be used for the VAS current source; good current sharing is now inherent.

Taking the MJE340 route means that power dissipation is much less of a problem. A heat-sink will probably not be required though the transistor will not of course dissipate anything like 20 W without one. The distortion from the VAS stage will almost certainly be higher.

The VAS current-source load naturally dissipates as much power as the VAS as they carry the same current, and must be treated accordingly.

The Importance of Voltage Drive

As explained above, it is fundamental to linear VAS operation that the collector impedance is high, and not subject to external perturbations, thus permitting a large amount of local negative feedback. A Class-B output stage, with large input impedance variations around the crossover point, is about the worst thing you could connect to it, and it is a tribute to the general robustness of the conventional amplifier configuration that it can handle this internal unpleasantness gracefully, 100 W/8 Ω distortion typically degrading only from 0.0008% to 0.0017% at 1 kHz, assuming that the avoidable distortions have been eliminated. Note, however, that the degradation becomes greater as the global feedback factor is reduced. There is little deterioration at HF, where other distortions dominate. To the best of my knowledge I first demonstrated this in Ref.[10]; if someone feels that I am wrong then I have no doubt I shall soon hear about it.

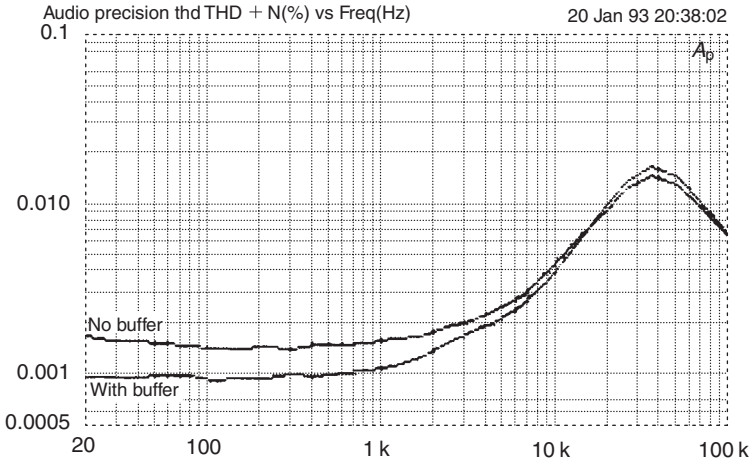


Figure 5.8: The beneficial effect of using a VAS buffer in a full-scale Class-B amplifier. Note that the distortion needs to be low already for the benefit to be significant

The VAS buffer is most useful when LF distortion is already low, as it removes Distortion 4, which is (or should be) only visible when grosser nonlinearities have been seen to. Two equally effective ways of buffering are shown in Figure 5.4e and f.

There are other potential benefits to VAS buffering. The effect of beta mismatches in the output stage halves is minimized[5]. Voltage drive also promises the highest fT from the output devices, and therefore potentially greater stability, though I have no data of my own to offer on this point. It is right and proper to feel trepidation about inserting another stage in an amplifier with global feedback, but since this is an emitter-follower its phase shift is minimal and it works well in practice.

If we have a VAS buffer then, providing we put it the right way up, we can implement a form of DC-coupled bootstrapping that is electrically very similar to providing the VAS with a separate current source (see Figure 5.4f). This variation may look a little unlikely, but I have used it in a commercial amplifier that was made in its tens of thousands, so I can assure the doubtful that it works as advertised.

The use of a buffer is essential if a VAS cascode is to do some good. Figure 5.8 shows before/after distortion for a full-scale power amplifier with cascode VAS driving 100 W into 8 Ω.

The use of a VAS buffer is not the only solution to the problem of nonlinear loading. Using an emitter-follower enhanced VAS (as shown in Figure 5.4c) also reduces the impedance at the VAS output because of the increased local feedback around the VAS.

The Push–Pull VAS

In previous editions, VAS configurations that had a signal-varying operating current (as opposed to a fixed operating current set by a constant-current source) were referred to a ‘balanced VAS’ but on mature consideration I have decided that the phrase ‘push–pull VAS’ is more accurate and more descriptive, and I have changed the nomenclature in this section accordingly.

When we are exhorted to 'make the amplifier linear before adding negative feedback', one of the few specific recommendations made is usually the use of a push–pull VAS – sometimes combined with a double input stage consisting of two differential amplifiers, one complementary to the other. The latter seems to have little to recommend it, as you cannot balance a stage that is already balanced, but a push–pull (and, by implication, more linear) VAS appears to have its attractions. However, as explained above, the distortion contribution from a properly designed VAS is negligible under most circumstances, and therefore there seems to be little to be gained.

Most amplifier configurations using a push–pull VAS generate two signals, one to drive the active element at the top of the VAS and one to drive that at the bottom. There are broadly two methods of doing this. In the first case there is one input stage differential pair, and the signals split off from the two collectors. In the second case, there are *two* input stages, and the output from each one drives the top or bottom of the VAS structure. The circuitry involved in both methods is described below.

A push–pull amplifier stage does not necessarily have to have two external drive signals; some configurations generate one of their drive signals by sensing their own internal current flow. An example is the Class-A output stage used in the model amplifier for input stage common-mode distortion investigations. This is described in Chapter 4. This might be a fruitful path of inquiry, as such a push–pull VAS would require only one drive signal and so simplify the input stage design.

The High-Current Capability VAS

A push–pull circuit is generally regarded as the most efficient available in terms of current delivery, unless you give up the linear operation of the circuit devices and opt for some form of Class-AB operation. This implies inferior distortion performance unless the high-current mode is strictly reserved for slew-rate testing and not used during the normal operation of the amplifier. An excellent starting point for the study of this sort of stage is Giovanni Stochino's fine article in *Electronics World*[6], in which he described input stages and a VAS that gave very high slew rates by operating in Class-AB.

Single Input Stages

Two possible versions of the single-input-stage configuration are shown in Figure 5.9. Here the VAS itself is also a differential pair, driven by both outputs of the first pair. This sort, which I will call Type 1, gives approximately 10 dB more O/L gain than the standard amplifier configuration, which naturally requires an increase in C_{dom} if the same stability margins are to be maintained. In a model amplifier, where C_{dom} need not be increased to achieve stability, any improvement in linearity can be wholly explained by this increase in open-loop gain, so if we are seeking greater open-loop linearity, this seems (not unexpectedly) an unpromising approach. Also, as Linsley-Hood has pointed out[7], the standing current through the output stage bias generator is ill-defined compared with the usual current-source VAS; this is of vital importance because in any Class-B output stage the accuracy of the bias voltage is critical in the minimization of crossover

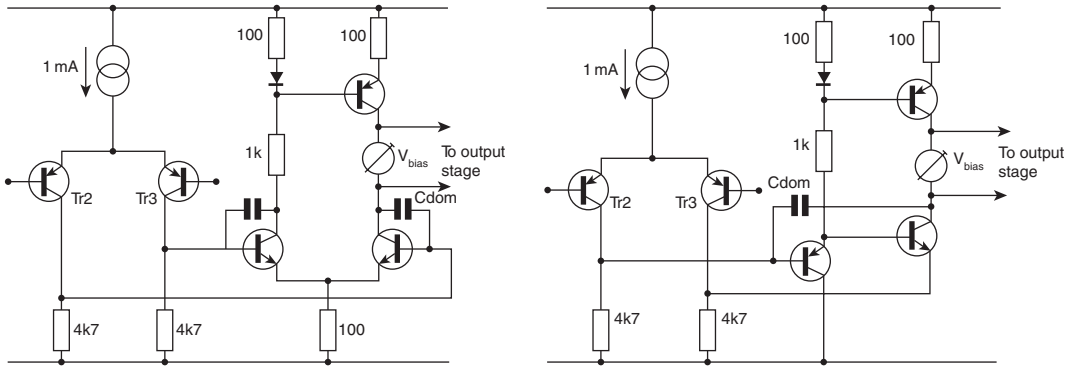


Figure 5.9: Two kinds of push–pull VAS. Type 1 gives more open-loop gain, but no better open-loop linearity. Type 2 is the circuit originated by Lender

distortion. While an ideal bias generator would show a zero effective series resistance – in other words the voltage across it would not vary as the current through it changed – in the real world bias generators are usually simple one-transistor circuits that fall some way short of this ideal (see Chapter 15 for more on this).

Similarly the balance of the input pair is likely to be poor compared with the current-mirror version. A further difficulty is that there are now two signal paths from the input stage to the VAS output, and it is difficult to ensure that these have exactly the same bandwidth; if they do not then a pole-zero doublet is generated in the open-loop gain characteristic that will markedly increase settling time after a transient. This seems likely to apply to all balanced VAS configurations, as they must have two signal paths in one way or another. Whether this is in any way audible is another matter – it seems most unlikely. If you want to dig deeper into the matter of frequency doublets – which have no connection with medieval clothing – then Dostal[8] is an excellent reference.

The exact origin of the Type 1 configuration is hard to pin down, but it certainly became popular in 1977 as the VAS stage to drive Hitachi MOSFETs when they were introduced.

Type 2 is attributed to Borbely and Lender[9]. Figure 5.9 shows one version, with a quasi-balanced drive to the VAS transistor, via both base and emitter. This configuration does not give good balance of the input pair, as this is at the mercy of the tolerances of the input stage collector resistors, the V_{be} of the VAS, and so on. Borbely has advocated using two complementary versions of this, and this approach is dealt with in the next section.

Another circuit variation using a single input stage is shown in Figure 5.10. This is usually called the folded-cascode configuration, because Q4, Q5 are effectively cascoding the collectors of input stage Q2, Q3. While it has been used extensively in op-amps, it has only rarely been applied to audio power amplifiers. The distinguishing characteristic of this configuration is that the two transistors Q4, Q5 are common-base stages and their emitters are driven from the collectors of input stage Q2, Q3. In op-amp usage the two resistors R1, R2 are normally replaced by constant-current sources. Q1 is a cascode transistor for the collector of Q4 (sort of a cascode of a cascode); it is not an essential part of

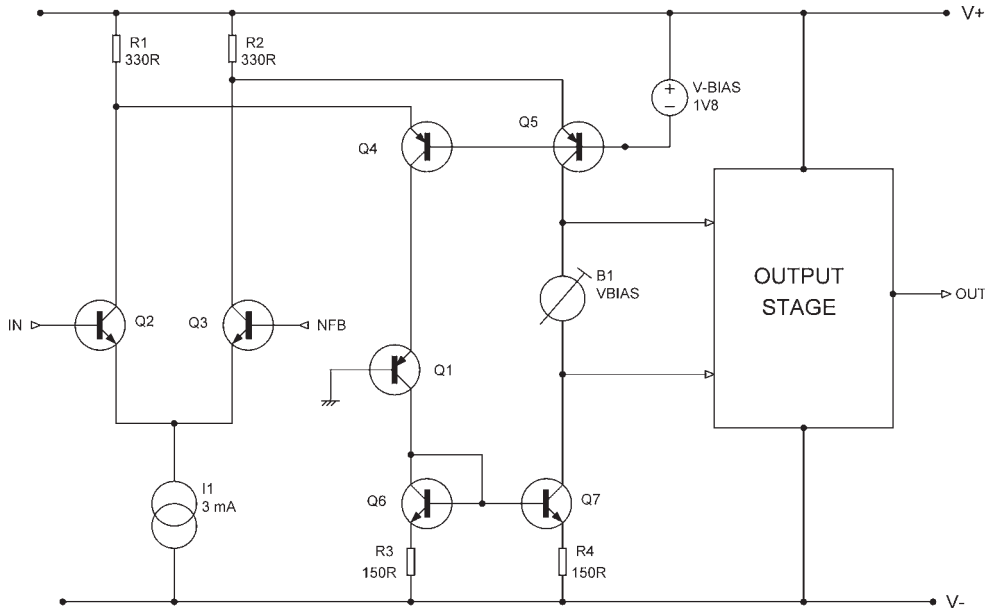


Figure 5.10: Folded-cascode configuration giving drive to top and bottom of the push–pull VAS stage Q5, Q7 via the current-mirror Q6, Q7

the folded-cascode concept. The current output of Q1 is bounced off the V− rail by the current-mirror Q6, Q7 and provides the lower part of the push–pull drive to the VAS stage Q5, Q7.

I might as well come clean at once and admit that I have no practical experience with power amplifiers using this configuration, but a few thoughts do occur. Firstly, there would appear to be a lack of overall open-loop gain because the common-base stages Q4, Q5 do not give any current gain. Secondly, there is no obvious way to apply the Miller dominant-pole compensation that is so very useful in linearizing a VAS.

I am not sure if any commercial amplifiers have been built using the folded-cascode structure, but at least one such design has been published for amateur construction by Michael Bittner, and the circuit values here are derived from this. To the best of my knowledge no performance figures have been published.

Double Input Stages

If two differential input stages are used, with one the complement of the other, then the two VAS drive signals are conveniently referenced to the top and bottom supply rails. The basic circuit is shown in Figure 5.11, with representative component values and operating currents.

Note that in Figure 5.11 the differential pairs have their collector currents approximately balanced by correct resistor values, but not held exactly correct by current-mirrors. This omission is just to simplify the diagram. For the same reason no degeneration resistors have been placed in series with the input device emitters, and the compensation components are also omitted.

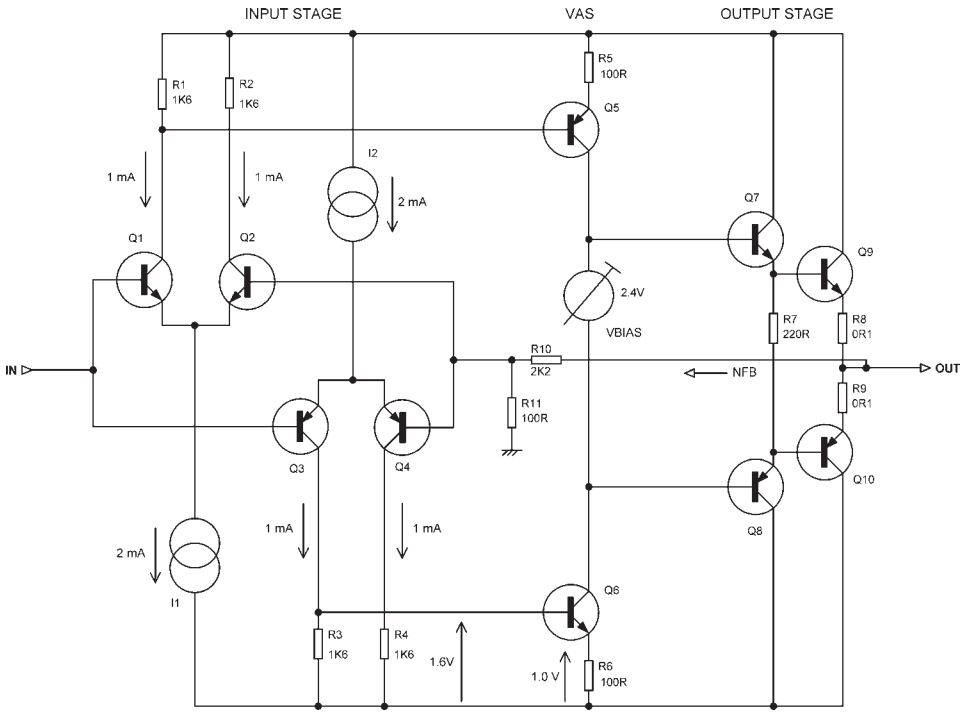


Figure 5.11: Double differential input configuration giving drive to top and bottom of push–pull VAS

As noted above, any push–pull VAS scheme has to be carefully examined to see exactly how accurately its standing current is defined, as this passes through the output stage bias generator. The current through the VAS is now not fixed directly by the VAS collector load-current source, but indirectly by the input pair tail-current sources. These two sources will not be of exactly the same value, but only one value of current can flow through the VAS (the base currents of the output stage drivers are assumed to be negligible compared with the VAS standing current when the amplifier is quiescent).

The use of two input differential pairs rather than one does not give the same dramatic reduction in input stage distortion that we get when going from a single transistor input to a differential pair; in that move we canceled out the second-order nonlinearities of the input stage and they cannot be canceled out twice over. However, there may be something to be gained in terms of input stage distortion; if the drive signals to the VAS are correctly proportioned then it should be possible to have each input differential pair working only half as hard as a single one. This would halve the voltage seen by each pair, reducing its distortion (which is effectively all third-order) by a factor of 4.

There is another possible advantage to the use of double input stages. If we can assume that the gain of the two signal paths is simply summed, with equal contributions from each, then noise from the input stage should be reduced by 3 dB as two uncorrelated noise sources add by rms summation rather than simple addition.

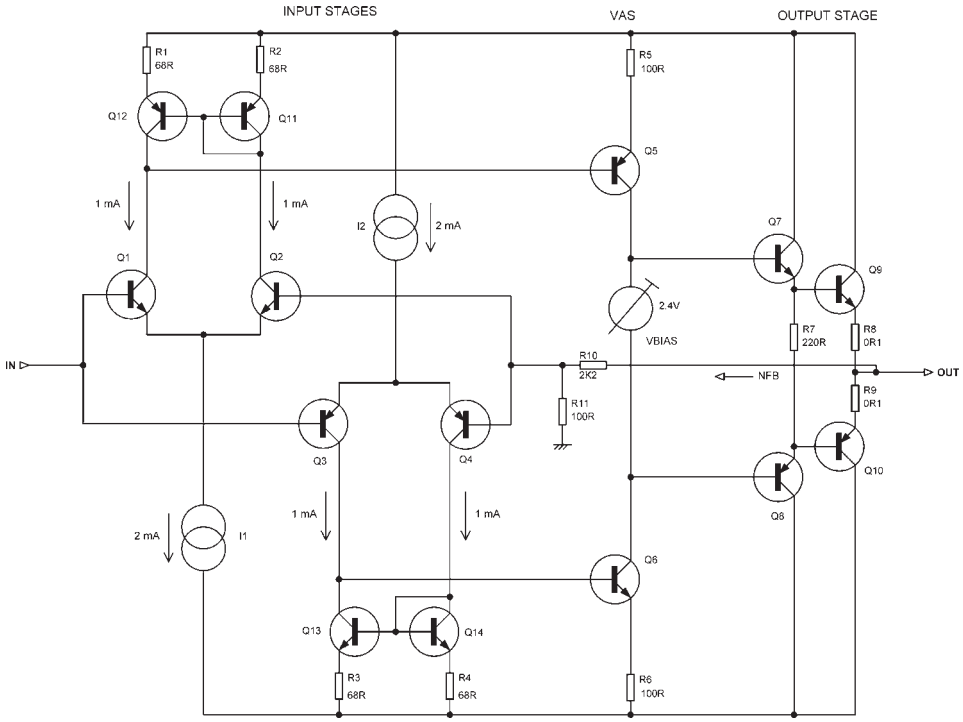


Figure 5.12: Double differential input configuration with current-mirrors

One downside of the double input stage philosophy is that the increase in complexity is significant. The increased current consumption is trivial, and the component cost small, but the extra complexity may make fault-finding a bit more difficult, and of course more PCB area is required.

As I said earlier, there is no reason not to adopt current-mirrors when using a double input stage and every reason to do so. This gives us the configuration in Figure 5.12.

As with the conventional single-ended VAS, an emitter-follower can be added within the Miller compensation loop to increase local negative feedback and so improve linearity. If we apply this to the configuration of Figure 5.11 then we get the circuit shown in Figure 5.13. It would of course be possible to add current-mirrors also. This may seem like an awful lot of transistors, but they are all small-signal types of low cost and high reliability.

A third method of generating push–pull drive signals is possible, and has been used occasionally in commercial equipment. Here there are still two input transistors, but they are complementary rather than the same type, and are connected in series rather than in the parallel format of the conventional long-tailed pair. The two output signals are once again conveniently referenced to the top and bottom supply rails (see Figure 5.14, which includes representative component values and operating currents). Note that two complementary level-shifting emitter-followers Q1, Q3 are required at the input, so that Q2 and Q4 have enough V_{ce} to operate. The version shown here has the collector currents from the input emitter-followers fed into the emitters of Q2, Q4.

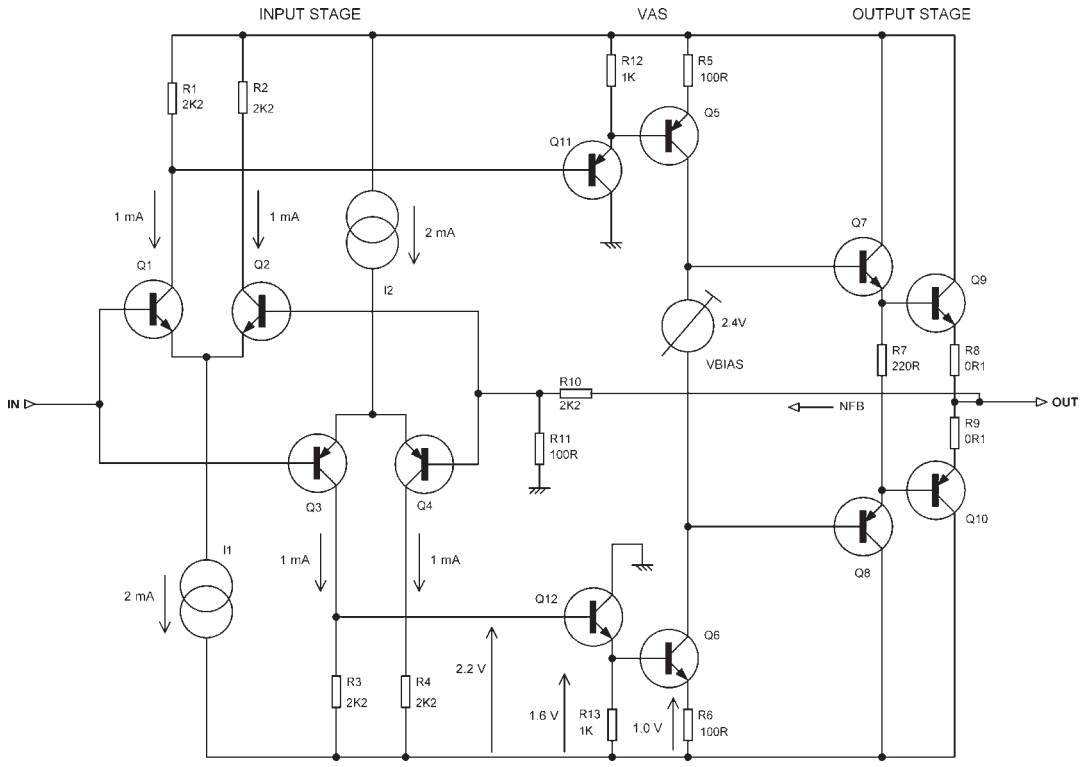


Figure 5.13: Double differential input configuration with emitter-follower VAS enhancement

A serious objection to this configuration is that it tries to cancel nonlinearity and temperature effects in two transistors that are not of the same type. Even so-called complementary pairs are not exact mirror images of each other. It is significant that in every example of this configuration that I have seen, a DC servo has been fitted to give an acceptable output offset voltage. The complementary emitter-followers in front of the gain devices are also expected to cancel the V_{be} values of the latter, which introduces more questions about accuracy.

This approach presents some interesting problems with the definition of the operating conditions. Note the degeneration resistors R8, R9. These are essential to define the collector current passing through Q2, Q4; the current through the series input stage depends on a relatively small voltage established across these two low-value resistors. In contrast, in a conventional differential pair the value of the emitter-degeneration resistors has no effect at all on the operating current, which is set by the tail-current source.

It is not easy to assess linearity of this configuration as it does not give a single current output, but two that are combined at the output of the push–pull VAS stage. It is therefore more difficult to separate the nonlinearities of the input and VAS stages in practical measurement. It should be easier in SPICE simulation as the two collector currents can be subtracted mathematically.

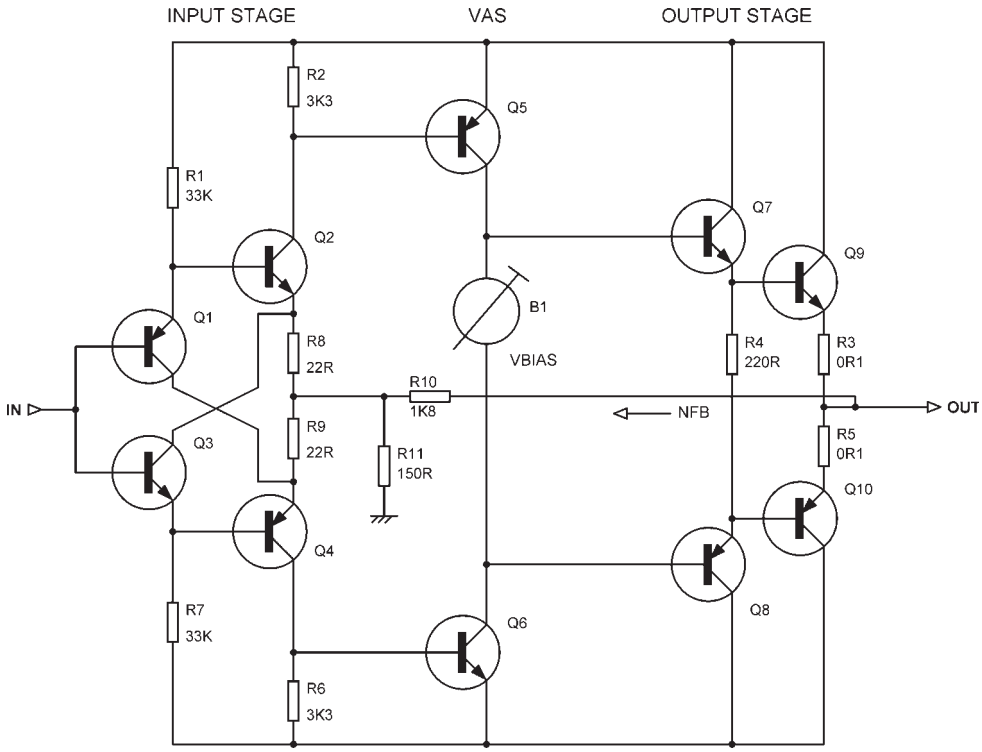


Figure 5.14: Series differential input configuration

This can be only a brief examination of push–pull VAS stages; many configurations are possible, and a comprehensive study of them all would be a major undertaking. All seem to be open to the objection that the standing current through the bias generator is not well defined. In some versions the vital balance of the input pair is not guaranteed. However, one advantage would seem to be the potential for sourcing and sinking large currents into C_{dom}, which might improve the ultimate slew rate and HF linearity of a very fast amplifier.

Manipulating Open-Loop Bandwidth

Acute marketing men will by now have realized that reducing the LF O/L gain, leaving HF gain unchanged, must move the *P*1 frequency upwards, as shown in Figure 5.15. 'Open-loop gain is held constant up to 2 kHz' sounds so much better than 'the open-loop bandwidth is restricted to 20 Hz', although these two statements could describe near-identical amplifiers, except that the first has plenty of open-loop gain at LF while the second has even more than that. Both amplifiers have the same feedback factor at HF, where the amount available has a direct effect on distortion performance, and could easily have the same slew rate. Nonetheless the second amplifier somehow reads as sluggish and indolent, even when the truth of the matter is known.

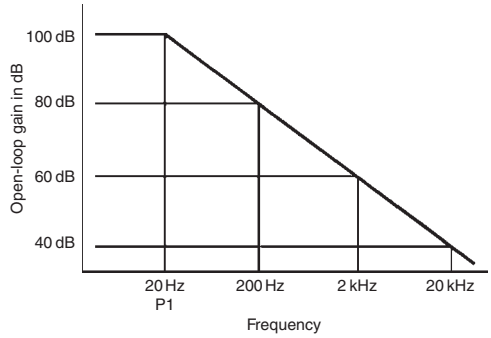


Figure 5.15: How dominant-pole frequency *P*1 can be altered by changing the LF open-loop gain; the gain at HF, which determines Nyquist stability and HF distortion, is unaffected

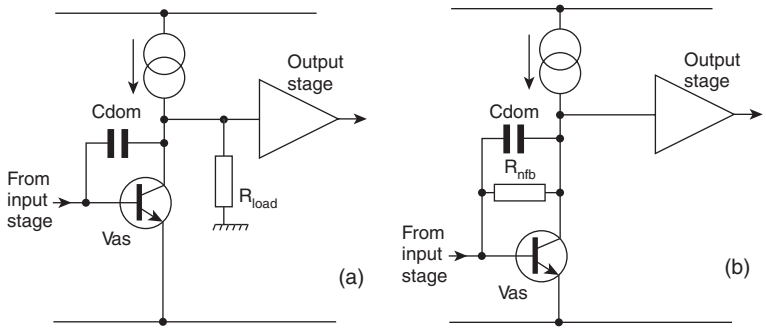


Figure 5.16: Two ways to reduce O/L gain. (a) By simply loading down the collector. This is a cruel way to treat a VAS; current variations cause extra distortion. (b) Local NFB with a resistor in parallel with C_{dom}. This looks crude, but actually works very well

It therefore follows that reducing the LF O/L gain may be of interest to commercial practitioners. Low values of open-loop gain also have their place in the dogma of the subjectivist, and the best way to bring about this state of affairs is worth examining, always bearing in mind that:

1. there is no engineering justification for it;
2. reducing the NFB factor will reveal more of the output stage distortion; since in general NFB is the only weapon we have to deal with this, blunting its edge seems ill-advised.

It is of course simple to reduce O/L gain by degenerating the input pair, but this diminishes it at HF as well as LF. To alter it at LF only it is necessary to tackle the VAS instead, and Figure 5.16 shows two ways to reduce its gain. Figure 5.16a reduces gain by reducing the value of the collector impedance, having previously raised it with the use of a current-source collector load. This is no way to treat a gain stage; loading resistors low enough to have a significant effect cause unwanted current variations in the VAS as well as shunting its high collector impedance, and serious LF distortion appears. While this sort of practice has been advocated in the past[10], it seems to have nothing to recommend it as it degrades VAS linearity at the same time as siphoning off the

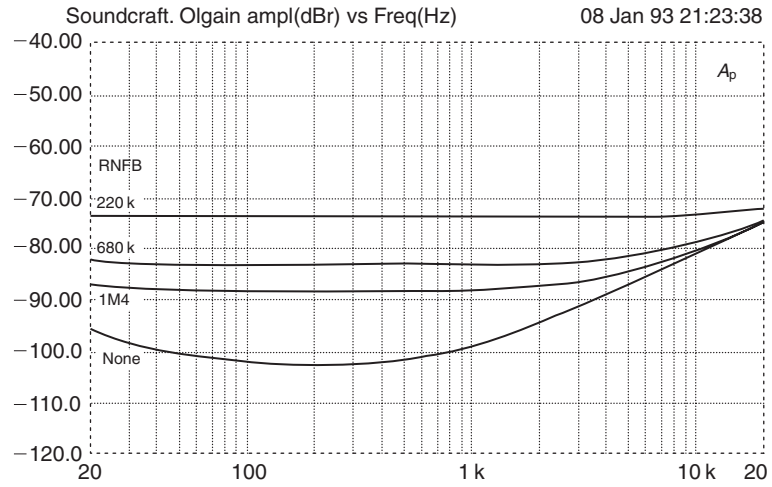


Figure 5.17: The result of VAS gain reduction by local feedback; the dominant-pole frequency is increased from about 800 Hz to about 20 kHz, with high-frequency gain hardly affected

feedback that would try to minimize the harm. Figure 5.16b also reduces overall O/L gain, but by adding a frequency-insensitive component to the local shunt feedback around the VAS. The value of R_{nfb} is too high to load the collector significantly and therefore the full gain is available for local feedback at LF, even before C_{dom} comes into action.

Figure 5.17 shows the effect on the open-loop gain of a model amplifier for several values of R_{nfb}; this plot is in the format described in Chapter 3, where error voltage is plotted rather than gain directly, and so the curve once more appears upside down compared with the usual presentation. Note that the dominant-pole frequency is increased from 800 Hz to above 20 kHz by using a 220 k value for R_{nfb}; however, the gain at higher frequencies is unaffected and so is the stability. Although the amount of feedback available at 1 kHz has been decreased by nearly 20 dB, the distortion at +16 dBu output is only increased from less than 0.001% to 0.0013%; most of this reading is due to noise.

In contrast, reducing the open-loop gain even by 10 dB by loading the VAS collector to ground requires a load of 4 k7, which under the same conditions yields distortion of more than 0.01%.

If the value of R_{nfb} required falls below about 100 k, then the standing current flowing through it can become large enough to upset the amplifier operating conditions (Figure 5.16b). This is revealed by a rise in distortion above that expected from reducing the feedback factor, as the input stage becomes unbalanced as a result of the global feedback straightening things up. This effect can be simply prevented by putting a suitably large capacitor in series with R_{nfb}. A 2 µ2 non-electrolytic works well, and does not cause any strange response effects at low frequencies.

An unwelcome consequence of reducing the global negative feedback is that power-supply rejection is impaired (see Chapter 9 on PSRR). To prevent negative supply-rail ripple reaching the output it is necessary to increase the filtering of the V-rail that powers the input stage and the VAS. Since the voltage drop in an RC filter so used detracts directly from the output voltage swing, there

are severe restrictions on the highest resistor value that can be tolerated. The only direction left to go is increasing C, but this is also subject to limitations as it must withstand the full supply voltage and rapidly becomes a bulky and expensive item.

That describes the 'brawn' approach to improving PSRR. The 'brains' method is to use the input cascode compensation scheme described in Chapter 9. This solves the problem by eliminating the change of reference at the VAS, and works extremely well with no compromise on HF stability. No filtering at all is now required for the V-supply rail – it can feed the input stage and VAS directly.

Conclusions

This chapter showed how the strenuous efforts of the input circuitry can be best exploited by the voltage-amplifier stage following it. At first it appears axiomatic that the stage providing all the voltage gain of an amplifier, at the full voltage swing, is the prime suspect for generating a major part of its nonlinearity. In actual fact, this is unlikely to be true, and if we select for an amplifier a cascode VAS with current-source collector load and buffer it from the output stage, or use a beta-enhancer in the VAS, the second of our eight distortions is usually negligible.

References

[1] P.P. Gray, R.G. Meyer, Analysis and Design of Analog Integrated Circuits, Wiley, 1984, p. 251 (VAS transfer characteristic).

[2] Antognetti (Ed.), Power Integrated Circuits, McGraw-Hill, 1986, p. 9.31.

[3] P.P. Gray, R.G. Meyer, Analysis and Design of Analog Integrated Circuits, Wiley, 1984, p. 252 (R_{co} limit on VAS gain).

[4] M. Hawksford, Reduction of transistor slope impedance dependent distortion in large-signal amplifiers, JAES 36 (4) (April 1988) (enhanced cascode VAS).

[5] B. Oliver, Distortion in complementary-pair Class-B amplifiers, Hewlett-Packard Journal (February 1971) p. 11.

[6] G. Stochino, Ultra-fast amplifier, Electronics & Wireless World (October 1996) p. 835.

[7] J. Linsley-Hood, Solid state audio power – 3, Electronics & Wireless World (January 1990) p. 16.

[8] J. Dostal, Operational Amplifiers, Butterworth-Heinemann, 1993, p. 195.

[9] E. Borbely, A 60W MOSFET power amplifier, Audio Amateur (2) (1982) p. 9.

[10] J. Hefley, High fidelity, low feedback, 200W, Electronics & Wireless World (June 1992) p. 454.

CHAPTER 6

The Output Stage

Classes and Devices

The almost universal choice in semiconductor power amplifiers is for a unity-gain output stage, and specifically a voltage-follower. Output stages with gain are not unknown – see Mann[1] for a design with 10 times gain in the output section – but they have significantly failed to win popularity. Most people feel that controlling distortion while handling large currents is quite hard enough without trying to generate gain at the same time. Nonetheless, I have now added a section on output stages with gain to this chapter.

In examining the small-signal stages of a power amplifier, we have so far only needed to deal with one kind of distortion at a time, due to the monotonic transfer characteristics of such stages, which usually (but not invariably[2]) work in Class-A. Economic and thermal realities mean that most output stages are Class-B, and so we must now also consider crossover distortion (which remains the thorniest problem in power amplifier design) and HF switch-off effects.

We must also decide what *kind* of active device is to be used; JFETs offer few if any advantages in the small-current stages, but power FETs in the output appear to be a real possibility, providing that the extra cost proves to bring with it some tangible benefits.

The most fundamental factor in determining output stage distortion is the class of operation. Apart from its inherent inefficiency, Class-A is the ideal operating mode, because there can be no crossover or switch-off distortion. However, of those designs that have been published or reviewed, it is notable that the large-signal distortion produced is still significant. This looks like an opportunity lost, as of the distortions enumerated in Chapter 3, we now only have to deal with Distortion 1 (input stage), Distortion 2 (VAS), and Distortion 3 (output stage large-signal nonlinearity). Distortions 4–7, as mentioned earlier, are direct results of Class-B operation and therefore can be thankfully disregarded in a Class-A design. However, Class-B is overwhelmingly of greater importance, and is therefore dealt with in detail below.

Class-B is subject to much misunderstanding. It is often said that a pair of output transistors operated without any bias are 'working in Class-B' and therefore 'generate severe crossover distortion'. In fact, with no bias each output device is operating for slightly less than half the time, and the question arises as to whether it would not be more accurate to call this Class-C and reserve Class-B for that condition of quiescent current which eliminates, or rather minimizes, the crossover artefacts.

There is a further complication; it is not generally appreciated that moving into what is usually called Class-AB, by increasing the quiescent current, does *not* make things better. In fact, if the output power is above the level at which Class-A operation can be sustained, the THD reading

will certainly increase as the bias control is advanced. This is due to what is usually called g_m-*doubling* (i.e. the voltage-gain increase caused by both devices conducting simultaneously in the centre of the output-voltage range – that is, in the Class-A region) putting edges into the distortion residual that generate high-order harmonics much as underbiasing does. This vital fact seems almost unknown, presumably because the g_m-doubling distortion is at a relatively low level and is completely obscured in most amplifiers by other distortions.

This phenomenon is demonstrated in Figure 6.1a–c, which shows spectrum analysis of the distortion residuals for underbiasing, optimal, and overbiasing of a 150 W/8 Ω amplifier at 1 kHz. As before, all nonlinearities except the unavoidable Distortion 3 (output stage) have been effectively eliminated. The over-biased case had the quiescent current increased until the g_m-doubling edges in the residual had an approximately 50:50 mark/space ratio, and so it was in Class-A about half the time, which represents a rather generous amount of quiescent current for Class-AB. Nonetheless, the higher-order odd harmonics in Figure 6.1c are at least 10 dB greater in amplitude than those for the optimal Class-B case, and the third harmonic is actually higher than for the underbiased case as well. However, the underbiased amplifier, generating the familiar sharp spikes on the residual, has a generally greater level of high-order odd harmonics above the fifth, about 8 dB higher than the AB case.

Since high-order odd harmonics are generally considered to be the most unpleasant, there seems to be a clear case for avoiding Class-AB altogether, as it will always be less efficient and generate more high-order distortion than the equivalent Class-B circuit as soon as it leaves Class-A. Class distinction seems to resolve itself into a binary choice between A or B.

It must be emphasized that these effects are only visible in an amplifier where the other forms of distortion have been properly minimized. The RMS THD reading for Figure 6.1a was 0.00151%, for Figure 6.1b 0.00103%, and for Figure 6.1c 0.00153%. The tests were repeated at the 40 W power level with very similar results. The spike just below 16 kHz is interference from the test-gear VDU.

This is complex enough, but there are other and deeper subtleties in Class-B, which are dealt with below.

The Distortions of the Output

I have called the distortion produced directly by the output stage Distortion 3 (see Chapters 3 and 6) and this can now be subdivided into three categories. Distortion 3a describes the large-signal distortion that is produced by both Class-A and -B, ultimately because of the large current swings in the active devices; in bipolars, but not FETs, large collector currents reduce the beta, leading to drooping gain at large output excursions. I shall use the term ‘LSN’ for large-signal nonlinearity, as opposed to crossover and switch-off phenomena that cause trouble at all output levels.

These other two contributions to Distortion 3 are associated with Class-B and -AB only; Distortion 3b is classic crossover distortion, resulting from the non-conjugate nature of the output characteristics, and is essentially non-frequency dependent. In contrast, Distortion 3c is switch-off distortion, generated by the output devices failing to turn off quickly and cleanly at

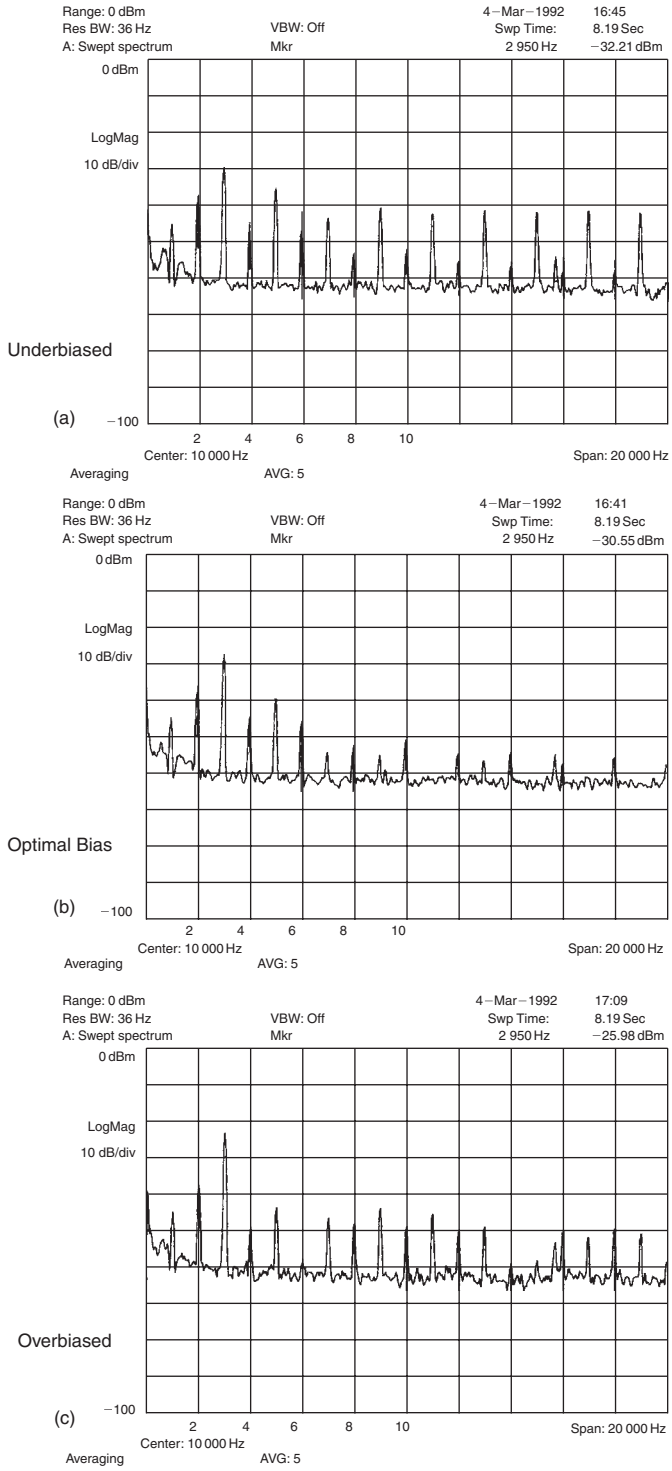


Figure 6.1: Spectrum analysis of Class-B and AB distortion residual

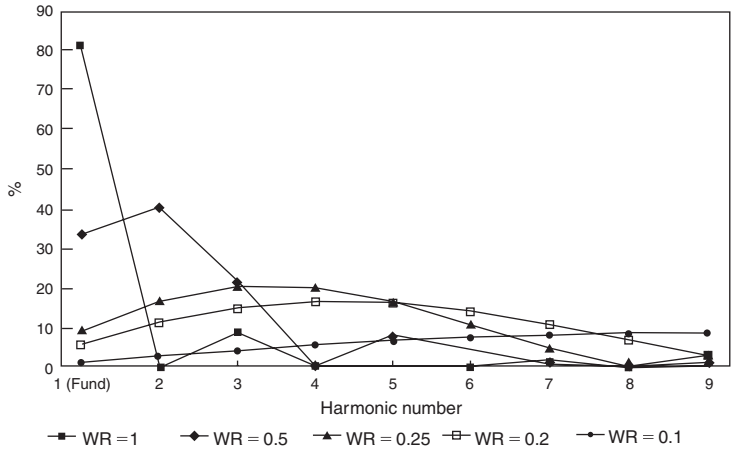


Figure 6.2: The amplitude of each harmonic changes with *WR*; as the error waveform gets narrower, energy is transferred to the higher harmonics

high frequencies, and is very strongly frequency-dependent. It is sometimes called 'switching distortion', but this allows room for confusion, as some writers use the term 'switching distortion' to cover crossover distortion as well; hence I have used the term 'switch-off distortion' to refer specifically to charge-storage turn-off troubles. Since Class-B is almost universal, and regrettably introduces all three kinds of nonlinearity, in this chapter we will concentrate on this kind of output stage.

Harmonic Generation by Crossover Distortion

The usual nonlinear distortions generate most of their unwanted energy in low-order harmonics that NFB can deal with effectively. However, crossover and switching distortions that warp only a small part of the output swing tend to push energy into high-order harmonics, and this important process is demonstrated here, by Fourier analysis of a SPICE waveform.

Taking a sine-wave fundamental, and treating the distortion as an added error signal E, let the ratio WR describe the proportion of the cycle where E is non-zero. If this error is a triangular wave extending over the whole cycle ($WR = 1$) this would represent large-signal nonlinearity, and Figure 6.2 shows that most of the harmonic energy goes into the third and fifth harmonics; the even harmonics are all zero due to the symmetry of the waveform.

Figure 6.3 shows how the situation is made more like crossover or switching distortion by squeezing the triangular error into the centre of the cycle so that its value is zero elsewhere; now E is non-zero for only half the cycle (denoted by $WR = 0.5$) and Figure 6.2 shows that the even harmonics are no longer absent. As WR is further decreased, the energy is pushed into higher-order harmonics, the amplitude of the lower falling.

The high harmonics have roughly equal amplitude, spectrum analysis (see Figure 6.1) confirming that even in a Blameless amplifier driven at 1 kHz, harmonics are freely generated from the seventh to the 19th at an equal level to a decibel or so. The 19th harmonic is only 10 dB below the third.

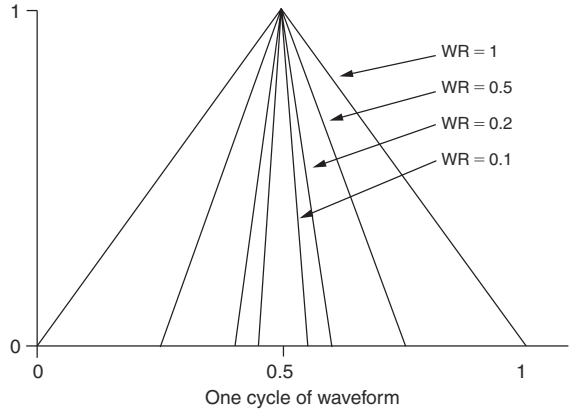


Figure 6.3: Diagram of the error waveform *E* for some values of *WR*

Thus, in an amplifier with crossover distortion, the order of the harmonics will decrease as signal amplitude reduces and *WR* increases; their lower frequencies allow them to be better corrected by the frequency-dependent NFB. This effect seems to work *against* the commonly assumed rise of percentage crossover distortion as level is reduced.

Comparing Output Stages

One of my aims in this book is to show how to isolate each source of distortion so that it can be studied (and hopefully reduced) with a minimum of confusion and perplexity. When investigating output behaviour, it is perfectly practical to drive output stages open-loop, providing the driving source impedance is properly specified; this is difficult with a conventional amplifier, as it means the output must be driven from a frequency-dependent impedance simulating that at the VAS collector, with some sort of feedback mechanism incorporated to keep the drive voltage constant.

However, if the VAS is buffered from the output stage by some form of emitter-follower, as advocated in Chapter 5, it makes things much simpler, a straightforward low-impedance source (e.g. 50 Ω) providing a good approximation of conditions in a VAS-buffered closed-loop amplifier. The VAS buffer makes the system more designable by eliminating two variables – the VAS collector impedance at LF, and the frequency at which it starts to decrease due to local feedback through C_{dom}. This markedly simplifies the study of output stage behavior.

The large-signal linearity of various kinds of open-loop output stage with typical values are shown in Figures 5.6–5.16. These diagrams were all generated by SPICE simulation, and are plotted as incremental output gain against output voltage, with the load resistance stepped from 16 to 2 Ω, which I hope is the lowest impedance that feckless loudspeaker designers will throw at us. They have come to be known as *wingspread* diagrams, from their vaguely bird-like appearance. The power devices are MJ802 and MJ4502, which are more complementary than many so-called pairs, and minimize distracting large-signal asymmetry. The quiescent conditions are in each case set to minimize the peak deviations of gain around the crossover point for 8 Ω loading; for the moment it is assumed that you can set this accurately and keep it where you want it. The difficulties in actually doing this will be examined later.

Table 6.1: Configurations of output stages

Configuration	No. of types	Illustration/description
Emitter-follower	3	Figure 6.4
Complementary feedback pair	1	Figure 6.5
Quasi-complementary	2	Figure 6.5
Output triples	At least 7	Figure 6.6
Power FET	3	Chapter 14

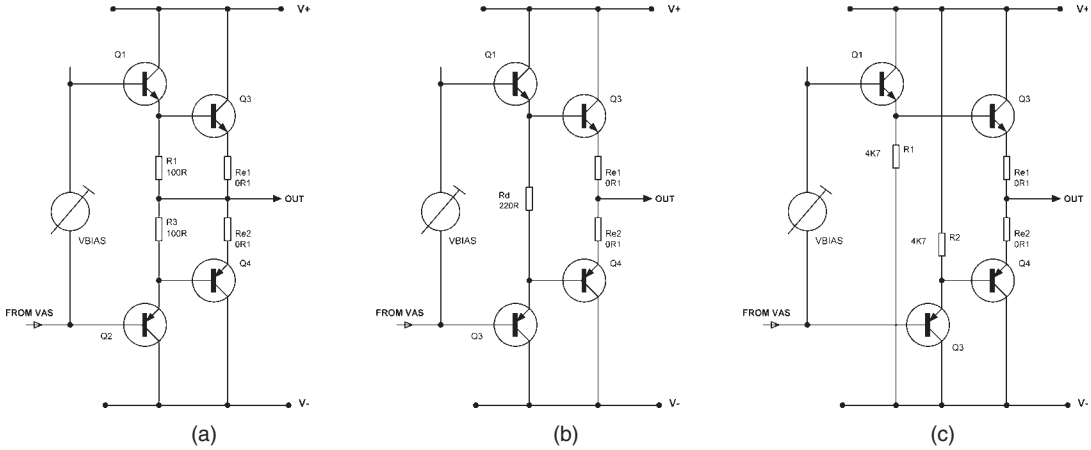


Figure 6.4: Three types of emitter-follower output stages

If we confine ourselves to the most straightforward output stages, there are at least 16 distinct configurations, without including error-correcting[3], current-dumping[4], or Blomley[5] types. These are summarized in Table 6.1.

The Emitter-Follower (EF) Output

Three versions of the most common type of output stage are shown in Figure 6.4; this is the double-emitter-follower, where the first follower acts as driver to the second (output) device. I have deliberately called this an emitter-follower (EF) rather than a Darlington configuration, as this latter implies an integrated device that includes driver, output, and assorted emitter resistors in one ill-conceived package (ill-conceived for this application because the output devices heat the drivers, making thermal stability worse). As for all the circuitry here, the component values are representative of real practice. Important attributes of this topology are:

1. The input is transferred to the output via two base–emitter junctions in series, with no local feedback around the stage (apart from the very local 100% voltage feedback that makes an EF what it is).

2. There are two dissimilar base–emitter junctions between the bias voltage and the emitter resistor Re, carrying different currents and at different temperatures. The bias generator must

attempt to compensate for both at once, though it can only be thermally coupled to one. The output devices have substantial thermal inertia, and so any thermal compensation can only be a time-average of the preceding conditions. Figure 6.4a shows the most prevalent version (Type I), which has its driver emitter resistors connected to the output rail.

The Type II EF configuration in Figure 6.4b is at first sight merely a pointless variation on Type I, but in fact it has a valuable extra property. The shared driver emitter resistor Rd, with no output-rail connection, allows the drivers to reverse-bias the base–emitter junction of the output device being turned off. Assume that the output voltage is heading downwards through the crossover region; the current through Re1 has dropped to zero, but that through Re2 is increasing, giving a voltage drop across it, so Q4 base is caused to go more negative to get the output to the right voltage. This negative excursion is coupled to Q3 base through Rd, and with the values shown can reverse-bias it by up to $-0.5\,\text{V}$, increasing to $-1.6\,\text{V}$ with a $4\,\Omega$ load. A speed-up capacitor Cs connected across Rd markedly improves this action, preventing the charge-suckout rate being limited by the resistance of Rd. While the Type I circuit has a similar voltage drop across Re2, the connection of the mid-point of R1, R2 to the output rail prevents this from reaching Q3 base; instead Q1 base is reverse-biased as the output moves negative, and since charge storage in the drivers is usually not a problem, this does little good. In Type II, the drivers are never reverse-biased, though they do turn off. The Type II EF configuration has of course the additional advantage that it saves a resistor!

The important issue of output turn-off and switching distortion is further examined in Chapter 6.

The Type III topology shown in Figure 6.4c maintains the drivers in Class-A by connecting the driver Re resistors to the opposite supply rail, rather than the output rail. It is a common misconception[6] that Class-A drivers somehow maintain better low-frequency control over the output devices, but I have yet to locate any advantage myself. The driver dissipation is of course substantially increased, and nothing seems to be gained at LF as far as the output transistors are concerned, for in both Type I and Type II the drivers are still conducting at the moment the outputs turn off, and are back in conduction before the outputs turn on, which would seem to be all that matters. Type III is equally good as Type II at reverse-biasing the output bases, and may give even cleaner HF turn-off as the carriers are being swept from the bases by a higher resistance terminated in a higher voltage, approximating constant-current drive; this remains to be determined by experiment. The Type III topology is used in the Lohstroh and Otala amplifier described in Chapter 2.

The large-signal linearity of these three versions is virtually identical; all have the same feature of two base–emitter junctions in series between input and load. The gain/output voltage plot is shown in Figure 6.7; with BJTs the gain reduction with increasing loading is largely due to the Re resistors. Note that the crossover region appears as a relatively smooth wobble rather than a jagged shape. Another major feature is the gain-droop at high output voltages and low loads, and this gives us a clue that high collector currents are the fundamental cause of this. A close-up of the crossover region gain for $8\,\Omega$ loading only is shown in Figure 6.8; note that no V_{bias} setting can be found to give a constant or even monotonic gain; the double-dip and central gain peak are characteristic of optimal adjustment. The region extends over an output range of about $\pm 5\,\text{V}$.

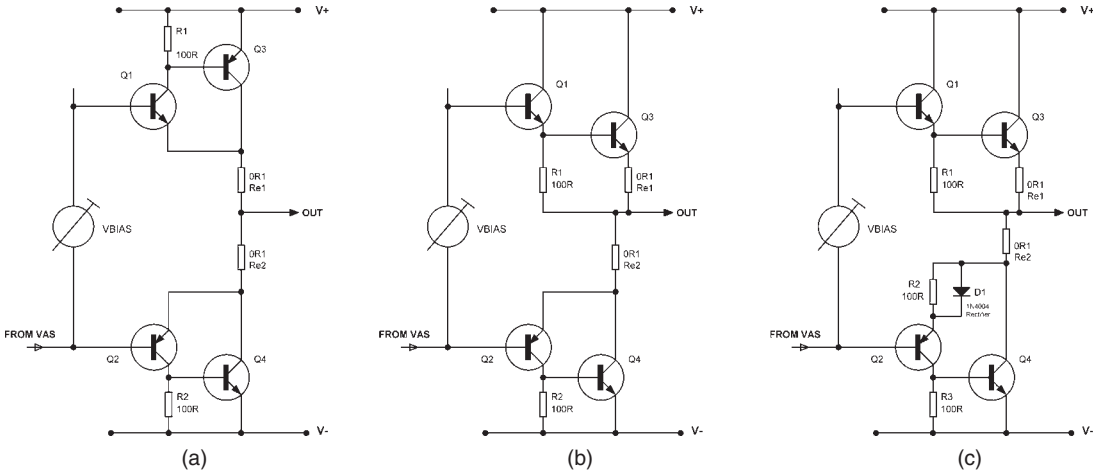


Figure 6.5: CFP circuit and quasi-complementary stages

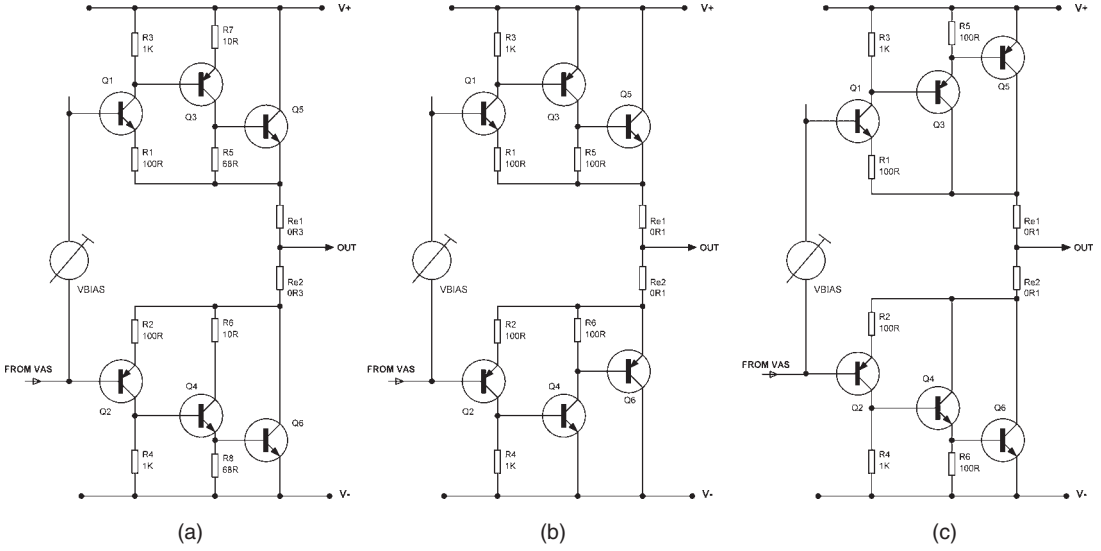


Figure 6.6: Three of the possible output-triple configurations

As the power output required from an amplifier increases, a point is reached when a single pair of output devices is no longer adequate for reliable operation. Multiple output devices also reduce large-signal nonlinearity (Distortion 3a) as described below. Adding parallel output devices to an EF stage is straightforward, as shown in Figure 6.9, which is configured as a Type II EF stage. The only precaution required is to ensure there is proper sharing of current between the output devices. If they were simply connected in parallel at all three terminals, the V_{be} tolerances could lead to unequal current-sharing and consequent over-dissipation of one or more devices. This is a potentially unstable situation as the V_{be} of the hottest device will fall and it will take an even bigger share of the current until something bad happens. It is therefore essential to give each transistor its own emitter resistor for local DC feedback; I have found 0R1 to be large enough in all circumstances; as described below in the section on crossover distortion, the value needs to be kept

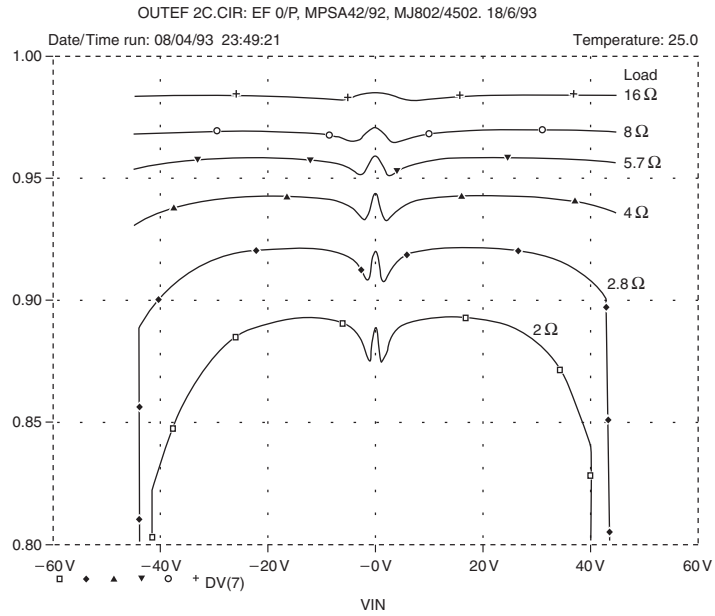


Figure 6.7: Emitter-follower large-signal gain versus output

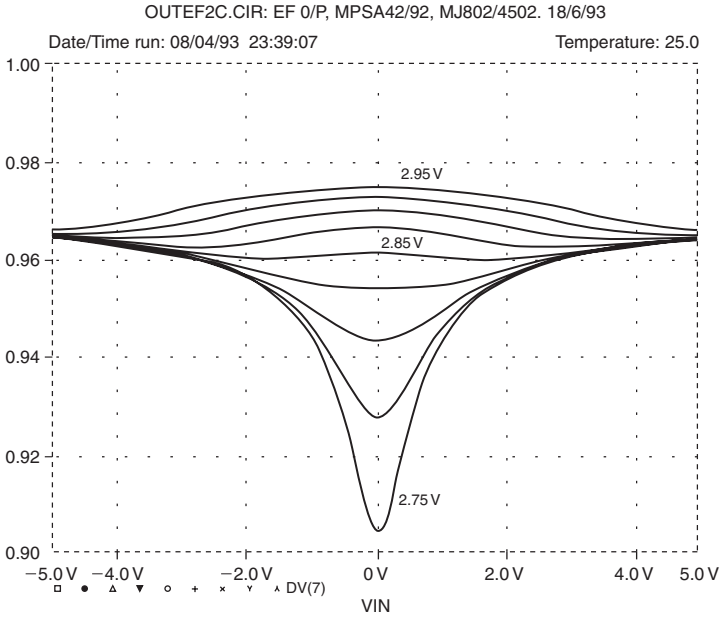


Figure 6.8: EF crossover region gain deviations, ±5 V range

as low as possible to minimize crossover nonlinearities. However, if you have an eccentric heat-sink design that does not keep all the output devices at the same temperature, it might be necessary to increase the value to give good current-sharing.

A triple-based EF output stage with three output pairs is shown in Figure 6.18 below.

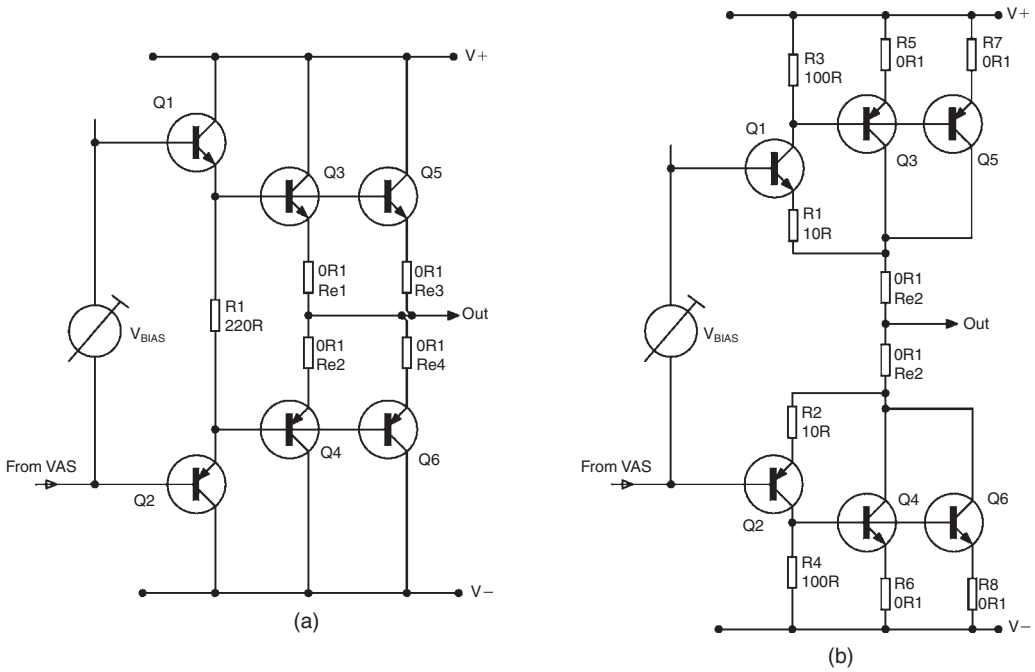


Figure 6.9: Using multiple output devices in the EF and CFP configurations

The Complementary Feedback Pair (CFP) Output

The other major type of bipolar complementary output is that using two CFPs. These are sometimes called Sziklai pairs or conjugate pairs. The output stage can be seen in Figure 6.5a. There seems to be only one popular configuration, though versions with gain are possible and have been used occasionally. The driver transistors are now placed so that they compare the output voltage with that at the input. Thus wrapping the outputs in a local NFB loop promises better linearity than emitter-follower versions with 100% feedback applied separately to driver and output transistors.

The CFP topology is generally considered to show better thermal stability than the EF, because the V_{be} of the output devices is inside the local NFB loop, and only the driver V_{be} has a major effect on the quiescent conditions. The true situation is rather more complex, and is explored in Chapter 15.

In the CFP output, like the EF, the drivers are conducting whenever the outputs are, so special arrangements to keep them in Class-A seem pointless. The CFP stage, like EF Type I, can only reverse-bias the driver bases, and not the output bases, unless extra voltage rails outside the main ones are provided.

The output gain plot is shown in Figure 6.10; Fourier analysis of this shows that the CFP generates less than half the LSN of an emitter-follower stage (see Table 6.2). Given also the greater quiescent stability, it is hard to see why this topology is not more popular. One possible reason is that it can be more prone to parasitic oscillation.

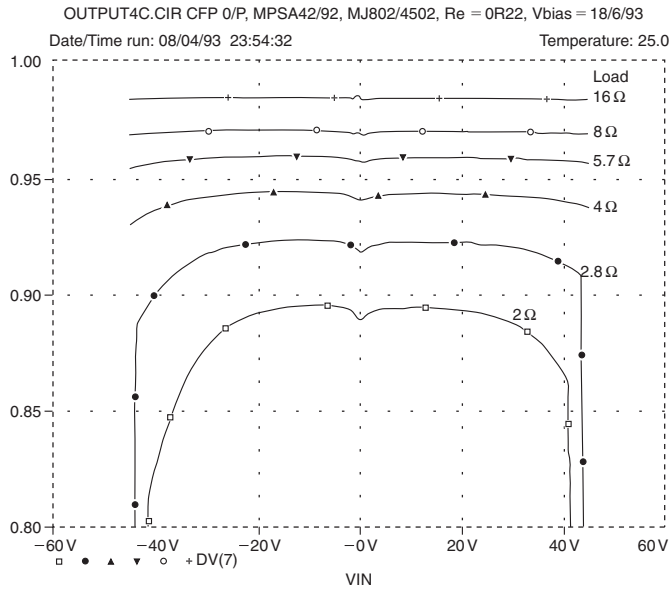


Figure 6.10: CFP gain versus output

Table 6.2: Summary of output distortion

	Emitter-follower	CFP	Quasi simple	Quasi Bax	Triple Type 1	Simple MOSFET	Quasi MOSFET	Hybrid MOSFET
8 Ω THD	0.031%	0.014%	0.069%	0.44%	0.13%	0.47%	0.44%	0.052%
Gain	0.97	0.97	0.97	0.96	0.97	0.83	0.84	0.97
4 Ω THD	0.042%	0.030%	0.079%	0.84%	0.60%	0.84%	0.072%	0.072%
Gain	0.94	0.94	0.94	0.94	0.92	0.72	0.73	0.94

Table 6.2 summarizes the SPICE curves for 4 and 8 Ω loadings; FET results from Chapter 14 are included for comparison; note the low gain for these. Each gain plot was subjected to Fourier analysis to calculate THD percentage results for a ±40 V input.

The crossover region is much narrower, at about ±0.3 V (Figure 6.11). When underbiased, this shows up on the distortion residual as narrower spikes than an emitter-follower output gives. The bad effects of g_m-doubling as V_{bias} increases above optimal (here 1.296 V) can be seen in the slopes moving outwards from the center.

Adding parallel output devices for increased output to a CFP stage is straightforward, as shown in Figure 6.9b, but extra current-sharing resistors R5, R7 and R6, R8 must be inserted in the output device emitter circuits. As for the EF output configuration, 0R1 is large enough in almost all circumstances. Note that the emitter resistors Re1, Re2 are still required and will need to be uprated to cope with the increased output permitted by the increased number of output devices.

Small emitter degeneration resistors R1, R2 are shown; these will not inevitably be required to ensure stability with this configuration, but in a CFP stage parasitic oscillation is more likely with multiple output devices, so it's best to make provision for them.

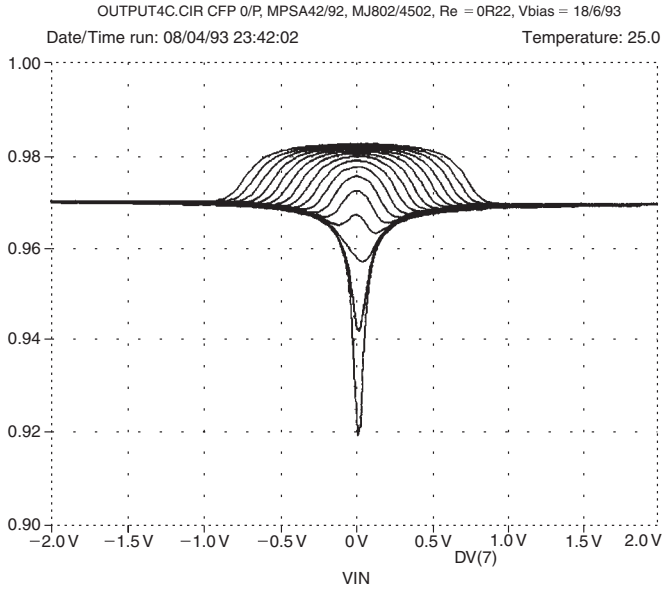


Figure 6.11: CFP crossover region ±2 V, V_{bias} as a parameter

As for the EF output stage, multiple output devices not only increase output capability but also reduce large-signal nonlinearity (Distortion 3a), as described in its own section below. However, multiple output devices in the CFP configuration do not necessarily decrease crossover distortion and can in some circumstances increase it.

Output Stages with Gain

It was explained at the start of this chapter that almost all output stages have a gain of unity, or to be precise, slightly less than unity. This is because, firstly, there are voltage losses in the emitter resistors Re, which form the upper arm of a potential divider with the external load as the lower arm. Thus if you assume an amplifier stage with resistors Re of 0R1, and an instantaneous operating point well away from the crossover region, you get a gain of 0.988 times with an 8 Ω load, reducing to 0.976 for 4 Ω loads. Secondly, in the case of EF-type output stages, the gain of an emitter-follower is always slightly less than 1.

Output stages with significant gain (typically two times) have been advocated on the grounds that the lower voltage swing required to drive the stage would reduce VAS distortion. This is much misguided, for as we have seen in earlier chapters, the distortion produced by the small-signal stages can be made very low by simple methods, and there is no pressing need to seek radical ways of reducing it further. On the other hand, distortion in the output stage is a much more difficult problem, so making things worse by seeking voltage gain is not the way forward.

A slightly better justification for seeking voltage gain in the output stage is that it would allow more output voltage swing from the same supply rails, improving efficiency. The VAS stage will have some saturation voltages, so it cannot swing fully between the rails (this point is looked at in detail

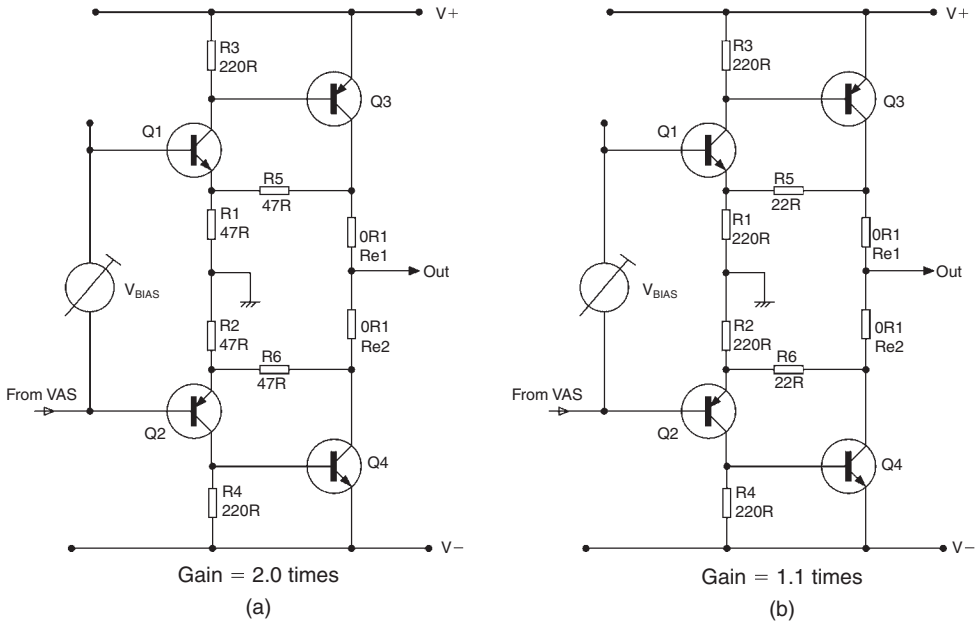


Figure 6.12: Examples of CFP output stages configured to give a voltage gain of 2 times (a) and a voltage gain of 1.1 times (b), by the addition of potential dividers R5, R1 and R6, R2 in the local feedback to the driver emitters

in the chapter on Class-A amplifiers, where you need to squeeze out every watt you can) and so a little gain afterwards, say 1.1 times, will allow the maximum swing the output stage can provide. However, even allowing for the fact that output power in watts, which is the figure everyone looks at, goes up with the square of voltage, the advantage to be gained is small compared with the extra difficulties you are likely to get into in the output stage. Figure 6.12 shows a CFP output stage with a gain of 2. For obvious reasons you cannot make an EF stage with gain – it is composed of emitter-followers that all have sub-unity gain.

The circuit in Figure 6.12 gives gain because two potential dividers R5, R1 and R6, R2 have been inserted in the local feedback path to the driver emitters; as you might expect, equal resistor values top and bottom give a gain of 2. The value of these resistors is problematic. If they are too large, the source impedance seen by the driver emitter is unduly increased and this local degeneration reduces the loop gain in the CFP output structure, and distortion will increase. If the divider resistors are kept low to avoid this, they are going to dissipate a lot of power, as they are effectively connected between the amplifier output and ground. The value of 47 Ω shown here in Figure 6.12a is a reasonable compromise, giving the driver stage an open-loop voltage gain of 10 times, while keeping the divider values up. However, a 100 W/8 Ω amplifier at full throttle is still going to dissipate 4.2 W in each of the 47 Ω divider resistors, requiring some hefty resistors that take up a lot of PCB space, and drawing in total an extra 16.8 W from the amplifier output.

If you are seeking just a small amount of gain such as 1.1 times, to maximize the output swing, things are slightly easier. The example in Figure 6.12b has a source impedance of 47/2 = 23.5 Ω

seen by the driver emitter; if we stick roughly to this figure the divider values for a gain of 1.1 times become R5 = 22 Ω and R1 = 220 Ω, the divider R6, R2 in the lower half of the output stage having corresponding values; this gives an impedance at the driver emitter of 20 Ω. For the same 100 W/8 Ω amplifier, this reduces the dissipation in R5 to 298 mW, and in R1 to 2.98 W; the total extra power drawn from the amplifier output is reduced to 6.56 W, which is a bit more manageable.

Output stages with gain can be made to work, but ultimately my advice would be that you probably do not want to go this way.

Quasi-complementary Outputs

Originally, the quasi-complementary configuration[7] was virtually mandatory, as it was a long time before PNP silicon power transistors were available in anything approaching complements of the NPN versions. The standard quasi-complementary circuit shown in Figure 6.5b is well known for poor symmetry around the crossover region, as shown in Figure 6.13. Figure 6.14 zooms in to show that the crossover region is a kind of unhappy hybrid of the EF and CFP, as might be expected, and that there is no setting of V_{bias} that can remove the sharp edge in the gain plot.

A major improvement to symmetry can be made by using a Baxandall diode[8], as shown in Figure 6.5c. Placing a diode in the driver circuit of the CFP (lower) part of the output stage gives it a more gradual turn-on, approximating to the EF section in the upper half of the output stage. This stratagem yields gain plots very similar to those for the true complementary EF in Figures 6.7 and 6.8, though in practice the crossover distortion seems rather higher. When this quasi-Baxandall stage is used closed-loop in an amplifier in which Distortions 1, 2, and 4–7 have been properly eliminated, it is capable of much better performance than is commonly believed; for example, 0.0015% (1 kHz) and 0.015% (10 kHz) at 100 W is straightforward to obtain from an amplifier with a moderate NFB factor of about 34 dB at 20 kHz.

Peter Baxandall introduced the concept of a diode in the CFP part of the output stage in response to an earlier proposal by Shaw[9], which put a power diode in series with the output of the CFP stage, as shown in Figure 6.15, with the same intention of making it turn on more slowly. A serious disadvantage of the Shaw scheme is that the added diode passes the full output stage current and therefore needs to be a hefty component. The Baxandall diode only passes the driver current and can therefore be a small part.

I received this communication[10] from Peter Baxandall, written not long before his untimely death:

> It is slightly preferable to use a transdiode (a transistor with collector connected to base) rather than an ordinary diode such as 1N4148, since the transdiode follows the transistor equation much more accurately, matching better the V_{be} characteristic of the top driver transistor. As you probably know, most diodes follow, over a moderate current range, the transistor equation but with mkT in place of kT, where m is a constant in the region of 1.8, though varying somewhat with the type of diode. Consequently whereas the voltage across a transdiode at fairly small currents varies, at 20°C, by a remarkably accurate 58 mV per decade of current change, that across an ordinary diode is nearer 100 mV per decade or just over.

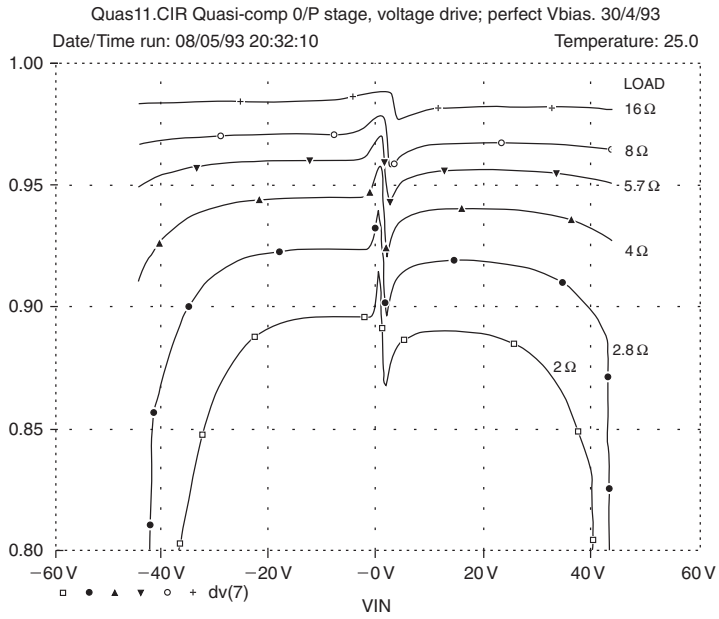


Figure 6.13: Quasi-complementary large-signal gain versus output

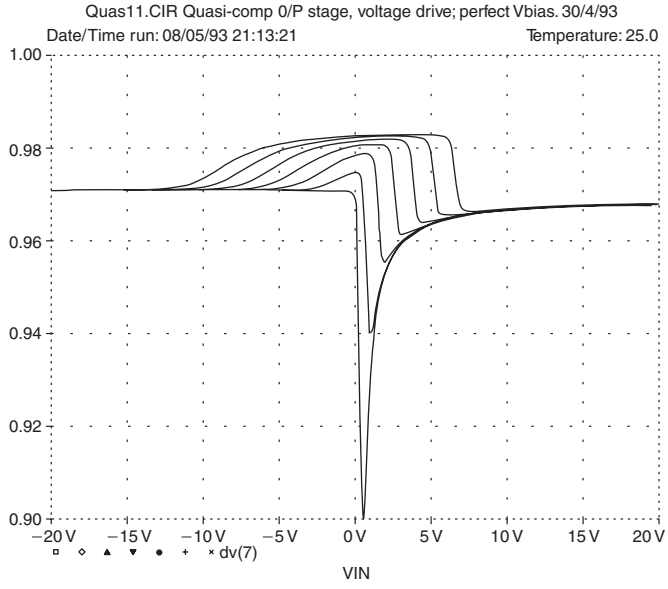


Figure 6.14: Quasi-crossover region ±20 V, V_{bias} as parameter

The transistor equation is the well-known fundamental relationship that describes how transistors work. It is:

$$I_c = I_o \times e^{q \times V_{be}/kT} - 1 \qquad \text{Equation 6.1}$$

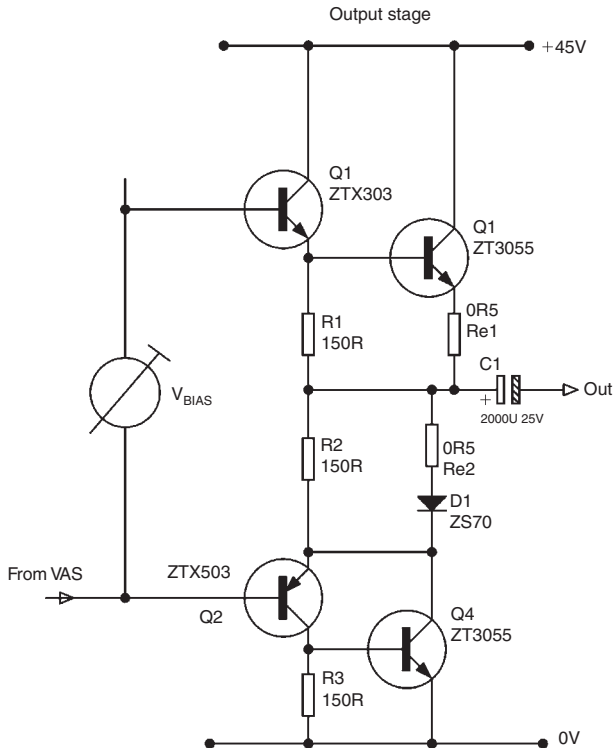


Figure 6.15: Quasi-complementary output stage with Shaw diode added to improve symmetry. The component values and transistor types are as used in Shaw's original circuit

where I_c is the collector current and I_o is the saturation current, q the charge on an electron, k is Boltzmann's constant, and T is absolute temperature in degrees kelvin. kT/q is often called V_t, the 'thermal voltage'. It is 25.3 mV at 20°C.

A Baxandall output stage with transdiode is shown in Figure 6.16. Peter did not discuss with me the type of transistor to be used as a transdiode, but I would guess that it should be the same as the lower driver Q2. He recommended the use of a transdiode rather than an ordinary diode in this reference[11].

Fully complementary output devices have been available for many years now, and you may be wondering why it is worth examining configurations that are obsolete. The answer is that in the world of hi-fi, no circuit concept seems to ever quite die, and at the time of writing (mid-2008) at least one well-known company is still producing quasi-complementary power amplifiers. The best reason to use the quasi-Baxandall approach today is to save a little money on output devices, as PNP power BJTs remain somewhat pricier than NPNs.

Given the tiny cost of a Baxandall diode or transdiode, and the absolutely dependable improvement it gives, there seems no reason why anyone should ever use the standard quasi circuit. My experiments show that the value of R1 in Figure 6.5c is not critical; making it about the same as R_c seems to work well.

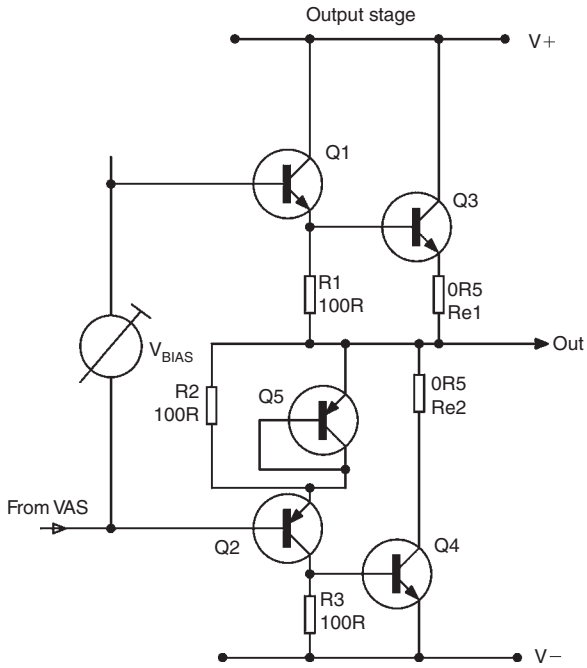


Figure 6.16: Quasi-complementary output stage with added Baxandall transdiode Q5 to give better symmetry than a simple diode. R1, R2, and R3 have the values used by Peter with the diode version

Triple-Based Output Configurations

If we allow the use of three rather than two bipolar transistors in each half of an output stage, the number of circuit permutations leaps upwards, and I cannot provide even a rapid overview of every possible configuration in the space available. Here are some of the possible advantages if output triples are used correctly:

1. Better linearity at high output voltages and currents, due to increased local feedback in the triple loop.
2. More stable quiescent setting as the pre-drivers can be arranged to handle very little power indeed, and to remain almost cold in use. This means they can be low-power TO92-type devices with superior beta, which enhances the local loop gain.
3. The extra current gain allows greater output power without undesirable increases in the operating currents of the VAS.

However, triples do not abolish crossover distortion, and they are, as usually configured, incapable of reverse-biasing the output bases to improve switch-off. Figure 6.6 shows three of the more useful ways to make a triple output stage – all of those shown have been used in commercial designs so they must be considered as practical in use. This is an important proviso as it is not hard to make a triple output stage that cannot be made to give reliable freedom from oscillation in the triple loop.

The most straightforward triple-based output stage is the triple-EF configuration, which adds to the output stage of Figure 6.4b a pair of pre-driver emitter-followers. This is dealt with in its own section below.

Figure 6.6a is the Quad-303 quasi-complementary triple. The Quad 303 amplifier was introduced in 1967, when complementary silicon output transistors were not yet a practical proposition. This configuration uses the extra local negative feedback of the triple stage to give much better linearity than a conventional quasi-complementary output stage. Note that the Re resistors here are shown as 0R3, which was the value used in the original Quad 303 circuit.

The output stage in Figure 6.6b is in contrast a fully complementary output stage. Its top half consists of Q1 and Q3 configured as common-emitter voltage amplifiers, while output device Q5 is a common-collector emitter-follower. The local negative-feedback loop is closed by connecting the emitter of Q1 to the top of Re1, making the top triple effectively a 'super emitter-follower' with high loop gain and a high degree of negative feedback, which gives it a very high input impedance, a low output impedance, and a gain of very nearly unity. The resistors R1, R2 limit the internal loop gain of the triple, by applying what might be called 'very local feedback' or emitter degeneration to the emitter of Q1; in my experience this is absolutely essential if anything like reliable stability is to be obtained with this configuration. In some versions the driver stages Q3, Q4 also have small resistors in their emitter circuits (typically 10 Ω) to give more control of loop gain. The bottom half of the output stage works in exactly the same way as the top.

The output stage in Figure 6.6c is another variation on the triple output. In this case only the pre-driver transistor Q1 is configured as a common-emitter voltage amplifier, with the driver and output transistors being connected as cascaded emitter-followers. This gives less voltage gain inside the triple loop, less local feedback, and hence less chance of local oscillation. Note that Q1 and Q2 still have emitter resistors R1, R2 to control the voltage gain of the pre-driver stages. The design and testing of triple-based output stages demands care, as the possibility of local HF instability in each output half is very real.

Given the number of possibilities for triple-based output stages, it might be useful to have a concise notation to describe them. The output stage in Figure 6.6b is composed of two common-emitter voltage amplifiers followed by a common-collector emitter-follower, making up a single local negative-feedback loop. It could be written as a CE–CE–EF triple output stage. Likewise, the output stage in Figure 6.6c has common-emitter pre-driver, with both the driver and output transistors connected as emitter-followers, so it could be described as a CE–EF–EF configuration.

Quasi-complementary stages like that of Figure 6.6a are a bit less straightforward, but if we adopt the convention that the top half the output stage is always described first, it could be written as CE–CE–EF/CE–EF–EF.

Some more triple output stages are shown in Figure 6.17a. These have the feature that the local negative feedback is only closed around two of the three devices in the triple. The first one is an emitter-follower feeding a CFP stage (which is in turn composed of two CE voltage-amplifier stages), which could be written EF–CE–CE, but EF–CFP is rather more indicative of its structure

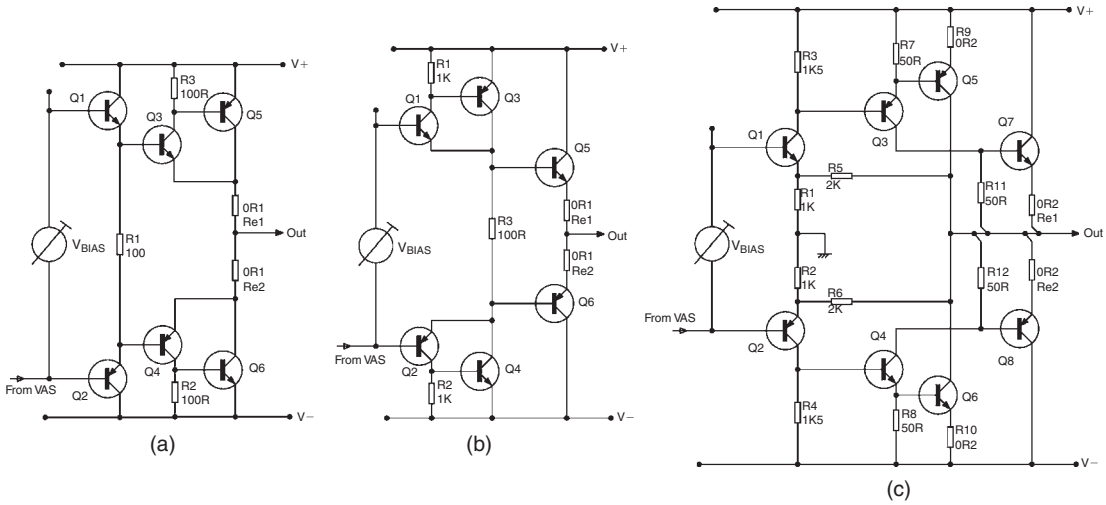


Figure 6.17: (a, b) Two more possible types of output triple. (c) The Bryston output stage

and operation. It can be regarded as a simple emitter-follower feeding a compound output device. This configuration has the potential disadvantage that the pre-driver emitter-follower is outside the local NFB loop; the same naturally applies to the EF–EF–EF triple-emitter-follower output stage described in the next section.

Figure 6.17b shows another variation on the triple theme. This time we have a CFP stage feeding an emitter-follower. Once again it could be written CE–CE–EF, but CFP–EF is more instructive. The importance of this configuration is that it looks promising for reducing the effects of large-signal nonlinearity when driving low impedances.

An unconventional triple output stage used by Bryston is shown in Figure 6.17c. As with other triple outputs, the pre-driver stage Q1 is run at low power so it stays cool and gives good bias stability. The driver stage Q3 is now a phase splitter; the output from its emitter drives a CE stage Q5 that feeds current directly into the output rail, while the output from its collector drives an EF stage Q7 that feeds current into the output rail via two emitter resistors Re1, Re2. An important feature of this stage is that it has a voltage gain of 3, set up by the potential dividers R5, R1 and R6, R2.

The Bryston configuration has the further interesting property that it completely defeats the output stage notation suggested only a few paragraphs ago. According to Bryston's own publicity material, their output stage configuration requires close matching of output transistor betas, not only between similar types, but between complementary devices. They say that since Bryston products are hand-built from selected components anyway, this is not a serious disadvantage in production.

Triple-EF Output Stages

Sometimes it is necessary to use a triple output stage simply because the currents flowing in the output stage are too big to be handled by two transistors in cascade. If you are driving 2 Ω or 1 Ω loads, then typically there will be multiple output devices in parallel. Providing the base current

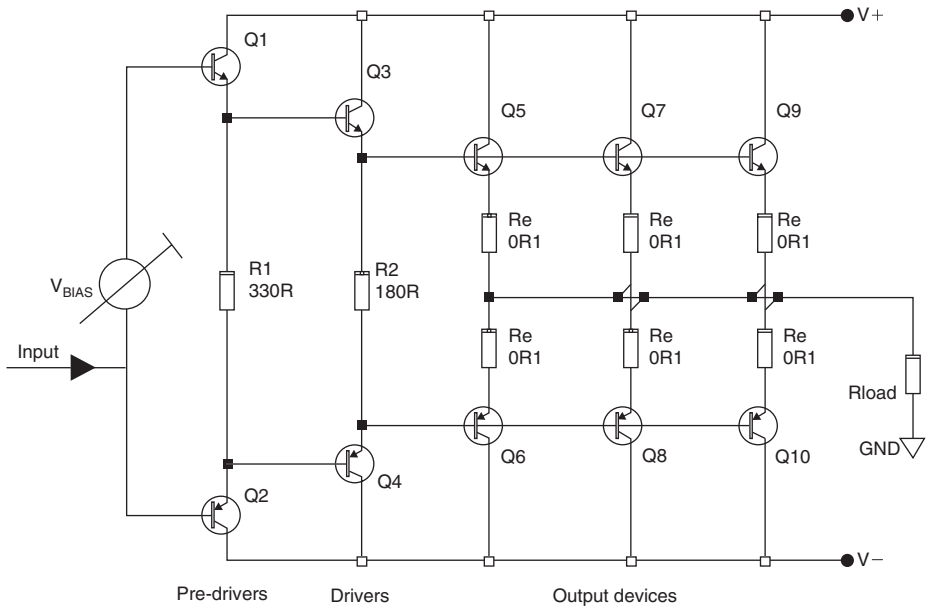


Figure 6.18: Triple-EF output stage. Both pre-drivers and drivers have emitter resistors

for five or more output transistors, with their relatively low beta, will usually be beyond the power capability of normal driver types, and it is common to use another output device as the driver. This will hopefully have the power-handling capability, but with this comes low beta once again. This means that the driver base currents in turn become too large for a normal VAS stage to source. There are two solutions – make the VAS capable of sourcing hundreds of mA, or insert another stage of current gain between VAS and drivers. The latter is much easier, and the usual choice. These extra transistors are usually called the pre-drivers (see Figure 6.18).

In this circuit the pre-drivers dissipate relatively little power, and providing they are medium-power devices such as those in a TO220 package it is unlikely that they will need heat-sinking to cope with the demands made on them. There is, however, another reason to fit pre-drive heat-sinks – or at least make room at the layout stage so you have the option.

In Figure 6.18 there is about 1.2 V across R2, so Q3, Q4 have to supply a standing current of about 7 mA. This has no effect on the drivers as they are likely to be well cooled to deal with normal load demands. However, the voltage across R1 is two V_{be} values higher, at 2.4 V, so the standing current through it is actually higher at 7.3 mA. (The exact figures naturally depend on the values for R1, R2 that are chosen, but it is difficult to make them much higher than shown here without compromising the speed of high-frequency turn-off.) The pre-drivers are usually small devices, and so they are likely to get warm, and this leads to drift in the bias conditions after switch-on. Adding heat-sinks cannot eliminate this effect, but does usefully reduce it.

In a triple-EF output stage like this the V_{bias} generator must produce enough voltage to turn on six base–emitter junctions, plus the small standing voltage V_q across the emitter resistors, totaling about

3.9V in practice. The V_{be} of the bias transistor is therefore being multiplied by a larger factor, and V_{bias} will drop more for the same temperature rise. This should be taken into account, as it is easy with this kind of output stage to come up with a bias generator that is overcompensated for temperature.

Quadruple Output Stages

If three transistors in a triple output stage can be useful, it is only human (if you're a designer or engineer, anyway) to ponder if four transistors would be even better. As I have stressed above, triple stages where all three transistors are configured in a single local negative loop can be difficult to stabilize. Four must be worse, and while I have not tried the experiment, it seems highly unlikely that a quadruple output stage with a single loop could be made reliably stable.

What looks a good deal more promising is a combination of the EF–CFP and CFP–EF output structures described above, which would give CFP–CFP – in other words two cascaded local feedback loops, with each loop only encompassing two transistors. A possible arrangement of this is shown in Figure 6.19.

This configuration could be regarded as an enhancement of an EF output stage, in that instead of two cascaded emitter-followers, there are two cascaded 'super-emitter-followers' in the form of CFP stages, which will be a good deal more linear than simple emitter-followers because each has their own local feedback loop.

I was going to call this a 'quad output stage', but you can see the opportunity for confusion there. It seems best to stick with 'quadruple output stage'.

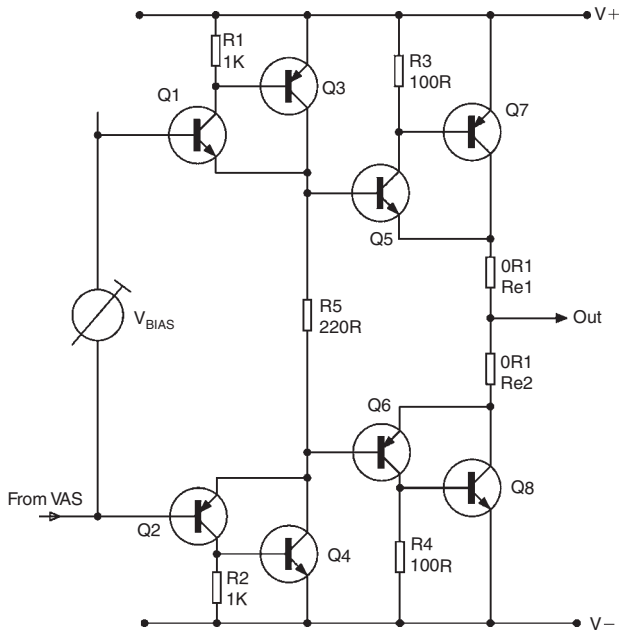


Figure 6.19: Quadruple CFP–CFP output stage

Output Stage Distortions and their Mechanisms

Subdividing Distortion 3 into large-signal nonlinearity (LSN), crossover, and switch-off distortion provides a basis for judging which output stage is best. The LSN is determined by both circuit topology and device characteristics, crossover distortion is critically related to quiescent conditions' stability, and switch-off distortion depends strongly on the output stage's ability to remove carriers from power BJT bases. I now look at how these shortcomings can be improved, and the effect they have when an output stage is used closed-loop.

In Chapters 4 and 5 it was demonstrated that the distortion from the small-signal stages can be kept to very low levels that will prove to be negligible compared with closed-loop output stage distortion, by the adroit use of relatively conventional circuitry. Likewise, Chapters 6 and 7 will reveal that Distortions 4–11 can be effectively eliminated by lesser-known but straightforward methods. This leaves Distortion 3, in its three components, as the only distortion that is in any sense unavoidable, as Class-B stages completely free from crossover artefacts are so far beyond us.

This is therefore a good place to review the concept of a 'Blameless' amplifier, introduced in Chapter 3, one designed so that all the easily defeated distortion mechanisms have been rendered negligible. (Note that the word 'Blameless' has been carefully chosen not to imply perfection.) Distortion 1 cannot be totally eradicated, but its onset can be pushed well above 20 kHz. Distortion 2 can be effectively eliminated by cascoding, and Distortions 4–7 can be made negligible by simple measures to be described later. This leaves Distortion 3, which includes the knottiest Class-B problems, i.e. crossover distortion (Distortion 3b) and HF switch-off difficulties (Distortion 3c).

The design rules presented here will allow the routine design of Blameless amplifiers. However, this still leaves the most difficult problem of Class-B unsolved, so it is too early to conclude that as far as amplifier linearity is concerned, history is over . . .

Large-Signal Distortion (Distortion 3a)

Amplifiers always distort more with heavier loading. This is true without exception so far as I am aware. Why? Is there anything we can do about it?

A Blameless Class-B amplifier typically gives an 8 Ω distortion performance that depends very little on variable transistor characteristics such as beta. At this load impedance output stage nonlinearity is almost entirely crossover distortion, which is a voltage-domain effect.

As the load impedance of the amplifier is decreased from infinity to 4 Ω, distortion increases in an intriguing manner. The unloaded THD is not much greater than that from the AP System-1 test oscillator, but as loading increases crossover distortion rises steadily (see Figure 6.25). When the load impedance falls below about 8 Ω, a new distortion begins to appear, overlaying the existing crossover nonlinearities. It is essentially third harmonic. In Figure 6.20 the upper trace shows the 4 Ω THD is consistently twice that for 8 Ω, once it appears above the noise floor.

I label this Distortion 3a, or large-signal nonlinearity (LSN), where 'large' refers to currents rather than voltages. Unlike crossover Distortion 3b, the amount of LSN generated is highly dependent

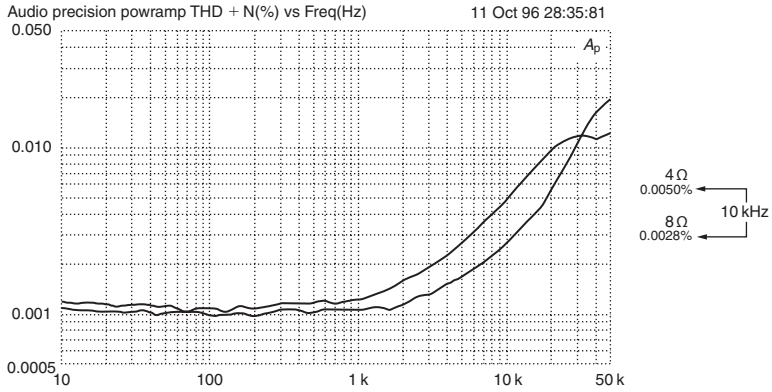


Figure 6.20: Upper trace shows distortion increase due to LSN as load goes from 8 to 4Ω. Blameless amplifier at 25W/8Ω

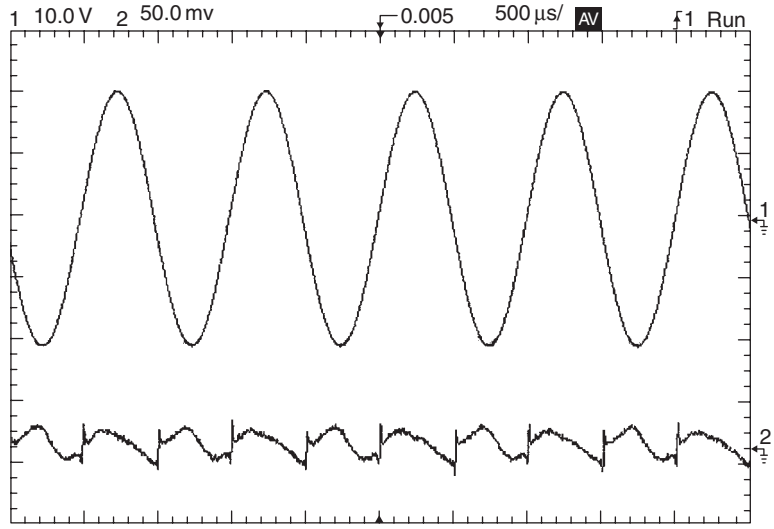


Figure 6.21: Large-signal nonlinearity, driving 50W into 4Ω and averaged 64 times

on device characteristics. The distortion residual is basically third order because of the symmetric and compressive nature of the output stage gain characteristic, with some second harmonic because the beta loss is component-dependent and not perfectly symmetrical in the upper and lower output stage halves. Figure 6.21 shows a typical THD residual for large-signal nonlinearity, driving 50W into 4Ω. The residual is averaged 64 times to reduce noise.

LSN occurs in both emitter-follower (EF) and complementary feedback pair (CFP) output configurations; this section concentrates on the CFP version, as shown in Figure 6.5a. Figure 6.22 shows the incremental gain of a simulated CFP output stage for 8 and 4Ω; the lower 4Ω trace has greater downward curvature, i.e. a greater fall-off of gain with increasing current. Note that this fall-off is steeper in the negative half, so the THD generated will contain even as well as odd harmonics. The simulated EF behavior is very similar.

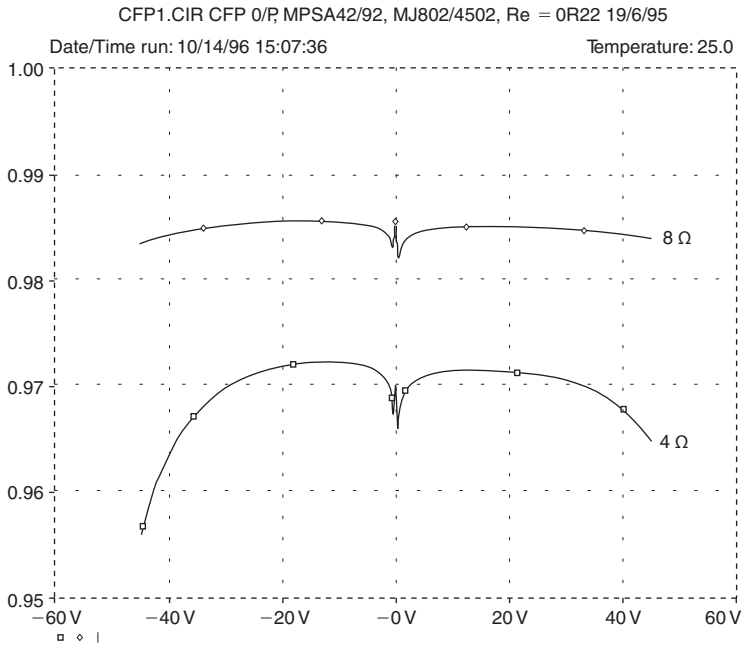


Figure 6.22: The incremental gain of a standard CFP output stage. The 4 Ω trace droops much more as the gain falls off at higher currents. PSPICE simulation

As it happens, an 8 Ω nominal impedance is a reasonably good match for standard power BJTs, though 16 Ω might be better for minimizing LSN if loudspeaker technology permits. It is coincidental that an 8 Ω nominal impedance corresponds approximately with the heaviest load that can be driven without LSN appearing, as this value is a legacy from valve technology. LSN is an extra distortion component laid on top of others, and usually dominating them in amplitude, so it is obviously simplest to minimize the 8 Ω distortion first; 4 Ω effects can then be seen more or less in isolation when load impedance is reduced.

The typical result of 4 Ω loading was shown in Figure 6.20, for the modern MJ15024/25 complementary pair from Motorola. Figure 6.23 shows the same diagram for one of the oldest silicon complementary pairs, the 2N3055/2955. The 8 Ω distortion is similar for the different devices, but the 4 Ω THD is 3.0 times worse for the venerable 2N3055/2955. Such is progress.

Such experiments with different output devices throw useful light on the Blameless concept – from the various types tried so far it can be said that Blameless performance, whatever the output device type, should not exceed 0.001% at 1 kHz and 0.006% at 10 kHz, when driving 8 Ω. The components existed to build sub-0.001% THD amplifiers in mid-1969, but not the knowledge.

Low-impedance loads have other implications beyond worse THD. The requirements for sustained long-term 4 Ω operation are severe, demanding more heat-sinking and greater power supply capacity. For economic reasons the peak/average ratio of music is usually fully exploited, though this can cause real problems on extended sine-wave tests, such as the FTC 40%-power-for-an-hour preconditioning procedure.

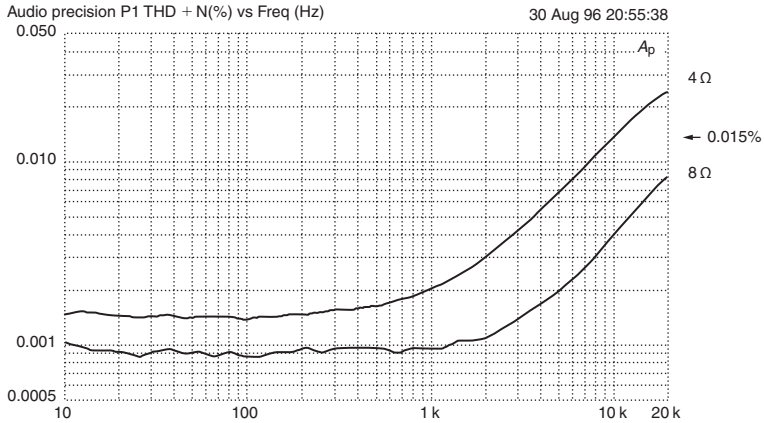


Figure 6.23: 4 Ω distortion is three times greater than 8 Ω for 2N3055/2955 output devices (compare Figure 6.14)

The focus of this section is the extra distortion generated in the output stage itself by increased loading, but there are other ways in which linearity may be degraded by the higher currents flowing. Of the amplifier distortion mechanisms (see Chapter 3), Distortions 1, 2, and 8 are unaffected by output stage current magnitudes. Distortion 4 might be expected to increase, as increased loading on the output stage is reflected in increased loading on the VAS. However, both the beta-enhanced EF and buffered-cascode methods of VAS linearization deal effectively with sub-8 Ω loads, and this does not seem to be a problem.

When a 4 Ω load is driven, the current taken from the power supply is greater, potentially increasing the rail ripple, which could worsen Distortion 5. However, if the supply reservoir capacitances have been sized to permit greater power delivery, their increased capacitance reduces ripple again, so this effect tends to cancel out. Even if rail ripple doubles, the usual RC filtering of bias supplies should keep it out of the amplifier, preventing intrusion via the input pair tail, and so on.

Distortion 6 could worsen as the half-wave currents flowing in the output circuitry are twice as large, with no counteracting mechanism. Distortion 7, if present, will be worse due to the increased load currents flowing in the output stage wiring resistances.

Of those mechanisms above, Distortion 4 is inherent in the circuit configuration (though easily reducible below the threshold of measurement) while Distortions 5–7 are topological, in that they depend on the spatial and geometrical relationships of components and wiring. The latter three distortions can therefore be completely eliminated in both theory and practice. This leaves only the LSN component, otherwise known as Distortion 3a, to deal with.

The Load-Invariant Concept

In an ideal amplifier the extra LSN distortion component would not exist. Such an amplifier would give no more distortion into 4 than 8 Ω and could be called 'load invariant to 4 Ω'. The minimum load qualification is required because it will be seen that the lower the impedance, the greater the difficulties in aspiring to load invariance. I assume that we start out with an amplifier that is

Blameless at 8 Ω; it would be logical but quite pointless to apply the term 'load invariant' to an ill-conceived amplifier delivering 1% THD into both 8 and 4 Ω.

The LSN Mechanism

When the load impedance is reduced, the voltage conditions are essentially unchanged. LSN is therefore clearly a current-domain effect, a function of the magnitude of the signal currents flowing in drivers and output devices.

A 4 Ω load doubles the output device currents, but this does not in itself generate significant extra distortion. The crucial factor appears to be that the current drawn from the drivers by the output device bases *more* than doubles, due to beta fall-off in the output devices as collector current increases.

It is this *extra* increase of current that causes almost all the additional distortion. The exact details of this have not been completely clarified, but it seems that this 'extra current' due to beta fall-off varies very nonlinearly with output voltage, and combines with driver nonlinearity to reinforce it rather than cancel. Beta-droop is ultimately due to high-level injection effects, which are in the province of semiconductor physics rather than amplifier design. Such effects vary greatly with device type, so when output transistors are selected, the likely performance with loads below 8 Ω must be considered.

There is good simulator evidence that LSN is entirely due to beta-droop causing extra current to be drawn from the drivers. To summarize:

- Simulated output stages with output devices modified to have no beta-droop (by increasing SPICE model parameter IKF) do not show LSN. It appears to be specifically that extra current taken due to beta-droop causes the extra nonlinearity.
- Simulated output devices driven with zero-impedance voltage sources instead of the usual transistor drivers exhibit no LSN. This shows that LSN does not occur in the outputs themselves, and so it must be happening in the driver transistors.
- Output stage distortion can be treated as an error voltage between input and output. The double emitter-follower (EF) stage error is therefore: driver V_{be} + output V_{be} + Re drop. A simulated EF output stage with the usual drivers shows that it is primarily nonlinearity increases in the driver V_{be} rather than in the output V_{be}, as load resistance is reduced. The voltage drop across the emitter resistors Re is essentially linear.

The knowledge that beta-droop caused by increased output device I_c is at the root of the problem leads to some solutions. Firstly, the per-device I_c can be reduced by using parallel output devices. Alternatively I_c can be left unchanged and output device types selected for those with the least beta-droop.

There is the possibility that increasing the current drawn from the drivers will in turn increase the current that they draw from the VAS, compromising its linearity. The investigations recorded here show that to be a very minor effect, if it exists at all. However, it is a possibility worth bearing in mind.

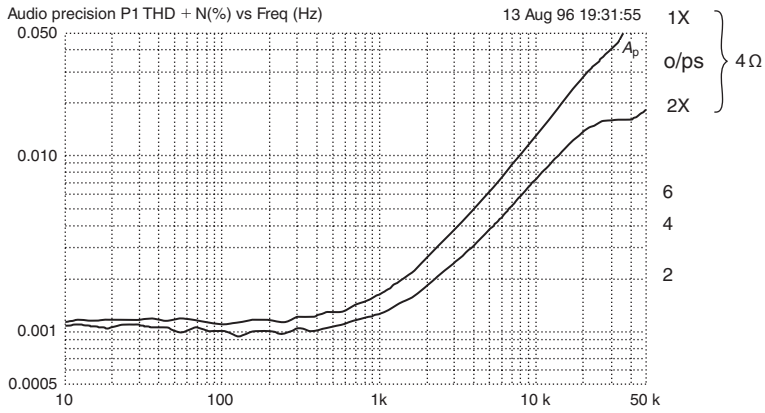


Figure 6.24: 4Ω distortion is reduced by 1.9 times upon doubling standard (MJ15024/15025) output transistors 30W/8Ω

Doubled Output Devices

LSN can be effectively reduced by doubling the output devices, when this is quite unnecessary for handling the rated power output. The fall-off of beta depends on collector current, and if two output devices are connected in parallel, the collector current divides in two between them. Beta-droop is much reduced.

From the above evidence, I predicted that this doubling ought to reduce LSN – and when measured, indeed it does. Such reality checks must never be omitted when using circuit simulators. Figure 6.24 compares the 4Ω THD at 60W for single and double output devices, showing that doubling reduces distortion by about 1.9 times, which is a worthwhile improvement.

The output transistors used for this test were modern devices, the Motorola MJ15024/15025. The much older 2N3055/2955 complementary pair give a similar halving of LSN when their number is doubled, though the initial distortion is three times higher into 4Ω. 2N3055 specimens with an H suffix show markedly worse linearity than those without.

No explicit current-sharing components were added when doubling the devices, and this lack seemed to have no effect on LSN reduction. There was no evidence of current hogging, and it appears that the circuit cabling resistances alone were sufficient to prevent this.

Doubling the number of power devices naturally increases the power output capability, though if this is exploited LSN will tend to rise again, and you are back where you started. Opting for increased power output will also make it necessary to uprate the power supply, heat-sinks, and so on. The essence of this technique is to use parallel devices to reduce distortion long before power handling alone compels you to do so.

Better Output Devices

The 2SC3281/2SA1302 complementary pair are plastic TO3P devices with a reputation in the hi-fi industry for being ‘more linear’ than the general run of transistors. Vague claims of this sort

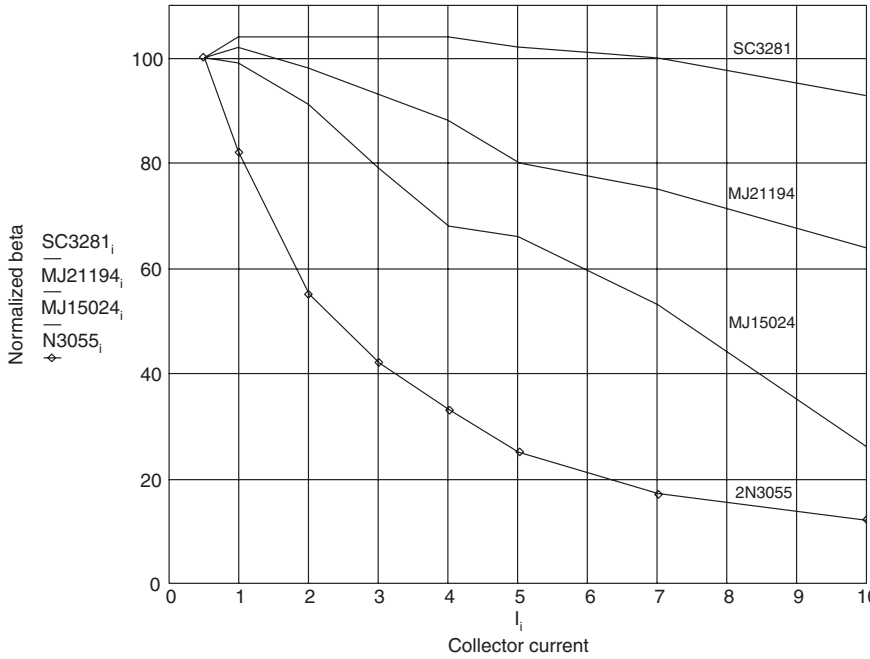


Figure 6.25: Power transistor beta falls as collector current increases. Beta is normalized to 100 at 0.5 A (from manufacturers' data sheets)

arouse the deepest of suspicions; compare the many assertions of superior linearity for power FETs, which is the exact opposite of reality. However, in this case the core of truth is that 2SC3281 and 2SA1302 show much less beta-droop than average power transistors. These devices were introduced by Toshiba; Motorola versions are MJL3281A, MJL1302A, also in the TO3P package. Figure 6.25 shows beta-droop for the various devices discussed here, and it is clear that more droop means more LSN.

The 3281/1302 pair are clearly in a different class from conventional transistors, as they maintain beta much more effectively when collector current increases. There seems to be no special name for this class of BJTs, so I have called them 'sustained-beta' devices here.

The THD into 4 and 8 Ω for single 3281/1302 devices is shown in Figure 6.26. Distortion is reduced by about 1.4 times compared with the standard devices of Figure 6.20, over the range 2–8 kHz. Several pairs of 3281/1302 were tested and the 4 Ω improvement is consistent and repeatable.

The obvious next step is to combine these two techniques by using doubled sustained-beta devices. The doubled-device results are shown in Figure 6.27, where the distortion at 80 W/4 Ω (15 kHz) is reduced from 0.009% in Figure 6.20 to 0.0045% – in other words, halved. The 8 and 4 Ω traces are now very close together, the 4 Ω THD being only 1.2 times higher than in the 8 Ω case.

There are other devices showing less beta-droop than standard. In a very quick survey I unearthed the MJ21193, MJ21194 pair (TO3 package) and the MJL21193, MJL21194 pair (TO3P package), both from Motorola. These devices show beta maintenance intermediate between the 'super'

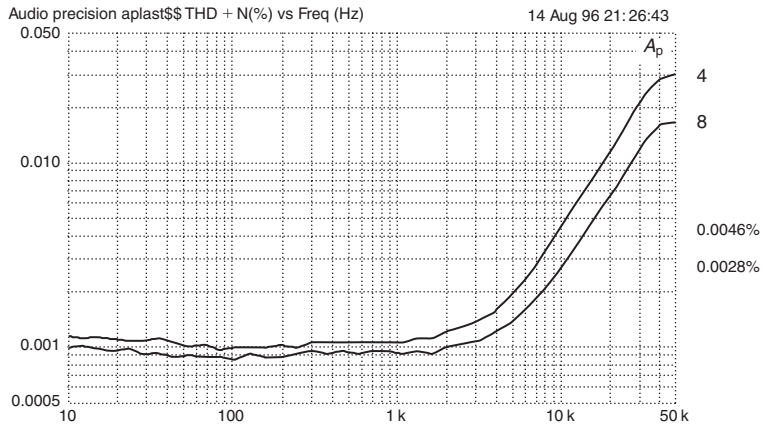


Figure 6.26: THD at 40 W/8 Ω and 80 W/4 Ω with single 3281/1302 devices

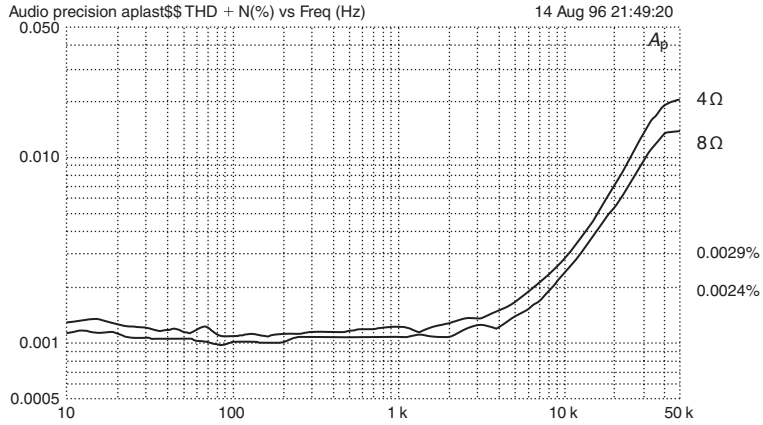


Figure 6.27: THD at 40 W/8 Ω and 80 W/4 Ω with doubled 3281/1302 output transistors. 4 Ω THD has been halved compared with Figure 6.12

3281/1302 and 'ordinary' MJ15024/25, so it seemed likely that they would give less LSN than ordinary power devices, but more than the 3281/1302. This prediction was tested and duly fulfilled.

It could be argued that multiplying output transistors is an expensive way to solve a linearity problem. To give this perspective, in a typical stereo power amplifier the total cost including heat-sink, metal work and mains transformer will only increase by about 5% when the output devices are doubled.

Feedforward Diodes

The first technique I tried to reduce LSN was the addition of power diodes across OR22 output emitter resistors. The improvement was only significant for high power into sub-3 Ω loading, and was of rather doubtful utility for hi-fi. Feedforward diodes treat the symptoms (by attempting distortion cancelation) rather than the root cause, so it is not surprising this method is of limited effectiveness (see Figure 6.28).

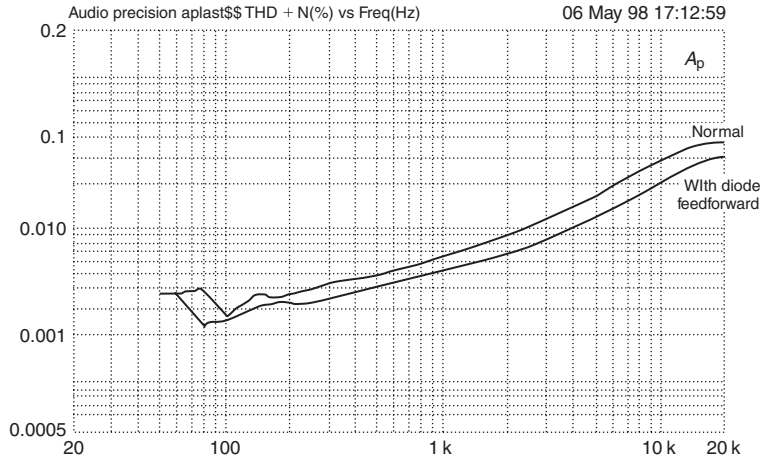


Figure 6.28: Simple diode feedforward reduces distortion with sub-8 Ω loads. Measured at 210 W into 2.7 Ω

It is my current practice to set the output emitter resistors Re at 0.1 Ω, rather than the more common OR22. This change both improves voltage-swing efficiency and reduces the extra distortion generated if the amplifier is erroneously biased into Class-AB. As a result even low-impedance loads give a relatively small voltage drop across Re, which is insufficient to turn on a silicon power diode at realistic output levels.

Schottky diodes have a lower forward voltage drop and might be useful here. Tests with 50 A diodes have been made but have so far not been encouraging in the distortion reduction achieved. Suitable Schottky diodes cost at least as much as an output transistor, and two will be needed.

Trouble with Triples

In electronics, as in many fields, there is often a choice between applying brawn (in this case multiple power devices) or brains to solve a given problem. The 'brains' option here would be a clever circuit configuration that reduced LSN without replication of expensive power silicon, and the obvious place to look is the output-triple approach. Note 'output triples' here refers to pre-driver, driver, and output device all in one local NFB loop, rather than three identical output devices in parallel, which I would call 'tripled outputs'. Getting the nomenclature right is a bit of a problem.

In simulation, output-triple configurations do reduce the gain-droop that causes LSN. There are many different ways to configure output triples, and they vary in their linearity and immunity to LSN. The true difficulty with this approach is that three transistors in a tight local loop are very prone to parasitic and local oscillations. This tendency is exacerbated by reducing the load impedances, presumably because the higher collector currents lead to increased device transconductance. This sort of instability can be very hard to deal with, and in some configurations appears almost insoluble. At present this approach has not been studied further.

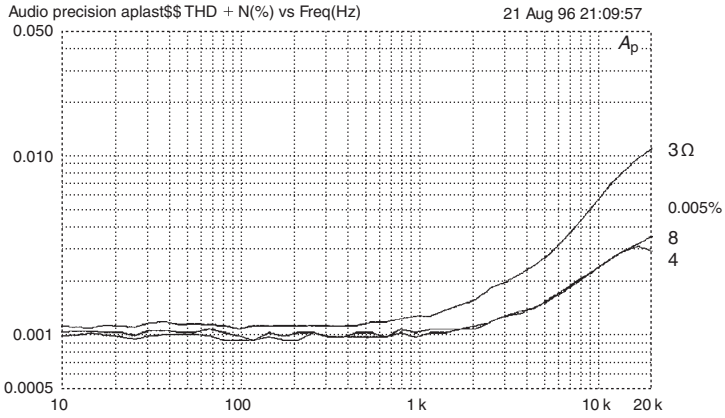


Figure 6.29: Distortion for 3, 4, and 8Ω loads, single 3281/1302 devices. 20W/8Ω, 40W/4Ω, and 60W/3Ω

Loads Below 4Ω

So far I have concentrated on 4Ω loads; loudspeaker impedances often sink lower than this, so further tests were done at 3Ω. One pair of 3281/1302 devices will give 50W into 3Ω for THD of 0.006% (10kHz), as shown in Figure 6.29. Two pairs of 3281/1302 reduce the distortion to 0.003% (10kHz), as in Figure 6.30. This is an excellent result for such simple circuitry, and may well be a record for 3Ω linearity.

It appears that whatever the device type, doubling the outputs halves the THD percentage for 4Ω loading. This principle can be extended to 2Ω operation, but tripled devices are required for sustained operation at significant powers. The resistive losses will be serious, so 2Ω power output may be little greater than that into 4Ω.

Better 8Ω Performance

It was not expected that the sustained-beta devices would also show lower crossover distortion at 8Ω, but they do, and the effect is once more repeatable. It may be that whatever improves the beta characteristic also somewhat alters the turn-on law so that crossover distortion is reduced; alternatively traces of LSN, not visible in the THD residual, may have been eliminated. The latter is probably the more likely explanation.

The plot in Figure 6.30 shows the improvement over the MJ15024/25 pair; compare the 8Ω line in Figure 6.20. The 8Ω THD at 10kHz is reduced from 0.003% to 0.002%, and with correct bias adjustment, the crossover artefacts are invisible on the 1kHz THD residual. Crossover artefacts are only just visible in the 4Ω case, and to get a feel for the distortion being produced, and to set the bias optimally, it is necessary to test at 5kHz into 4Ω.

A Practical Load-Invariant Design

Figure 6.31 is the circuit of a practical Load-Invariant amplifier designed for 8Ω nominal loads with 4Ω impedance dips, not for speakers that start out at 4Ω nominal and plummet from there.

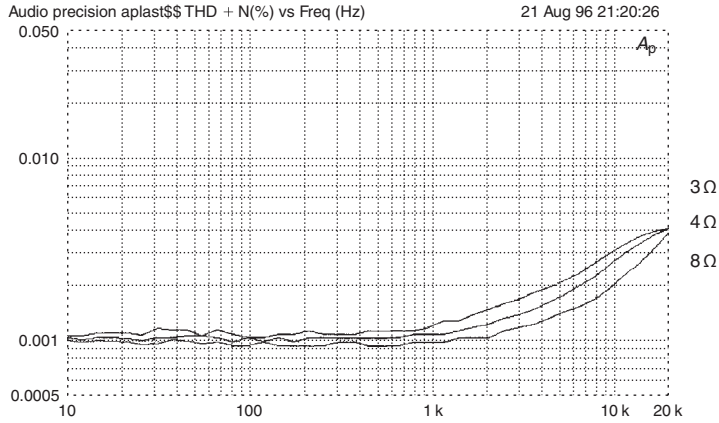


Figure 6.30: Distortion for 3, 4, and 8 Ω load, double 3281/1302 devices. Power as in Figure 6.22

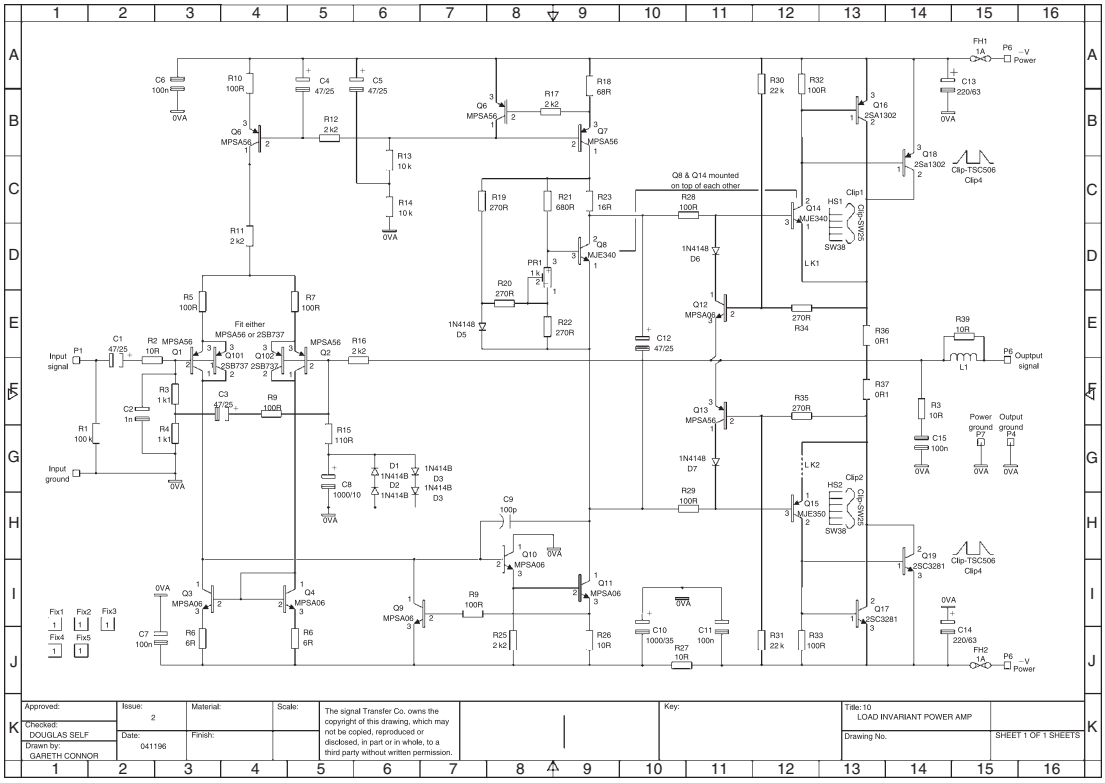


Figure 6.31: Circuit diagram of the Load-Invariant power amplifier

The distortion performance is shown in Figures 6.26–6.28 for various fitments of output device. The supply voltage can be from ±20 to ±40 V; checking power capability for a given output device fit must be left to the constructor.

Apart from load invariance, the design also incorporates two new techniques from the Thermal Dynamics section of this book.

The first technique greatly reduces time lag in the thermal compensation. With a CFP output stage, the bias generator aims to shadow driver junction temperature rather than the output junctions. A much faster response to power dissipation changes is obtained by mounting bias generator transistor TR8 on top of driver TR14, rather than on the other side of the heat-sink. The driver heat-sink mass is largely decoupled from the thermal compensation system, and the response is speeded up by at least two orders of magnitude.

The second innovation is a bias generator with an increased temperature coefficient, to reduce the static errors introduced by thermal losses between driver and sensor. The bias generator tempco is increased to −4.0 mV/°C. D5 also compensates for the effect of ambient temperature changes.

This design is not described in detail because it closely resembles the Blameless Class-B amp described elsewhere. The low-noise feedback network is taken from the Trimodal amplifier in Chapter 10; note the requirement for input bootstrapping if a 10k input impedance is required. Single-slope VI limiting is incorporated for overload protection, implemented by TR12, TR13. The global NFB factor is once more a modest 30 dB at 20 kHz.

More on Multiple Output Devices

I have done some further experiments with multiple devices, using three, four, five and six in parallel. The 2SC2922/2SA1612 complementary pair was used. In this case the circuit used was somewhat different (see Figure 6.32). With a greater number of devices I was now more concerned about proper current-sharing, and so each device has its own emitter resistor. This makes it look much more like a conventional paralleled output stage, which essentially it is. This time I tried both double and the triple-EF output configurations, as I wished to prove:

(a) that LSN theory worked for both of the common configurations EF and CFP – it does;

(b) that LSN theory worked for both double and triple versions of the EF output stage – it does.

For reasons of space only the triple-EF results are discussed here.

Figure 6.33 shows the measured THD results for one complementary pair of output devices in the triple-EF circuit of Figure 6.32. Distortion is slightly higher, and the noise floor relatively lower, than in previous graphs because of the higher output power of 50 W/8 Ω. Figure 6.34 shows the same except there are now two pairs of output devices. Note that THD has halved at both 8 and 4 Ω loads; this is probably due to the larger currents taken by 8 Ω loads at this higher power. Figure 6.35 shows the result for six devices; 8 Ω distortion has almost been abolished and the 4 Ω result is almost as good. It is necessary to go down to a 2 Ω load to get the THD clear of the noise so it can be measured accurately. With six outputs, driving a substantial amount of power into this load is not a problem.

On a practical note, the more output devices you have, the harder the amplifier may be to purge of parasitic oscillations in the output stage. This is presumably due to the extra raw transconductance available, and can be a problem even with the triple-EF circuit, which has no local NFB loops. I do not pretend to be able to give a detailed explanation of this effect at the moment.

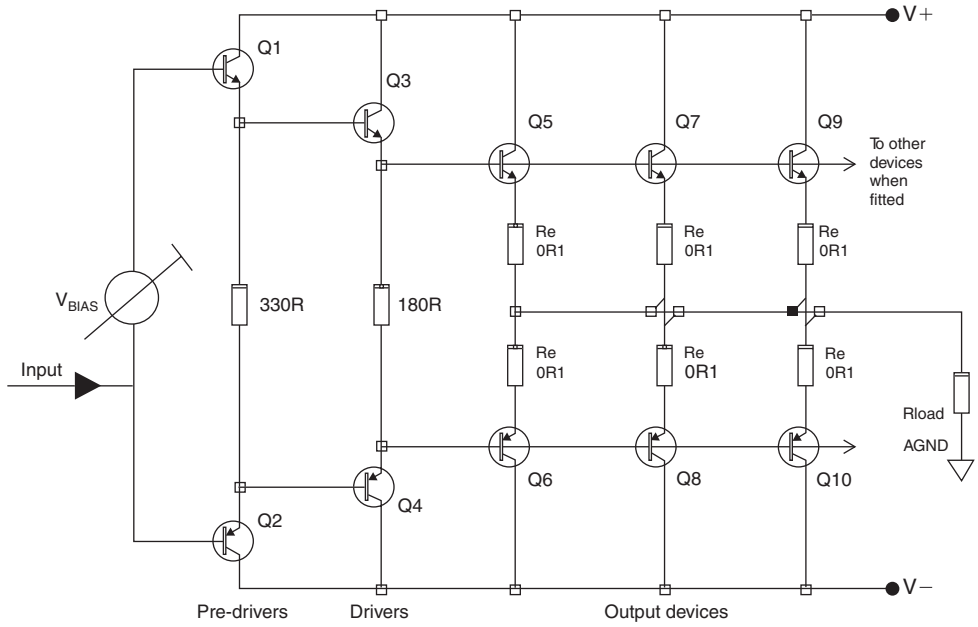


Figure 6.32: The triple-EF output stage used for the measurements described below. 'Triple' refers to the fact that there are three transistors from input to output, rather than the fact that there happen to be three output devices in parallel

Having demonstrated that sustained-beta output devices not only reduce LSN but also unexpectedly reduce crossover distortion, it seemed worth checking if using multiple output devices would give a similar reduction at light loading. I was rather surprised to find they did.

Adding more output devices in parallel, while driving an 8 Ω load, results in a steady reduction in distortion. Figures 6.33–6.35 show how this works in reality. The SPICE simulations in Figure 6.36 reveal that increasing the number *N* of output devices not only flattens the crossover gain wobble, but spreads it out over a greater width. This spreading effect is an extra bonus because it means that lower-order harmonics are generated, and at lower frequencies there will be more negative feedback to linearize them. (Bear in mind also that a triple-EF output has an inherently wider gain wobble than the double-EF.) Taking the gain wobble width as the voltage between the bottoms of the two dips, this appears to be proportional to *N*. The amount of gain wobble, as measured from top of the peak to bottom of the dips, appears to be proportional to 1/*N*.

This makes sense. We know that crossover distortion increases with heavier loading, i.e. with greater currents flowing in the output devices, but under the same voltage conditions. It is therefore not surprising that reducing the device currents by using multiple devices has the same effect as reducing loading. If there are two output devices in parallel, each sees half the current variations and crossover nonlinearity is reduced. The voltage conditions are the same in each half and so are unchanged. This offers us the interesting possibility that crossover distortion – which has hitherto appeared inescapable – can be reduced to an arbitrary level simply by paralleling enough output transistors. To the best of my knowledge this is a new insight.

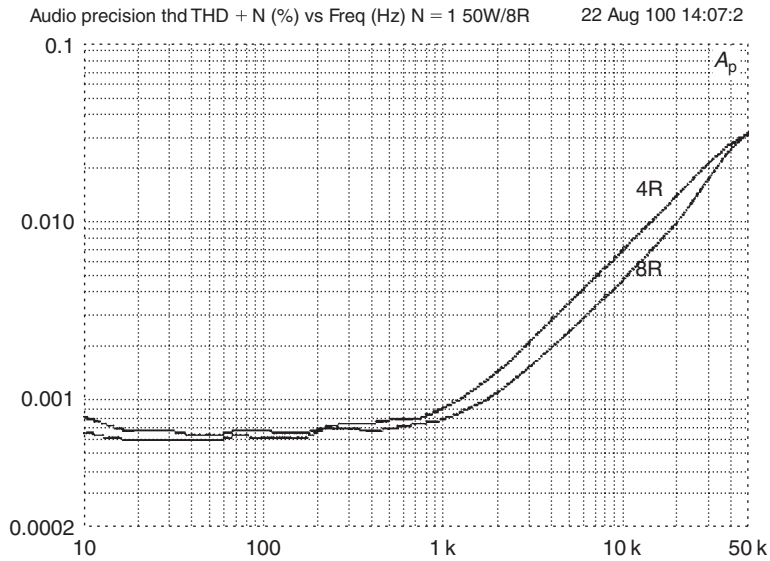


Figure 6.33: THD for one pair ($N = 1$) of output devices, at 50 W/8 R and 100 W/4 R

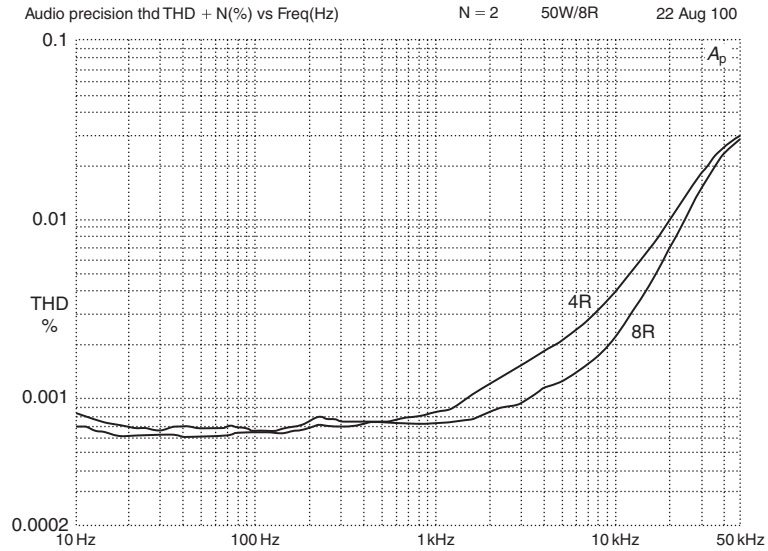


Figure 6.34: THD for two pairs ($N = 2$) of output devices, at 50 W/8 R and 100 W/4 R – a definite improvement

Load Invariance: Summary

In conventional amplifiers, reducing the 8 Ω load to 4 Ω increases the THD by 2–3 times. The figure attained by the Load-Invariant amplifier presented here is 1.2 times, and the ratio could be made even closer to unity by tripling or further multiplying the output devices.

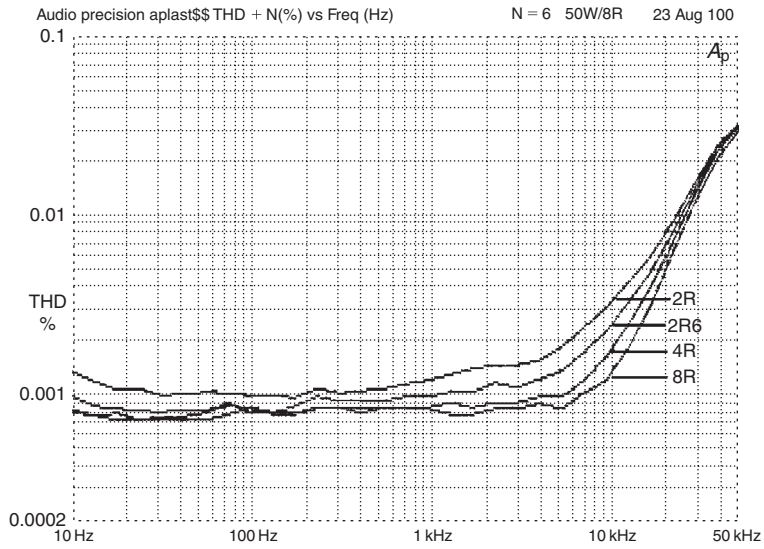


Figure 6.35: THD for six pairs ($N = 6$) of output devices, at 50 W/8 R, 100 W/4 R, and 200 W/2 R. Note very low distortion at 8 Ω

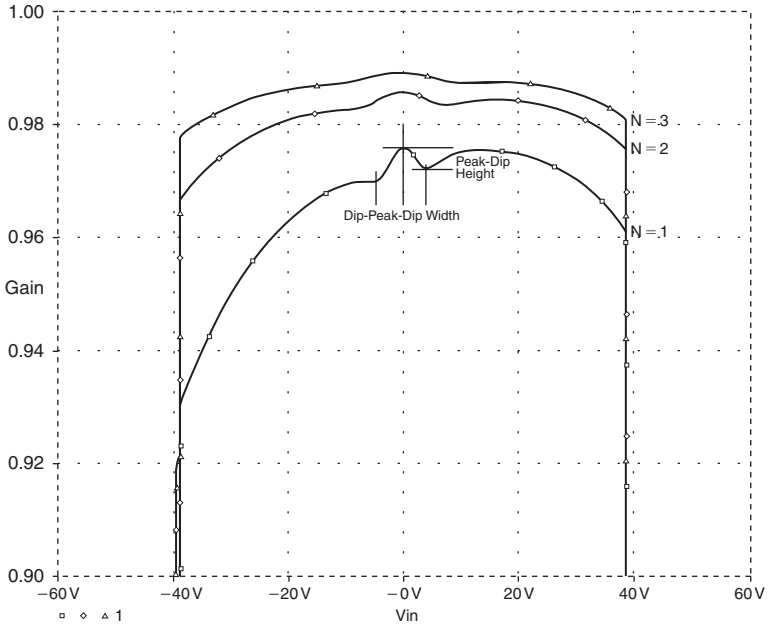


Figure 6.36: SPICE simulation of triple-EF output with $N = 1$, 2, and 3. As N increases the crossover gain wobble becomes flatter and more spread out laterally

Crossover Distortion (Distortion 3b)

In a field like Audio, where consensus of any sort is rare, it is a truth universally acknowledged that crossover distortion is the worst problem that can afflict Class-B power amplifiers. The problem is

the crossover region, where control of the output voltage must be handed over from one device to another. Crossover distortion is rightly feared as it generates unpleasant high-order harmonics, with at least the potential to increase in percentage as signal level falls.

The pernicious nature of crossover distortion is partly because it occurs over a small part of the signal swing, and so generates high-order harmonics. Worse still, this small range over which it does occur is at the zero-crossing point, so not only is it present at all levels and all but the lightest loads, but it is generally believed to increase as output level falls, threatening very poor linearity at the modest listening powers that most people use.

There is a consensus that crossover caused the 'transistor sound' of the 1960s, though to the best of my knowledge this has never actually been confirmed by the double-blind testing of vintage equipment.

The V_{be}/I_c characteristic of a bipolar transistor is initially exponential, blending into linear as the internal emitter resistance r_e comes to dominate the transconductance. The usual Class-B stage puts two of these curves back to back, and Peter Blomley has shown[5] that these curves are non-conjugate, i.e. there is no way they can be shuffled sideways so they will sum to a completely linear transfer characteristic, whatever the offset between them imposed by the bias voltage. This can be demonstrated quickly and easily by SPICE simulation (see Figure 6.37). There is at first sight not much you can do except maintain the bias voltage, and hence quiescent current, at some optimal level for minimum gain deviation at crossover; quiescent-current control is a complex subject that could fill a big book in itself, and is considered in some detail in Chapter 15.

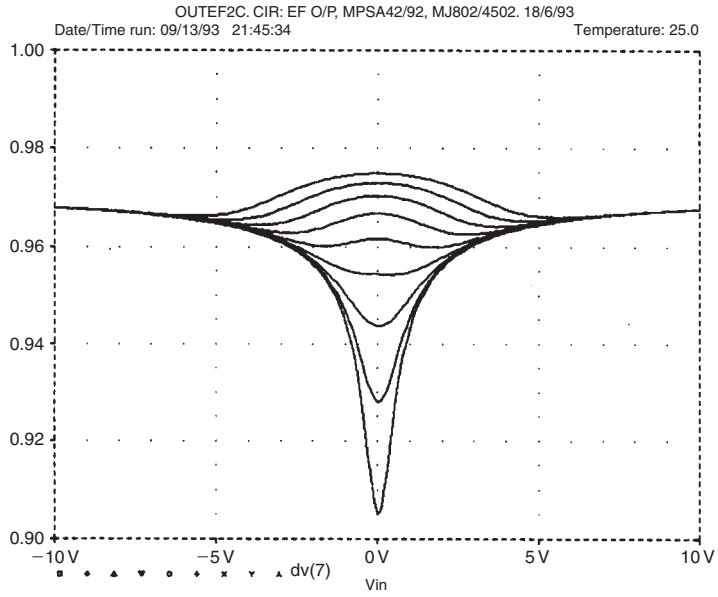


Figure 6.37: Gain/output voltage plot for an EF output shows how non-conjugate transistor characteristics at the crossover region cannot be blended into a flat line at any bias voltage setting. Bias varies from 2.75 to 2.95 V in 25 mV steps, from too little to too much quiescent

It should be said that the crossover distortion levels generated in a Blameless amplifier can be very low up to around 1 kHz, being barely visible in residual noise and only measurable with a spectrum analyzer. As an instructive example, if a Blameless closed-loop Class-B amplifier is driven through a TL072 unity-gain buffer the added noise from this op-amp will usually submerge the 1 kHz crossover artefacts into the noise floor, at least as judged by the eye on the oscilloscope. (It is most important to note that Distortions 4–7 create disturbances of the THD residual at the zero-crossing point that can be easily mistaken for crossover distortion, but the actual mechanisms are quite different.) However, the crossover distortion becomes obvious as the frequency increases, and the high-order harmonics benefit less from NFB.

It will be seen later that in a Blameless amplifier driving 8 Ω the overall linearity is dominated by crossover distortion, even with a well-designed and optimally biased output stage. There is an obvious incentive to minimize this distortion mechanism, but there seems no obvious way to reduce crossover gain deviations by tinkering with any of the relatively conventional stages considered so far.

Figure 6.38 shows the signal waveform and THD residual from a Blameless power amplifier with optimal Class-B bias. Output power was 25 W into 8 Ω, or 50 W into 4 Ω (i.e. the same output voltage) as appropriate, for all the residuals shown here. The figure is a record of a single sweep so the residual appears to be almost totally random noise; without the visual averaging that occurs when we look at an oscilloscope the crossover artefacts are much less visible than in real time.

In Figure 6.39, 64 times averaging is applied, and the disturbances around crossover become very clear. There is also revealed a low-order component at roughly 0.0004%, which is probably due to very small amounts of Distortion 6 that were not visible when the amplifier layout was optimized.

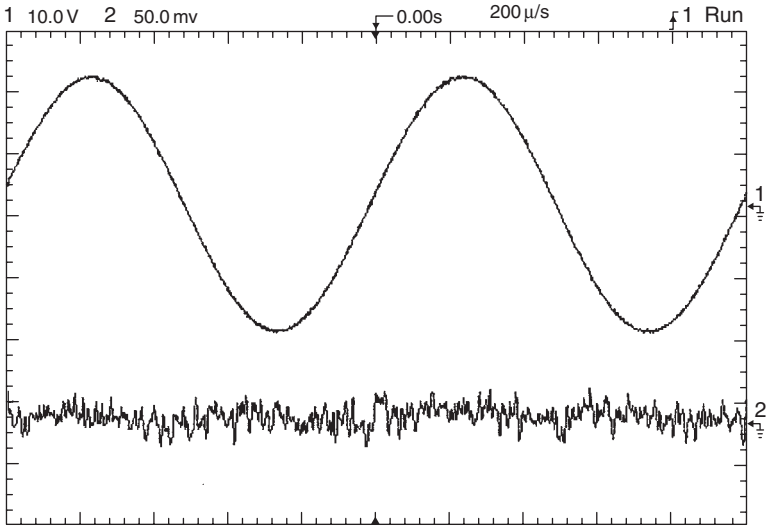


Figure 6.38: The THD residual from an optimally biased Blameless power amplifier at 1 kHz, 25 W/8 Ω is essentially white noise. There is some evidence of artefacts at the crossover point, but they are not measurable. THD 0.00097%, 80 kHz bandwidth

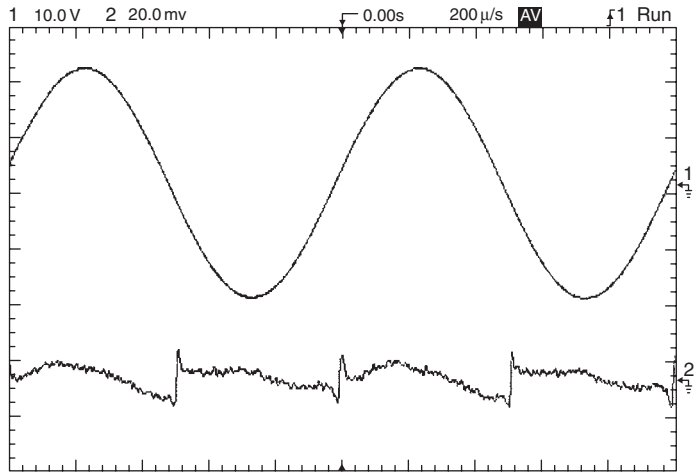


Figure 6.39: Averaging Figure 6.2 residual 64 times reduces the noise by 1 8 dB, and crossover discontinuities are now obvious. The residual has been scaled up by 2.5 times from Figure 6.38 for greater clarity

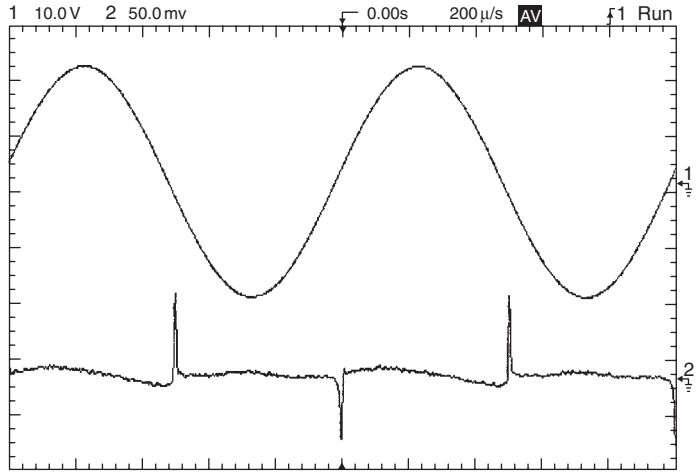


Figure 6.40: The results of mild underbias in Class-B

Figure 6.40 shows Class-B slightly underbiased to generate crossover distortion. The crossover spikes are very sharp, so their height in the residual depends strongly on measurement bandwidth. Their presence warns immediately of underbiasing and avoidable crossover distortion.

In Figure 6.41 an optimally biased amplifier is tested at 10 kHz. The THD increases to approximately 0.004%, as the amount of global negative feedback is 20 dB less than at 1 kHz. The timebase is faster so crossover events appear wider than in Figure 6.39. The THD level is now higher and above the noise so the residual is averaged eight times only. The measurement bandwidth is still 80 kHz, so harmonics above the eighth are now lost. This is illustrated in Figure 6.42, which is Figure 6.41 rerun with a 500 kHz bandwidth. The distortion products now look much more jagged.

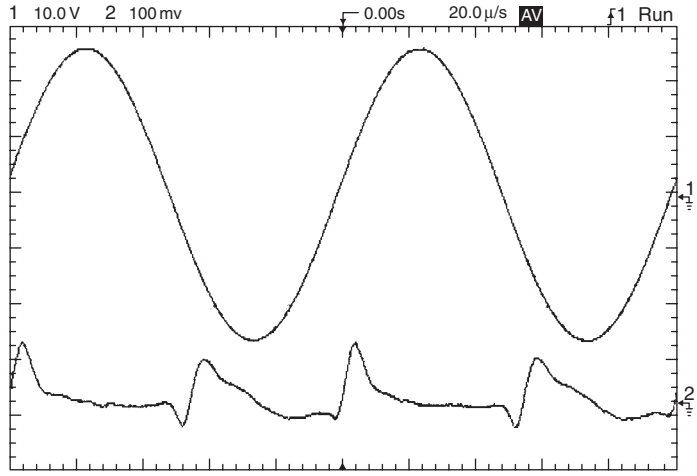


Figure 6.41: An optimally biased Blameless power amplifier at 10 kHz. THD approximately 0.004%, bandwidth 80 kHz. Averaged 8 times

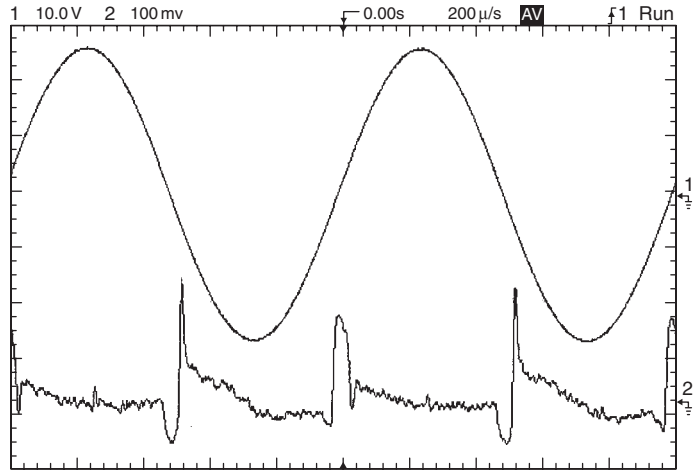


Figure 6.42: As in Figure 6.6, but in 500 kHz bandwidth. The distortion products look quite different

Figure 6.43 shows the gain-step distortion introduced by Class-AB. The undesirable edges in the residual are no longer in close pairs that partially cancel, but are spread apart on either side of the zero crossing. No averaging is used here as the THD is higher (see Chapter 10 for more on Class-AB distortion).

It is commonplace in Audio to discover that a problem like crossover distortion has been written about and agonized over for decades, but the amount of technical investigation that has been done (or at any rate published) is disappointingly small. I had to do some basic investigations myself.

I first looked to see if crossover distortion really *did* increase with decreasing output level in a Blameless amplifier; to attempt its study with an amplifier contaminated with any of the avoidable distortion mechanisms is completely pointless. One problem is that a Blameless amplifier has such

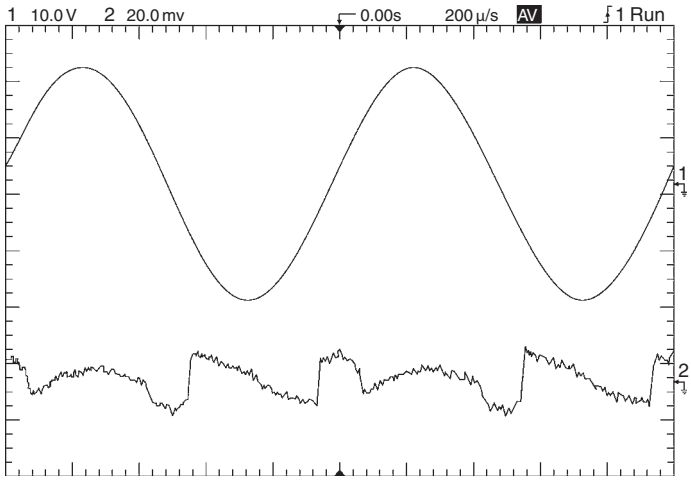


Figure 6.43: The g_m-doubling distortion introduced by Class-AB. The edges in the residual are larger and no longer at the zero crossing, but displaced either side of it

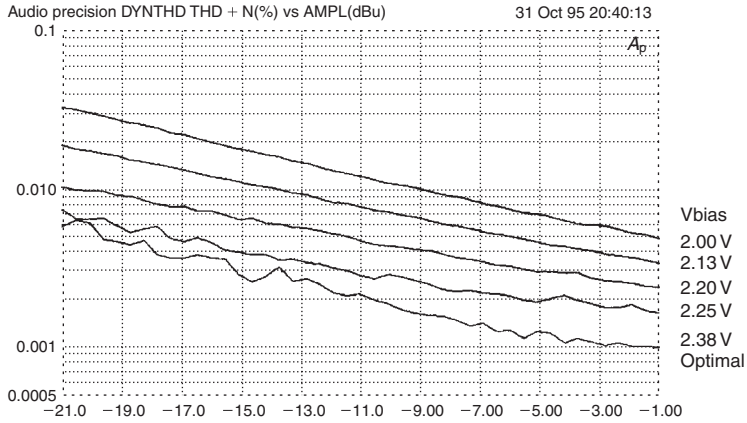


Figure 6.44: How crossover distortion rises slowly as output power is reduced from 25 W to 250 mW (8 Ω) for optimal bias and increasingly severe underbias (upper lines). This is an EF-type output stage. Measurement bandwidth 22 kHz

a low level of distortion at 1 kHz (0.001% or less) that the crossover artefacts are barely visible in circuit noise, even if low-noise techniques are used. The measured percentage level of the noise-plus-distortion residual is bound to rise with falling output, because the noise voltage remains constant; this is the lowest line in Figure 6.44. To circumvent this, the amplifier was deliberately underbiased by varying amounts to generate ample crossover spikes, on the assumption that any correctly adjusted amplifier should be less barbarous than this.

The answer from Figure 6.44 is that the THD percentage does increase as level falls, but relatively slowly. Both EF and CFP output stages give similar diagrams to Figure 6.38, and whatever the degree of underbias, THD increases by about 1.6 times as the output voltage is halved. In other words, reducing the output power from 25 W to 250 mW, which is pretty drastic, only increases

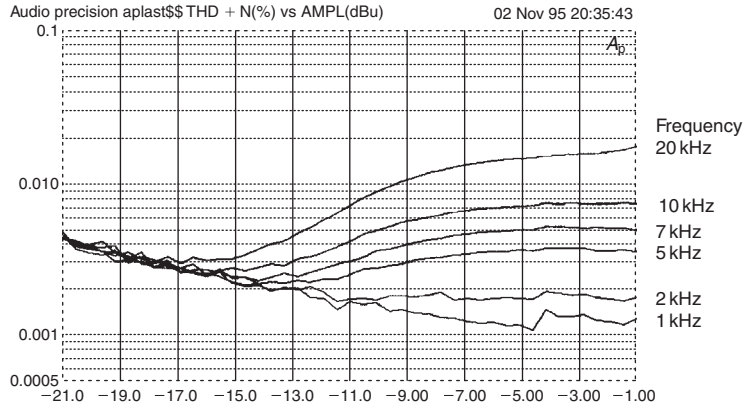


Figure 6.45: Variation of crossover distortion with output level for higher frequencies. Optimally biased EF output stage. Bandwidth 80 kHz

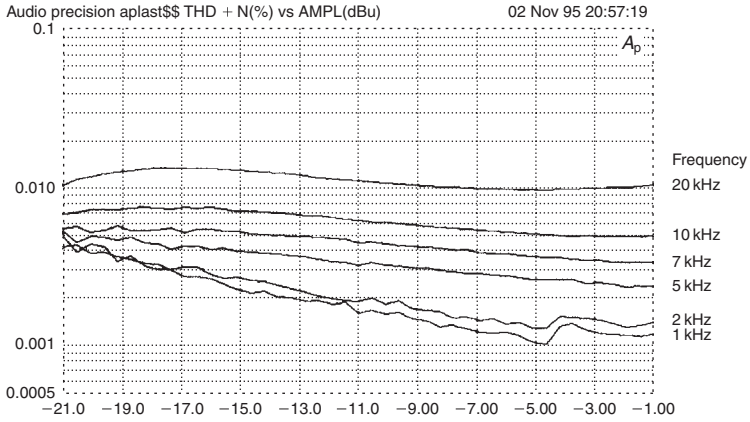


Figure 6.46: Variation of distortion with level for higher frequencies. Optimally biased CFP output stage. Bandwidth 80 kHz

THD percentage by six times, and so it is clear that the *absolute* (as opposed to percentage) THD level in fact falls slowly with amplitude, and therefore probably remains imperceptible. This is something of a relief; but crossover distortion remains a bad thing to have.

Distortion versus level was also investigated at high frequencies, i.e. above 1 kHz, where there is more THD to measure and optimal biasing can be used. Figure 6.45 shows the variation of THD with level for the EF stage at a selection of frequencies; Figure 6.46 shows the same for the CFP. Neither shows a significant rise in percentage THD with falling level, though it is noticeable that the EF gives a good deal less distortion at lower power levels around 1 W. This is an unexpected observation, and possibly a new one.

To further get the measure of the problem, Figure 6.47 shows how HF distortion is greatly reduced by increasing the load resistance, providing further confirmation that almost all the 8 Ω distortion originates as crossover in the output stage.

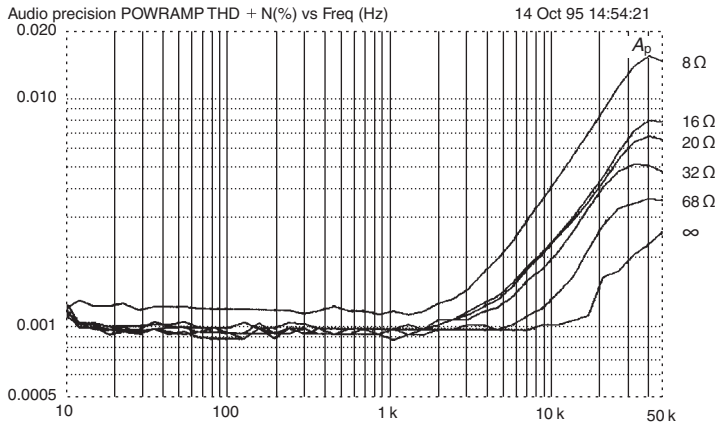


Figure 6.47: How crossover distortion is reduced with increasing load resistance. 20 W into 8 Ω, 80 kHz bandwidth

Crossover distortion, unlike some more benign kinds of signal-warping, is unanimously agreed to be something any amplifier could well do without. The amount of crossover distortion produced depends strongly on optimal quiescent adjustment, so the thermal compensation used to stabilize this against changes in temperature and power dissipation must be accurate.

This section deals with the crossover region and its quiescent conditions, and the specific issues of the effectiveness of the thermal compensation for temperature effects are dealt with in detail in Chapter 15.

Output Stage Quiescent Conditions

Figure 6.48 shows the two most common types of output stage: the EF and CFP configurations. The manifold types of output stage based on triples will have to be set aside for the moment. The two circuits shown have few components, and there are equally few variables to explore in attempting to reduce crossover distortion.

To get the terminology straight: here, as in my previous writings, V_{bias} refers to the voltage set up across the driver bases by the V_{be}-multiplier bias generator, and is in the range 1–3 V for Class-B operation. V_q is the quiescent voltage across the two emitter resistors (hereafter Re) alone, and is between 5 and 50 mV, depending on the configuration chosen. Quiescent current I_q refers only to that flowing in the output devices, and does not include driver standing currents.

I have already shown that the two most common output configurations are quite different in behavior, with the CFP being superior on most criteria. Table 6.3 shows that crossover gain variation for the EF stage is smoother (being some 20 times wider) but of four times higher amplitude than for the CFP version. It is not immediately obvious from this which stage will generate the least HF THD, bearing in mind that the NFB factor falls with frequency.

Table 6.3 also emphasizes that a little-known drawback of the EF version is that its quiescent dissipation may be far from negligible.

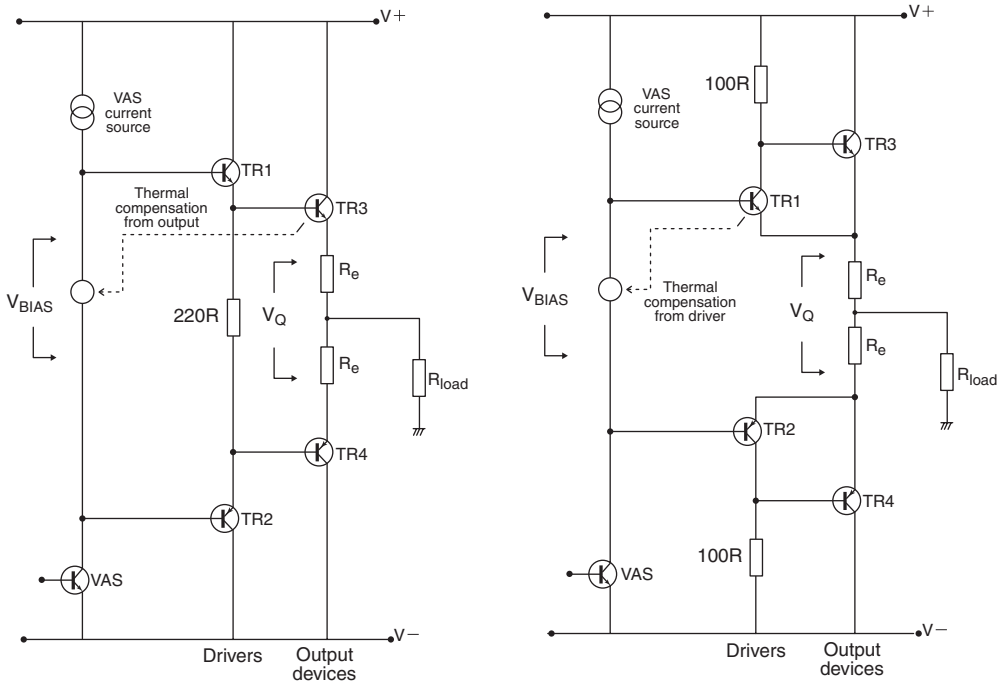


Figure 6.48: The two most popular kinds of output stage: the emitter-follower (EF) and complementary feedback pair (CFP). V_{bias} and V_q are identified

Table 6.3: Quiescent conditions compared

	EF	CFP
V_{bias} (V)	2.930	1.297
V_q (mV)	50	5
I_q (mA)	114	11
P_q per O/P device (W)	4.6	0.44
Average gain	0.968	0.971
Peak gain deviation from average (%)	0.48	0.13
Crossover width (V)*	±12	±0.6

For Re = OR22, 8 Ω load, and ±40V supply rails.

*Crossover width is the central region of the output voltage range over which crossover effects are significant; I have rather arbitrarily defined it as the ± output range over which the incremental gain curves diverge by more than 0.0005 when V_{bias} is altered around the optimum value. This is evaluated here for an 8 Ω load only.

An Experiment on Crossover Distortion

Looking hard at the two output stage circuit diagrams, intuition suggests that the value of emitter resistor Re is worth experimenting with. Since these two resistors are placed between the output devices, and alternately pass the full load current, it seems possible that their value could be critical in mediating the handover of output control from one device to the other. Re was therefore stepped

from 0.1 to 0.47 Ω, which covers the practical range. V_{bias} was re-optimized at each step, though the changes were very small, especially for the CFP version.

Figure 6.49 shows the resulting gain variations in the crossover region for the EF stage, while Figure 6.50 shows the same for the CFP configuration. Table 6.4 summarizes some numerical results for the EF stage, and Table 6.5 for the CFP.

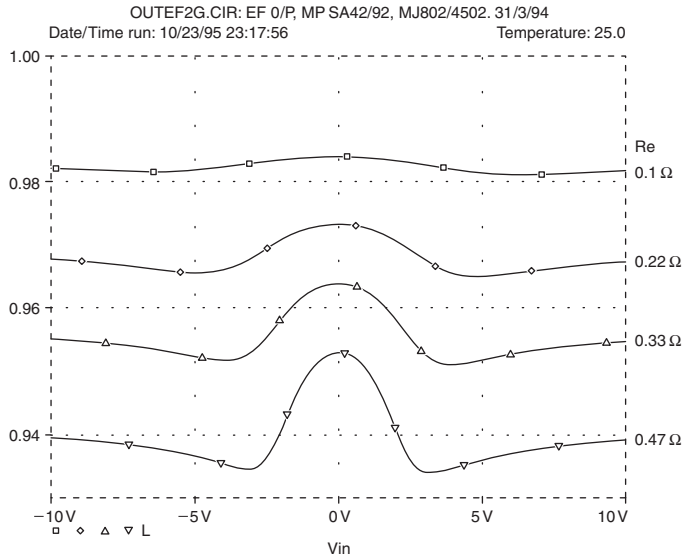


Figure 6.49: Output linearity of the EF output stage for emitter resistor Re between 0.1 and 0.47 Ω

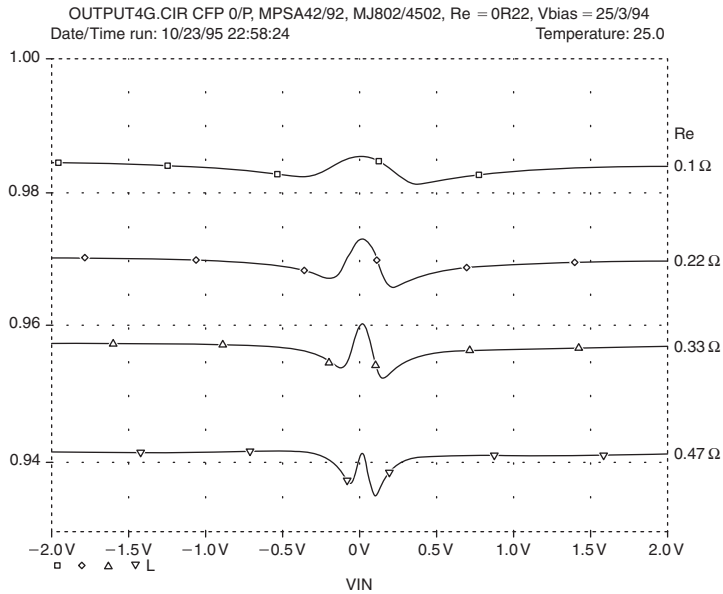


Figure 6.50: Output linearity of the CFP output stage for emitter resistor Re between 0.1 and 0.47 Ω

Table 6.4: Emitter-follower output (Type 1): data for 8 Ω load and EF O/P stage

Re (Ω)	Optimal V_{bias} (V)	Optimal V_q (mV)	I_q (mA)	X-Width (V)	Average gain ratio
0.1	2.86	42.6	215	18	0.982
0.22	2.87	46.2	107	12	0.968
0.33	2.89	47.6	74	9	0.955
0.47	2.93	54.8	59	7	0.939

As Re is varied, V_q varies by only 29%, while I_q varies by 365%.

Table 6.5: CFP output: data for 8 Ω load and CFP O/P stage

Re (Ω)	Optimal V_{bias} (V)	Optimal V_q (mV)	I_q (mA)	X-Width (V)	Average gain ratio
0.1	1.297	3.06	15.3	1.0	0.983
0.22	1.297	4.62	11.5	0.62	0.971
0.33	1.297	5.64	8.54	0.40	0.956
0.47	1.298	7.18	7.64	0.29	0.941

There are some obvious features. First, Re is clearly not critical in value as the gain changes in the crossover region are relatively minor. Reducing the Re value allows the average gain to approach unity more closely, with a consequent advantage in output power capability. Similarly, reducing Re widens the crossover region for a constant load resistance, because more current must pass through one Re to generate enough voltage drop to turn off the other output device. This implies that as Re is reduced, the crossover products become lower order and so of lower frequency. They should be better linearized by the frequency-dependent global NFB, and so overall closed-loop HF THD should be lower.

The simulated crossover distortion experiment described earlier in this chapter showed that as the crossover region was made narrower, the distortion energy became more evenly spread over higher harmonics. A wider crossover region implies energy more concentrated in the lower harmonics, which will receive the benefit of more negative feedback. However, if the region is made wider, but retains the same amount of gain deviation, it seems likely that the total harmonic energy is greater, and so there are two opposing effects to be considered.

I conclude that selecting Re = OR1 for maximum efficiency is probably the overriding consideration. This has the additional benefit that if the stage is erroneously overbiased into Class-AB, the resulting g_m-doubling distortion will only be half as bad as if the more usual OR22 values had been used for Re.

It would be easy to assume that higher values of Re must be more linear, because of a vague feeling that there is more local feedback, but this cannot be true as an emitter-follower already has 100% voltage feedback to its emitter, by definition. Changing the value of Re alters slightly the total resistive load seen by the emitter itself, and this does seem to have a small but measurable effect on linearity.

As Re is varied, V_q varies by 230% while I_q varies by 85%. However, the absolute V_q change is only 4 mV, while the sum of V_{be} values varies by only 0.23%. This makes it pretty plain that the voltage domain is what counts, rather than the absolute value of I_q.

The first surprise from this experiment is that in the typical Class-B output stage, quiescent current as such does not matter a great deal. This may be hard to believe, particularly after my repeated statements that quiescent conditions are critical in Class-B, but both assertions are true. The data for both the EF and CFP output stages show that changing Re alters I_q considerably, but the optimal values of V_{bias} and V_q barely change.

The voltage across the transistor base–emitter junctions and Re resistors seems to be what counts, and the actual value of current flowing as a result is not in itself of much interest. However, the V_{bias} setting remains critical for minimum distortion; once the Re value is settled at the design stage, the adjustment procedure for optimal crossover is just as before.

The irrelevance of quiescent current was confirmed by the Trimodal amplifier, which was designed after the work described here was done, and where I found that changing the output emitter resistor value Re over a 5:1 range required no alteration in V_{bias} to maintain optimal crossover conditions.

The critical factor is therefore the voltages across the various components in the output stage. Output stages get hot, and when the junction temperatures change, both experiment and simulation show that if V_{bias} is altered to maintain optimal crossover, V_q remains virtually constant. This confirms the task of thermal compensation is solely to cancel out the V_{be} changes in the transistors; this may appear to be a blinding glimpse of the obvious, but it was worth checking as there is no inherent reason why the optimal V_q should not be a function of device temperature. Fortunately it is not, for thermal compensation that also dealt with a need for V_q to change with temperature might be a good deal more complex.

V_q as the Critical Quiescent Parameter

The recognition that V_q is the critical parameter has some interesting implications. Can we immediately start setting up amplifiers for optimal crossover with a cheap DVM rather than an expensive THD analyzer? Setting up quiescent current with a milliammeter has often been advocated, but the direct measurement of this current is not easy. It requires breaking the output circuit so a meter can be inserted, and not all amplifiers react favorably to so rude an intrusion. (The amplifier must also have near-zero DC offset voltage to get any accuracy.) Measuring the total amplifier consumption is not acceptable because the standing current taken by the small-signal and driver sections will, in the CFP case at least, swamp the quiescent current. It is possible to determine quiescent current indirectly from the V_q drop across the Re resistors (still assuming zero DC offset) but this can never give a very accurate current reading as the tolerance of low-value Re resistors is unlikely to be better than $\pm 10\%$.

However, if V_q is the real quantity we need to get at, then Re tolerances can be blissfully ignored. This does not make THD analyzers obsolete overnight. It would be first necessary to show that V_q was always a reliable indicator of crossover setting, no matter what variations occurred in driver or output transistor parameters. This would be a sizeable undertaking.

There is also the difficulty that real-life DC offsets are not zero, though this could possibly be sidestepped by measuring V_q with the load disconnected. A final objection is that without THD

analysis and visual examination of the residual, you can never be sure an amplifier is free from parasitic oscillations and working properly.

I have previously demonstrated that the distortion behavior of a typical amplifier is quite different when driving 4 Ω rather than 8 Ω loads. This is because with the heavier load, the output stage gain behavior tends to be dominated by beta loss in the output devices at higher currents, and consequent extra loading on the drivers, giving third-harmonic distortion. If this is to be reduced, which may be well worthwhile as many loudspeaker loads have serious impedance dips, then it will need to be tackled in a completely different way from crossover distortion.

It is disappointing to find that no manipulation of output stage component values appears to significantly improve crossover distortion, but apart from this one small piece of (negative) information gained, we have in addition determined the following:

1. Quiescent current as such does not matter; V_q is the vital quantity.
2. A perfect thermal compensation scheme, which was able to maintain V_q at exactly the correct value, requires no more information than the junction temperatures of the driver and output devices. Regrettably none of these temperatures are actually accessible, but at least we know what to aim for. The introduction of the Sanken and ONsemi ThermalTrak transistors with integral temperature-sense diodes (see Chapter 15) opens possibilities in this direction but it remains to be seen how best to exploit this new technology.

As an aside, there is anecdotal evidence that back when transistors were made of germanium, crossover distortion was less of a problem because germanium transistors turn on more gradually. I have no idea if this is true or not, and making a germanium-device power amplifier nowadays is hardly practical, but it is an interesting point.

Switching Distortion (Distortion 3c)

This depends on several variables, notably the speed characteristics of the output devices and the output topology. Leaving aside the semiconductor physics and concentrating on the topology, the critical factor is whether or not the output stage can reverse-bias the output device base–emitter junctions to maximize the speed at which carriers are sucked out, so the device is turned off quickly. The only conventional configuration that can reverse-bias the output base–emitter junctions is the EF Type II, described earlier.

A second influence is the value of the driver emitter or collector resistors; the lower they are, the faster the stored charge can be removed. Applying these criteria can reduce HF distortion markedly, but of equal importance is that it minimizes overlap of output conduction at high frequencies, which if unchecked results in an inefficient and potentially destructive increase in supply current[12]. To illustrate this, Figure 6.51 shows a graph of current consumption versus frequency for varying driver collector resistance, for a CFP-type output.

Figure 6.52 shows the reduction of HF THD by adding a speed-up capacitor across the common driver resistor of an EF Type II. At LF the difference is small, but at 40 kHz THD is halved, indicating much cleaner switch-off. There is also a small benefit over the range 300 Hz–8 kHz.

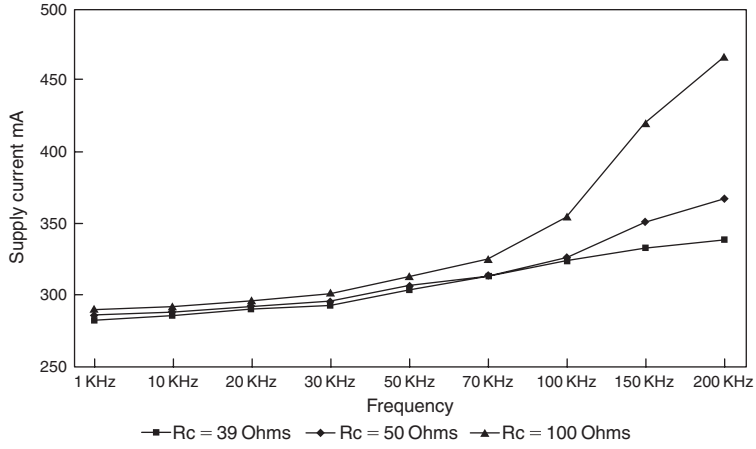


Figure 6.51: Power supply current versus frequency, for a CFP output with the driver collector resistors varied. There is little to be gained from reducing R_c below 50 Ω

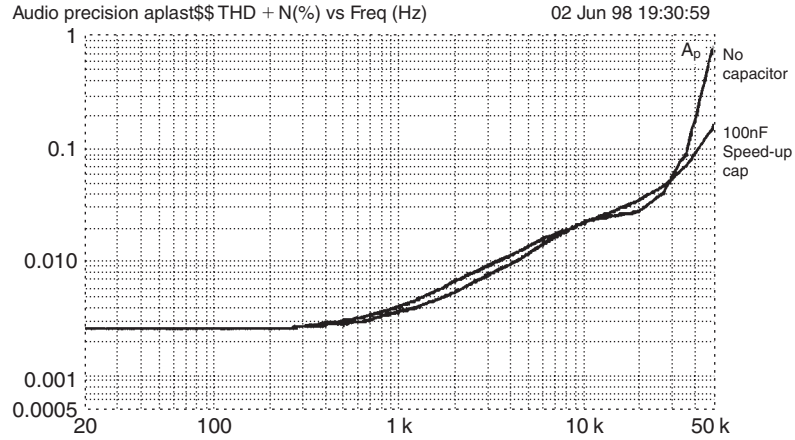


Figure 6.52: HF THD reduction by adding speed-up capacitance across the common driver resistance of a Type II EF output stage

Thermal Distortion

Thermal distortion is that caused by cyclic temperature changes at signal frequency, causing corresponding modulation of device parameters. While it is certainly a real problem in IC op-amps, which have input and output devices in very close thermal proximity, the situation in a normal discrete-component power amplifier is quite different, and thermal distortion cannot be detected. Having studied in detail distortion mechanisms that are all too real, it comes as some relief to find that one prospective distortion is illusory. Some writers appear to take it as given that such a distortion mechanism exists in power amplifiers, but having studied the subject in some depth I have yet to see the effect, and quite frankly I do not think it exists.

While now and again there have been odd mentions of thermal distortion in power amps in some of the hi-fi press, you will never find:

1. any explanation of how it might work;
2. any estimate of the magnitude of the effect;
3. a circuit that will demonstrate its production.

In the usual absence of specific theories, one can only assume that the alleged mechanism induces parameter changes in semiconductors whose power dissipation varies over a cycle. If this were to happen, it would presumably manifest itself as a rise in second- or third-harmonic distortion at very low frequencies, but this simply does not happen. The largest effects would be expected in Class-B output stages where dissipation varies wildly over a cycle; the effect is still wholly absent.

One reason for this may be that drivers and output devices have relatively large junctions with high thermal inertia – a few seconds with a hammer and chisel revealed that an MJE340 driver has a chip with four times the total area of a TL072. Given this thermal mass, parameters presumably cannot change much even at 10 Hz. Low frequencies are also where the global NFB factor is at its maximum; it is perfectly possible to design an amplifier with 100 dB of feedback at 10 Hz, though much more modest figures are sufficient to make distortion unmeasurably low up to 1 kHz or so. Using my design methodology a Blameless amplifier can be straightforwardly designed to produce less than 0.0006% THD at 10 Hz (150 W/8 Ω) without even considering thermal distortion; this suggests that we have here a non-problem.

I accept that it is not uncommon to see amplifier THD plots that rise at low frequencies, but whenever I have been able to investigate this, the LF rise could be eliminated by attending to either defective decoupling or feedback-capacitor distortion. Any thermal distortion must be at a very low level as it is invisible at 0.0006%; remember that this is the level of a THD reading that is visually pure noise, though there are real amplifier distortion products buried in it.

I have therefore done some deeper investigation by spectrum analysis of the residual, which enables the harmonics to be extracted from the noise. The test amplifier was an optimally biased Class-B machine very similar to that in Figure 6.16, except with a CFP output. The Audio Precision oscillator is very, very clean but this amplifier tests it to its limits, and so Table 6.6 shows harmonics in a before-and-after amplifier comparison. The spectrum analyzer bandwidth was 1 Hz for 10 Hz tests and 4.5 Hz for 1 kHz, to discriminate against wideband noise.

This further peeling of the distortion onion shows several things: that the AP is a brilliant piece of machinery, and that the amplifier is really quite linear too. However, there is nothing resembling evidence for thermal distortion effects.

As a final argument, consider the distortion residual of a slightly underbiased power amp, using a CFP output configuration so that output device junction temperatures do not affect the quiescent current; it therefore depends only on the driver temperatures. When the amplifier is switched on and begins to apply sine-wave power to a load, the crossover spikes (generated by the deliberate

Table 6.6: Relative amplitude of distortion harmonics

	10 Hz AP out (%)	Amp out (%)	1 kHz AP out (%)	Amp out (%)
Fundamental	0.00013	0.00031	0.00012	0.00035
Second	0.00033	0.00092	0.00008	0.00060
Third	0.00035	0.000050	0.000013	0.00024
Fourth	<0.000002	0.00035	<0.000008	0.00048
Fifth	<0.00025	<0.00045	0.000014	0.00024
Sixth	<0.000006	0.00030	0.000008	0.00021
Seventh	<0.000006	<0.00008	0.000009	0.00009
Eighth	<0.000003	0.000003	0.000008	0.00016
Ninth	<0.000004	0.00011	0.000007	<0.00008
AP THD reading (80 kHz bandwidth)	0.00046	0.00095	0.00060	0.00117

NB: The rejection of the fundamental is not perfect, and this is shown as it contributes to the THD figure.

underbiasing) will be seen to slowly shrink in height over a couple of minutes as the drivers warm up. This occurs even with the usual temperature compensation system, because of the delays and losses in heating up the V_{be}-multiplier transistor.

The size of these crossover spikes gives in effect a continuous readout of driver temperature, and the slow variations that are seen imply time-constants measured in tens of seconds or more; this must mean a negligible response at 10 Hz.

There is no doubt that long-term thermal effects can alter Class-B amplifier distortion, because as I have written elsewhere, the quiescent current setting is critical for the lowest possible high-frequency THD. However, this is strictly a slow (several minutes) phenomenon, whereas enthusiasts for thermal distortion are thinking of the usual sort of per-cycle distortion.

The above arguments lead me to conclude that thermal distortion as usually described does not exist at a detectable level.

Thermal Distortion in a Power Amp IC

As explained above, thermal nonlinearities would presumably appear as second- or third-harmonic distortion rising at low frequencies, and the largest effects should be in Class-B output stages where dissipation varies greatly over a cycle. There is absolutely no such effect to be seen in discrete-component power amplifiers.

But thermal distortion certainly does exist in IC power amplifiers. Figure 6.53 is a distortion plot for the Philips TDA 1522Q power amp IC, which I believe shows the effect. The power level was 4.4 W into 8 Ω, 8 W into 4 Ω. As is usual for such amplifiers, the distortion is generally high, but drops into a notch at 40 Hz; the only feasible explanation for this is cancelation of distortion products from two separate distortion sources. At frequencies below this notch there is second-harmonic distortion rising at 12 dB/octave as frequency falls. The LF residual looks quite different from the midband distortion, which was a mixture of second- and third-harmonic plus crossover spikes.

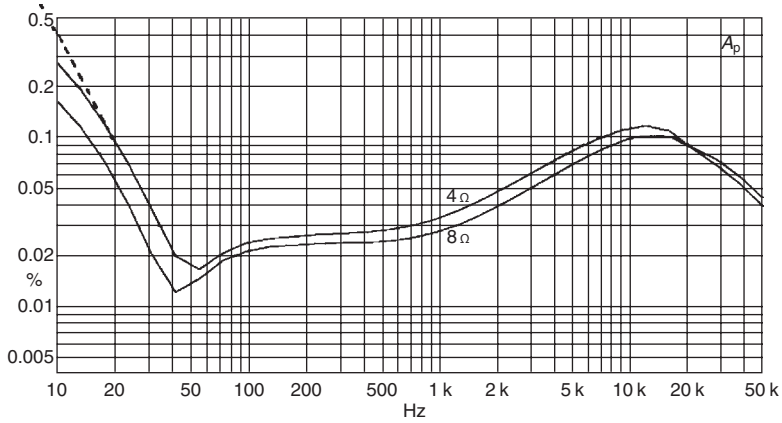


Figure 6.53: Distortion plot for the Philips TDA1522Q IC. Power out was 4.4 W rms into 8 Ω, 8 W rms into 4 Ω. The dotted line shows a 12 dB/octave slope

The THD figure falls above 10 kHz because of the 80 kHz bandwidth limitation on the residual, and the high-order nature of the harmonics that make up crossover distortion.

All other possible sources of an LF distortion rise, such as inadequate decoupling, were excluded. There was no output capacitor to introduce nonlinearity.

It seems pretty clear that the steep LF rise here is due to thermal distortion, in the form of feedback from the power output stage to earlier parts of the amplifier – probably the input stage. As would be expected, the effect is greater with a heavier load, which causes more heating; in fact halving the load doubles the THD reading below the 40 Hz notch.

Selecting an Output Stage

Even if we stick to the most conventional of output stages, there are still an embarrassingly large number to choose form. The cost of a complementary pair of power FETs is currently at least twice that of roughly equivalent BJTs, and taken with the poor linearity and low efficiency of these devices, the use of them may require a marketing rather than a technical motivation.

Turning to BJTs, I conclude that there are the following candidates for Best Output Stage:

1. The EF Type II output stage is the best at coping with switch-off distortion but the quiescent-current stability needs careful consideration.
2. The CFP topology has good quiescent stability and low large-signal nonlinearity; it has the drawback that reverse-biasing the output device bases for fast switch-off is impossible without additional HT rails.
3. The quasi-complementary-with-Baxandall-diode stage comes close to mimicking the EF-type stages in linearity, with a potential for some cost savings on output devices. Quiescent stability is not as good as the CFP configuration.

Closing the Loop: Distortion in Complete Amplifiers

In Chapters 4 and 5 it was shown how relatively simple design rules could ensure that the THD of the small-signal stages alone could be reduced to less than 0.001% across the audio band, in a thoroughly repeatable fashion, and without using frightening amounts of negative feedback. Combining this subsystem with one of the more linear output stages described in Chapter 4, such as the CFP version, which gives 0.014% THD open-loop, and bearing in mind that ample NFB is available, it seems we have all the ingredients for a virtually distortionless power amplifier. However, life is rarely so simple . . .

Figure 6.54 shows the distortion performance of such a closed-loop amplifier with an EF output stage, Figure 6.55 showing the same with a CFP output stage. Figure 6.56 shows the THD of a quasi-complementary stage with Baxandall diode. In each case Distortions 1, 2, and 4–7 have been eliminated, by methods described in past and future chapters, to make the amplifier Blameless.

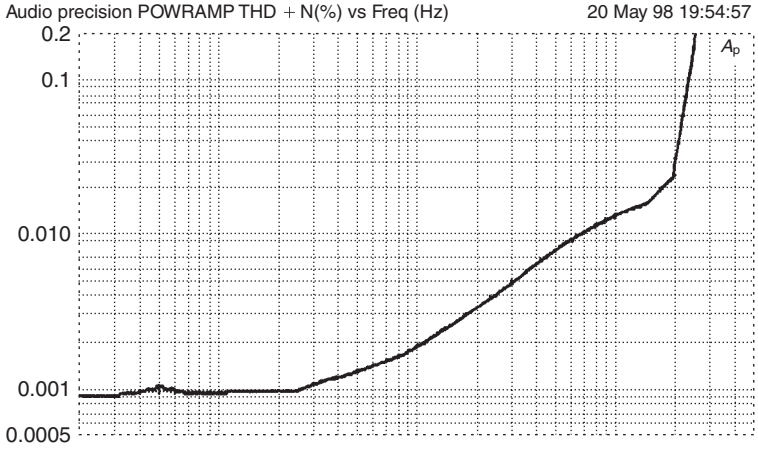


Figure 6.54: Closed-loop amplifier performance with emitter-follower output stage, 100 W into 8 Ω

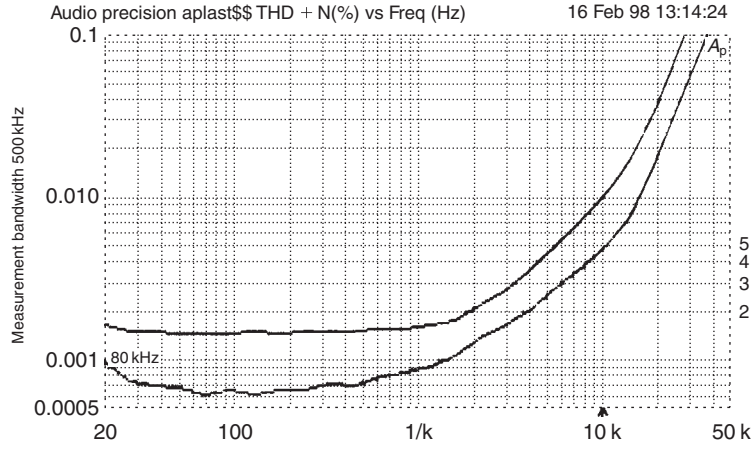


Figure 6.55: Closed-loop amplifier performance with CFP output, 100 W into 8 Ω

(Note: the AP plots in Figures 6.54–6.56 were taken at 100 W rms into 8 Ω, from an amplifier with an input error of −70 dB at 10 kHz and a C/L gain of 27 dB, giving a feedback factor of 43 dB at this frequency. This is well above the dominant-pole frequency and so the NFB factor is dropping at 6 dB/octave and will be down to 37 dB at 20 kHz. My experience suggests that this is about as much feedback as is safe for general hi-fi usage, assuming an output inductor to improve stability with capacitive loads. Sadly, published data on this touchy topic seems to be nonexistent.)

It will be seen at once that these amplifiers are not distortionless, though the performance is markedly superior to the usual run of hardware. THD in the LF region is very low, well below a noise floor of 0.0007%, and the usual rise below 100 Hz is very small indeed. However, above 2 kHz, THD rises with frequency at between 6 and 12 dB/octave, and the distortion residual in this region is clearly time-aligned with the crossover region, and consists of high-order harmonics rather than second or third. It is intriguing to note that the quasi-Baxandall output gives about the same HF THD as the EF topology, which confirms my earlier statement that the addition of a Baxandall diode essentially turns a conventional quasi-complementary stage with serious crossover asymmetry into a reasonable emulation of a complementary EF stage. There is less HF THD with a CFP output; this cannot be due to large-signal nonlinearity as this is negligible with an 8 Ω load for all three stages, and so it must be due to high-order crossover products (see Table 6.7).

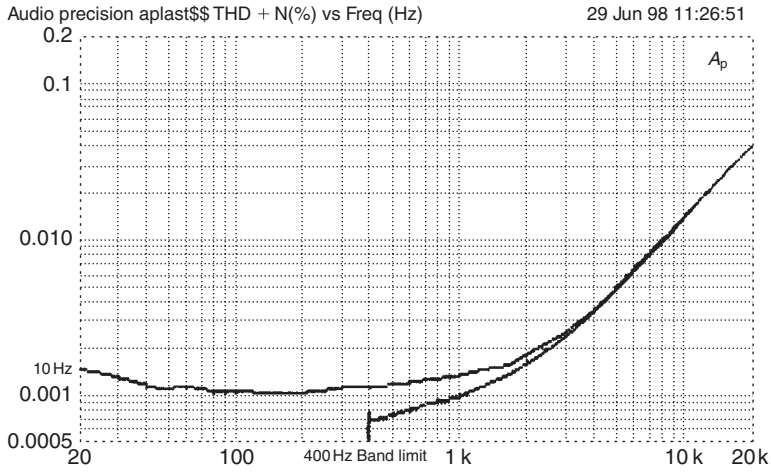


Figure 6.56: Closed-loop amplifier performance. Quasi-complementary output stage with Baxandall diode, 100 W into 8 Ω

Table 6.7: Summary of closed-loop amp performance

	1 kHz (%)	10 kHz (%)
EF	0.0019	0.013
CFP	0.0008	0.005
Quasi Bax	0.0015	0.015

The distortion figures given in this book are rather lower than usual. I would like to emphasize that these are not freakish or unrepeatable figures; they are the result of attending to all of the major sources of distortion, rather than just one or two. I have, at the time of writing, personally built 12 models of the CFP version, and performance showed little variation.

Here the closed-loop distortion is much greater than that produced by the small-signal stages alone; however, if the input pair is badly designed its HF distortion can easily exceed that caused by the output stage.

Our feedback factor here is a minimum of 70 times across the band (being much higher at LF) and the output stages examined above are mostly capable of less than 0.1% THD open-loop. It seems a combination of these should yield a closed-loop distortion at least 70 times better, i.e. below 0.001% from 10 Hz to 20 kHz. This happy outcome fails to materialize, and we had better find out why . . .

First, when an amplifier with a frequency-dependent NFB factor generates distortion, the reduction is not that due to the NFB factor at the fundamental frequency, but the amount available at the frequency of the harmonic in question. A typical amplifier with O/L gain rolling off at 6 dB/octave will be half as effective at reducing fourth-harmonic distortion as it is at reducing the second harmonic. LSN is largely third (and possibly second) harmonic, and so NFB will deal with this effectively. However, both crossover and switch-off distortions generate high-order harmonics significant up to at least the nineteenth and these receive much less linearization. As the fundamental moves up in frequency the harmonics do too, and benefit from even less feedback. This is the reason for the 'differentiated' look to many distortion residuals; higher harmonics are emphasized at the rate of 6 dB/octave.

Here is a real example of the inability of NFB to cure all possible amplifier ills. To reduce this HF distortion we must reduce the crossover gain deviations of the output stage before closing the loop. There seems no obvious way to do this by minor modifications to any of the conventional output stages; we can only optimize the quiescent current.

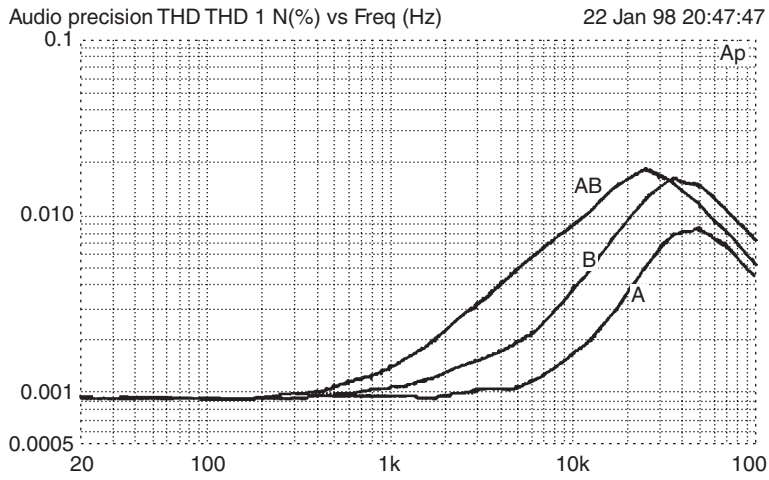


Figure 6.57: Closed-loop CFP amp. Setting quiescent for Class-AB gives more HF THD than either Class-A or - B

As I stated earlier in this chapter, Class-AB is generally not a good thing, as it gives more distortion than Class-B, rather than less, and so will not help us. Figure 6.57 makes this very clear for the closed-loop case; Class-AB clearly gives the worst performance. (As before, the AB quiescent was set for 50:50 m/s ratio of the g_m-doubling artefacts on the residual.)

Conclusions

1. Class-AB is best avoided. Use pure Class-A or -B, as AB will always have more distortion than either.
2. FET outputs offer freedom from some BJT problems, but in general have poorer linearity and cost more.
3. The distortion generated by a Blameless amplifier driving an 8 Ω load is almost wholly due to the effects of crossover and switching distortion. This does not hold for 4 Ω or lower loads, where third harmonic on the residual shows the presence of large-signal nonlinearity, caused by beta loss at high output currents.

References

[1] R. Mann, The Texan 20 + 20 watt stereo amplifier, Practical Wireless (May 1972) p. 48 (output stage with gain).

[2] S. Takahashi, Design and construction of high slew rate amplifiers, Preprint No. 1348 (A-4) for 60th AES Convention, 1978 (Class-B small-signal stages).

[3] M. Hawksford, Distortion correction in audio power amplifiers, JAES (January–February 1981) p. 27 (error correction).

[4] P. Walker, Current-dumping audio amplifier, Wireless World (1975) pp. 560–562.

[5] P. Blomley, New approach to Class-B, Wireless World (February 1971) p. 57. (March 1971) pp. 127–131.

[6] J. Lohstroh, M. Otala, An audio power amplifier for ultimate quality requirements, IEEE Trans. Audio and Electroacoustics (December 1973) p. 548.

[7] H. Lin, Quasi complementary transistor amplifier, Electronics (September 1956) p. 173–175 (quasi-comp).

[8] P. Baxandall, Symmetry in Class B (Letters), Wireless World (September 1969) p. 416 (Baxandall diode).

[9] I.M. Shaw, Quasi-complementary output stage modification, Wireless World (June 1969) p. 265.

[10] P. Baxandall, Private communication, 1995.

[11] P. Baxandall, in: Amos (Ed.), Radio, TV & Audio Technical Reference Book, 1977.

[12] J. Alves, Power bandwidth limitations in audio amplifiers, IEEE Trans. Broadcast and TV (March 1973) p. 79.

CHAPTER 7

More Distortion Mechanisms

Distortion 4: VAS-Loading Distortion

Distortion 4 is that which results from the loading of the voltage-amplifier stage (VAS) by the nonlinear input impedance of a Class-B output stage. This was looked at in Chapter 4 from the point of view of the VAS, where it was shown that since the VAS provides all the voltage gain, its collector impedance tends to be high. This renders it vulnerable to nonlinear loading unless it is buffered or otherwise protected.

The VAS is routinely (though usually unknowingly) linearized by applying local negative feedback via the dominant-pole Miller capacitor C_{dom}, and this is a powerful argument against any other form of compensation. If VAS distortion still adds significantly to the amplifier total, then the local open-loop gain of the VAS stage can be raised to increase the local feedback factor. The obvious method is to raise the impedance at the VAS collector, and thus the gain, by cascoding. However, if this is done without buffering the output stage loading will render the cascoding almost completely ineffective. Using a VAS buffer eliminates this problem.

As explained in Chapter 4, the VAS collector impedance, while high at LF compared with other circuit nodes, falls with frequency as soon as C_{dom} takes effect, and so Distortion 4 is usually only visible at LF. It is also often masked by the increase in output stage distortion above dominant-pole frequency $P1$ as the amount of global NFB reduces.

The fall in VAS impedance with frequency is demonstrated in Figure 7.1, obtained from the SPICE conceptual model in Chapter 4, but with values appropriate to real-life components; the input stage transconductance is set at 3 mA/V, and the VAS beta is assumed to be constant at 350. The LF impedance is basically that of the VAS collector resistance, but halves with each octave once $P1$ is reached. By 3 kHz the impedance is down to 1 kΩ, and still falling. Nevertheless, it usually remains high enough for the input impedance of a Class-B output stage to significantly degrade linearity, the actual effect being shown in Figure 7.2.

In Chapter 4, it was shown that as an alternative to cascoding, an effective means of linearizing the VAS is to add an emitter-follower within the VAS local feedback loop, increasing the local NFB factor by raising effective beta rather than the collector impedance. As well as good VAS linearity, this establishes a much lower VAS collector impedance across the audio band, and is much more resistant to Distortion 4 than the cascode version. VAS buffering is not required, so this method has a lower component count. The only drawback is a greater tendency to parasitic oscillation near negative clipping, when used with a CFP output stage.

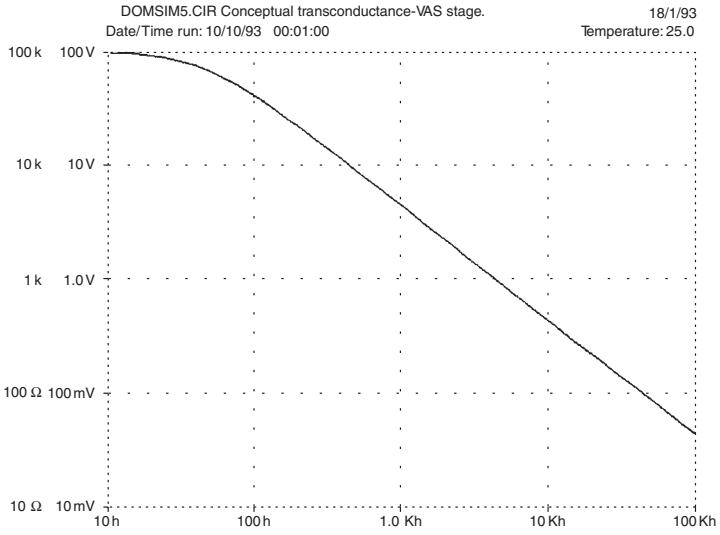


Figure 7.1: Distortion 4. The impedance at the VAS collector falls at 6 dB/octave with frequency

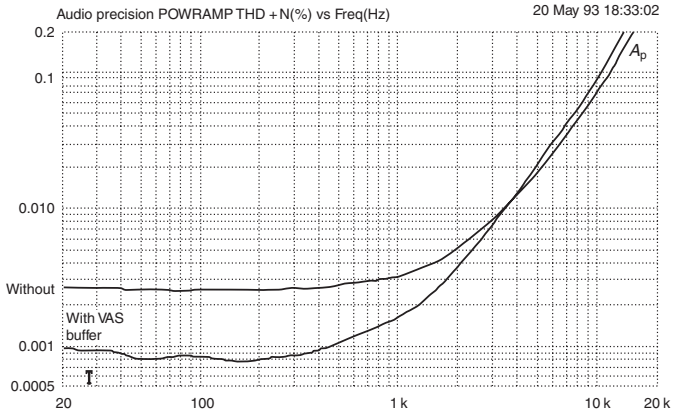


Figure 7.2: Distortion 4 in action. The lower trace shows the result of its elimination by the use of a VAS buffer

Figure 7.3 confirms that the input impedance of a conventional EF Type I output stage is highly nonlinear; the data is derived from a SPICE output stage simulation with optimal I_q. Even with an undemanding 8 Ω load, the impedance varies by 10:1 over the output voltage swing. The Type II EF output (using a shared drive emitter resistance) has a 50% higher impedance around crossover, but the variation ratio is rather greater. CFP output stages have a more complex variation that includes a precipitous drop to less than 20 kΩ around the crossover point. With all types underbiasing produces additional sharp impedance changes at crossover.

Distortion 5: Rail-Decoupling Distortion

Almost all amplifiers have some form of rail decoupling apart from the main reservoir capacitors; this is usually required to guarantee HF stability. Standard decoupling arrangements include

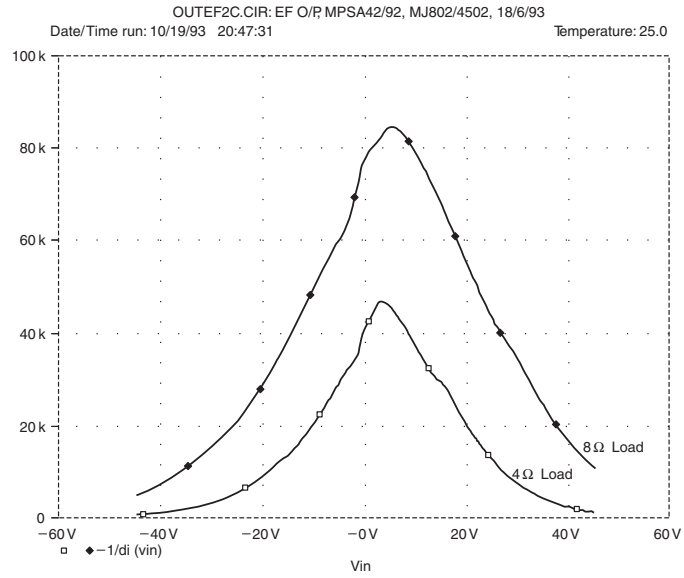


Figure 7.3: Distortion 4 and its root cause: the nonlinear input impedance of an EF Class-B output stage

small to medium-sized electrolytics (say 10–470 μF) connected between each rail and ground, and an inevitable consequence is that rail-voltage variations cause current to flow into the ground connection chosen. This is just one mechanism that defines the power-supply rejection ratio (PSRR) of an amplifier, but it is one that can seriously damage linearity.

If we use an unregulated power supply (and there are almost overwhelming reasons for using such a supply, detailed in Chapter 9) comprising transformer, bridge rectifier, and reservoir capacitors, then these rails have a non-zero AC impedance and their voltage variations will be due to amplifier load currents as well as 100 Hz ripple. In Class-B, the supply-rail currents are half-wave-rectified sine pulses with strong harmonic content, and if they contaminate the signal then distortion is badly degraded; a common route for interaction is via decoupling grounds shared with input or feedback networks, and a separate decoupler ground is usually a complete cure. This point is easy to overlook, and attempts to improve amplifier linearity by laboring on the input pair, VAS, etc. are doomed to failure unless this distortion mechanism is eliminated first. As a rule it is simply necessary to take the decoupling ground separately back to the ground star-point, as shown in Figure 7.4. (Note that the star-point A is defined on a short spur from the heavy connection joining the reservoirs; trying to use B as the star-point will introduce ripple due to the large reservoir-charging current pulses passing through it.)

Figure 7.5 shows the effect on an otherwise Blameless amplifier handling 60 W/8 Ω, with 220 μF rail-decoupling capacitors; at 1 kHz distortion has increased by more than 10 times, which is quite bad enough. However, at 20 Hz the THD has increased at least 100-fold, turning a very good amplifier into a profoundly mediocre one with one misconceived connection.

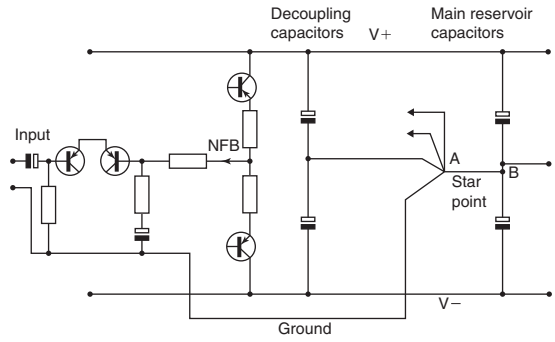


Figure 7.4: Distortion 5. The correct way to route decouple grounding to the star-point

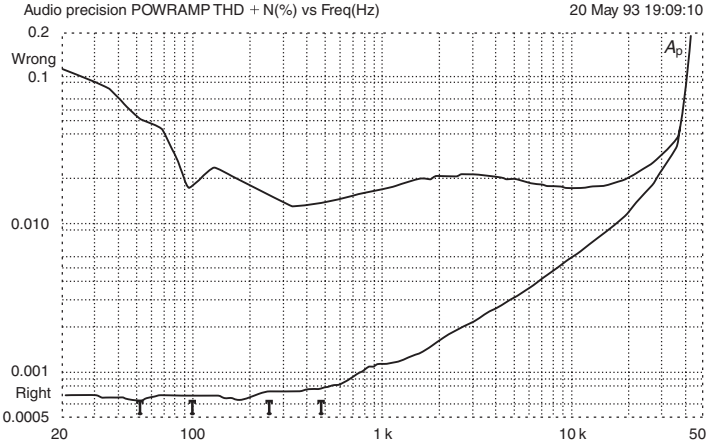


Figure 7.5: Distortion 5 in action. The upper trace was produced simply by taking the decoupler ground from the star-point and connecting it via the input ground line instead

When the waveform on the supply rails is examined, the 100 Hz ripple amplitude will usually be found to exceed the pulses due to Class-B signal current, and so some of the 'distortion' on the upper curve of the plot is actually due to ripple injection. This is hinted at by the phase crevasse at 100 Hz, where the ripple happened to partly cancel the signal at the instant of measurement. Below 100 Hz the curve rises as greater demands are made on the reservoirs, the signal voltage on the rails increases, and more distorted current is forced into the ground system.

Figure 7.6 shows a typical Distortion 5 residual, produced by deliberately connecting the negative supply-rail decoupling capacitor to the input ground instead of properly giving it its own return to the far side of the star-point. THD increased from 0.00097% to 0.008%, appearing mostly as second harmonic. Distortion 5 is usually easy to identify as it is accompanied by 100 Hz power-supply ripple; Distortions 6 and 7 introduce no extra ripple. The ripple contamination here – the two humps at the bottom – is significant and contributes to the THD reading.

As a general rule, if an amplifier is made free from ripple injection under drive conditions, demonstrated by a THD residual without ripple components, there will be no distortion from the power-supply rails, and the complications and inefficiencies of high-current rail regulators are quite unnecessary.

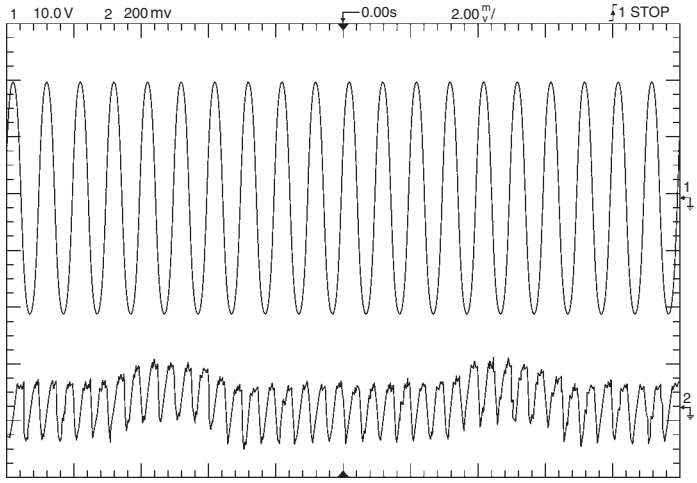


Figure 7.6: Distortion 5 revealed. Connecting the rail decoupler to input ground increases THD eight-fold from 0.00097% to 0.008%, mostly as second harmonic. 100 Hz ripple is also visible. No averaging

There has been much discussion of PSRR-induced distortion in the literature recently, e.g. Greg Ball[1]. I part company with some writers at the point where they assume a power amplifier is likely to have 25 dB PSRR, making an expensive set of HT regulators the only answer. Greg Ball also initially assumes that a power amp has the same PSRR characteristics as an op-amp, i.e. falling steadily at 6 dB/octave. There is absolutely no need for this to be so, given a little RC decoupling, and Ball states at the end of his article that 'a more elegant solution . . . is to depend on a high PSRR in the amplifier proper'. Quite so. This issue is dealt with in detail in Chapter 9.

Distortion 6: Induction Distortion

The existence of this distortion mechanism, like Distortion 5, stems directly from the Class-B nature of the output stage. With a sine input, the output hopefully carries a good sine wave, but the supply-rail currents are half-wave-rectified sine pulses, which will readily crosstalk into sensitive parts of the circuit by induction. This is very damaging to the distortion performance, as Figure 7.7 shows.

The distortion signal may intrude into the input circuitry, the feedback path, the output inductor, or even the cables to the output terminals. The result is a kind of sawtooth on the distortion residual that is very distinctive, and a large extra distortion component that rises at 6 dB/octave with frequency.

A Distortion 6 residual is displayed in Figure 7.8. The V-supply rail was routed parallel to the negative-feedback line to produce this diagram. THD is more than doubled, but is still relatively low at 0.0021%; 64 times averaging is used. Distortion 6 is easily identified if the DC supply cables are movable, for altering their run will strongly affect the quantity generated.

This inductive effect appears to have been first publicized by Cherry[2], in a paper that deserves more attention. The effect has, however, been recognized and avoided by some practitioners for

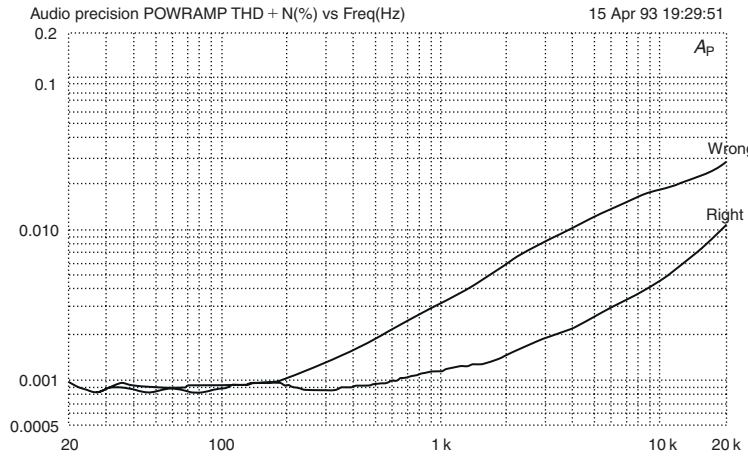


Figure 7.7: Distortion 6 exposed. The upper trace shows the effects of Class-B rail induction into signal circuitry

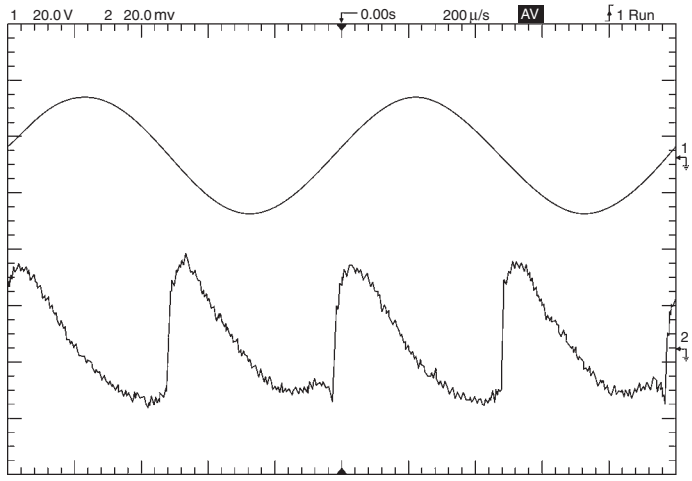


Figure 7.8: Distortion 6. Induction of half-wave signal from the negative supply rail into the NFB line increases THD to 0.0021%. Averaged 64 times

many years[3]. However, having examined many power amplifiers with varying degrees of virtue, I feel that this effect, being apparently unknown to most designers, is probably the most widespread cause of unnecessary distortion.

The contribution of Distortion 6 can be reduced below the measurement threshold by taking sufficient care over the layout of supply-rail cabling relative to signal leads, and avoiding loops that will induce or pick up magnetic fields. I wish I could give precise rules for layout that would guarantee freedom from the problem, but each amplifier has its own physical layout, and the cabling topology has to take this into account. However, here are some guidelines.

Firstly, implement rigorous minimization of loop area in the input and feedback circuitry, keeping each signal line as close to its ground return as possible. Secondly, minimize the ability of the

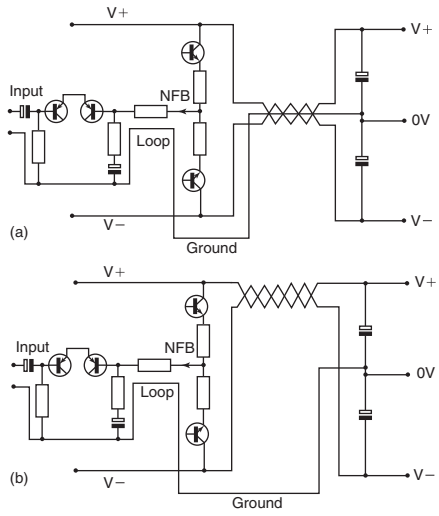


Figure 7.9: Distortion 6. Countermeasures against the induction of distortion from the supply rails. (b) is usually the more effective

supply wiring to establish magnetic fields in the first place. Thirdly, put as much distance between these two areas as you can. Fresh air beats shielding on price every time.

Figure 7.9 shows one straightforward approach to solving the problem; the supply and ground wires are tightly twisted together to reduce radiation. In practice this does not seem to be effective, for reasons that are not wholly clear, but seem to involve the difficulty of ensuring exactly equal coupling between three twisted conductors. In Figure 7.9, the supply rails are twisted together but kept well away from the ground return; this will allow field generation, but if the currents in the two rails butt together to make a nice sine wave at the output, then they should do the same when the magnetic fields from each rail sum. There is an obvious risk of interchannel crosstalk if this approach is used in a stereo amplifier, but it does deal effectively with the induced distortion problem in some layouts.

It is difficult to overemphasize the importance of keeping a good lookout for this form of distortion when evaluating prototype amplifiers; trying to remove it by any method other than correcting the physical layout is quite futile, and I cannot help wondering how many unhappy man-hours have been spent trying to do just that. The sawtooth-like distortion residual is a dead give-away; another simple test is to move around the power-supply cables if it is possible to do so, and see if the distortion residual varies. The output inductor is inherently fairly sensitive to unwanted magnetic fields, and you may have to change its orientation to avoid picking them up. In a recent case, an experimental amplifier was giving an excessive 0.0075% THD at 10 kHz (25 W/8 Ω) and squashing the output coil flat with an authoritative thumb reduced this at once to a fairly Blameless 0.0026%.

In cases of difficulty with this problem, a powerful tool is a small search coil connected to an audio analyzer input (a spare output inductor works very well for this); it can be moved around to look for unsuspected current paths carrying half-wave-rectified sine pulses.

This distortion mechanism does not of course trouble Class-A amplifiers.

Distortion 7: NFB Take-Off Point Distortion

It has become a tired old truism that negative feedback is a powerful technique, and like all such, must be used with care if you are to avoid tweeter-frying HF instability.

However, there is another and much more subtle trap in applying global NFB. Class-B output stages are a maelstrom of high-amplitude half-wave-rectified currents, and if the feedback take-off point is in slightly the wrong place, these currents contaminate the feedback signal, making it an inaccurate representation of the output voltage, and hence introducing distortion; Figure 7.10 shows the problem. At the current levels in question, all wires and PCB tracks must be treated as resistances, and it follows that point C is not at the same potential as point D whenever TR1 conducts. If feedback is taken from D, then a clean signal will be established here, but the signal at output point C will have a half-wave-rectified sine-wave added to it, due to the resistance C–D. The actual output will be distorted but the feedback loop will do nothing about it as it does not know about the error.

Figure 7.11 shows the practical result for an amplifier driving 100 W into 8 Ω, with the extra distortion interestingly shadowing the original curve as it rises with frequency. The resistive path C–D that did the damage was a mere 6 mm length of heavy-gauge wire-wound resistor lead.

Figure 7.12 shows a THD residual for Distortion 7, introduced by deliberately taking the NFB from the wrong point. The THD rose from 0.00097% to 0.0027%, simply because the NFB feed was taken from the wrong end of the leg of one of the output emitter resistors Re. Note this is not the wrong side of the resistor, or the distortion would have been gross, but a mere 10 mm along a very thick resistor leg from the actual output junction point.

Of the distortions that afflict generic Class-B power amplifiers, Distortions 5–7 all look rather similar when they appear in the THD residual, which is perhaps not surprising since all result from adding half-wave disturbances to the signal.

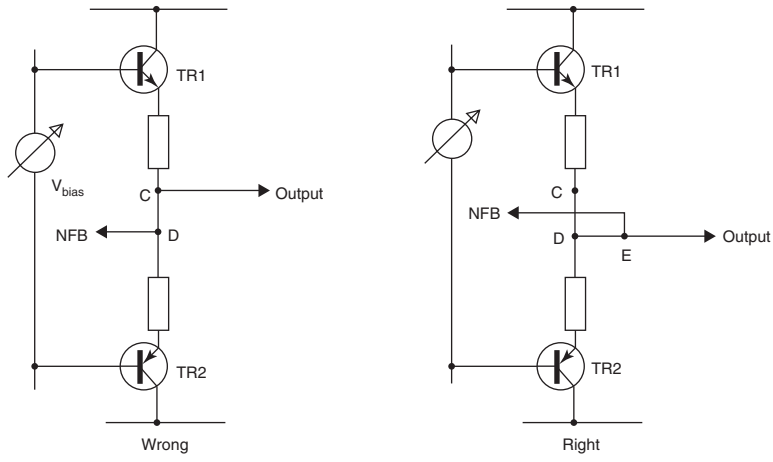


Figure 7.10: Distortion 7. Wrong and right ways of arranging the critical negative-feedback take-off point

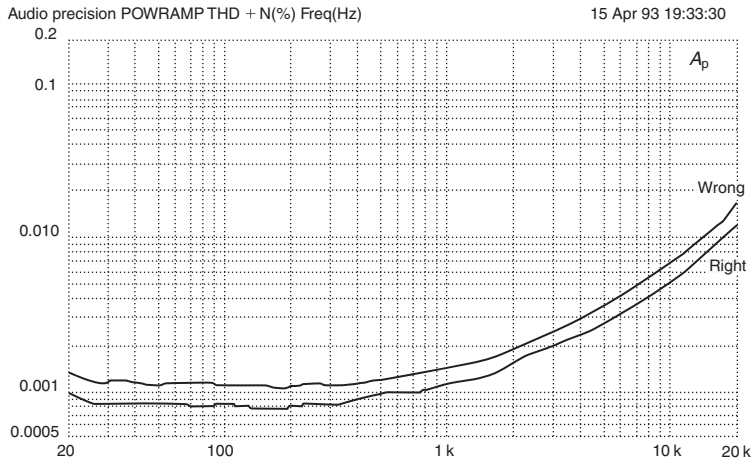


Figure 7.11: Distortion 7 at work. The upper (WRONG) trace shows the result of a mere 6 mm of heavy-gauge wire between the output and the feedback point

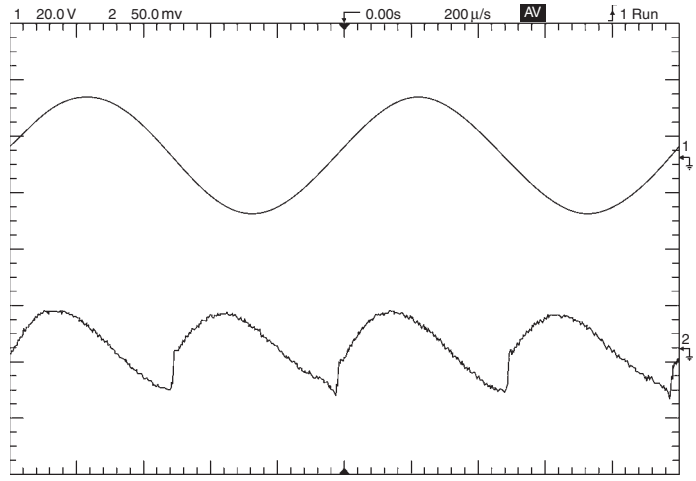


Figure 7.12: Distortion 7, caused by choosing an NFB take-off point inside the Class-B output stage rather than on the output line itself. THD is increased from 0.00097% to 0.0027%, by taking the NFB from the wrong end of 10 mm of very thick resistor leg. Averaged 64 times

To eliminate this distortion is easy, once you are alert to the danger. Taking the NFB feed from D is not advisable as D is not a mathematical point, but has a physical extent, inside which the current distribution is unknown. Point E on the output line is much better, as the half-wave currents do not flow through this arm of the circuit.

Distortion 8: Capacitor Distortion

When I wrote the original series on amplifier distortion[4], I listed seven types of distortion that defined an amplifier's linearity. The number has since grown, and Distortion 8 refers to capacitor distortion. This has nothing to do with subjectivist hypotheses about mysterious non-measurable effects; this

phenomenon is all too real, though for some reason it seems to be almost unknown – or at any rate not talked about – amongst audio designers. Clearly this is the distortion that dare not speak its name.

It is, however, a sad fact that both electrolytic and non-electrolytic capacitors generate distortion whenever they are used in such a fashion that a significant AC voltage develops across them.

Standard aluminum electrolytics create distortion when they are used for coupling and DC blocking, while driving a significant resistive load. Figure 7.13 is the test circuit; Figure 7.14 shows the resulting distortion for a 47 μF, 25 V capacitor driving +20 dBm (7.75 V rms) into a 680 Ω load, while Figure 7.15 shows how the associated LF roll-off has barely begun. The distortion is a mixture of second and third harmonic, and rises rapidly as frequency falls, at something between 12 and 18 dB/octave.

The great danger of this mechanism is that serious distortion begins while the response roll-off is barely detectable; here the THD reaches 0.01% when the response has only fallen by 0.2 dB. The

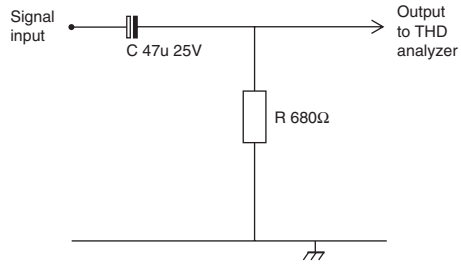


Figure 7.13: A very simple circuit to demonstrate electrolytic capacitor distortion. Measurable distortion begins at 100 Hz

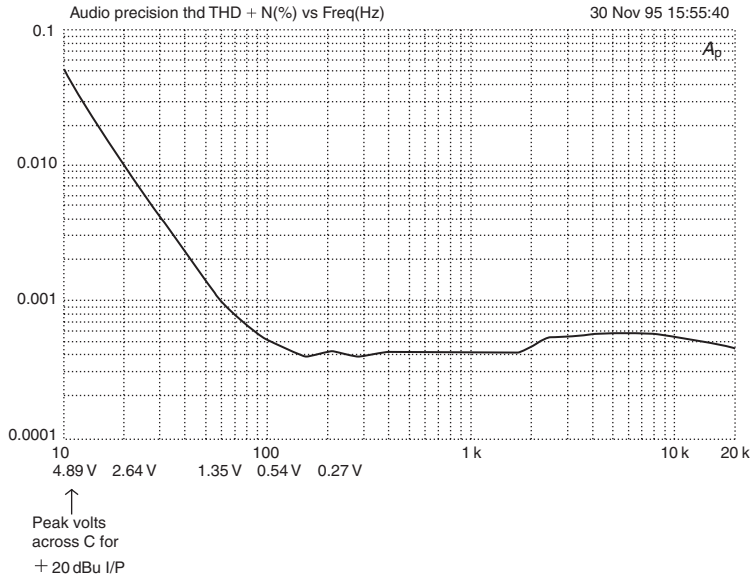


Figure 7.14: Capacitor distortion versus frequency, showing the rapid rise in THD once the distortion threshold is reached

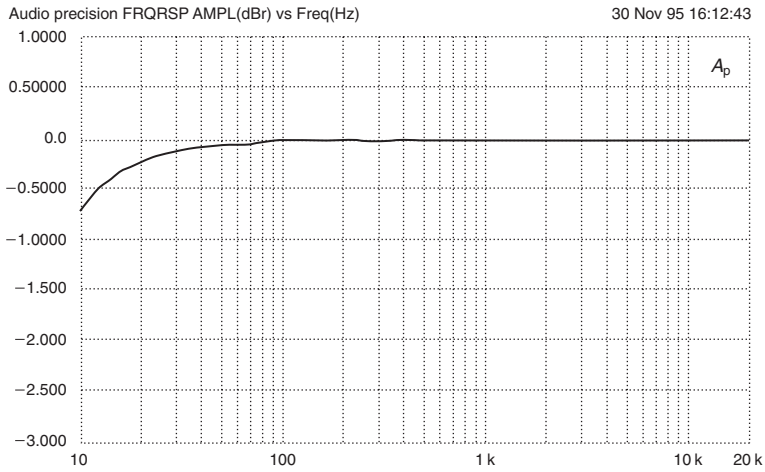


Figure 7.15: The small amount of LF roll-off associated with the distortion rise in Figure 7.14

voltage across the capacitor is 2.6 V peak, and this voltage is a better warning of danger than the degree of roll-off.

Further tests showed that the distortion roughly triples as the applied voltage doubles; this factor seems to vary somewhat between different capacitor rated voltages.

The mechanism by which capacitors generate this distortion is unclear. Dielectric absorption appears to be ruled out as this is invariably (and therefore presumably successfully) modeled by adding linear components, in the shape of resistors and capacitors, to the basic capacitor model. Reverse-biasing is not the problem, for capacitors DC biased by up to +15 V show slightly increased, not reduced, distortion. Nonpolarized electrolytics show the same effect but at a much greater AC voltage, typically giving the same distortion at one-tenth the frequency of a conventional capacitor with the same time-constant; the cost and size of these components generally rules out their use to combat this effect. Usually the best solution is simply to keep increasing the capacitor value until the LF distortion rise disappears off the left of the THD graph. Negligible roll-off in the audio band is not a sufficient criterion.

Electrolytics are therefore best reserved for DC filtering, and for signal coupling where the AC voltage across them will be negligible. If a coupling capacitor does have AC voltage across it, and drives the usual resistive load, then it must be acting as a high-pass filter. This is never good design practice, because electrolytics have large tolerances and make inaccurate filters; it is now clear they generate distortion as well.

It is therefore most undesirable to define the lower bandwidth limit simply by relying on the high-pass action of electrolytics and circuit resistances; it should be done with a non-electrolytic capacitor, made as large as possible economically in order to reduce the value of the associated resistance and so keep down circuit impedances, thus minimizing the danger of noise and crosstalk.

Capacitor distortion in power amplifiers is most likely to occur in the feedback network blocking capacitor, assuming it is a DC-coupled amplifier; if it is AC-coupled the output capacitor may

generate serious distortion, as described in Chapter 2. The input blocking capacitor usually feeds a high impedance, but the feedback arm must have the lowest possible resistances to minimize both noise and DC offset. The feedback capacitor therefore tends to be relatively large, and if it is not quite large enough the THD plot of the amplifier will show the characteristic kick-up at the LF end. An example of this is dealt with in detail in Chapter 4.

It is common for amplifiers to show a rise in distortion at the LF end, but there is no reason why this should ever occur. Capacitor distortion is usually the reason, but Distortion 5 (rail-decoupling distortion) can also contribute. These two mechanisms can be distinguished because Distortion 5 typically rises by only 6 dB/octave as frequency decreases, rather than the 12–18 dB/octave of capacitor distortion.

Amplifiers with AC-coupled outputs are now fairly rare, and one reason may be that distortion in the output capacitor is a major problem, occurring in the mid-band as well as at LF. The reason for this mid-band problem is not obvious; probably it is due to the much higher levels of current passing through the output capacitor activating distortion mechanisms that are not otherwise visible. If an amplifier is driving 50 W into an 8 Ω load, and has a feedback resistor of 2k2 (which is probably about as low as is likely) then the peak current through the output capacitor will be 3.5 A, while the peak current through the feedback capacitor at the bottom of the feedback network is only 12.7 mA (see the section on AC-coupled amplifiers in Chapter 2 for more details of output capacitor distortion).

Non-electrolytic capacitors of middling value (say 10–470 nF) also generate distortion when operated with significant signal voltages across them, but this typically occurs when they are used to realize time-constants in filter and equalization circuits, and this is not relevant to power amplifier design.

However, the linearity of non-electrolytic capacitors of small value (say 10–220 pF) is very much of interest to the designer, as they are used for compensation and RF filtering purposes. This is of particular relevance to the capacitor C_{dom} used for dominant-pole Miller compensation, which stabilizes the overall feedback loop by converting global feedback into local feedback around the VAS transistor. It has the full output voltage of the amplifier impressed across it, and it is therefore vital that it is a completely linear component.

Its size, usually around 100 pF, means that it will almost certainly be a ceramic type. It is essential that a type with C0G or NP0 dielectric is used. These have the lowest capacitance/temperature dependence (NP0 stands for negative–positive zero) and the lowest losses, but for our purposes the important point is that they have the lowest capacitance/voltage coefficients. It is generally known that ceramics with X7R dielectrics have large capacitance/voltage coefficients and are quite unsuitable for any application where linearity matters; their value is that they pack a lot of capacitance into a small space, which makes them useful for decoupling jobs. In general, all that is required is to specify a C0G or NP0 type; but this can go wrong, as I will now relate.

Figure 7.16 shows the THD plot for a Blameless amplifier delivering 180 W into 8 Ω. It had the usual mirrored input pair, an emitter-follower-enhanced VAS, and the EF output configuration with three pairs of output devices in parallel. This multiple-output approach can give excellent

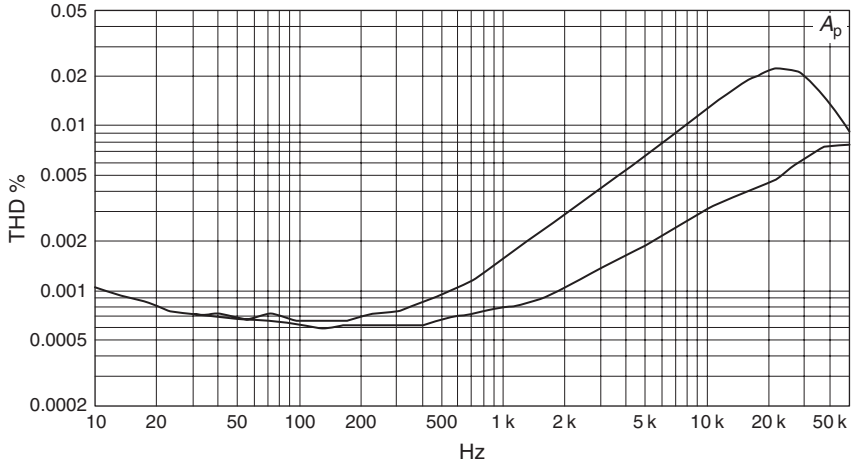


Figure 7.16: The upper trace shows the excess distortion generated by a substandard NP0 Miller compensation capacitor. The lower trace is the result with a good component (180 W into 8 Ω)

distortion performance, as shown in the lower trace. What, however, we actually got was the upper trace; between 1 and 20 kHz there is about three times more distortion than there should be. Given that the product was on the very threshold of mass production, there was alarm, consternation, and worse. The culprit was quickly shown to be the 100 pF dominant-pole Miller compensation capacitor, a Chinese-sourced component that in theory, but not in practice, was an NP0 part. Replacing it with an identically specified part from a more reputable Chinese manufacturer cured the problem at once, yielding the expected lower trace in Figure 7.16.

There are several interesting points here; the extra distortion was fairly pure second harmonic and, as the plot shows, the amount is rising at a steady 6 dB/octave. 'Bad NP0 distortion' is here shown in action for the first time, I believe. Note that the capacitor was not an X7R type by mistake – if that had been the case the distortion would have been gross. What we had was an attempt at making an NP0 capacitor that failed.

Distortion 9: Magnetic Distortion

This arises when a signal at amplifier output level is passed through a ferromagnetic conductor. Ferromagnetic materials have a nonlinear relationship between the current passing through them and the magnetic flux it creates, and this induces voltages that add distortion to the signal. The effect has been found in some types of output relays where the signal being switched passes through the soft-iron frame that makes up part of the magnetic circuit. That particular manifestation is dealt with in detail in Chapter 17, where output relays are examined.

The problem has also been experienced with loudspeaker terminals. The terminal pair in question was a classy-looking Chinese item with all its metal parts gold-plated, and had proved wholly satisfactory at the prototype stage. Once again the product involved was trembling on the brink of mass production, and once again the pre-production batch showed more distortion than expected. The THD residual showed third-harmonic distortion that had certainly not been there before. Some

rapid investigation revealed the hitherto unknown concept of nonlinear loudspeaker terminals. The metal parts of the terminals appeared to be made of gold-plated brass (as they were in all the prototype samples) but were actually gold-plated steel, which is of course a cheaper material – brass has copper in it, and copper is expensive. Although the amplifier output currents were only passing through about 10 mm of steel (the current went through that length twice, on go and return), the nonlinear magnetic effects were sufficient to increase the output distortion from 0.00120% to 0.00227% at 100 W into 8 Ω at 1 kHz. In other words distortion nearly doubled. It is, however, highly likely that if the offending terminals had been used with a non-Blameless amplifier having rather more distortion of its own the extra nonlinearity would have gone completely unnoticed, and I can only presume that this was what the manufacturer hoped and expected. Parts incorrectly made from steel can of course be readily detected by the application of a small magnet.

It might be thought that ferromagnetic distortion might be most likely to affect the only part of the signal path that is deliberately inductive – the output coils. Amplifier chassis are very often made of steel, and the output coil is usually close to a large ferromagnetic component in the shape of the output relay.

While it is certainly good practice to keep the output coil away from ferrous metals as far as practical, in fact there is very little to worry about; the effect of adjacent steel or iron parts on the coil is not as large as you might think. To put it into perspective, a little experiment was performed. A Blameless power amplifier driving 115 W into an 8 Ω load was yielding 0.00080% THD at 1 kHz. Inserting a steel screwdriver shaft 6 mm in diameter into the output coil, which consisted of 10 turns of heavy copper wire 24 mm in diameter, only degraded the THD to 0.0094%, which while clearly undesirable is not exactly a dramatic change. When the screwdriver shaft was replaced with a complete small-signal relay tucked wholly inside the coil, the THD only worsened slightly to 0.0011%. The effect across the audio frequency band is seen in Figure 7.17; the worst effect is at about 7 kHz, where

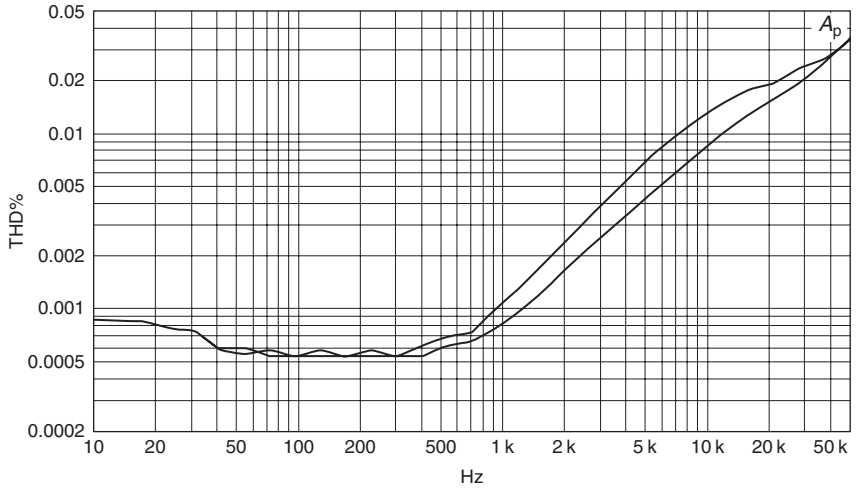


Figure 7.17: The not very dramatic effect of placing a complete relay inside the output coil (115 W into 8 Ω)

0.006% is degraded to 0.010%. These tests put gross amounts of ferromagnetic material right inside the coil, so it is safe to assume (and, I hasten to add, further experiments prove) that metal chassis sections some centimeters away from the coil are going to have no detectable effect.

Some further tests showed that mounting an output relay (which contains substantially more ferrous metal than the small-signal relay alluded to above) so that the end of the output coil was in contact with the plastic relay casing caused no detectable degradation from 0.00080% THD under the same conditions. However, not everybody seems prepared to believe this, and there is a wide consensus that it 'looks wrong' so it's best avoided if humanly possible.

Other output coil issues – such as the crosstalk between two coils in a stereo amplifier – are dealt with in Chapter 8.

I don't want you to think that I am prejudiced against Chinese electronic components. I have used them extensively, and providing you take due care with suppliers there are few difficulties. The worst problems I have had with components – none of them Chinese – were thus:

1. Defective electrolytic capacitors that generated their own DC voltage. Short them out and it would disappear; remove the short and it would slowly return, like dielectric absorption only much, much worse. The result: big mixing consoles where every switch clicked when operated – not good.
2. Batches of IC power amplifiers that died after a few weeks of normal domestic use. This caused mayhem. Every possible design-based reason was investigated, without result, and it took the manufacturer (the very well-known, apparently thoroughly reputable manufacturer) something like nine months to admit that they had made a large batch of thoroughly defective ICs.
3. IC voltage regulators with nonfunctional overload protection. The application was a power supply that could quite easily be short-circuited by the user, so it did matter. The manufacturer's response was not to offer to replace the parts, but to fly in a team of four people from another European country to convince us that we didn't *really* need overload protection after all. I need hardly say we remained unconvinced, and years after the event I'm still wondering about the mental state of whoever decided that was an appropriate reaction to the problem.

I won't tell you the manufacturers involved, as this might turn historical technical problems into contemporary legal ones, but the capacitors came from Japan and the ICs both came from very big Western semiconductor manufacturers.

Distortion 10: Input Current Distortion

This distortion is caused when an amplifier input is driven from a significant source impedance. The input current taken by the amplifier is nonlinear, even if the output of the amplifier is distortion free, and the resulting voltage drop in the source impedance introduces distortion.

This mechanism is dealt with in detail in Chapter 4, as it relates closely to the design of the amplifier input stage.

Distortion 11: Premature Overload Protection

The most common method of overload protection of a power amplifier is the use of VI limiters that shunt signal current away from the inputs to the output stage. In their most common form these come into operation relatively gradually as their threshold is exceeded, and start introducing distortion into the signal long before they close it down entirely. This problem is made more serious because the simplest and most used VI limiter circuits show significant temperature sensitivity, coming into action sooner as they warm up in the internal environment of the amplifier. It is therefore vital to design an adequate safety margin into the output stage so that the VI limiters need never be near activation during normal use. This issue is examined more closely in Chapter 17.

Design Example – A 50 W Class-B Amplifier

Figure 7.18 shows a design example of a Class-B amplifier, intended for domestic hi-fi applications. Despite its relatively conventional appearance, the circuit parameters selected give much better than a conventional distortion performance; this is potentially a Blameless design, but only if due care is given to wiring topology and physical layout will this be achieved.

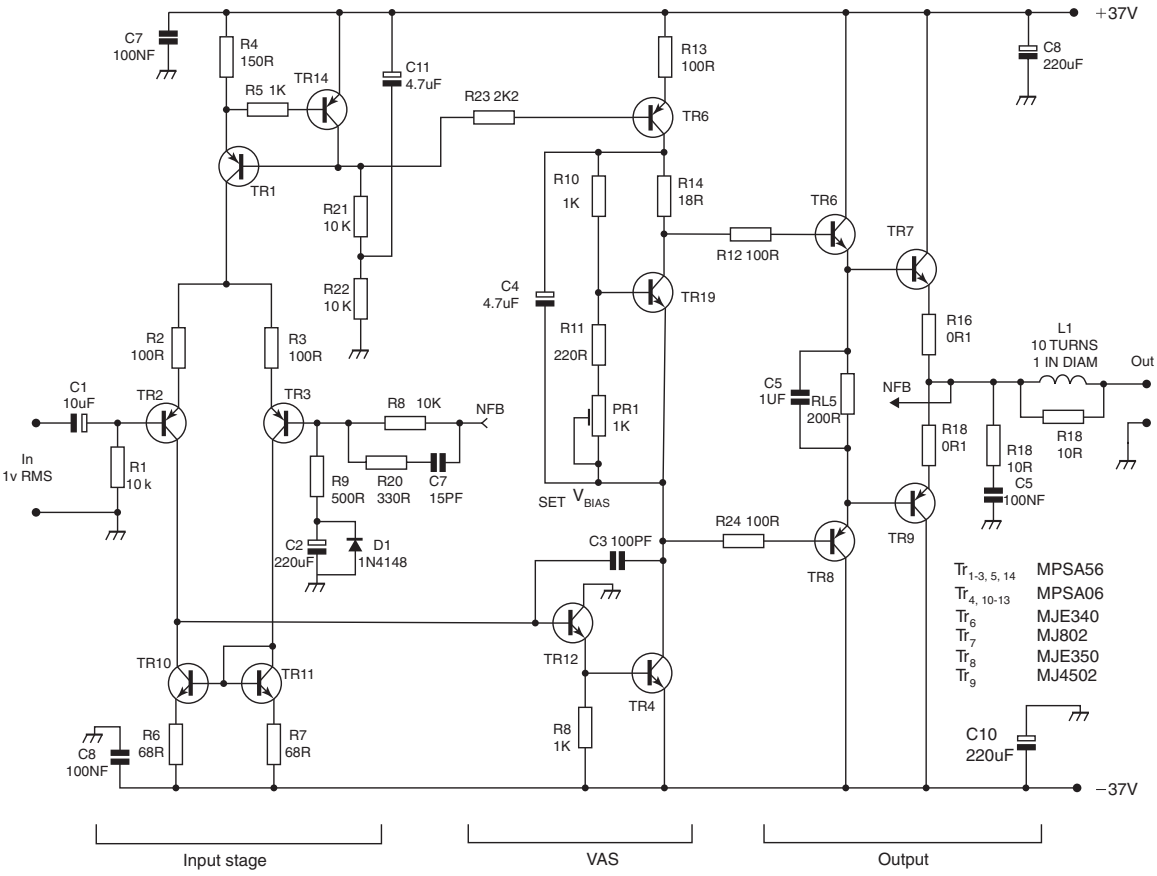


Figure 7.18: 50 W Class-B amplifier circuit diagram. Transistor numbers correspond with the generic amplifier in Chapter 3

With the supply voltages and values shown it gives 50 W into 8 Ω, for 1 V rms input. In earlier chapters, I have used the word *Blameless* to describe amplifiers in which all distortion mechanisms, except the apparently unavoidable ones due to Class-B, have been rendered negligible. This circuit has the potential to be Blameless (as do we all), but achieving this depends on care in cabling and layout. It does not aim to be a cookbook project; for example, overcurrent and DC-offset protection are omitted.

In Chapter 14, output topologies are examined, and the conclusion drawn that power-FETs are disappointingly expensive, inefficient, and nonlinear. Therefore, bipolars it is. The best BJT configurations were the emitter-follower (EF) Type II, with least output switch-off distortion, and the complementary feedback pair (CFP), giving the best basic linearity.

The output configuration chosen is the emitter-follower Type II, which has the advantage of reducing switch-off nonlinearities (Distortion 3c) due to the action of R15 in reverse-biasing the output base–emitter junctions as they turn off. A possible disadvantage is that quiescent stability might be worse than for the CFP output topology, as there is no local feedback loop to servo out V_{be} variations in the hot output devices. Domestic ambient temperature changes will be small, so that adequate quiescent stability can be attained by suitable heat-sinking and thermal compensation.

A global NFB factor of 30 dB at 20 kHz was chosen, which should give generous HF stability margins. The input stage (current source TR1 and differential pair TR2, TR3) is heavily degenerated by R2, R3 to delay the onset of third-harmonic Distortion 1, and to assist this the contribution of transistor internal r_e variation is minimized by using the unusually high tail current of 4 mA. TR11, TR12 form a degenerated current-mirror that enforces accurate balance of the TR2, TR3 collector currents, preventing the generation of second-harmonic distortion. Tail source TR1, TR14 has a basic PSRR 10 dB better than the usual two-diode version, though this is academic when C11 is fitted.

Input resistor R1 and feedback arm R8 are made equal and kept as low as possible, consistent with a reasonably high-input impedance, so that base-current mismatch caused by beta variations will give a minimal DC offset; this does not affect TR2–TR3 V_{be} mismatches, which appear directly at the output, but these are much smaller than the effects of I_b. Even if TR2, TR3 are high-voltage types with low beta, the output offset should be within ±50 mV, which should be quite adequate, and eliminates balance presets and DC servos. A low value for R8 also gives a low value for R9, which improves the noise performance.

The value of C2 shown (220 μF) gives an LF roll-off with R9 that is −3 dB at 1.4 Hz. The aim is not an unreasonably extended sub-bass response, but to prevent an LF rise in distortion due to capacitor nonlinearity; 100 μF degraded the THD at 10 Hz from less than 0.0006% to 0.0011%, and I judge this unacceptable aesthetically if not audibly. Band-limiting should be done earlier, with non-electrolytic capacitors. Protection diode D1 prevents damage to C2 if the amplifier suffers a fault that makes it saturate negatively; it looks unlikely but causes no measurable distortion[5]. C7 provides some stabilizing phase advance and limits the closed-loop bandwidth; R20 prevents it upsetting TR3.

The VAS stage is enhanced by an emitter-follower inside the Miller compensation loop, so that the local NFB that linearizes the VAS is increased by augmenting total VAS beta, rather than by increasing the collector impedance by cascoding. This extra local NFB effectively eliminates Distortion 2 (VAS nonlinearity). Further study has shown that thus increasing VAS beta gives a much lower collector impedance than a cascode stage, due to the greater local feedback, and so a

VAS buffer to eliminate Distortion 4 (loading of VAS collector by the nonlinear input impedance of the output stage) appears unnecessary. C_{dom} is relatively high at 100 pF, to swamp transistor internal capacitances and circuit strays, and make the design predictable. The slew rate calculates as 40 V/μs. The VAS collector load is a standard current source, to avoid the uncertainties of bootstrapping.

Since almost all the THD from a Blameless amplifier is crossover, keeping the quiescent conditions optimal is essential. Quiescent stability requires the bias generator to cancel out the V_{be} variations of four junctions in series: those of two drivers and of two output devices. Bias generator TR8 is the standard V_{be}-multiplier, modified to make its voltage more stable against variations in the current through it. These occur because the biasing of TR5 does not completely reject rail variations; its output current also drifts initially due to heating and changes in TR5 V_{be}. Keeping Class-B quiescent stable is hard enough at the best of times, and so it makes sense to keep these extra factors out of the equation. The basic V_{be}-multiplier has an incremental resistance of about 20 Ω; in other words its voltage changes by 1 mV for a 50 μA drift in standing current. Adding R14 converts this to a gently peaking characteristic that can be made perfectly flat at one chosen current (see Figure 7.19). Setting R14 to 22 Ω makes the voltage peak at 6 mA, and standing current now must deviate from this value by more than 500 μA for a 1 mV bias change. The R14 value needs to be altered if TR15 is run at a different current; for example, 16 Ω makes the voltage peak at 8 mA instead. If TO-3 outputs are used the bias generator should be in contact with the top or can of one of the output devices, rather than the heat-sink, as this is the fastest and least attenuated source for thermal feedback.

The output stage is a standard double emitter-follower apart from the connection of R15 between the driver emitters without connection to the output rail. This gives quicker and cleaner switch-off of the outputs at high frequencies; switch-off distortion may significantly degrade THD from 10 kHz upwards, dependent on transistor type. Speed-up capacitor C4 noticeably improves the switch-off action, though I should say at this point that its use has been questioned because of the possibility of unhelpful charges building up on it during asymmetrical clipping. C6, R18 form the Zobel network (sometimes confusingly called a Boucherot cell) while L1, damped by R19, isolates the amplifier from load capacitance.

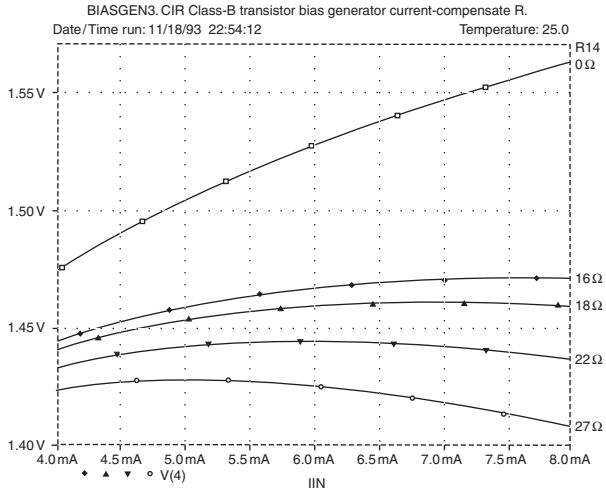


Figure 7.19: SPICE plot of the voltage-peaking behavior of a current-compensated bias generator

Figure 7.20 shows the 50 W/8 Ω distortion performance: about 0.001% at 1 kHz and 0.006% at 10 kHz (see Table 7.1). The measurement bandwidth makes a big difference to the appearance, because what little distortion is present is crossover-derived, and so high order. It rises at 6 dB/octave, at the rate the feedback factor falls, and it is instructive to watch the crossover glitches emerging from the noise, like Grendel from the marsh, as the test frequency increases above 1 kHz. There is no precipitous THD rise in the ultrasonic region.

The zigzags on the LF end of the plot are measurement artefacts, apparently caused by the Audio Precision system trying to winkle out distortion from visually pure white noise. Below 700 Hz the residual was pure noise with a level equivalent to approximately 0.0006% (yes, three zeros) at 30 kHz bandwidth; the actual THD here must be microscopic. This performance can only be obtained if all seven of the distortion mechanisms are properly addressed; Distortions 1–4 are determined by the circuit design, but the remaining three depend critically on physical layout and grounding topology.

It is hard to beat a well-gilded lily, and so Figure 7.21 shows the startling results of applying two-pole compensation to the basic amplifier; C3 remains 100 pF, while CP2 was 220 pF and Rp 1 k (see Figure 8.1d). The extra global NFB does its work extremely well, the 10 kHz THD dropping to 0.0015%,

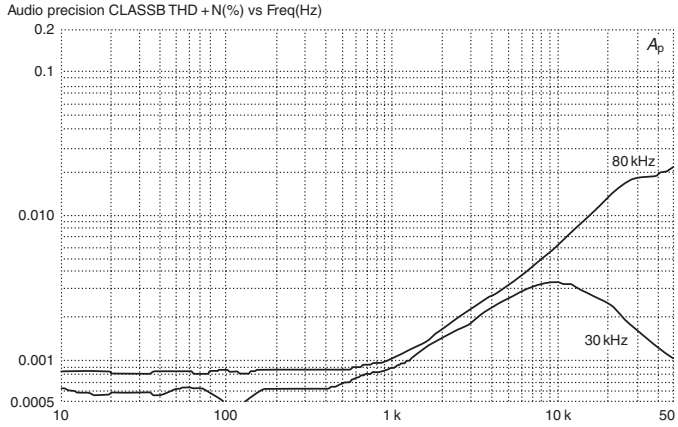


Figure 7.20: Class-B amplifier: THD performance at 50 W/8 Ω; measurement bandwidths 30 and 80 kHz

Table 7.1: Class-B amplifier performance

Power output	50 W rms into 8 Ω
Distortion	Below 0.0006% at 1 kHz and 50 W/8 Ω
	Below 0.006% at 10 kHz
Slew rate	Approximately 35 V/μs
Noise	−91 dBu at the output
EIN	−117 dBu (referred to input)
Frequency response	+0, −0.5 dB over 20 Hz–20 kHz

Most of the AP plots in this book were obtained from an amplifier similar to that in Figure 7.18, though with higher supply rails and so greater power capability. The main differences were the use of a cascode-VAS with a buffer, and a CFP output to minimize distracting quiescent variations. Measurements at powers above 100 W/8 Ω used a version with two paralleled output devices.

while the 1 kHz figure can only be guessed at. There were no unusual signs of instability, but as always unusual compensation schemes require careful testing. It does appear that a Blameless amplifier with two-pole compensation takes us close to the long-sought goal of the Distortionless amplifier.

The basic Blameless EF amplifier was experimentally rebuilt with three alternative output stages: the simple quasi-complementary, the quasi-Baxandall, and the CFP. The results for both single- and two-pole compensation are shown in Figures 7.22–7.24. The simple quasi-complementary generates more crossover distortion, as expected, and the quasi-Baxandall version is not a lot better, probably due to remaining asymmetries around the crossover region. The CFP gives even lower distortion than the original EF II output, with Figure 7.21 showing only the result for single-pole compensation; in this case the improvement with two-pole was marginal and the trace is omitted for clarity.

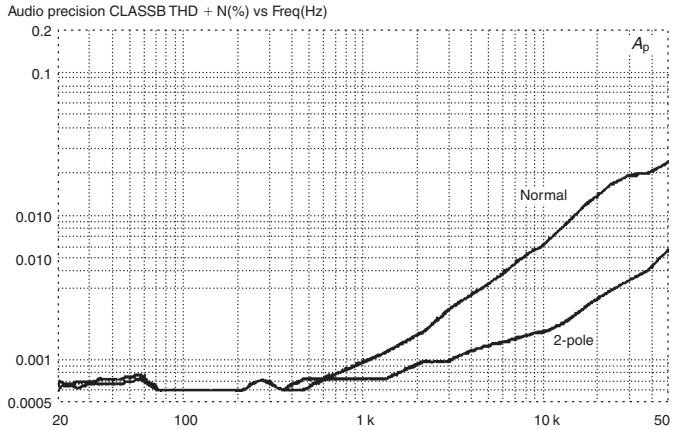


Figure 7.21: The dramatic THD improvement obtained by converting the Class-B amplifier to two-pole compensation

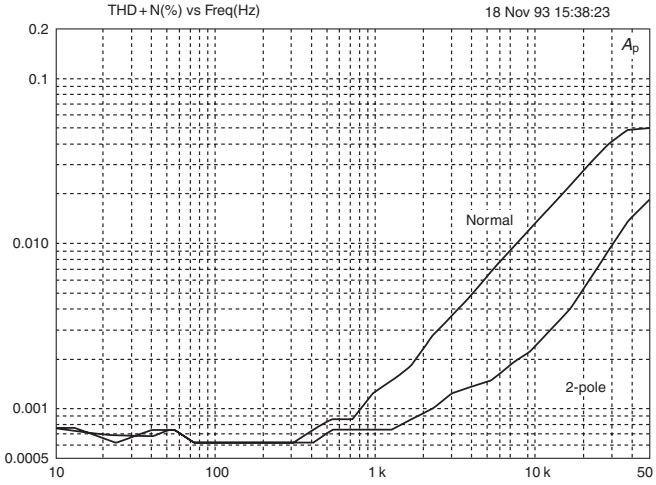


Figure 7.22: Class-B amplifier with simple quasi-complementary output. The lower trace is for two-pole compensation

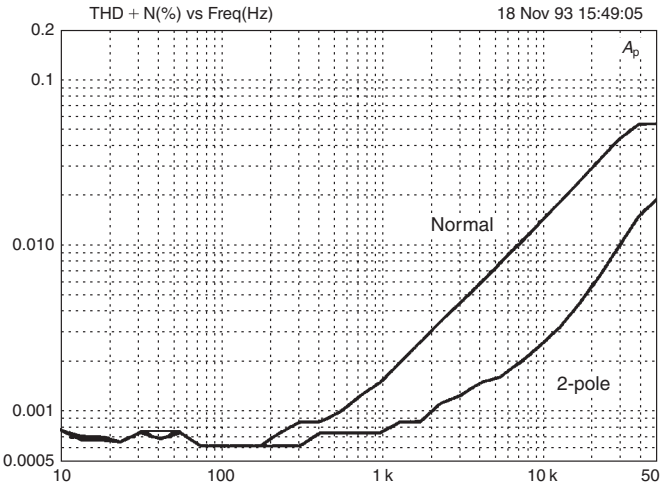


Figure 7.23: Class-B amplifier with quasi-complementary plus Baxandall diode output. The lower trace is the two-pole case

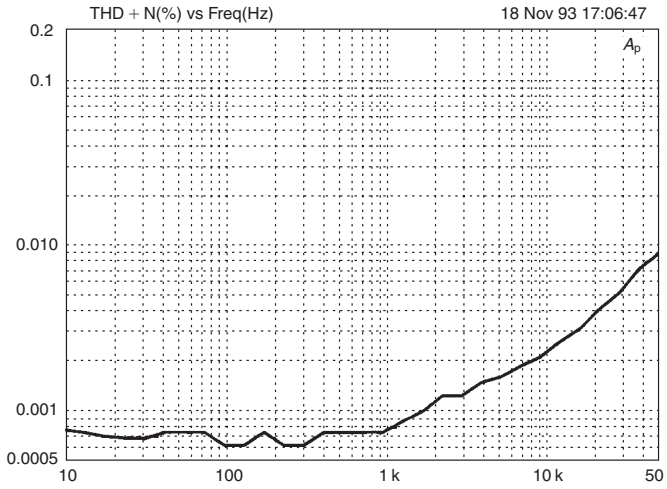


Figure 7.24: Class-B amplifier with complementary feedback pair (CFP) output stage. Normal compensation only

References

[1] G. Ball, Distorting power supplies, Electronics & Wireless World (December 1990) p. 1084.

[2] E. Cherry, A new distortion mechanism in Class-B amplifiers, JAES (May 1981) p. 327.

[3] P. Baxandall, Private communication, 1995.

[4] D. Self, Distortion in power amplifiers, Series in Electronics & Wireless World (August 1993 to March 1994).

[5] D. Self, An advanced preamplifier, Wireless World (November 1976) p. 43.

CHAPTER 8

Compensation, Slew Rate, and Stability

Frequency Compensation in General

The compensation of an amplifier is the tailoring of its open-loop gain and phase characteristics so that it is dependably stable when the global feedback loop is closed.

It must be said straight away that 'compensation' is a thoroughly misleading word to describe the subject of this chapter. It implies that one problematic influence is being balanced out by another opposing force, when in fact it means the process of tailoring the open-loop gain and phase of an amplifier so that it is satisfactorily stable when the global feedback loop is closed. The derivation of the word is historical, going back to the days when all servomechanisms were mechanical, and usually included an impressive Watt governor pirouetting on top of the machinery.

An amplifier requires compensation because its basic open-loop gain is still high at frequencies where the internal phase shifts are reaching 180°. This turns negative feedback into positive at high frequencies, and causes oscillation, which in audio amplifiers can be very destructive. The way to prevent this is to ensure that the loop gain falls to below unity before the phase shift reaches 180°; oscillation therefore cannot develop. Compensation is therefore vital simply because it makes the amplifier stable; there are other considerations, however, because the way in which the compensation is applied has a major effect on the closed-loop distortion behavior.

The distortion performance of an amplifier is determined not only by open-loop linearity, but also the negative-feedback factor applied when the loop is closed; in most practical circumstances doubling the NFB factor halves the distortion. So far I have assumed that open-loop gain falls at 6 dB/octave due to a single dominant pole, with the amount of NFB permissible at HF being set by the demands of HF stability. We have seen that this results in the distortion from a Blameless amplifier consisting almost entirely of crossover artefacts, because of their high order and hence high frequency. Audio amplifiers using more advanced compensation are rather rare. However, certain techniques do exist, and are described later.

This book concentrates on conventional topologies, because even apparently commonplace circuitry has proven to have little-known aspects, and to be capable of remarkable linearity. This means the classical three-stage architecture circuit with transconductance input, transimpedance VAS, and unity-gain output stage. Negative feedback is applied globally, but is smoothly transferred by C_{dom} to be local solely to the VAS as frequency increases. Other configurations are possible; a two-stage amplifier with transconductance input and unity-gain output is an intriguing possibility – this is common in CMOS op-amps – but is probably ill-suited to power-amp

impedances. Four-stage amplifiers are described in Chapter 2; the best known is probably that by Otala[1], a four-stage amplifier with a low open-loop gain of 52 dB (due to the dogged use of local feedback) and only 20 dB of global feedback. Most of this chapter relates only to the conventional three-stage structure.

Dominant-Pole Compensation

Dominant-pole compensation is the simplest kind, though its action is subtle. Simply take the lowest pole to hand ($P1$) and make it dominant, i.e. so much lower in frequency than the next pole $P2$ that the total loop gain (i.e. the open-loop gain as reduced by the attenuation in the feedback network) falls below unity before enough phase shift accumulates to cause HF oscillation. With a single pole, the gain must fall at 6 dB/octave, corresponding to a constant 90° phase shift. Thus the phase margin will be 90°, giving good stability.

Figure 8.1a shows the traditional Miller method of creating a dominant pole. The collector pole of TR4 is lowered by adding the external Miller capacitance C_{dom} to that which unavoidably exists as the internal C_{bc} of the VAS transistor. However, there are some other beneficial effects; C_{dom} causes *pole-splitting,* in which the pole at TR2 collector is pushed up in frequency as $P1$ is moved down – most desirable for stability. Simultaneously the local NFB through C_{dom} linearizes the VAS.

Assuming that input stage transconductance is set to a plausible 5 mA/V, and stability considerations set the maximal 20 kHz open-loop gain to 50 dB, then from Equations 3.1–3.3 in Chapter 3, C_{dom} must be 125 pF. This is more than enough to swamp the internal capacitances of the VAS transistor, and is a practical real-life value.

The peak current that flows in and out of this capacitor, for an output of 20 V rms at 20 kHz, is 447 μA. Since the input stage must sink C_{dom} current while the VAS collector load sources it, and likewise the input stage must source it while the VAS sinks it, there are four possible ways in which slew rate may be limited by inadequate current capacity; if the input stage is properly designed

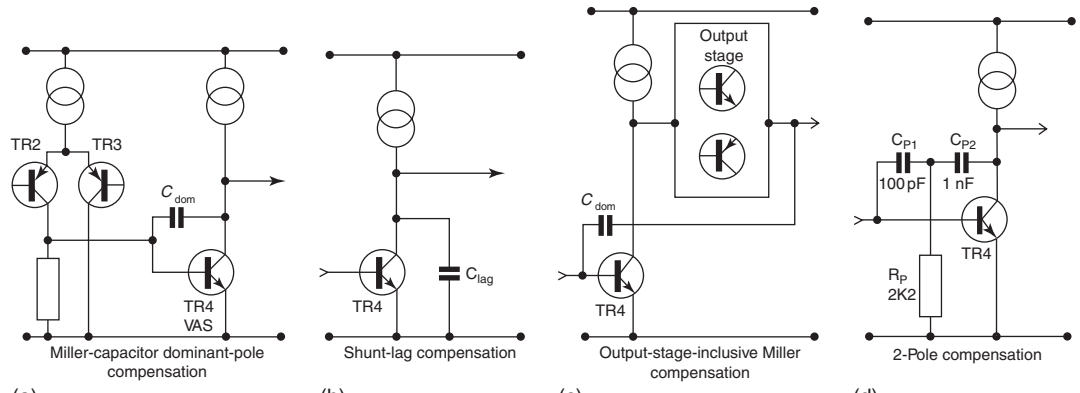


Figure 8.1: (a) The traditional Miller method of making a dominant pole. (b) Shunt compensation shows a much less satisfactory method – the addition of capacitance to ground from the VAS collector. (c) Inclusive Miller compensation. (d) Two-pole compensation

then the usual limiting factor is VAS current-sourcing. In this example a peak current of less than 0.5 mA should be easy to deal with, and the maximum frequency for unslewed output will be comfortably above 20 kHz.

Lag Compensation

Figure 8.1b shows a much less satisfactory method of compensation – the addition of capacitance to ground from the VAS collector. This is usually called shunt or lag compensation, but is sometimes called parallel compensation. As Peter Baxandall[2] aptly put it, 'The technique is in all respects suboptimal'. We have already seen in Chapter 5 that loading the VAS collector resistively to ground is a very poor option for reducing LF open-loop gain, and a similar argument shows that capacitive loading to ground for compensation purposes is an even worse idea. To reduce open-loop gain at 20 kHz to 50 dB as before, the shunt capacitor C_{lag} must be 43.6 nF, which is a whole different order of things from 125 pF. The current in and out of C_{lag} at 20 V rms, 20 kHz is 155 mA peak, which is going to require some serious electronics to provide it. This important result is yielded by simple calculation, confirmed by SPICE simulation. The input stage no longer constrains the slew-rate limits, which now depend entirely on the VAS.

A VAS working under these conditions will have poor linearity. The I_c variations in the VAS, caused by the heavy extra loading, produce more distortion and there is no local NFB through a Miller capacitor to correct it. To make matters worse, the dominant pole $P1$ will probably need to be set to a lower frequency than for the Miller case, to maintain the same stability margins, as there is now no pole-splitting action to increase the frequency of the pole at the input stage collector. Hence C_{lag} may have to be even larger than 43 nF, requiring yet higher peak currents. The bad effect of adding much small shunt capacitances than this to a VAS collector is illustrated below in Figure 8.7.

Takahashi et al.[3] have produced a fascinating paper on this approach, showing one way of generating the enormous compensation currents required for good slew rates. The only thing missing is an explanation of why shunt compensation was chosen in the first place.

The use of a *small* capacitor (say 33 pF) from the VAS collector to ground is often useful in suppressing output stage parasitics; this has nothing to do with the amplifier compensation. This handy fix is discussed in detail later in this chapter.

Including the Output Stage: Output-Inclusive Miller Compensation

Miller dominant-pole compensation elegantly solves several problems at once, and the decision to adopt it is simple. However, the question of whether to include the output stage in the Miller feedback loop is less easy. Such inclusion (see Figure 8.1c) presents the alluring possibility that local feedback could linearize both the VAS and the output stage, with just the input stage left out in the cold as frequency rises and global NFB falls. This idea is most attractive as it would greatly increase the total feedback available to linearize a distortive Class-B output stage.

There is certainly some truth in this, as I have shown[4], where applying C_{dom} around the output as well as the VAS reduced the peak (not rms) 1 kHz THD from 0.05% to 0.02%. However, I must

say that the output stage was deliberately underbiased to induce crossover spikes, because with optimal bias the improvement, although real, was too small to be either convincing or worthwhile. A vital point is that this demonstration used a model amplifier with TO92 'output' transistors, because in my experience the technique just does not work well with real power bipolars, tending to intractable HF oscillation. There is evidence that inclusive compensation, when it can be made stable, is much less effective at dealing with ordinary crossover distortion than with the spikes produced by deliberate underbiasing.

The use of local NFB to linearize the VAS demands a tight loop with minimal extra phase shift beyond that inherent in the C_{dom} dominant pole. It is permissible to insert a cascode or a small-signal emitter-follower into this local loop, but a slow output stage with all sorts of complexities in its frequency response seems to be pushing luck too far; the output stage poles are now included in the loop, which loses its dependable HF stability. Bob Widlar[5] stated that output stage behavior must be well controlled up to 100 MHz for the technique to be reliable; this would appear to be virtually impossible for discrete power stages with varying loads.

However, I have recently done some work that shows it is possible to partly include the output stage in the Miller loop, and it does give significant advantages in normal operation. This is a rather exciting development, but still under study and I am not ready to disclose it here. The circuit of Figure 8.1c was not used.

Other Forms of Inclusive Compensation

Other forms of inclusive compensation have been put forward, which have as their purpose the inclusion of the input stage rather than the output stage in the Miller loop.

The form of compensation shown in Figure 8.2 has frequently been advocated, notably by the late John Linsley-Hood[6]. It was his contention that this configuration prevented input-device overload (i.e. slew-limiting) on fast transients. There certainly seems to be no need to include the input stage to reduce its distortion, as it has been conclusively shown in Chapter 4 that this can be reduced as much as required by straightforward circuit techniques. If another stage can be safely incorporated in the Miller loop, it makes much more sense for it to be the output stage, which makes most of the distortion. My experience with this configuration was that it was unstable, and any advantages it might have had were therefore irrelevant. I corresponded with JLH on this matter in 1994, hoping to find exactly how it was supposed to work, but we were unable to reach any consensus on the matter.

A very similar compensation configuration was put forward by Marshall Leach[7], where he described it as a form of feedforward compensation. I must admit I find this description puzzling, but the paper is certainly worth reading.

Two-Pole Compensation

Two-pole compensation is well known as a technique for squeezing the best performance from an op-amp[8,9], but it has rarely been applied to power amplifiers; the only example I know is found in Widlar[5]. An extra HF time-constant is inserted in the C_{dom} path, giving an open-loop gain curve

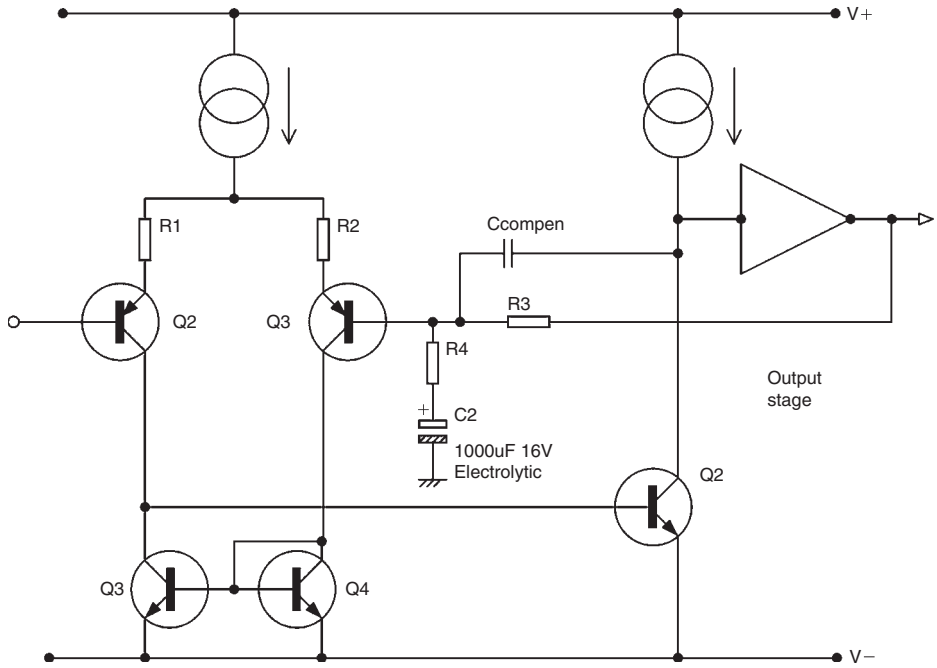


Figure 8.2: Returning the compensation capacitor to the inverting input instead of the VAS base

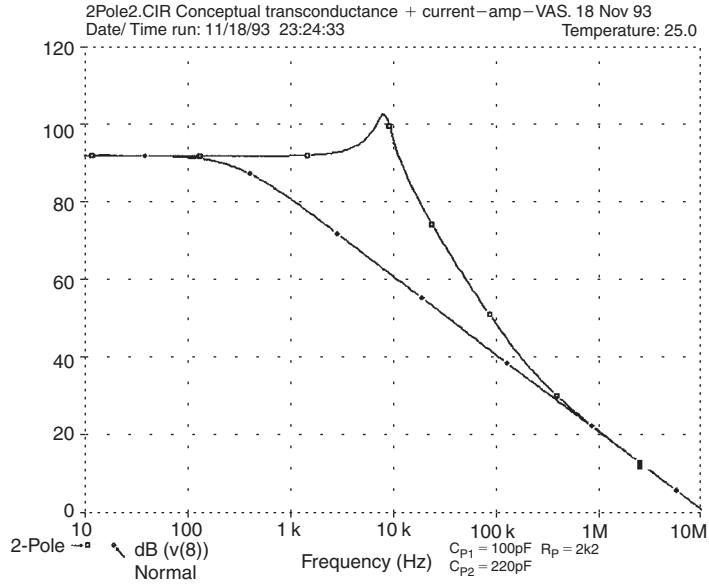


Figure 8.3: The open-loop gain plot for two-pole compensation with realistic component values

that initially falls at almost 12 dB/octave, but which gradually reverts to 6 dB/octave as frequency continues to increase, as in Figure 8.3. This reversion is arranged to happen well before the unity loop-gain line is reached, and so stability should be the same as for the conventional dominant-pole scheme, but with increased negative feedback over part of the operational frequency range.

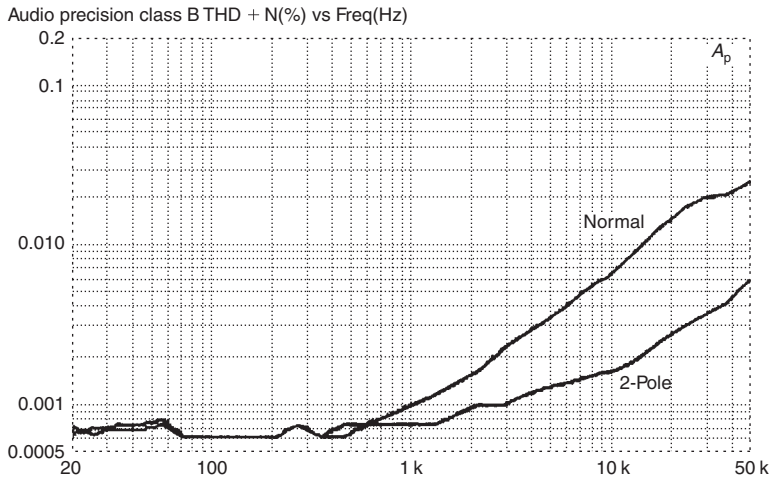


Figure 8.4: Distortion reduction with two-pole compensation

The faster gain roll-off means that the maximum amount of feedback can be maintained up to a higher frequency. There is no measurable mid-band peak in the closed-loop response.

It is right to feel nervous about any maneuver that increases the NFB factor; power amplifiers face varying conditions and it is difficult to be sure that a design will always be stable under all circumstances. This makes designers rather conservative about compensation, and I approached this technique with some trepidation. However, results were excellent with no obvious reduction in stability (see Figure 8.4 for the happy result of applying this technique to the Class-B amplifier of Figure 8.5).

The simplest way to implement two-pole compensation is shown in Figure 8.1d, with typical values. C_{p1} should have the same value as it would for stable single-pole compensation, and C_{p2} should be at least twice as big; R_p is usually in the region 1 k–10 k. At intermediate frequencies C_{p2} has an impedance comparable with R_p, and the resulting extra time-constant causes the local feedback around the VAS to increase more rapidly with frequency, reducing the open-loop gain at almost 12 dB/octave. At HF the impedance of R_p is high compared with C_{p2}, the gain slope asymptotes back to 6 dB/octave, and then operation is the same as conventional dominant pole, with C_{dom} equal to the series capacitance combination. So long as the slope returns to 6 dB/octave before the unity loop-gain crossing occurs, there seems no obvious reason why the Nyquist stability should be impaired. Figure 8.3 shows a simulated two-pole open-loop gain plot for realistic component values; C_{p2} should be at least twice C_{p1} so the gain falls back to the 6 dB/octave line before the unity loop-gain line is crossed. The potential feedback factor has been increased by more than 20 dB from 3 kHz to 30 kHz, a region where THD tends to increase due to falling NFB. The open-loop gain peak at 8 kHz looks extremely dubious, but I have so far failed to detect any resulting ill-effects in the closed-loop behavior. Peter Baxandall[10] pointed out to me, and demonstrated mathematically, that the open-loop gain peak has no repercussions at all in the closed-loop gain plot. It is not a question of a resonance being heavily suppressed – it simply does not exist.

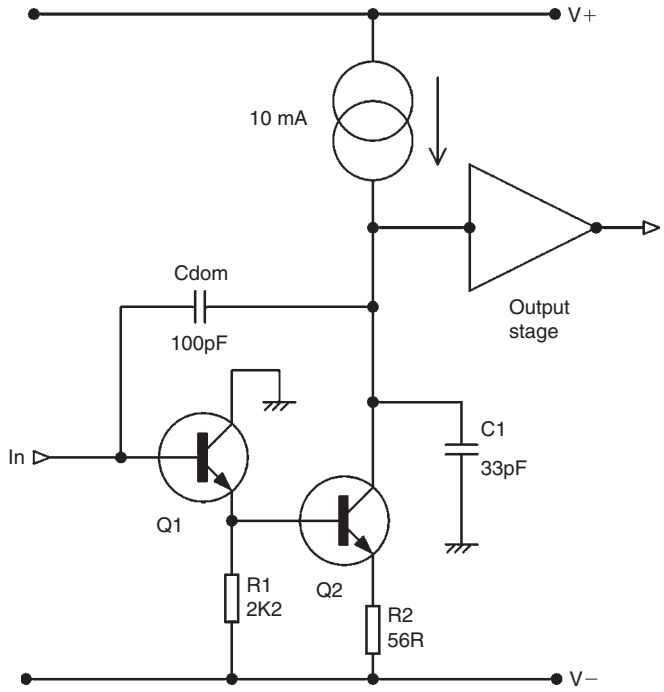


Figure 8.5: Adding a small shunt capacitor C1 from the VAS collector to ground can be very helpful in obtaining dependable HF stability

There is, however, a snag to the approach shown here, which reduces the linearity improvement. Two-pole compensation may decrease open-loop linearity at the same time as it raises the feedback factor that strives to correct it. At HF, C_{p2} has low impedance and allows R_p to directly load the VAS collector to ground, which could worsen VAS linearity.

However, if C_{p2} and R_p are correctly proportioned the overall reduction in distortion is dramatic and extremely valuable. When two-pole compensation was added to the amplifier circuit shown in Figure 8.9, the crossover glitches on the THD residual almost disappeared, being partially replaced by low-level second harmonic which almost certainly results from VAS loading. The positive slew rate will also be slightly reduced.

After the publication of the original two-pole material in *Wireless World*, Peter Baxandall pointed out to me[10] that exactly the same response can be obtained by making C_{p1} larger than C_{p2}, providing the value of the series combination remains the same; I have confirmed this myself in SPICE. This has the great advantage that the VAS loading is reduced while everything else stays the same. In tests on an experimental amplifier based on the Load-Invariant design but not fully optimized, I started off with C_{p1} at 100 pF and C_{p2} at 1000 pF; the THD at 10 kHz was 0.0043% (25 W/8 Ω). Swapping the capacitors dropped it to 0.00317%, due to reduced VAS loading.

Two-pole compensation looks like an attractive technique, as it can be simply applied to an existing design by adding two inexpensive components; adding/removing R_p allows instant comparison between the two kinds of compensation. Be warned that if an amplifier is prone to HF parasitics

then two-pole compensation may worsen them; as always, if you decide to use unconventional compensation then you need to allow plenty of time for assessing HF stability.

Stability and VAS-Collector-to-Ground Capacitance

In the search for HF stability, a capacitor from the VAS collector to ground can be a very present help in time of trouble (see C1 in Figure 8.5). I will be the first to admit that this is a strictly empirical modification that looks a bit suspect, but the fact is that it works. It is especially useful if there are stability issues with capacitive loads. Note that the shunt capacitor is very small in value, often 10 pF; the largest value I have so far used is 33 pF. The value is not critical.

The basic function of this component is the suppression of parasitic oscillation in the output stage. The exact theoretical mechanism is not fully known, but the key point appears to be that the impedance seen at the VAS collector is prevented from becoming inductive at very high frequencies.

This expedient is *not* the same as lag compensation, which was roundly condemned earlier in this chapter. C1 does not replace the dominant-pole capacitor, which remains at its original value, and C1 is orders of magnitude smaller in value than a typical lag capacitor. Obviously if C1 is too big there may be effects on both linearity and maximum slew rate; if it needs to be larger than say 47 pF, there may be something wrong with the output stage or output network design.

Figure 8.6 shows that small values of shunt capacitor C1 can be added without significantly affecting a good distortion performance. The amplifier used was one of my more recent designs (2008).

With this sort of measure, it is always worth enquiring as to how far it can be taken before things go wrong; this will help you avoid picking a value that initially appears OK but is actually poised on the brink of disaster. The results of this enquiry are shown in Figure 8.7.

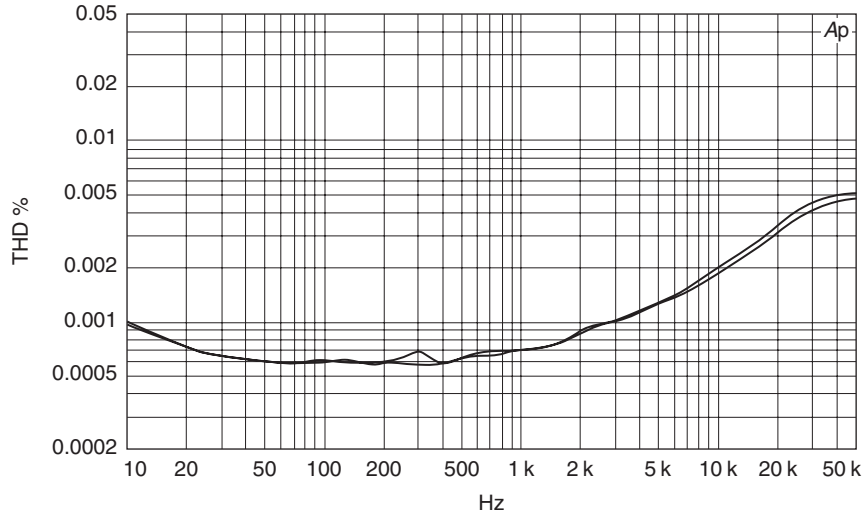


Figure 8.6: Adding C1 need not compromise a good distortion performance. Lower trace C1 = 10 pF; upper trace C1 = 37 pF. Power 180 W into 8 Ω

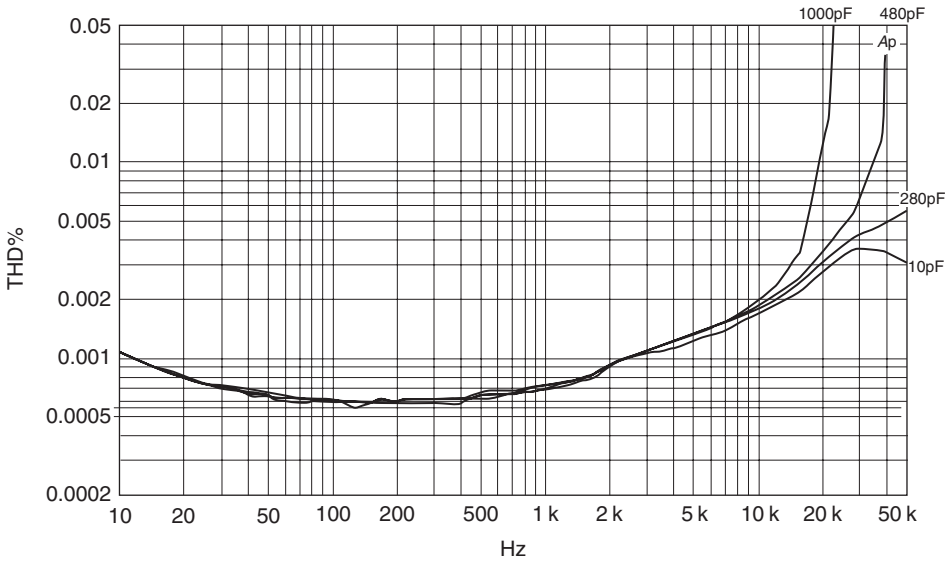


Figure 8.7: What happens to amplifier distortion when the over-large values of C1 shown are used. Power 180 W into 8 Ω

With C1 = 280 pF, the HF distortion is slightly worse, but only above 20 kHz where its effect is less important; 480 pF causes a sharp increase of distortion at 40 kHz, characteristic of slew-rate-limiting, but linearity is not much worse from 10 to 20 kHz. With 1000 pF, twice as large, we predictably get slew-limiting at 20 kHz, i.e. half the frequency. Output power was 180 W into 8 Ω (38 V rms), a large voltage swing on the VAS being chosen to bring out the possibility of slew-limiting.

I think these results confirm that small values of shunt capacitor can be used to improve stability without affecting the distortion performance.

Nested Feedback Loops

Nested feedback is a way to apply more NFB around the output stage without increasing the global feedback factor. The output has an extra voltage gain stage bolted on, and a local feedback loop is closed around these two stages. This NFB around the composite output block reduces output stage distortion and increases frequency response, to make it safe to include in the global NFB loop.

Suppose that block A1 (Figure 8.8a) is a Distortionless small-signal amplifier providing all the open-loop gain and so including the dominant pole. A3 is a unity-gain output stage with its own main pole at 1 MHz and distortion of 1% under given conditions; this 1 MHz pole puts a firm limit on the amount of global NFB that can be safely applied. Figure 8.8b shows a nested feedback version; an extra gain block A2 has been added, with local feedback around the output stage. A2 has the modest gain of 20 dB so there is a good chance of stability when this loop is closed to bring the gain of A3 + A2 back to unity. A2 now experiences 20 dB of NFB, bringing the distortion down to 0.1%, and raising the main pole to 10 MHz, which should allow the application of 20 dB more global NFB around the overall loop that includes A1. We have thus decreased the distortion

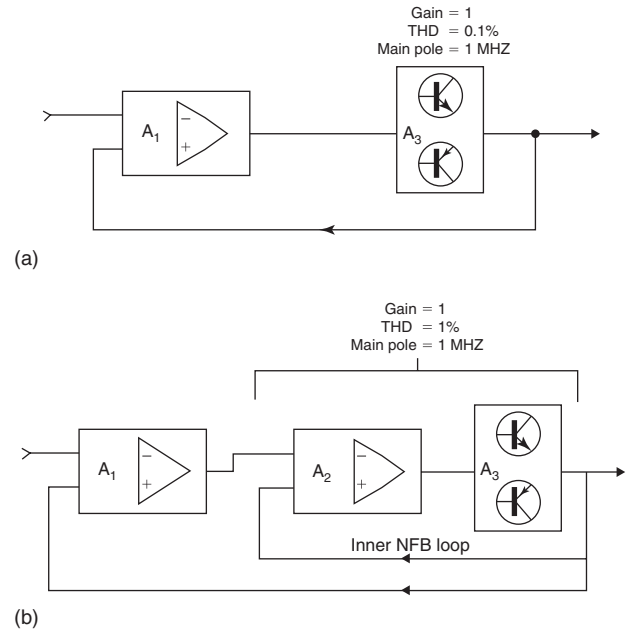


Figure 8.8: (a) Normal single-loop global negative feedback. (b) Nested feedback

that exists before global NFB is applied, and simultaneously increased the amount of NFB that can be safely used, promising that the final linearity could be very good indeed. For another theoretical example, see Pernici et al.[11].

Real-life examples of this technique in power amps are not easy to find, but it is widely used in op-amps. Many of us were long puzzled by the way that the much-loved 5534 maintained such low THD up to high frequencies. Contemplation of its enigmatic entrails appears to reveal a three-gain-stage design with an inner Miller loop around the third stage, and an outer Miller loop around the second and third stages; global NFB is then applied externally around the whole lot. Nested Miller compensation has reached its apotheosis in CMOS op-amps – the present record appears[11] to be three nested Miller loops plus the global NFB; do not try this one at home. More details on the theory of nested feedback can be found in Scott and Spears[12]; the treatment is wholly mathematical.

Output Networks

The usual output networks for a power amplifier are shown in Figure 8.10, with typical values. They comprise a shunt Zobel network, for stability into inductive loads, and a series output inductor/damping resistor for stability into capacitive loads.

Amplifier Output Impedance

The main effect of output impedance is usually thought to be its effect on damping factor. This is wrong, as explained in Chapter 1. Despite this demonstration of its irrelevance, I will refer to damping factor here, to show how an apparently impressive figure dwindles as more parts of the speaker–cable system are included.

Figure 8.10 shows a simplified amplifier with Zobel network and series output inductor, plus simple models of the connecting cable and speaker load. The output impedance of a solid-state amplifier is very low if even a modest amount of global NFB is used. I measured a Blameless Class-B amplifier similar to Figure 8.9 with the usual NFB factor of 29 dB at 20 kHz, increasing at 6 dB/octave as frequency falls. Figure 8.11 shows the output impedance at point B before the output inductor, measured by injecting a 10 mA signal current into the output via a 600 Ω resistance.

The low-frequency output impedance is approximately 9 mΩ (an 8 Ω damping factor of 890). To put this into perspective, one meter of thick 32/02 equipment cable (32 strands of 0.2 mm diameter) has a resistance of 16.9 mΩ. The internal cabling resistance in an amplifier can equal or exceed the output impedance of the amplifier itself at LF.

Output impedance rises at 6 dB/octave above 3 kHz, as global NFB falls off, reaching 36 mΩ at 20 kHz. The 3 kHz break frequency does not correspond with the amplifier dominant-pole frequency, which is much lower at around 10 Hz.

The closed-loop output impedance of any amplifier is set by the open-loop output impedance and the negative feedback factor. The output impedance is not simply the output impedance of the output stage alone, because the latter is driven from the VAS, so there is a significant and frequency-varying source impedance at point A in Figure 8.10.

When the standard EF and CFP stages are driven from a zero-impedance source, in both cases the raw output impedance is in the region of 150–180 mΩ. This assumes the emitter resistors Re are 0.1 Ω. Increasing Re to 0.22 Ω increases output impedance to the range 230–280 mΩ, showing that these resistors in fact make up most of the output impedance. The output devices and drivers have little influence.

If the average open-loop output impedance is 200 mΩ, and the NFB factor at 20 kHz is 29 dB, or 28 times, we would expect a closed-loop output impedance of approximately 200/28, which is 7 mΩ. Since it is actually about 33 mΩ at this frequency, there is clearly more going on than simple theory implies. In a real amplifier the output stage is not driven from a zero impedance, but a fairly high one that falls proportionally with frequency; for my Blameless Class-B design it falls from 3 kΩ at 1 kHz to about 220 Ω at 20 kHz. A 220 Ω source impedance produces an open-loop output impedance of about 1 Ω, which when reduced by a factor of 28 when global feedback is applied, gives 35 mΩ. This is close to the value measured at 20 kHz at point B in Figure 8.10.

All of these measured closed-loop output impedances are very low compared with the other impedances in the amp–cable–speaker system. It would appear they can in most cases be ignored.

The Blameless amplifier design has an output inductor of approximately 6 μH; the aim is absolutely guaranteed stability into all capacitive loads, and the inductance is therefore at the high end of the permissible range. This is limited by the HF roll-off into the lowest load resistance to be driven. This substantial component comprises 20 turns of 1.5-mm-diameter copper wire, wound in a 1-inch-diameter coil, and has a DC resistance of 19 mΩ. This small extra resistance raises the flat section of the impedance plot to 24 mΩ, and in fact dominates the LF output impedance as measured at the amplifier terminals (point C). It also sharply reduces the notional damping factor from 890 to 330.

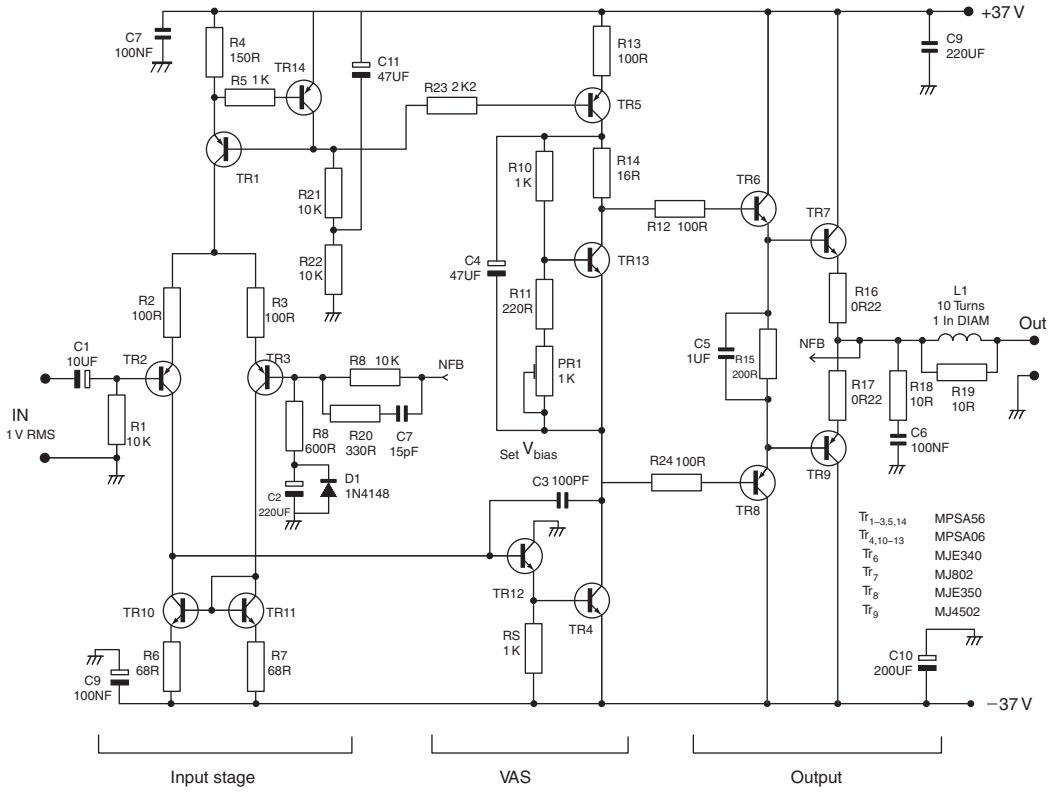


Figure 8.9: The Class-B amplifier from Chapter 7. At the simplest level the maximum slew rate is defined by the current source TR1 and the value of C_{dom}

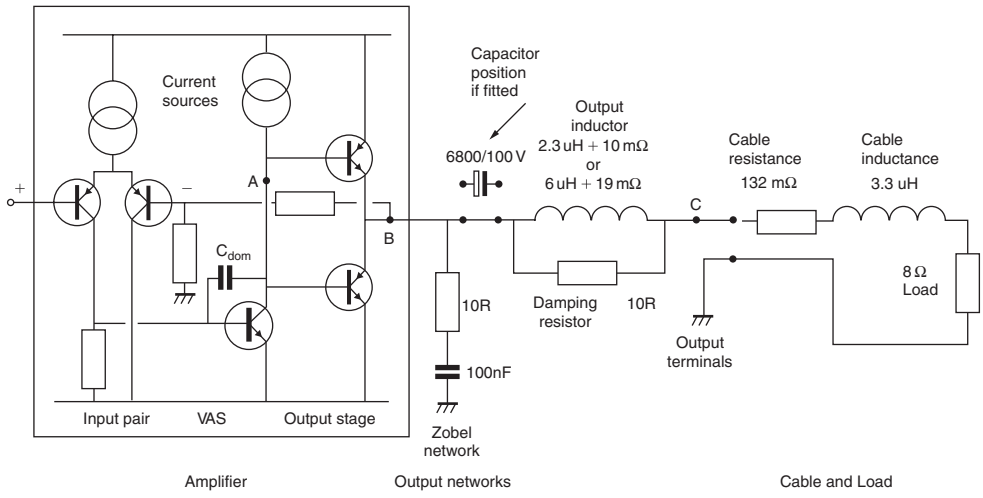


Figure 8.10: The amplifier–cable–speaker system. Simplified amplifier with Zobel network and damped output inductor, and a resistive load. Cable resistance and inductance values are typical for a 5 m length

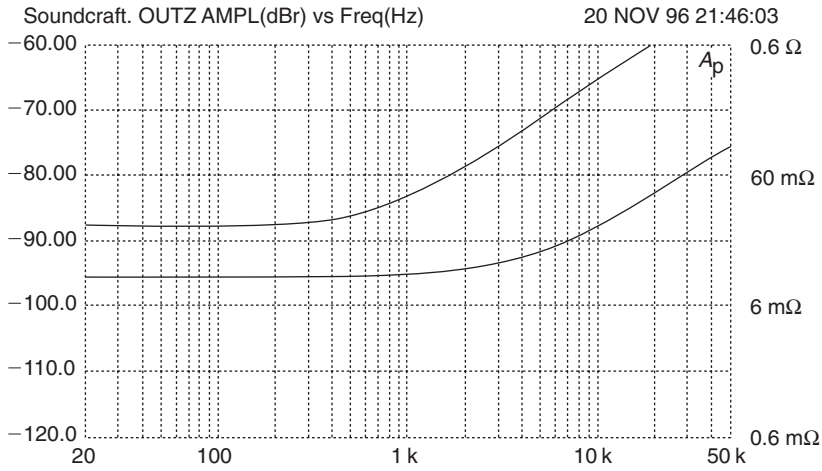


Figure 8.11: Output impedance of a Blameless amplifier, with and without 6 µH output inductor. Adding the inductor (upper trace) increases both the flat LF output impedance, due to its series resistance, and the rising HF impedance

Naturally the inductance of the coil pushes the rising portion of the impedance curve higher. The output impedance now starts to rise from 700 Hz, still at 6 dB/octave, reaching 0.6 Ω at 20 kHz (see Figure 8.11).

Minimizing Amplifier Output Impedance

This issue is worth considering, not because it optimizes speaker dynamics, which it does not, but because it minimizes frequency-response variations due to varying speaker impedance. There is also, of course, specmanship to be considered.

It is clear from Figure 8.11 that the output impedance of a generic amplifier will very probably be less than the inductor resistance, so the latter should be attended to first. Determine the minimum output inductance for stability with capacitive loads, because lower inductance means fewer turns of wire and less resistance. Some guidance on this is given in the next section. Note, however, that the inductance of the usual single-layer coil varies with the square of the number of turns, so halving the inductance only reduces the turns, and hence the series resistance, by root-2. The coil wire must be as thick as the cost/quality trade-offs allow.

It is also desirable to minimize the resistance of the amplifier internal wiring, and to carefully consider any extra resistance introduced by output relays, speaker switching, etc. When these factors have been reduced as far as cost and practicality allow, it is likely that the output impedance of the actual amplifier will still be the smallest component of the total.

Zobel Networks

All power amplifiers except for the most rudimentary kinds include a Zobel network in their arrangements for stability. This simple but somewhat enigmatic network comprises a resistor and capacitor in series from the amplifier output rail to ground. It is always fitted on the inside

(i.e. upstream) of the output inductor, though a few designs have a second Zobel network after the output inductor; the thinking behind this latter approach is obscure. The resistor approximates to the expected load impedance, and is usually between 4.7 and 10 Ω. The capacitor is almost invariably 100 nF, and these convenient values and their constancy in the face of changing amplifier design might lead one to suppose that they are not critical; in fact experiment suggests that the real reason is that the traditional values are just about right.

The function of the Zobel network (sometimes also called a Boucherot cell) is rarely discussed, but is usually said to prevent too inductive a reactance being presented to the amplifier output by a loudspeaker voice-coil, the implication being that this could cause HF instability. It is intuitively easy to see why a capacitive load on an amplifier with a finite output resistance could cause HF instability by introducing extra lagging phase shift into the global NFB loop, but it is less clear why an inductive load should be a problem; if a capacitive load reduces stability margins, then it seems reasonable that an inductive one would increase them.

At this point I felt some experiments were called for, and so I removed the standard 10 Ω/0.1 μF Zobel from a Blameless Class-B amplifier with CFP output and the usual NFB factor of 32 dB at 20 kHz. With an 8 Ω resistive load the THD performance and stability were unchanged. However, when a 0.47 mH inductor was added in series, to roughly simulate a single-unit loudspeaker, there was evidence of local VHF instability in the output stage; there was certainly no Nyquist instability of the global NFB loop.

I also attempted to reduce the loading placed on the output by the Zobel network. However, increasing the series resistance to 22 Ω still gave some evidence of stability problems, and I was forced to the depressing conclusion that the standard values are just about right. In fact, with the standard 10 Ω/0.1 μF network the extra loading placed on the amplifier at HF is not great; for a 1 V output at 10 kHz the Zobel network draws 6.3 mA, rising to 12.4 mA at 20 kHz, compared with 125 mA drawn at all frequencies by an 8 Ω resistor. These currents can be simply scaled up for realistic output levels, and this allows the Zobel resistor power rating to be determined. Thus an amplifier capable of 20 V rms output must have a Zobel resistor capable of sustaining 248 mA rms at 20 kHz, dissipating 0.62 W; a 1 W component could be chosen.

In fact, the greatest stress is placed on the Zobel resistor by HF instability, as amplifier oscillation is often in the range 50–500 kHz. It should therefore be chosen to withstand this for at least a short time, as otherwise fault-finding becomes rather fraught; ratings in the range 3–5 W are usual.

To conclude this section, there seems no doubt that a Zobel network is required with any load that is even mildly inductive. The resistor can be of an ordinary wire-wound type, rated to 5 W or more; this should prevent its burn-out under HF instability. A wire-wound resistor may reduce the effectiveness of the Zobel at VHF, but seems to work well in practice; the Zobel still gives effective stabilization with inductive loads.

Output Inductors

Only in the simplest kinds of power amplifier is it usual for the output stage to be connected directly to the external load. Direct connection is generally only feasible for amplifiers with low feedback factors, which have large safety margins against Nyquist instability caused by reactive loads.

When the stability of amplifiers into various loads is discussed, the phrase 'unconditional stability' is usually bandied about by people who are under the impression it means 'stable with any load you can think up'. Its original meaning, which comes from Control Theory, is quite different. In a normal dominant-pole compensated amplifier, reducing the loop gain (e.g. by reducing the amount of NFB) simply makes it more stable; this is the true meaning of 'unconditional stability'. If, however, you have a complicated compensation scheme, it is not hard to come up with an amplifier that becomes unstable when the loop gain is reduced, and this is called 'conditional stability'.

For many years designers have been wary of what may happen when a capacitive load is connected to their amplifiers, a fear that dates back to the introduction of the first practical electrostatic loudspeaker from Quad Acoustics, which was crudely emulated by adding a 2 μF capacitor in parallel to the usual 8 Ω resistive test load. The real load impedance presented by an electrostatic speaker is far more complex than this, largely as a result of the step-up transformer required to develop the appropriate drive voltages, but a 2 μF capacitor alone can cause instability in an amplifier unless precautions are taken.

When a shunt capacitor is placed across a resistive load in this way, and no output inductor is fitted, it is usually found that the value with the most destabilizing effect is nearer 100 nF than 2 μF.

The most effective precaution against this form of instability is a small air-cored inductor in series with the amplifier output. This isolates the amplifier from the shunt capacitance, without causing significant losses at audio frequencies. The value is normally in the region 1–7 μH, the upper limit being set by the need to avoid significant HF roll-off into a 4 Ω load. If 2 Ω loads are contemplated then this limit must be halved.

It is usual to test amplifier transient response with a square wave while the output is loaded with 8 Ω and 2 μF in parallel to simulate an electrostatic loudspeaker, as this is often regarded as the most demanding condition. However, there is an inductor in the amplifier output, and when there is significant capacitance in the load they resonate together, giving a peak in the frequency response at the HF end, and overshoot and ringing on fast edges.

This test therefore does not actually examine amplifier response at all, for the damped ringing that is almost universally seen during these capacitive loading tests is due to the output inductor resonating with the test load capacitance, and has nothing whatever to do with amplifier stability. The ringing is usually around 40 kHz or so, and this is much too slow to be blamed on any normally compensated amplifier. The output network adds ringing to the transient response even if the amplifier itself is perfect.

It is good practice to put a low-value damping resistor across the inductor; this reduces the Q of the output LC combination on capacitive loading, and thus reduces overshoot and ringing.

If a power amplifier is deliberately provoked by shorting out the output inductor and applying a capacitive load, then the oscillation is usually around 100–500 kHz and can be destructive of the output transistors if allowed to persist. It is nothing like the neat ringing seen in typical capacitive load tests. In this case there is no such thing as 'nicely damped ringing' because damped oscillation at 500 kHz probably means you are one bare step away from oscillatory disaster.

Attempts to test this on the circuit of Figure 8.9 were frustrated because it is actually rather resistant to capacitance-induced oscillation, probably because the level of global feedback is fairly modest; 100 nF directly across the output induced damped ringing at 420 kHz, while 470 nF gave ringing at 300 kHz, and 2 μF at 125 kHz.

While the 8 Ω/2 μF test described above actually reveals nothing about amplifier transient response, it is embedded in tradition, and it is too optimistic to expect its doubtful nature to be universally recognized. Minimizing output ringing is of some commercial importance. Several factors affect it, and can be manipulated to tidy up the overshoot and avoid deterring potential customers:

- The output inductance value. Increasing the inductance with all other components held constant reduces the overshoot and the amount of response peaking, but the peak moves downward in frequency so the rising response begins to invade the audio band (see Figures 8.12 and 8.13).
- The value of the damping resistor across the output coil. Reducing its value reduces the *Q* of the output LC tuned circuit, and so reduces overshoot and ringing. The resistor is usually 10 Ω, and can be a conventional wire-wound type without problems due to self-inductance; 10 Ω reduces the overshoot from 58% without damping to 48%, and much reduces ringing. Response peaking is reduced with only a slight effect on frequency (see Figures 8.14 and 8.15). The damping resistor can in fact be reduced to as low as 1 Ω, providing the amplifier stability into capacitance remains dependable, and this reduces the transient overshoot further from 48% to 19%, and eliminates ringing altogether; there is just a single overshoot. Whether this is more visually appealing to the potential customer is an interesting point.

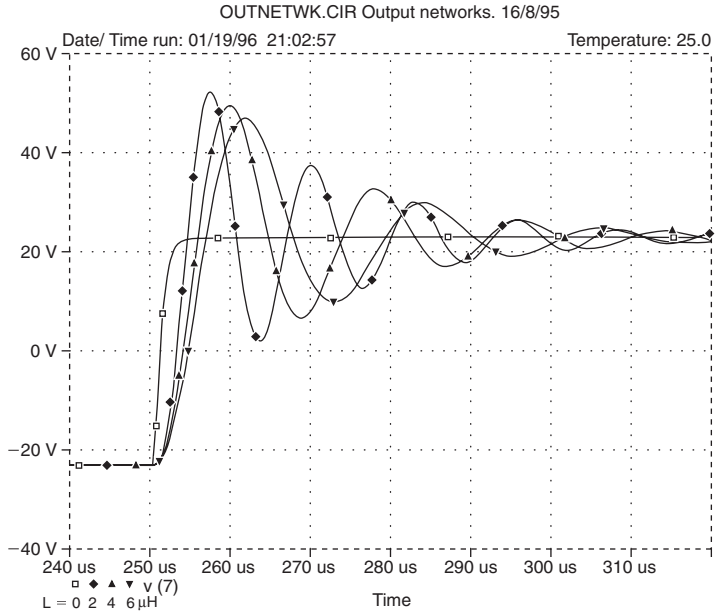


Figure 8.12: Transient response with varying output inductance; increasing *L* reduces ringing frequency without much effect on overshoot. Input rise time 1 μs

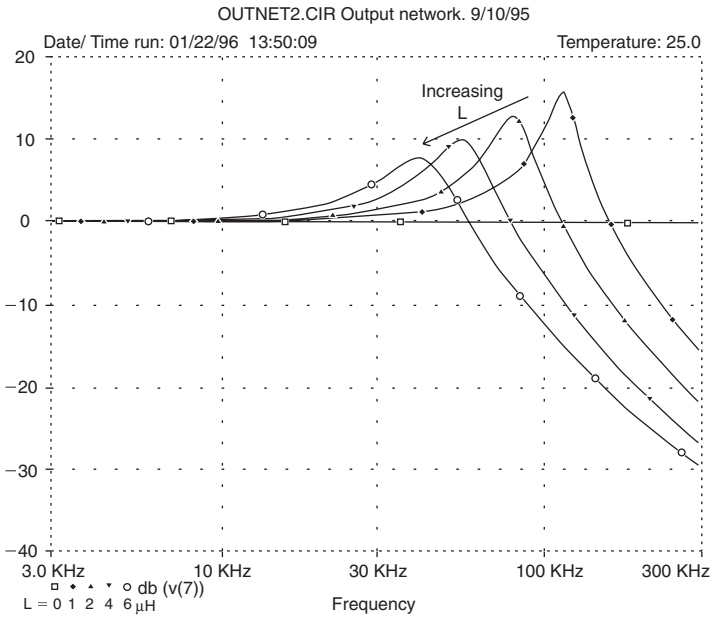


Figure 8.13: Increasing the output inductance reduces frequency response peaking and lowers its frequency

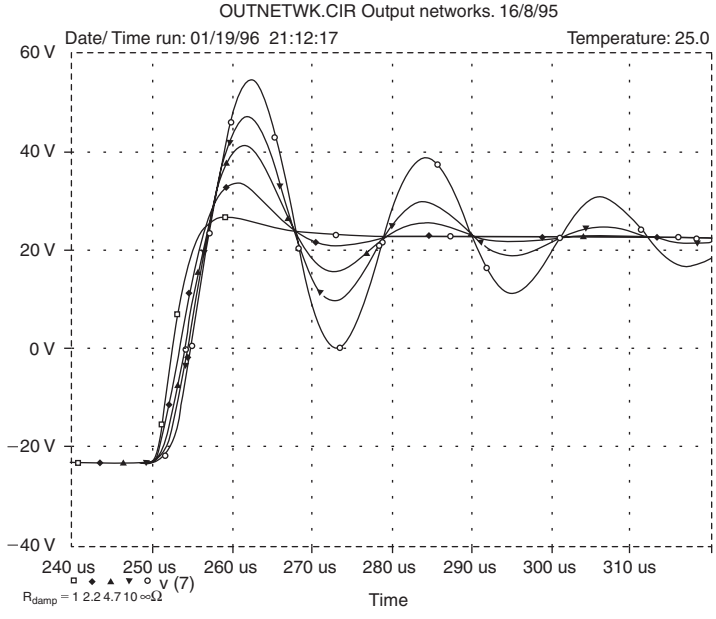


Figure 8.14: The effect of varying the damping resistance on transient response; 1 Ω almost eliminates overshoot

- The load capacitance value. Increasing this with the shunt resistor held at 8 Ω gives more overshoot and lower frequency ringing that decays more slowly. The response peaking is both sharper and lower in frequency, which is not a good combination. However, this component is part of the standard test load and is outside the designer's control (see Figures 8.16 and 8.17).

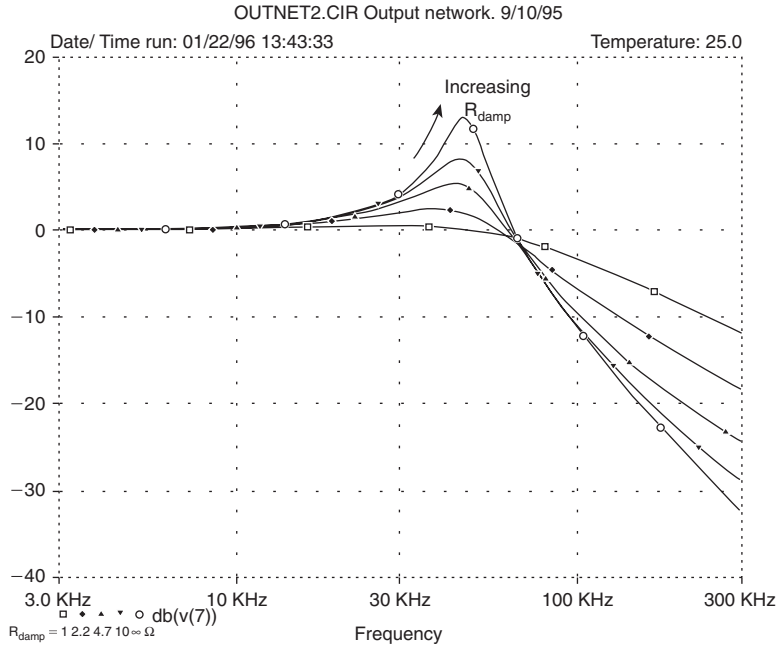


Figure 8.15: The effect of varying damping resistance on frequency response. Lower values reduce the peaking around 40 kHz

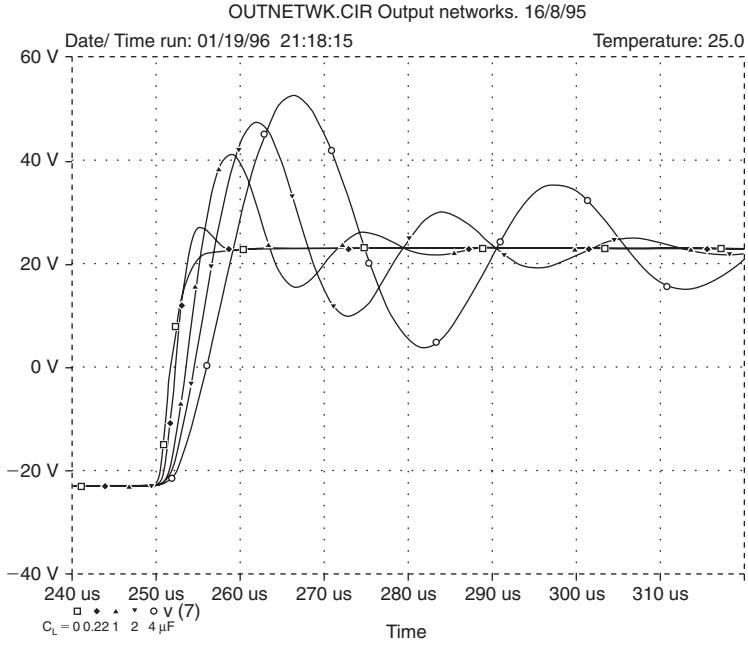


Figure 8.16: Increasing the load capacitance increases the transient overshoot, while lowering its frequency

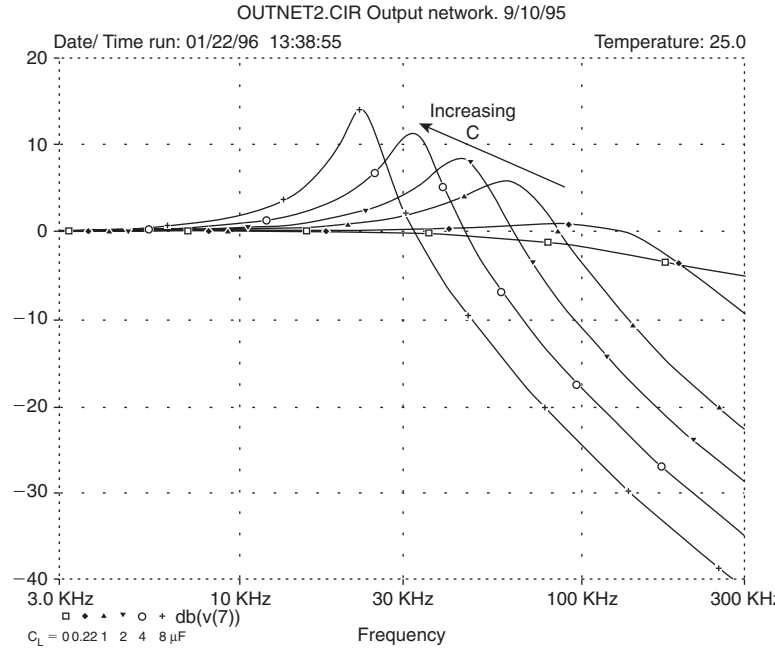


Figure 8.17: Increasing the load capacitance increases frequency-response peaking and lowers its frequency

- In actual fact, by far the most important factor affecting overshoot and ringing is the rise time of the applied square wave. This is yet another rather important audio fact that seems to be almost unknown. Figure 8.18 shows how the overshoot given by the circuit in Figure 8.10 is 51% for a 1 µs rise time, but only 12% for a 20 µs rise time. It is clear that the 'transient response' measured in this test may depend critically on the details of the test gear and the amplifier slew rate, and can be manipulated to give the result you want.

An output inductor should be air-cored to eliminate the possibility of extra distortion due to the saturation of magnetic materials. Ferrite-based VHF chokes give stable operation, but their linearity must be considered dubious. In the 1970s there was a fashion for using one of the big power-supply electrolytics as a coil-former, but this is not a good idea. The magnetic characteristics of the capacitor are unknown, and its lifetime may be reduced by the heat dissipated in the coil winding resistance.

The resistance of an air-cored 7 µH coil made from 20 turns of 1.5-mm-diameter wire (this is quite a substantial component 3 cm in diameter and 6 cm long) is enough to cause a measurable power loss into a 4 Ω load, and to dominate the output impedance as measured at the amplifier terminals. The coil wire should therefore be as thick as your cost/quality trade-offs allow.

The power rating for the damping resistor is assessed as follows. For a resistive 8 Ω load the voltage across the output inductor increases slowly with frequency, and the damping resistor dissipation only reaches 1.2 mW at 20 kHz for 1 V rms output. This assumes a normal 10 Ω damping resistor; if the value is reduced to 1 Ω to eliminate ringing into capacitive loads, as described above, then the dissipation is 10 times as great at 12 mW.

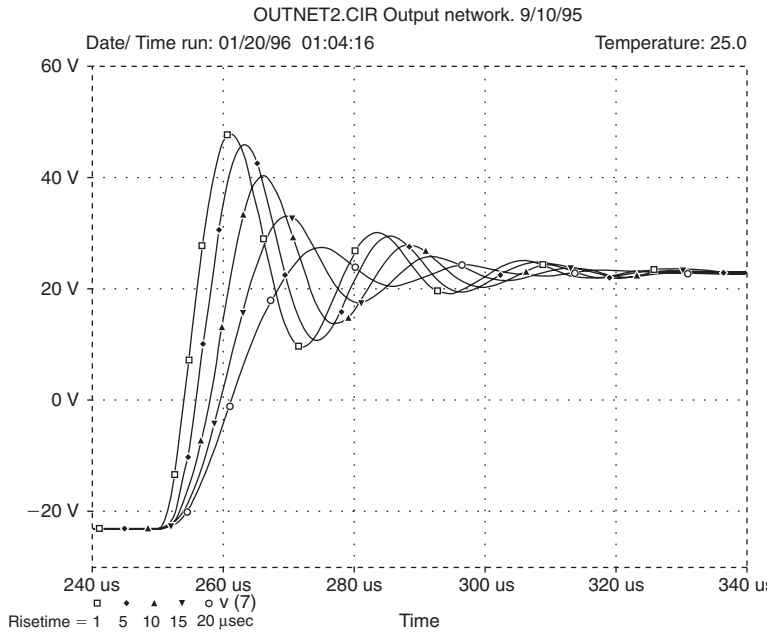


Figure 8.18: The most important factor in transient response is actually the rise time of the square-wave input, especially for overshoot percentage. The ringing frequency is unaffected

A much greater potential dissipation occurs when the load is the traditional 8Ω/2μF combination. The voltage across the output inductor peaks as it resonates with the load capacitance, and the power dissipated in a 10Ω damping resistor at resonance is 0.6W for 1V rms. This is, however, at an ultrasonic frequency (around 50kHz with a 7μH inductor) and is a fairly sharp peak, so there is little chance of musical signals causing high dissipation in the resistor in normal use. However, as for the Zobel network, some allowance must be made for sine-wave testing and oscillatory faults, so the damping resistor is commonly rated at between 1 and 5W. An ordinary wire-wound component works well with no apparent problems due to self-inductance.

The Output Inductor Value

As mentioned above, the output inductor for all my designs started out at 20 turns and approximately 6μH. In later tests the inductor was cut in half, now measuring 2.3μH inductance and 10.1mΩ DC resistance; this component was stable for all capacitor values, but has not had rigorous testing with real loudspeakers. It does now look more like an 'average' amplifier inductor, rather than an oversized one.

An alternative method of stabilization is to put in a small series resistor instead of the inductor; this approach has been used by at least one manufacturer. Even with 100nF loading, a 0R1 wire-wound output resistor completely removed ringing on the amplifier output. This is cheaper, but obviously less efficient than an inductor, as 100mΩ of extra resistance has been introduced instead of 10mΩ with the new 2.3μH inductor. The damping factor with 0R1 cannot exceed 80. A more important objection is that the 4Ω output power appears to be significantly reduced – a 200W/4Ω amplifier

is reduced to a 190 W unit, which does not look so good in the specs, even though the reduction in perceived loudness is negligible.

For the same reason – minimizing the resistive losses – output coils should be made from good thick copper wire.

Cable Effects

Looking at the amplifier–cable–load system as a whole, the amplifier and cable impedances have the following effects with an 8 Ω resistive load:

- A constant amplitude loss due to the cable resistance forming a potential divider with the 8 Ω load. The resistive component from the amplifier output is usually negligible.
- A high-frequency roll-off due to the cable inductance forming an LR low-pass filter with the 8 Ω load. The amplifier's output inductor (to give stability with capacitive loads) adds directly to this to make up the total series inductance. The shunt capacitance of any normal speaker cable is trivially small, and can have no significant effect on frequency response or anything else.

The main factors in speaker cable selection are therefore series resistance and inductance. If these parameters are below 100 mΩ and 3 μH, any effects will be imperceptible. This can be met by 13 A mains cable, especially if all three conductors are used.

If the amplifier is connected to a typical loudspeaker rather than a pure resistance the further effects are:

- The frequency response of the voltage at the loudspeaker terminals shows small humps and dips as the uneven speaker impedance loads the series combination of amplifier output impedance and cable resistance.
- The variable loading affects the amplifier distortion performance. HF crossover distortion reduces as load resistance increases above 8 Ω; even 68 Ω loading increases HF distortion above the unloaded condition. For heavier loading than 8 Ω, crossover may continue to increase, but this is usually masked by the onset of large-signal nonlinearity (see Chapter 6).
- Severe dips in impedance may activate the overload protection circuitry unexpectedly. Signal amplitudes are higher at LF so impedance dips here are potentially more likely to draw enough current to trigger protection.

Crosstalk in Amplifier Output Inductors

When designing a stereo power amplifier, the issue of interchannel crosstalk is always a concern. Now that amplifiers with up to seven channels for home theater are becoming more common, the crosstalk issue is that much more important, if only because the channels are likely to be more closely packed. Here I deal with one aspect of it. Almost all power amplifiers have output coils

to stabilize them against capacitive reactances, and a question often raised is whether inductive coupling between the two is likely to degrade crosstalk. It is sometimes suggested that the coils – which are usually in solenoid form, with length and diameter of the same order – should be mounted with their axes at right angles rather than parallel, to minimize coupling. But does this really work?

I think I am pretty safe in saying there is no published work on this, so it was time to make some. The coil coupling could no doubt be calculated (though not by me) but, as often in the glorious pursuit of electronics, it was quicker to measure it.

The coils I used were both of 14 turns of 1-mm-diameter copper wire, overall length 22 mm and diameter 20 mm. This has an inductance of about 2 μH, and is pretty much an 'average' output coil, suitable for stabilizing amplifiers up to about 150 W/8 Ω. Different coils will give somewhat different results, but extrapolation to whatever component you are using should be straightforward; for example, twice the turns on both coils means four times the coupling.

Figure 8.19a shows the situation in a stereo power amplifier. The field radiated due to the current in coil A is picked up by coil B and a crosstalk voltage added to the output signal at B.

Figure 8.19b shows the experimental setup. Coil A is driven from a signal generator with a source impedance of 50 Ω, set to 5 V rms. Virtually all of this is dropped across the source resistance, so coil A is effectively driven with a constant current of 100 mA rms.

Figure 8.21 shows the first result, taken with the coils coaxial and the ends touching, as in Figure 8.20. (This proved, as expected, to be the worst case for coupling.) The crosstalk rises at 6 dB/octave, because the voltage induced in coil B is proportional to the rate of change of flux, and the magnitude of peak flux is fixed. This is clearly not the same as conventional transformer action, where the frequency response is flat. In a transformer the primary inductance is much greater than the circuit series impedance, so the magnetic flux that couples with the secondary halves when the input frequency

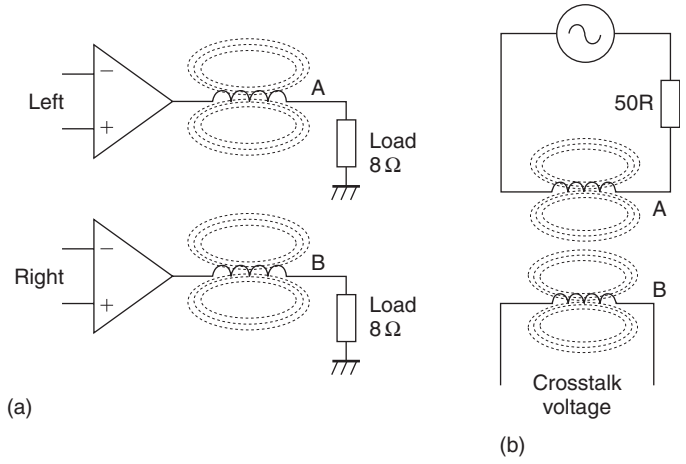


Figure 8.19: (a) The coupling of output coils in a stereo power amplifier. (b) The experimental circuit. The 'transmitting' coil A is driven with an effectively constant current, and the voltage across the 'receiving' coil B measured

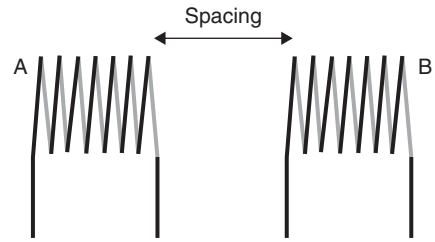


Figure 8.20: The physical coil configuration for the measurement of coaxial coils

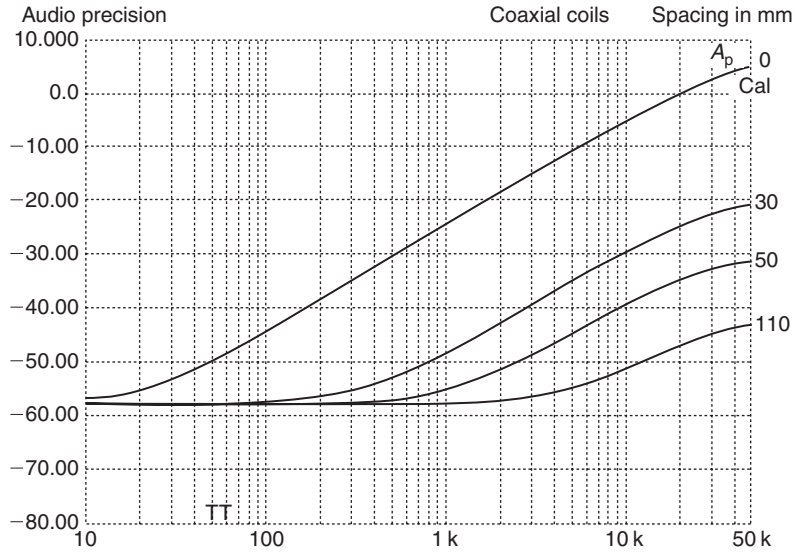


Figure 8.21: Crosstalk versus spacing for coaxial coils

doubles, and the voltage induced in the secondary is constant. The crosstalk at 20 kHz was taken as the 0 dB reference. This represented 2.4 mV rms across coil B; 100 mA rms in coil A corresponds to 800 mV rms across an 8 Ω load, so this gives a final crosstalk figure from channel to channel of −54 dB at 20 kHz. It carries on deteriorating above 20 kHz but no one can hear it. All crosstalk figures given below are at 20 kHz.

The coils were then separated 10 mm at a time, and with each increment the crosstalk dropped by 10 dB, as seen in Figure 8.21. At 110 mm spacing, which is quite practical for most designs, the crosstalk had fallen by 47 dB from the reference case, giving an overall crosstalk of −54 and −47 dB = −101 dB total. This is a very low level, and at the very top of the audio band. At 1 kHz, where the ear is much more sensitive, the crosstalk will be some 25 dB less, which brings it down to −126 dB total, which I can say with some confidence is not going to be a problem. This is obtained with what looks like the least favorable orientation of coils. Coil–coil coupling is −32 dB at 50 mm, and the figure at this spacing will be used to compare the configurations.

The next configuration tested was that of Figure 8.22, where the coils have parallel axes but are displaced to the side. The results are in Figure 8.23; the crosstalk is now −38 dB at 50 mm. With each 10 mm spacing increment the crosstalk dropped by 7 dB. This setup is worse than the crossed-axis version but better than the coaxial one.

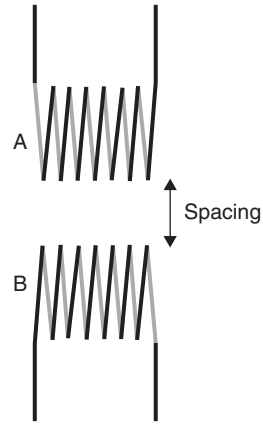


Figure 8.22: The coil configuration for non-coaxial parallel-axis coils

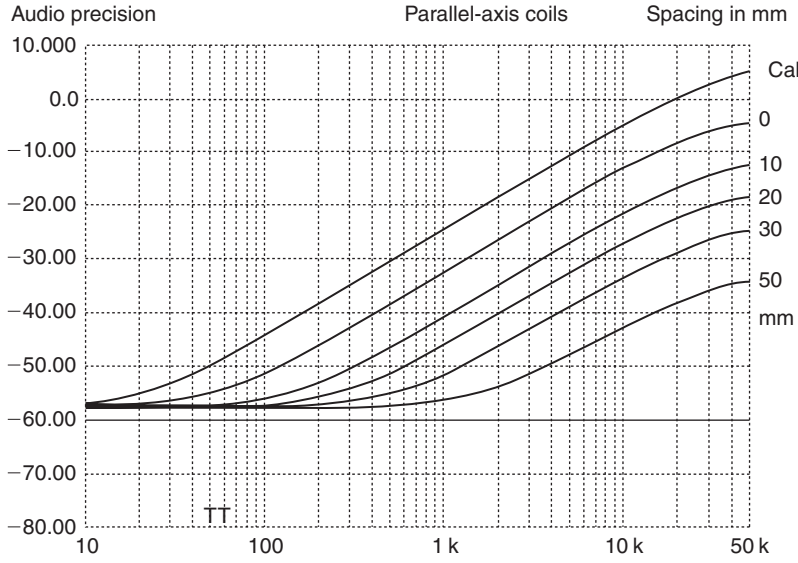


Figure 8.23: Crosstalk versus spacing for parallel-axis coils

The final configurations had the axes of the coils at 90°: the crossed-axis condition. The base position is with the corners of the coils touching (see Figure 8.24). When the coil is in the position X, still touching, crosstalk almost vanishes as there is a cancelation null. With the coils so close, this is a very sharp null and exploiting it in quantity production is quite impractical. The slightest deformation of either coil ruins the effect. Moving coil A away from B again gives the results in Figure 8.25. The crosstalk is now −43 dB at 50 mm, only an improvement of 11 dB over the coaxial case; turning coils around is clearly not as effective as might be supposed. This time, with each 10 mm spacing increment the crosstalk dropped by 8 dB rather than 10 dB.

The obvious next step is to try combining distance with cancelation, as in Figure 8.26. This can give a good performance even if a large spacing is not possible. Figure 8.27 shows that careful coil

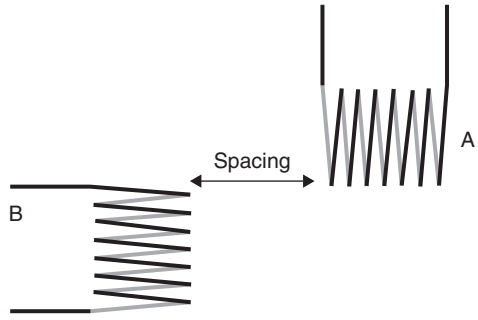


Figure 8.24: The coil configuration for crossed-axis measurements

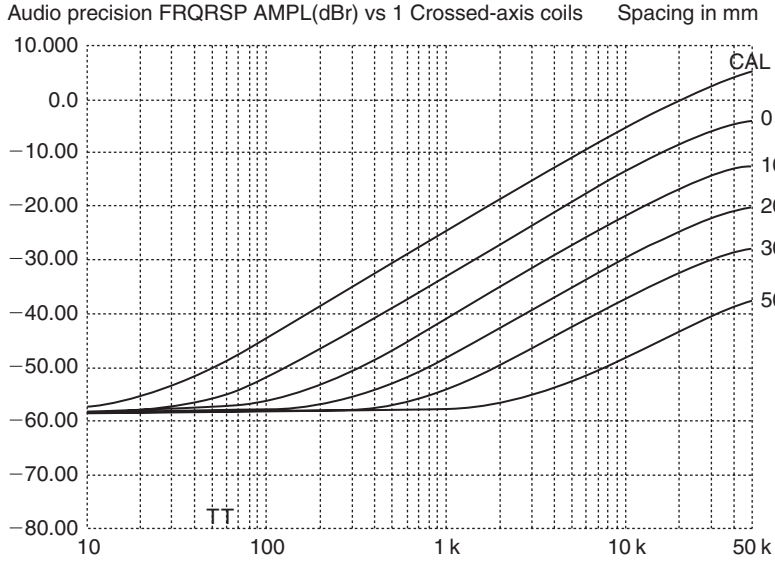


Figure 8.25: Crosstalk versus spacing for crossed-axis coils

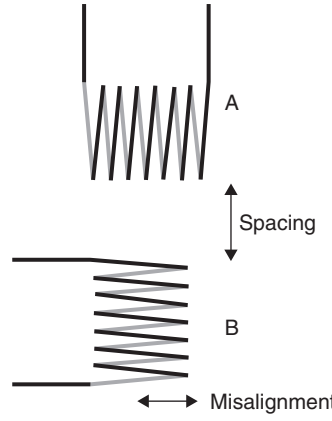


Figure 8.26: The coil configuration for crossed-axis with cancelation

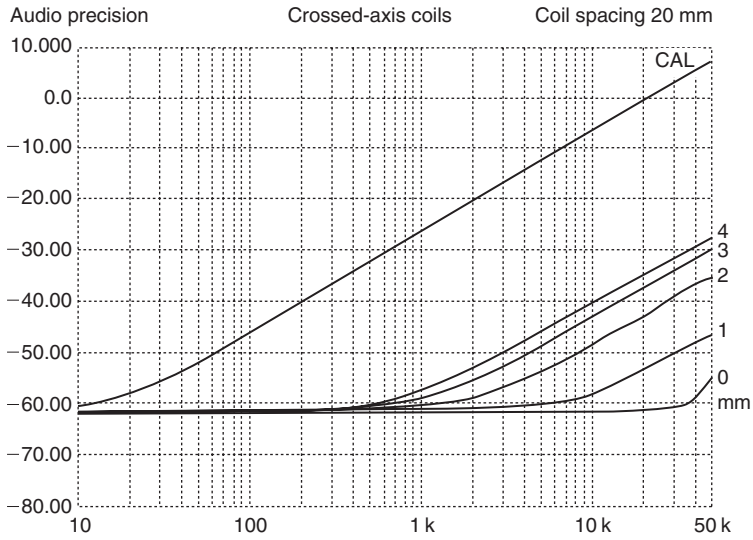


Figure 8.27: Crosstalk versus alignment for crossed-axis coils spaced at 20 mm, using cancelation

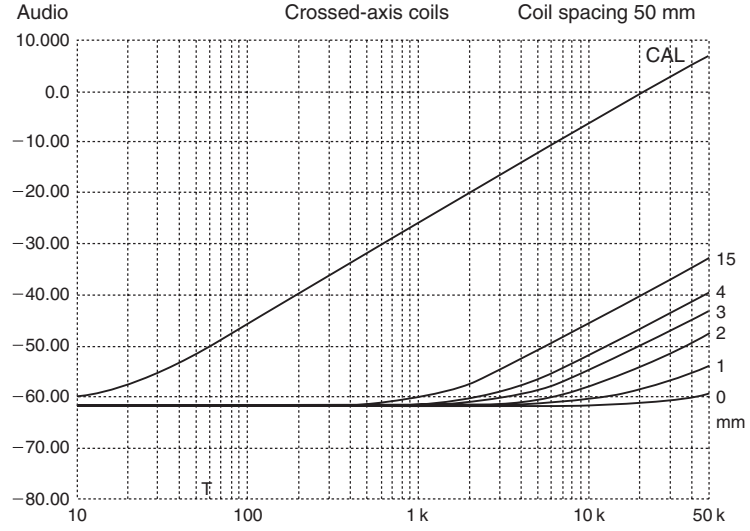


Figure 8.28: Crosstalk versus alignment for crossed-axis coils spaced at 50 mm, using cancelation

positioning can give crosstalk better than −60 dB (−114 dB total) across the audio band, although the spacing is only 20 mm. The other curves show the degradation of performance when the coil is misaligned by moving it bodily sideways by 1, 2, 3, and 4 mm; just a 2 mm error has worsened crosstalk by 20 dB at 20 kHz. Obviously in practice the coil PCB hole will not move – but it is very possible that coils will be bent slightly sideways in production.

Figure 8.28 gives the same results for a 50 mm spacing, which can usually be managed in a stereo design. The null position once more just gives the noise floor across the band, and a 2 mm

misalignment now only worsens things by about 5 dB. This is definitely the best arrangement if the spacing is limited.

Coil Crosstalk Conclusions

Coil orientation can help. Simply turning one coil through 90° gives an improvement of only 11 dB, but if it is aligned to cancel out the coupling, there is a big improvement. See how −38 dB in Figure 8.23 becomes −61 dB in Figure 8.28 at 20 kHz. On a typical stereo amplifier PCB, the coils are likely to be parallel – probably just for the sake of appearance – but their spacing is unlikely to be less than 50 mm unless the output components have been deliberately grouped together. As with capacitive crosstalk, physical distance is cheaper than anything else, and if the results are not good enough, use more of it. In this case the overall crosstalk at 20 kHz will be −54 + −38 dB = −92 dB total, which is probably already well below other forms of interchannel crosstalk. A quick quarter-turn of the coil improves this to at least −114 dB. It should do.

Reactive Loads and Speaker Simulation

Amplifiers are almost universally designed and tested running into a purely resistive load, although they actually spend their working lives driving loudspeakers, which contain both important reactive components and also electromechanical resonances. At first sight this is a nonsensical situation; however, testing into resistive loads is neither naive nor an attempt to avoid the issue of real loads; there is in fact little alternative.

Loudspeakers vary greatly in their design and construction, and this is reflected in variations in the impedance they present to the amplifier on test. It would be necessary to specify a *standard speaker* for the results from different amplifiers to be comparable. Second, loudspeakers have a notable tendency to turn electricity into sound, and the sine-wave testing of a 200 W amplifier would be a demanding experience for all those in earshot; soundproof chambers are not easy or cheap to construct. Third, such a standard test speaker would have to be capable of enormous power-handling if it were to be able to sustain long-term testing at high power; loudspeakers are always rated with the peak/average ratio of speech and music firmly in mind, and the lower signal levels at high frequencies are also exploited when choosing tweeter power ratings. A final objection is that loudspeakers are not noted for perfect linearity, especially at the LF end, and if the amplifier does not have a very low output impedance this speaker nonlinearity may confuse the measurement of distortion. Amplifier testing would demand a completely different sort of loudspeaker from that used for actually listening to music; the market for it would be very, very small, so it would be expensive.

Resistive Loads

Amplifiers are normally developed through 8 and 4 Ω testing, though intermediate values such as 5.66 Ω (the geometric mean of 8 and 4) are rarely explored considering how often they occur in real use. This is probably legitimate in that if an amplifier works well at 8 and 4 Ω it is most unlikely to give trouble at intermediate loadings. In practice few nominal 8 Ω speakers have impedance dips that go below 5 Ω, and design to 4 Ω gives a safety margin, if not a large one.

The most common elaboration on a simple resistive load is the addition of 2 μF in parallel with 8 Ω to roughly simulate an electrostatic loudspeaker; this is in fact not a particularly reactive load, for the impedance of a 2 μF capacitor only becomes equal to the resistance at 9.95 kHz, so most of the audio band is left undisturbed by phase shift. This load is in fact a worse approximation to a moving-coil speaker than is a pure resistance.

Modeling Real Loudspeaker Loading

The impedance curve of a real loudspeaker may be complex, with multiple humps and dips representing various features of the speaker. The resonance in the bass driver unit will give a significant hump in LF impedance, with associated phase changes. Reflex (ported enclosure) designs have a characteristic double-hump in the LF, with the middle dip corresponding to the port tuning. The HF region is highly variable, and depends in a complicated fashion on the number of drive units and their interactions with the crossover components.

Connection of an amplifier to a typical speaker impedance rather than a resistance has several consequences:

- The frequency response, measured in terms of the voltage across the loudspeaker terminals, shows small humps and bumps due to the uneven impedance loading the series combination of amplifier output impedance and connecting cable resistance.
- Severe dips in impedance may activate the overload protection circuitry prematurely. This has to be looked at in terms of probability, because a high amplitude in a narrow frequency band may not occur very often, and if it does it may be so brief that the distortion generated is not perceptible. Amplitudes are higher at LF and so impedance dips here are potentially more serious.
- The variable loading affects the distortion performance.

Figure 8.29 shows how the HF crossover distortion varies with load resistance for loads lighter than those usually considered. Even 68 Ω loading increases HF distortion.

Figure 8.30 shows an electrical model of a single full-range loudspeaker unit. While a single-driver design is unlikely to be encountered in hi-fi applications, many PA, disco, and sound reinforcement applications use full-range drive units, for which this is a good model. R_c and L_c represent the resistance and inductance of the voice-coil. L_r and C_r model the electromechanical resonance of the cone mass with the suspension compliance and air-spring of the enclosure, with R_r setting the damping; these last three components have no physical existence, but give the same impedance characteristics as the real resonance.

The input impedance magnitude this network presents to an amplifier is shown in Figure 8.31. The peak at 70 Hz is due to the cone resonance; without the sealed enclosure, the restoring force on the cone would be less and the free-air resonance would be at a lower frequency. The rising impedance above 1 kHz is due to the voice-coil inductance L_c.

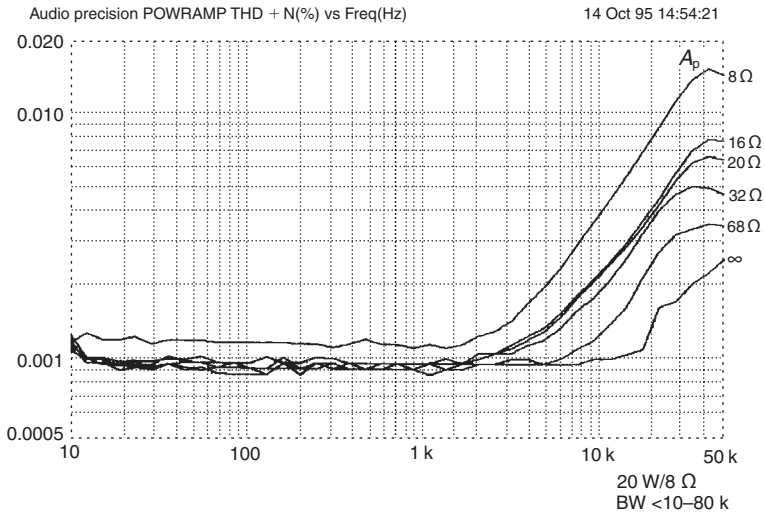


Figure 8.29: The reduction of HF THD as resistive amplifier loading is made lighter than 8 Ω

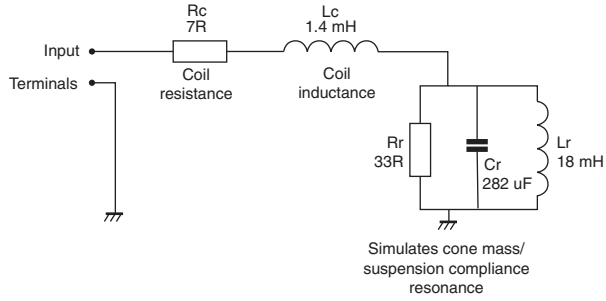


Figure 8.30: Electrical model of a single speaker unit in a sealed enclosure

When the electrical model of a single-unit load replaces the standard 8 Ω resistive load, something remarkable happens; HF distortion virtually disappears, as shown in Figure 8.32. This is because a Blameless amplifier driving 8 Ω only exhibits crossover distortion, increasing with frequency as the NFB factor falls, and the magnitude of this depends on the current drawn from the output stage; with an inductive load this current falls at high frequencies.

Most hi-fi amplifiers will be driving two- or three-way loudspeaker systems, and four-way designs are not unknown. This complicates the impedance characteristic, which in a typical two-way speaker looks something like Figure 8.33, though the rise above 10 kHz is often absent. The bass resonance remains at 70 Hz as before, but there are two drive units and hence two resonances. There is also the considerable complication of a crossover network to direct the HF to the tweeter and the LF to the low-frequency unit, and this adds several extra variables to the situation. In a bass reflex design the bass resonance hump may be supplemented by another LF resonant peak due to the port tuning. An attempt at a representative load simulator for a two-way infinite-baffle loudspeaker system is shown in Figure 8.34. This assumes a simple crossover network without

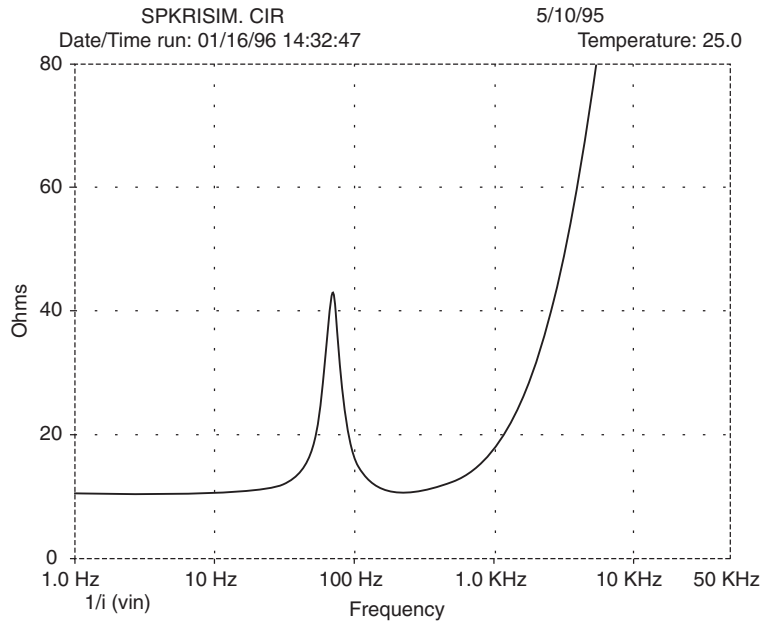


Figure 8.31: Input impedance of single speaker unit

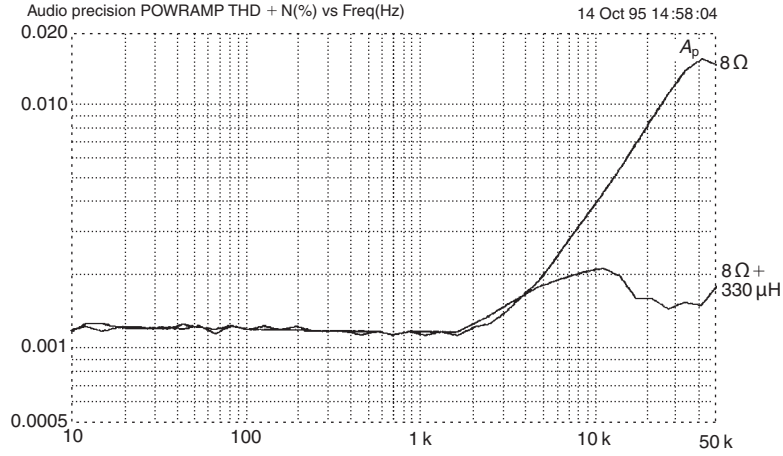


Figure 8.32: The reduction of HF THD with an inductive load; adding 330 µH in series with the 8 Ω reduces the 20 kHz THD by more than four times

compensation for rising tweeter coil impedance, and is partially based on a network proposed by Ken Kantnor in Atkinson[13].

Some loudspeaker crossover designs include their own Zobel networks, typically placed across the tweeter unit, to compensate for the HF rise in impedance due to the voice-coil inductance. If these Zobels are placed there to terminate the crossover circuitry in a roughly resistive load, then

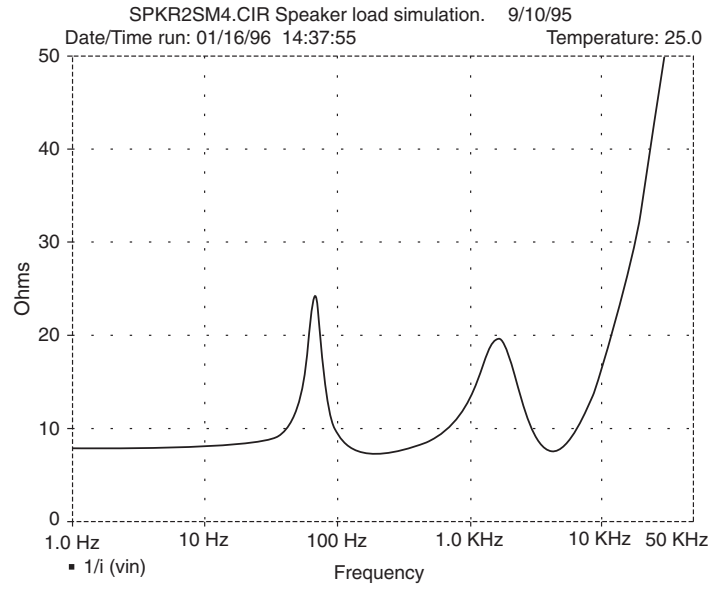


Figure 8.33: The impedance plot of the two-way speaker model

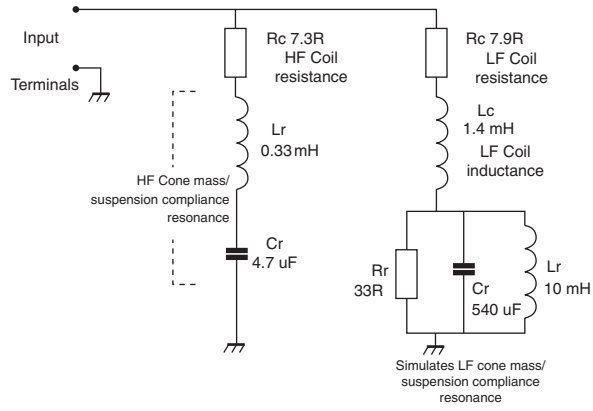


Figure 8.34: The circuit of the two-way speaker model

the loudspeaker designer has every right to do it; electroacoustic design is quite difficult enough without adding extra restrictions. However, if they are incorporated simply to make the impedance curve look tidier, and allow a claim that the load has been made easier for the amplifier to drive, then this seems misguided. The actual effect is the opposite; a typical amplifier has no difficulty driving an inductive reactance, and the HF crossover distortion can be greatly reduced when driving a load with an impedance that rises above the nominal value at HF.

This is only an introduction to the huge subject of real amplifier loads. More detailed information is given in Benjamin[14].

Loudspeaker Loads and Output Stages

There is a common assumption that any reactive load is more difficult for an amplifier to drive than a purely resistive one; however, it is devoutly to be wished that people would say what they mean by 'difficult'. It could mean that stability margins are reduced, or that the stresses on the output devices are increased. Both problems can exist, but I suspect that this belief is rooted in anthropomorphic thinking. It is easy to assume that if a signal is more complex to contemplate, it is harder for an amplifier to handle. This is not, however, true; it is not necessary to understand the laws of physics to obey them. Everything does anyway.

When solid-state amplifiers show instability it is always at ultrasonic frequencies, assuming we are not grappling with some historical curiosity that has AC-coupling in the forward signal path. It never occurs in the middle of the audio band, although many loudspeakers have major convulsions in their impedance curves in this region. Reactive loading can and does imperil stability at high frequencies unless precautions are taken, usually in the form of an output inductor. It does not cause oscillation or ringing mid-band.

Reactive loads do increase output device stresses. In particular, peak power dissipation is increased by the altered voltage/current phase relationships in a reactive load.

Single-Speaker Load

Considering a single speaker unit with the equivalent circuit of Figure 8.30, the impedance magnitude never falls below the 8 Ω nominal value, and is much greater in some regions; this suggests the overall amplifier power dissipation would be less than for an 8 Ω resistive load.

Unfortunately this is not so; the voltage/current phase relationship brought about by the reactive load is a critical factor. When a pure resistance is driven, the voltage across the output device falls as the current through it rises, and they never reach a maximum at the same time (see Figure 8.35, for Class-B with an 8 Ω resistive load). The instantaneous power is the product of instantaneous current and voltage drop, and in Class-B has a characteristic two-horned shape, peaking twice at 77 W during its conducting half-cycle.

When the single-speaker load is driven at 50 Hz, the impedance is a mix of resistive and inductive, at 8.12 + 3.9 jΩ. Therefore the current phase-lags the voltage, altering the instantaneous product of voltage and power to that shown in Figure 8.36. The average dissipation over the Class-B half-cycle is slightly reduced, but the peak instantaneous power increases by 30% due to the voltage/current phase shift. This could have serious results on amplifier reliability if not considered at the design stage. Note that this impedance is equivalent *at 50 Hz only* to 8.5 Ω in series with 10.8 mH. Trying to drive this replacement load at any other frequency, or with a non-sine waveform, would give completely wrong results. Not every writer on this topic appears to appreciate this.

Similarly, if the single-speaker load is driven at 200 Hz, on the other side of the resonance peak, the impedance is a combination of resistive and capacitive at 8.4 − 3.4 jΩ and the current leads the voltage. This gives much the same result as in Figure 8.36, except that the peak power now occurs in the first part of the half-cycle. The equivalent load *at 200 Hz only* is 10.8 Ω in parallel with 35 μF.

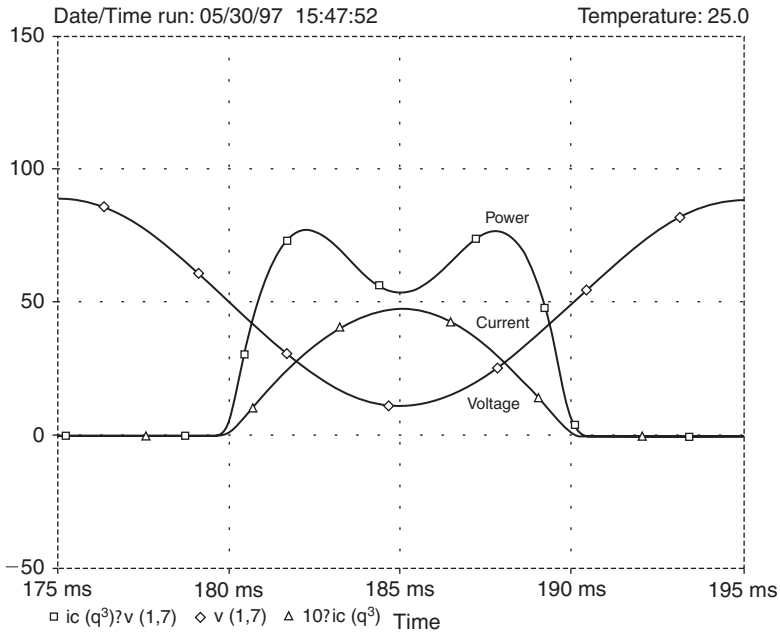


Figure 8.35: Instantaneous V_{ce}, I_c, and P_{diss} in an output transistor driving 8 Ω to 40 V peak at 50 Hz from ±50 V rails. Device dissipation peaks twice at 77 W in each half-cycle

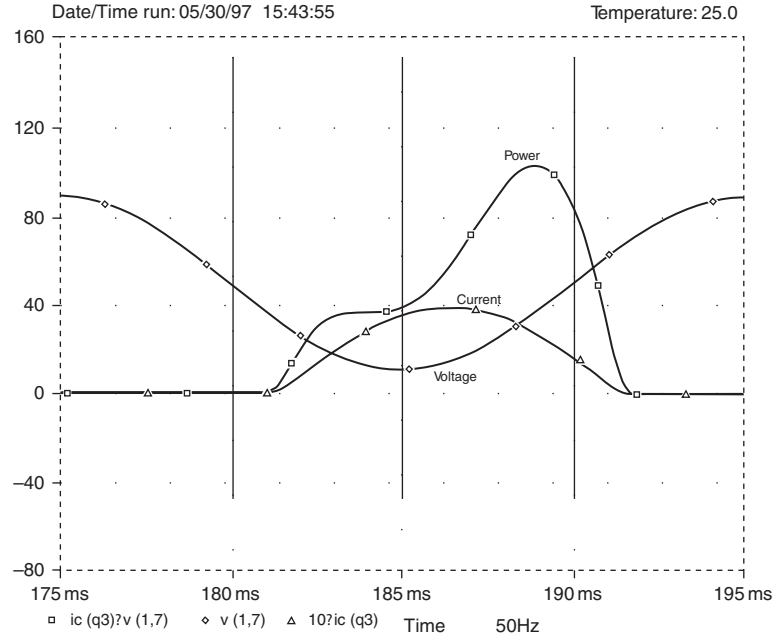


Figure 8.36: As for Figure 8.35, but driving 50 Hz into the single-speaker load. At this frequency the load is partly inductive so current lags voltage and the instantaneous power curve is asymmetrical, peaking higher at 110 W towards the end of the half-cycle

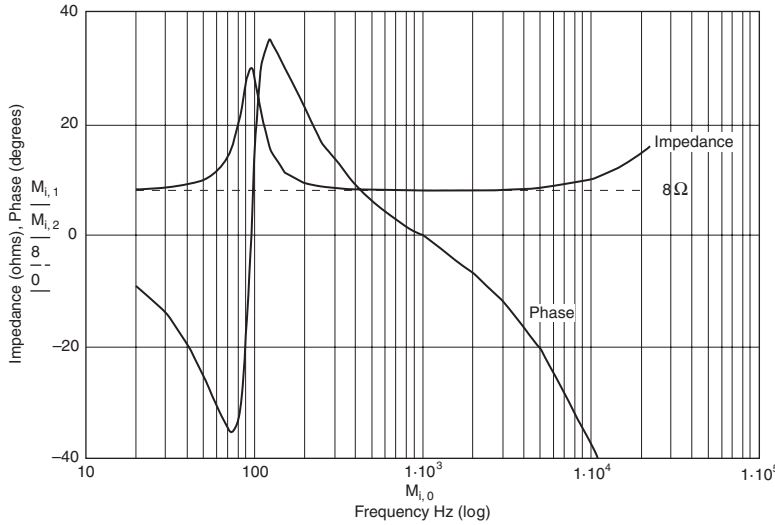


Figure 8.37: Impedance curve of the single-speaker model. The dotted line is 8Ω resistive

When designing output stages, there are four electrical quantities to accommodate within the output device ratings: peak current, average current, peak power, and average power. (Junction temperatures must of course also be considered at some point.) The critical quantities for semiconductor safety in amplifiers are usually the peak instantaneous values; for heat-sink design average power is what counts, while for the power-supply average current is the significant quantity.

To determine the effect of real speaker loads on device stress I simulated an EF output stage driving a single-speaker load with a 40V peak sine wave, powered from ±50V rails. The load was as in Figure 8.30 except for a reduction in the voice-coil inductance to 0.1mH; the resulting impedance curve is shown in Figure 8.37. Transient simulations over many cycles were done for 42 spot frequencies from 20Hz to 20kHz, and the peak and average quantities recorded and plotted. Many cycles must be simulated as the bass resonance in the impedance model takes time to reach steady state when a sine wave is abruptly applied; not everyone writing on this topic appears to have appreciated this point.

Steady sine-wave excitation was used as a practical approach to simulation and testing, and does not claim to be a good approximation to music or speech. Arbitrary noncyclic transients could be investigated by the same method, but the number of waveform possibilities is infinite. It would also be necessary to be careful about the initial conditions.

Figures 8.37–8.39 are the distilled results of a very large number of simulations. Figure 8.38 shows that the gentle foothills of the impedance peak at bass resonance actually increase the peak instantaneous power stress on the output devices by 30%, despite the reduced current drawn.

The most dangerous regions for the amplifier are the sides of a resonance hump where the phase shift is the greatest. Peak dissipation only falls below that for an 8Ω resistor (shown dotted) around the actual resonance peak, where it drops quickly to a quarter of the resistive case.

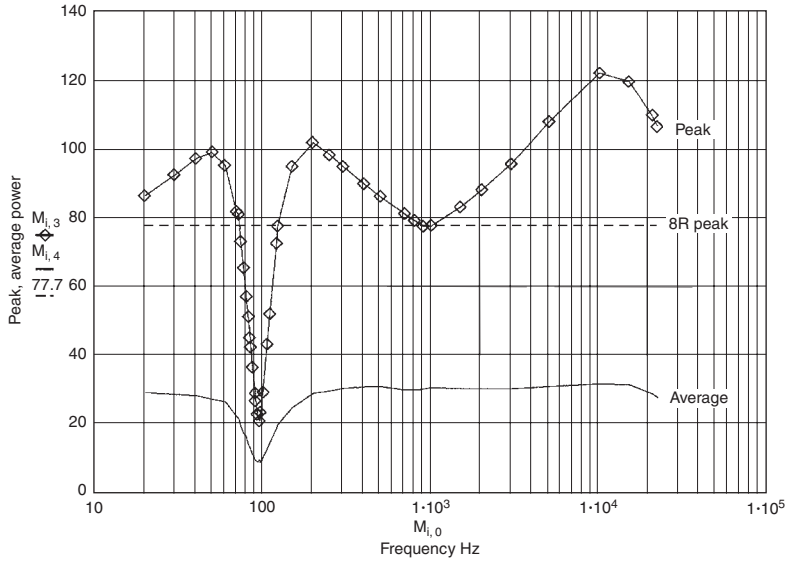


Figure 8.38: Peak and average output device power dissipation driving the single-unit speaker impedance as in Figure 8.33. The dotted line is peak power for 8 Ω resistive

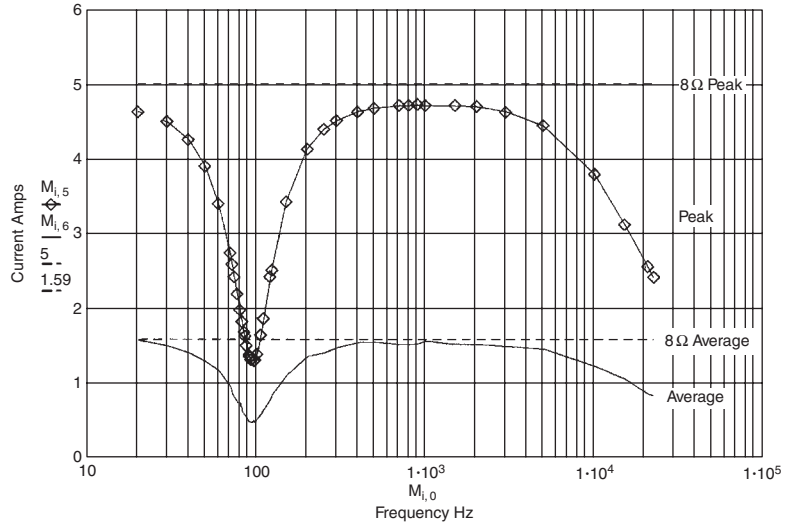


Figure 8.39: Peak and average output device current driving the single-unit speaker impedance. Dotted lines are peak and average current into 8 Ω

Likewise, the increase in impedance at the HF end of the spectrum, where voice-coil inductance is significant, causes a more serious rise in peak dissipation to 50% more than the resistive case. The conclusion is that, for peak power, the phase angle is far more important than the impedance magnitude.

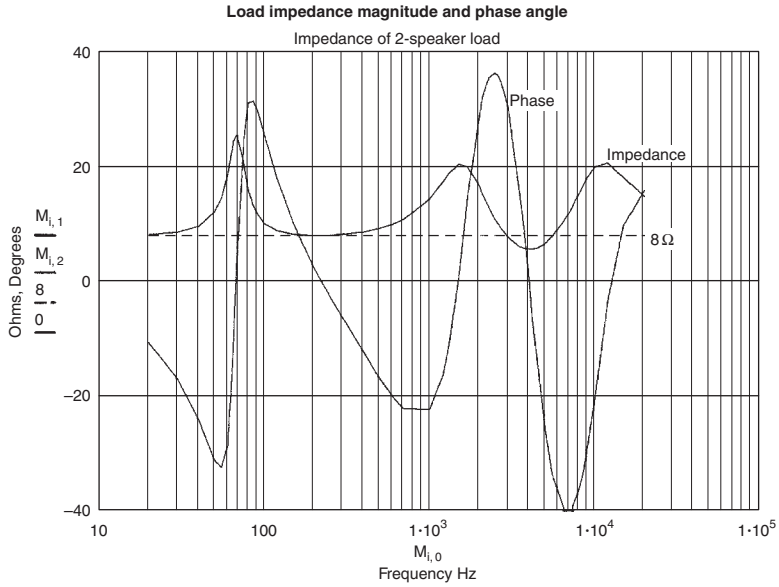


Figure 8.40: Impedance curve of model of the two-unit speaker model in Figure 8.34. Dotted line is 8Ω resistive

The effects on the average power dissipation, and on the peak and average device current in Figure 8.39, are more benign. With this type of load network, all three quantities are reduced when the speaker impedance increases, the voltage/current phase shifts having no effect on the current.

Two-Way Speaker Loads

The impedance plot for the simulated two-way speaker load of Figure 8.34 is shown in Figure 8.40 at 59 spot frequencies. The curve is more complex and shows a dip below the nominal impedance as well as peaks above; this is typical of multi-speaker designs. An impedance dip causes the maximum output device stress as it combines increased current demand with phase shifts that increase peak instantaneous dissipation.

In Figure 8.41 the impedance rise at bass resonance again causes increased peak power dissipation due to phase shifts; the other three quantities are reduced. In the HF region there is an impedance dip at 6kHz that nearly doubles peak power dissipation on its lower slopes, the effect being greater because both phase shift and increased current demand are acting. The actual bottom of the dip sharply reduces peak power where the phase angle passes through zero, giving the notch effect at the top of the peak.

Average power (Figure 8.41) and peak and average current (Figure 8.42) are all increased by the impedance dip, but to a more modest extent.

Peak power would appear to be the critical quantity. Power device ratings often allow the power and second-breakdown limits (and sometimes the bondwire current limit also) to be exceeded

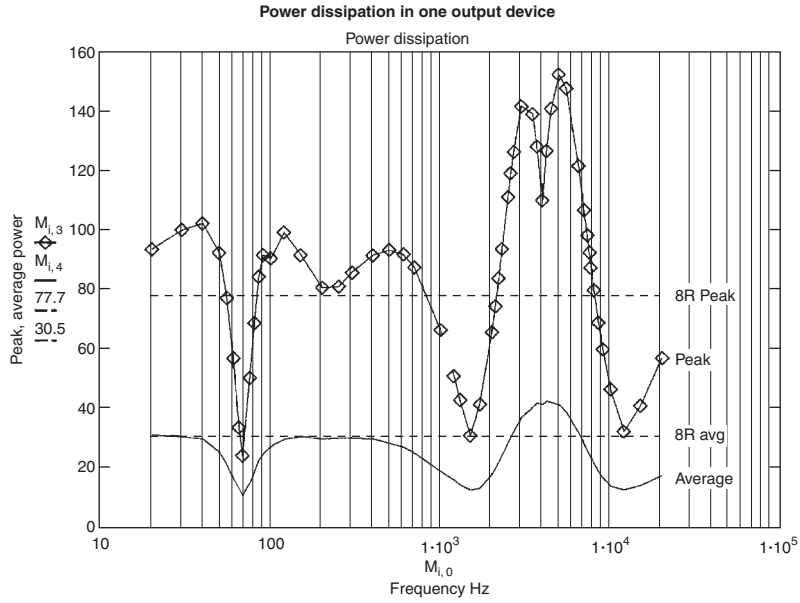


Figure 8.41: Peak and average output device power dissipation driving the two-way speaker model. Dotted lines are peak and average for 8 Ω

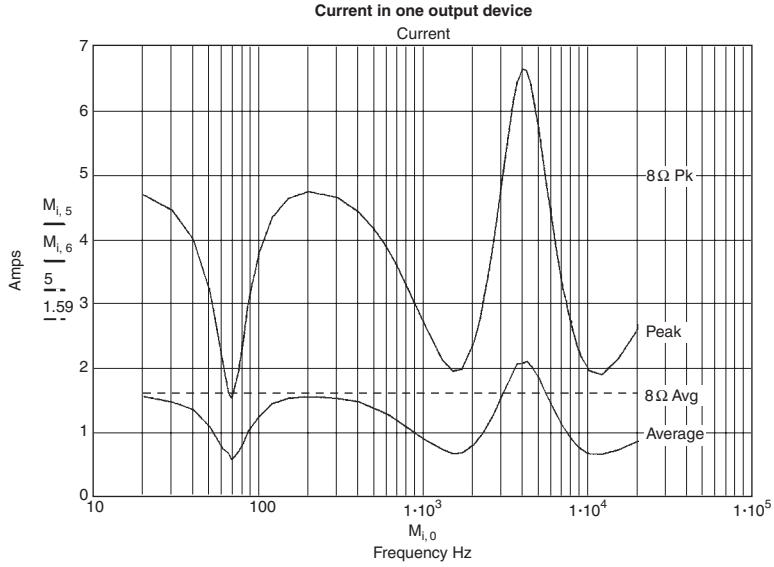


Figure 8.42: Peak and average output device current driving two-way speaker impedance as in Figure 8.34. Dotted lines are peak and average for 8 Ω

for brief periods. If you attempt to exploit these areas in an audio application, you are living very dangerously, as the longest excursion specified is usually 5 ms, and a half-cycle at 20 Hz lasts for 25 ms.

From this it can be concluded that a truly 'difficult' load impedance is one with lots of small humps and dips giving significant phase shifts and increased peak dissipation across most of the audio

band. Impedance dips cause more stress than peaks, as might be expected. Low impedances at the high-frequency end (above 5 kHz) are particularly undesirable as they will increase amplifier crossover distortion.

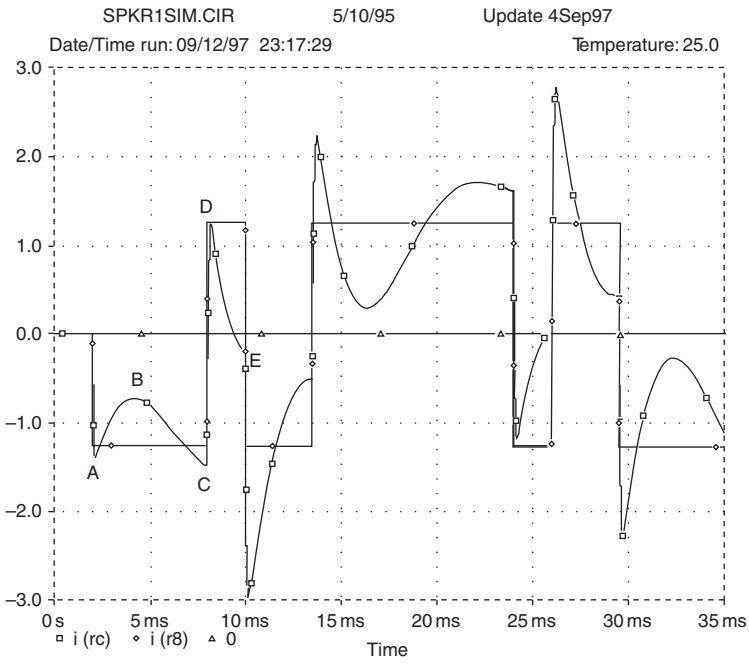


Figure 8.43: An asymmetrical waveform to generate enhanced speaker currents. The sequence ABCDE generates a negative current spike; to the right, the inverse sequence produces a positive spike. The rectangular waveform is the current through an 8 Ω resistive load

Enhanced Loudspeaker Currents

When amplifier current capability and loudspeaker loading are discussed it is often said that it is possible to devise special waveforms that cause a loudspeaker to draw more transient current than would at first appear to be possible. This is perfectly true. The issue was raised by Otala et al.[15], and expanded on in Otala and Huttunen[16]. The effect was also demonstrated by Cordell[17].

The effect may be demonstrated with the electrical analog of a single-speaker unit as shown in Figure 8.30. R_c is the resistance of the voice-coil and L_c its inductance. L_r and C_r model the cone resonance, with R_r controlling its damping. These three components simulate the impedance characteristics of the real electromechanical resonance. The voice-coil inductance is 0.29 mH and its resistance 6.8 Ω, typical for a 10 inch bass unit of 8 Ω nominal impedance. Measurements on this circuit cannot show an impedance below 6.8 Ω at any frequency, and it is easy to assume that the current demands can therefore never exceed those of a 6.8 Ω resistance. This is not so.

The secret of getting unexpectedly high currents flowing is to make use of the energy stored in the circuit reactances. This is done by applying an asymmetrical waveform with transitions carefully timed to match the speaker resonance. Figure 8.43 shows PSPICE simulation of the currents drawn

by the circuit of Figure 8.30. The rectangular waveform is the current in a reference 8 Ω resistance driven with the same waveform. A ±10 V output limit is used here for simplicity but this could obviously be much higher, depending on the amplifier rail voltages.

At the start of the waveform at A, current flows freely into C_r, reducing to B as the capacitance charges. Current is also slowly building up in L_r, causing the total current drawn to increase again to C. A positive transition to the opposite output voltage then takes the system to point D; this is not the same state as at A because energy has been stored in L_r during the long negative period.

A carefully timed transition is then made at E, at the lowest point in this part of the curve. The current change is the same amplitude as at D, but it starts off from a point where the current is already negative, so the final peak goes much lower to 2.96 A, 2.4 times greater than that drawn by the 8 Ω resistor. I call this the current timing factor, or CTF.

Otala and Huttunen[16] show that the use of multi-way loudspeakers, and more complex electrical models, allows many more degrees of freedom in maximizing the peak current. They quote a worst case CTF of 6.6 times. An amplifier driving 50 W into 8 Ω must supply a peak current into an 8 Ω resistance of 3.53 A; amplifiers are usually designed to drive 4 Ω or lower to allow for impedance dips and this means the peak current capability must be at least 7.1 A. However, a CTF of 6.6 implies that the peak capability should be at least 23 A. This peak current need only be delivered for less than a millisecond, but it could complicate the design of protection circuitry.

The vital features of the provocative waveform are the fast transitions and their asymmetrical timing. The optimal transition timing for high currents varies with the speaker parameters. The waveform in Figure 8.43 uses ramped transitions lasting 10 μs; if these transitions are made longer the peak currents are reduced. There is little change up to 100 μs, but with transitions lengthened to 500 μs the CTF is reduced from 2.4 to 2.1.

Without doing an exhaustive survey, it is impossible to know how many power amplifiers can supply six times the nominal peak current required. I suspect there are not many. Is this therefore a neglected cause of real audible impairment? I think not, because:

1. Music signals do not contain high-level rectangular waveforms, nor trapezoidal approximations to them. A useful investigation would be a statistical evaluation of how often (if ever) waveforms giving significant peak current enhancement occur. As an informal test, I spent some time staring at a digital scope connected to general-purpose rock music, and saw nothing resembling the test waveform. Whether the asymmetrical timings were present is not easy to say; however, the large-amplitude vertical edges were definitely not.
2. If an amplifier does not have a huge current-peak capability, then the overload protection circuitry will hopefully operate. If this is of a non-latching type that works cleanly, the only result will be rare and very brief periods of clipping distortion when the loudspeaker encounters a particularly unlucky waveform. Such infrequent transient distortion is known to be inaudible and this may explain why the current enhancement effect has attracted relatively little attention so far.

Amplifier Instability

Amplifier instability can be one of the more challenging areas of design. Instability can refer to unwanted oscillations at either HF or LF, but the latter is very rare in solid-state amplifiers, though still very much an issue for valve designers. Instability has to be taken very seriously, because it may not only destroy the amplifier that hosts it, but also damage the loudspeakers.

Instability at middle frequencies such as 1 kHz is virtually impossible unless you have a very eccentric design with roll-offs and phase shifts in the middle of the audio band.

HF Instability

HF instability is probably the most difficult problem that may confront the amplifier designer, and there are several reasons for this:

1. The most daunting feature of HF oscillation is that under some circumstances it can cause the destruction of the amplifier in relatively short order. It is often most inadvisable to let the amplifier sit there oscillating while you ponder its shortcomings.

BJT amplifiers will suffer overheating because of conduction overlap in the output devices; it takes time to clear the charge carriers out of the device junctions. Some designs deal with this better than others, but it is still true that subjecting a BJT design to prolonged sine-wave testing above 20 kHz should be done with great caution. Internal oscillations may of course have much higher frequencies than this, and in some cases the output devices may be heated to destruction in a few seconds. The resistor in the Zobel network will probably also catch fire.

FET amplifiers are less vulnerable to this overlap effect, due to their different conduction mechanism, but show a much greater tendency to parasitic oscillation at high frequencies, which can be equally destructive. Under high-amplitude oscillation plastic-package FETs may fail explosively; this is usually a prompt failure within a second or so and leaves very little time to hit the off switch.

2. Various subsections of the amplifier may go into oscillation on their own account, even if the global feedback loop is stable against Nyquist oscillation. Even a single device may go into parasitic oscillation (e.g. emitter-followers fed from inappropriate source impedances) and this is usually at a sufficiently high frequency that it either does not fight its way through to the amplifier output, or does not register on a 20 MHz scope. The presence of this last kind of parasitic is usually revealed by excessive and unexpected nonlinearity.
3. Another problem with HF oscillation is that it cannot in general be modeled theoretically. The exception to this is global Nyquist oscillation (i.e. oscillation around the main feedback loop because the phase shift has become too great before the loop gain has dropped below unity), which can be avoided by calculation, simulation, and design. The forward-path gain and the dominant-pole frequency are both easy to calculate, though the higher pole frequencies that cause phase shift to accumulate are usually completely mysterious; to the best of my knowledge virtually no work has been done on the frequency response of audio amplifier

output stages. Design for Nyquist stability therefore reduces to deciding what feedback factor at 20 kHz will give reliable stability with various resistive and reactive loads, and then apportioning the open-loop gain between the transconductance of the input stage and the transresistance of the VAS.

The other HF oscillations, however, such as parasitics and other more obscure oscillatory misbehavior, seem to depend on various unknown or partly known second-order effects that are difficult or impossible to deal with quantitatively and are quite reasonably left out of simulator device models. This means we are reduced to something not much better than trial and error when faced with a tricky problem.

The CFP output stage has two transistors connected together in a very tight 100% local feedback loop, and there is a clear possibility of oscillation inside this loop. When it happens, this tends to be benign, at a relatively high frequency (say 2–10 MHz) with a clear association with one polarity of half-cycle.

LF Instability

Amplifier instability at LF (motorboating) is largely a thing of the past now that amplifiers are almost invariably designed with DC-coupling throughout the forward and feedback paths. The theoretical basis for it is exactly as for HF Nyquist oscillation; when enough phase shift accumulates at a given frequency, there will be oscillation, and it does not matter if that frequency is 1 Hz or 1 MHz. It can be as destructive of bass drivers as HF oscillation is of tweeters, especially with bass reflex designs that impose no cone loading at subsonic frequencies.

At LF things are actually easier, because all the relevant time-constants are known, or can at least be pinned down to a range of values based on electrolytic capacitor tolerances, and so the system is designable, which is far from the case at high frequencies. The techniques for dealing with almost any number of LF poles and zeros were well known in the valve era, when AC-coupling between stages was usually unavoidable, because of the large DC voltage difference between the anode of one stage and the grid of the next.

The likeliest cause of LF instability is probably a misdesigned multi-pole DC servo (see Chapter 16 for more on this). Oscillation at LF is very unlikely to be provoked by awkward load impedances. This is not true at HF, where a capacitive load can cause serious instability. However, this problem at least is easily handled by adding an output inductor.

Speed and Slew Rate in Audio Amplifiers

It seems self-evident that a fast amplifier is a better thing to have than a slow one, but what is a fast amplifier? Closed-loop bandwidth is not a promising yardstick; it is virtually certain that any power amplifier employing negative feedback will have a basic closed-loop frequency response handsomely in excess of any possible aural requirements, even if the overall system bandwidth is defined at a lower value by earlier filtering.

There is always a lot of loose talk about the importance of an amplifier's open-loop bandwidth, much of it depressingly ill-informed. I demonstrated[18] that the frequency of the dominant pole *P*1 that sets the open-loop bandwidth is a variable and rather shifty quantity that depends on transistor beta and other ill-defined parameters. (I also showed how it can be cynically manipulated to make it higher by reducing open-loop gain below *P*1.) While *P*1 may vary, the actual gain at HF (say 20 kHz) is thankfully a much more dependable figure that is set only by frequency, input stage transconductance, and the value of C_{dom}. It is this which is the meaningful figure in describing the amount of NFB that an amplifier enjoys.

The most meaningful definition of an amplifier's 'speed' is its maximal slew rate. The minimum slew rate for a 100 W/8 Ω amplifier to cleanly reproduce a 20 kHz sine wave is easily calculated as 5.0 V/μs, so 10 V/μs is adequate for 400 W/8 Ω, a power level that takes us somewhat out of the realms of domestic hi-fi. A safety margin is desirable, and if we make this a bare factor of 2 then it could be logically argued that 20 V/μs is enough for any hi-fi application; there is in fact a less obvious but substantial safety margin already built in, as 20 kHz signals at maximum level are mercifully rare in music; the amplitude distribution falls off rapidly at higher frequencies.

Firm recommendations on slew rate are not common; Peter Baxandall made measurements of the slew rate produced by vinyl disk signals, and concluded that they could be reproduced by an amplifier with a slew limit corresponding to maximum output at 2.2 kHz. For the 100 W amplifier this corresponds to 0.55 V/μs[19].

Nelson Pass made similar tests, with a moving-magnet (MM) cartridge, and quoted a not dissimilar maximum of 1 V/μs at 100 W. A moving-coil (MC) cartridge doubled this to 2 V/μs, and Pass reported[20] that the absolute maximum, with an MM cartridge, and possible with a combination of direct-cut disks and MC cartridges, was 5 V/μs at 100 W. This is comfortably below the 20 V/μs figure arrived at theoretically above; Pass concluded that even if a generous 10:1 factor of safety was adopted, 50 V/μs would be the highest speed ever required from a 100 W amplifier.

However, in the real world we must also consider the numbers game; if all else is equal then the faster amplifier is the more saleable. As an example of this, it has been recently reported in the hi-fi press that a particular 50 W/8 Ω amplifier has been upgraded from 20 to 40 V/μs[21], and this is clearly expected to elicit a positive response from intending purchasers. This report is exceptional, for equipment reviews in the hi-fi press do not usually include slew-rate measurements. It is therefore difficult to get a handle on the state of the art, but a trawl through the accumulated data of years shows that the most highly specified equipment usually plumps for 50 V/μs – slew rates always being quoted in suspiciously round numbers. There was one isolated claim of 200 V/μs, but I must admit to doubts about the reality of this.

The Class-B amplifier shown in Figure 8.9 is that already described in Chapter 7; the same component numbers have been preserved. This generic circuit has many advantages, though an inherently good slew performance is not necessarily one of them. However, it remains the basis for the overwhelming majority of amplifiers so it seems the obvious place to start. I have glibly stated that its slew rate calculated at 40 V/μs, which by the above arguments is more than adequate. However, let us assume that a major improvement in slew rate is required to counter the

propaganda of the Other Amplifier Company down the road, and examine how it might be done. As in so many areas of life, things will prove much more complicated than expected.

The Basics of Amplifier Slew-Limiting

At the simplest level, slew rate in a conventional amplifier configuration like Figure 8.9 depends on getting current in and out of C_{dom} (C3) with the convenient relation:

$$\text{Slew rate} = \frac{I}{C_{dom}} \text{ V/}\mu\text{s, for } I \text{ in A and } C_{dom} \text{ in pF} \qquad \text{Equation 8.1}$$

The maximum output frequency for a given slew rate (*SR*) and voltage is:

$$\text{Freq. max} = \frac{SR}{2 \times \pi \times V_{pk}} = \frac{SR}{2 \times \pi \times \sqrt{2} \times \text{V rms}} \qquad \text{Equation 8.2}$$

So, for example, with a slew rate of 20 V/μs the maximum frequency at which 35 V rms can be sustained is 64 kHz, and if C_{dom} is 100 pF then the input stage must be able to source and sink 2 mA peak. Likewise, a sine wave of given amplitude and frequency has a maximum slew rate (at zero-crossing) of:

$$\begin{aligned} SR \text{ of sine wave} &= \frac{dV}{dT} = \varphi_{max} V_{pk} \\ &= 2 \times \pi \times \text{freq.} \times V_{pk} \end{aligned} \qquad \text{Equation 8.3}$$

For Figure 8.9, our slew-rate equation yields 4000/100, or about 40 V/μs, as quoted above, if we assume (as all textbooks do) that the only current limitation is the tail source of the input pair. If this differential pair has a current-mirror collector load – and there are pressing reasons why it should – then almost the full tail current is available to service C_{dom}. This seems very simple – to increase slew rate, increase the tail current. But . . .

The tail current is not the only limit on the slew current in C_{dom}. (This point was touched on by me in Ref. [22].) Figure 8.44 shows the current paths for positive and negative slew limit, and it can be seen at once that the positive current can only be supplied by the VAS current-source load. This will reduce the maximum positive rate, causing slew asymmetry, if the VAS current source cannot supply as much current as the tail source. In contrast, for negative slewing TR4 can turn on as much as required to sink the C_{dom} current, and the VAS collector load is not involved.

In most designs the VAS current-source value does not appear to be an issue, as the VAS is run at a higher current than the input stage to ensure enough pull-up current for the top half of the output stage; however, it will transpire that the VAS source can still cause problems.

Slew-Rate Measurement Techniques

Directly measuring the edge slopes of fast square waves from a scope screen is not easy, and without a delayed timebase it is virtually impossible. A much easier (and far more accurate)

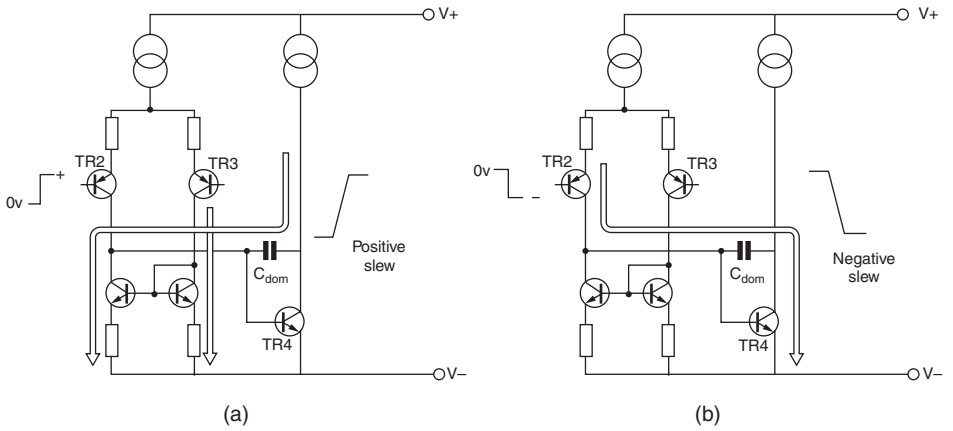


Figure 8.44: (a) The current path for positive slewing. At the limit all of the slewing current has to pass through the current-mirror, TR2 being cut off. (b) The current path at negative slew limit. TR2 is saturated and the current-mirror is cut off

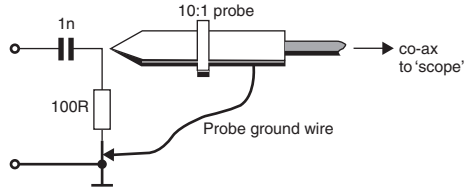


Figure 8.45: A simple (but very useful) differentiator. A local probe ground is essential for accuracy to exceed ±10%

method is to pass the amplifier output through a suitably scaled differentiator circuit; slew rate then becomes simple amplitude, which is much easier to read from a graticule. The circuit in Figure 8.45 gives a handy 100 mV output for each V/μs of slew; the RC time-constant must be very short for reasonable accuracy. The differentiator was driven directly by the amplifier, and *not* via an output inductor. Be aware that this circuit needs to be coupled to the scope by a proper ×10 probe; the capacitance of plain screened cable gives serious under-readings. We are dealing here with sub-microsecond pulse techniques, so bear in mind that waveform artefacts such as ringing are as likely to be due to test cabling as to the amplifier.

Applying a fast-edged square wave to an amplifier does not guarantee that it will show its slew-rate limits. If the error voltage so generated is not enough to saturate the input stage then the output will be an exponential response, without nonlinear effects. For most of the tests described here, the amplifier had to be driven hard to ensure that the true slew limits were revealed; this is due to the heavy degeneration that reduces the transconductance of the input pair. Degeneration increases the error voltage required for saturation, but does not directly alter slew limits.

Running a slew test on the amplifier of Figure 8.9, with an 8 Ω load, sharply highlights the inadequacies of simple theory. The differentiator revealed asymmetrical slew rates of +21 V/μs up and −48 V/μs down, which is both a letdown and a puzzle considering that the simple theory

promises 40 V/μs. To get results worse than theory predicts is merely the common lot of the engineer; to simultaneously get results that are *better* is grounds for the gravest suspicions.

Improving the Slew Rate

Looking again at Figure 8.9, the VAS current-source value is apparently already bigger than required to source the current C_{dom} requires when the input stage is sinking hard, so we confidently decrease R4 to 100R (to match R13) in a plausible attempt to accelerate slewing. With considerable disappointment we discover that the slew rate only changes to +21 V/μs, −62 V/μs; the negative rate still exceeds the new theoretical value of 60 V/μs. Just what is wrong here? Honesty compels us to use the lower of the two figures in our ads (doesn't it?) and so the priority is to find out why the positive slewing is so feeble.

At first it seems unlikely that the VAS current source is the culprit, as with equal-value R4 and R13, the source should be able to supply all the input stage can sink. Nonetheless, we can test this cherished belief by increasing the VAS source current while leaving the tail current at its original value. We find that R4 = 150R, R13 = 68R gives +23 V/μs, −48 V/μs, and this small but definite increase in positive rate shows clearly there is something non-obvious going on in the VAS source.

(This straightforward method of slew acceleration by increasing standing currents means a significant increase in dissipation for the VAS and its current source. We are in danger of exceeding the capabilities of the TO92 package, leading to a cost increase. The problem is less in the input stage, as dissipation is split between at least three devices.)

Simulating Slew-Limiting

When circuits turn truculent, it's time to simplify and simulate. The circuit was reduced to a *model* amplifier by replacing the Class-B output stage with a small-signal Class-A emitter-follower; this was then subjected to some brutally thorough PSPICE simulation, which revealed the various mechanisms described below.

Figure 8.46 shows the positive-going slew of this model amplifier, with both the actual output voltage and its differential, the latter suitably scaled by dividing by 10^6 so it can be read directly in V/μs from the same plot. Figure 8.47 shows the same for the negative-going slew. The plots are done for a series of changes to the resistors R4, R13 that set the standing currents.

Several points need to be made about these plots; first, the slew rates shown for the lower R4, R13 values are not obtainable in the real amplifier with output stage, for reasons that will emerge. Note that almost imperceptible wobbles in the output voltage put large spikes on the plot of the slew rate, and it is unlikely that these are being simulated accurately, if only because circuit strays are neglected. To get valid slew rates, read the flat portions of the differential plots.

Using this method, the first insight into slew-rate asymmetry was obtained. At audio frequencies, a constant current-source provides a fairly constant current and that is the end of the matter, making it the usual choice for the VAS collector load; as a result its collector is exposed to the full output swing and the full slew rate. When an amplifier slews rapidly, there is a transient feed-through from

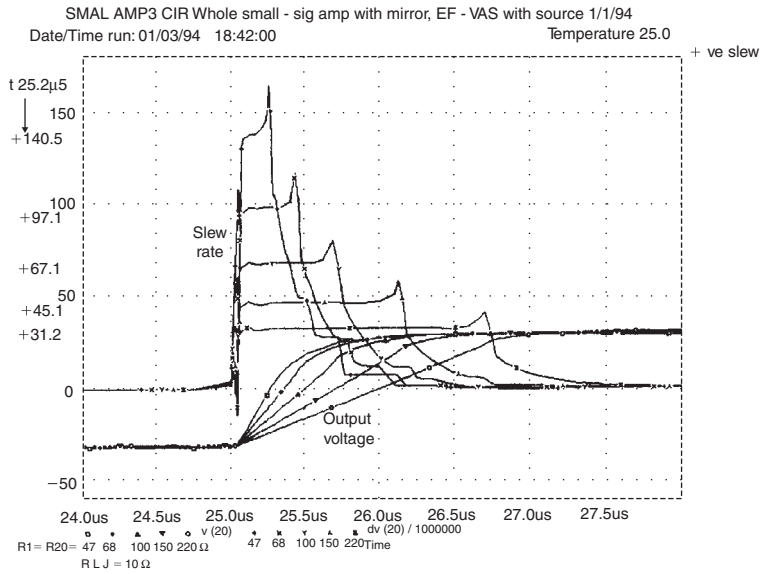


Figure 8.46: Positive slewing of simulated model amplifier. The lower traces show the amplifier output slewing from −30 to +30 V while the upper traces are the scaled differentiation

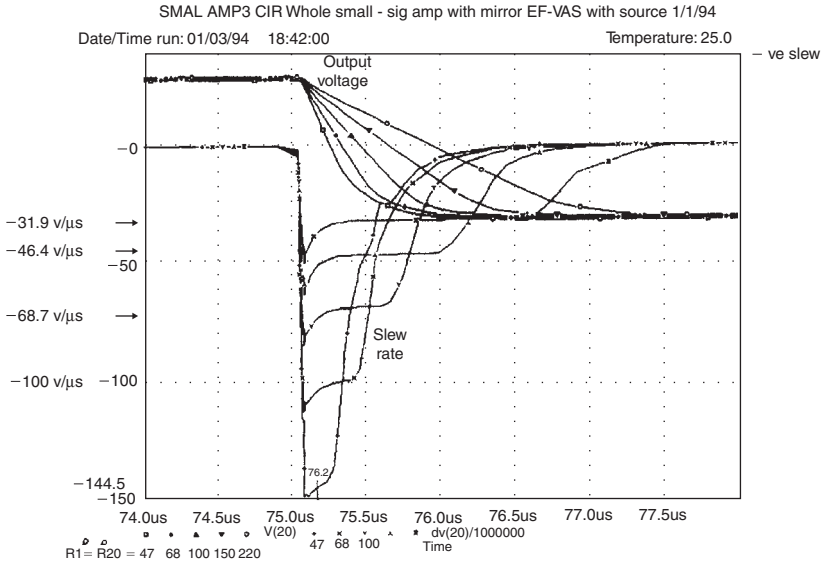


Figure 8.47: Negative slewing of simulated model amplifier. Increasing the slew-rate limit causes a larger part of the output transient to become exponential, as the input pair spends less time saturated. Thus the differential trace has a shorter flat period

the collector to the base (see Figure 8.48) via the collector–base capacitance. If the base voltage is not tightly fixed then fast positive slewing drives the base voltage upwards, reducing the voltage on the emitter and hence the output current. Conversely, for negative slew the current-source output briefly increases (see Erdi[23]). In other words, fast positive slewing itself reduces the current available to implement it.

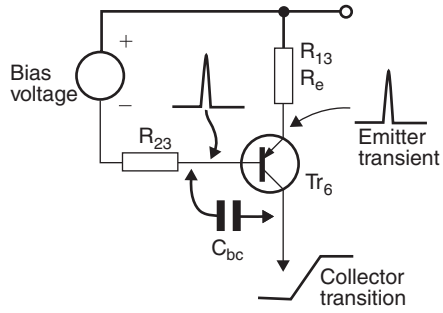


Figure 8.48: One reason why simple theory fails. Fast positive edges on the collector of the VAS source TR6 couple through the internal C_{bc} to momentarily reduce standing current

Having discovered this hidden constraint, the role of isolation resistor R23 feeding TR5 base immediately looks suspect. Simulation confirms that its presence worsens the feed-through effect by increasing the impedance of the reference voltage fed to TR5 base. As is usual, the input stage tail source TR1 is biased from the same voltage as TR5; this minor economy complicates things significantly, as the tail current also varies during fast transients, reducing for positive slew and increasing for negative.

Slewing Limitations in Real Life

Bias isolation resistors are not unique to the amplifier of Figure 8.9; they are very commonly used. For an example taken at random, see Meyer[24]. My own purpose in adding R23 was not to isolate the two current sources from each other at AC (something it utterly fails to do) but to aid fault-finding. Without this resistor, if the current in either source drops to zero (e.g. if TR1 fails open-circuit) then the reference voltage collapses, turning off both sources, and it can be time-consuming to determine which has died and which has merely come out in sympathy. Accepting this, we return to the original Figure 8.9 values and replace R23 with a link; the measured slew rates at once improve from +21, −48 to +24, −48 (from here on the V/μs is omitted). This is already slightly faster than our first attempt at acceleration, without the thermal penalties of increasing the VAS standing current.

The original amplifier used an active tail source, with feedback control by TR14; this was a mere whim, and a pair of diodes gave identical THD figures. It seems likely that reconfiguring the two current sources so that the VAS source is the active one would make it more resistant to feed-through, as the current-control loop is now around TR5 rather than TR1, with feedback applied directly to the quantity showing unwanted variations (see Figure 8.49). There is indeed some improvement, from +24, −48 to +28, −48.

This change seems to work best when the VAS current is increased, and R4 = 100 Ω, R13 = 68 Ω now gives us +37, −52, which is definite progress on the positive slewing. The negative rate has also slightly increased, indicating that the tail current is still being increased by feed-through effect.

It seems desirable to minimize this transient feed-through, as it works against us just at the wrong time. One possibility would be a cascode transistor to shield TR5 collector from rapid voltage

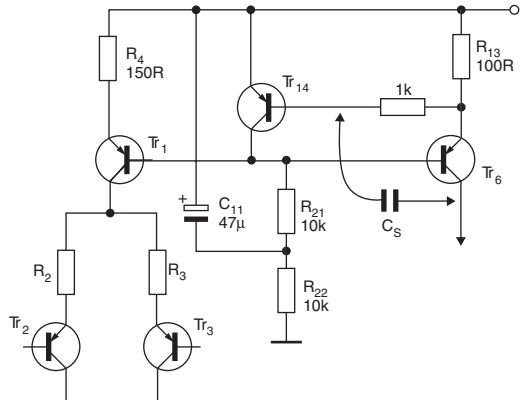


Figure 8.49: A modified biasing system that makes TR6 current the controlled variable and reduces the feed-through effect

changes; this would require more biasing components and would reduce the positive output swing, albeit only slightly.

Since it is the VAS current-source feed-through capacitance that causes so much grief, can we turn it against itself, so that an abrupt voltage transition increases the current available to sustain it, rather than reducing it? Oh yes we can, for if a small capacitance C_s is added between TR5 collector (carrying the full voltage swing) and the sensing point A of the active tail source, then as the VAS collector swings upward, the base of TR14 is also driven positive, tending to turn it off and hence increasing the bias applied to VAS source TR5 via R21. This technique is highly effective, but it smacks of positive feedback and should be used with caution; C_s must be kept small. I found 7.5 pF to be the highest value usable without degrading the amplifier's HF stability.

With R4 = 100 Ω, R13 = 68 Ω adding C_s = 6 pF takes us from +37, −52 to +42, −43, and the slew asymmetry that has dogged this circuit from the start has been corrected. Fine adjustment of this capacitance is needful if good slew symmetry is demanded.

Some Additional Complications

Some other unsuspected effects were uncovered in the pursuit of speed; it is not widely known that slew rate is affected both by output loading and the output stage operating class. For example, above we have noted that R4 = 100 Ω, R13 = 68 Ω yields +37, −52 for Class-B and an 8 Ω load. With 4 Ω loading this changes to +34, −58, and again the loss in positive speed is the most significant. If the output stage is biased into Class-A (for an 8 Ω load) then we get +35, −50. The explanation is that the output stage, despite the cascading of drivers and output devices, draws significant current from the VAS stage. The drivers draw enough base current in the 4 Ω case to divert extra current from C_{dom} and current is in shortest supply during positive slew. The effect in Class-A is more severe because the output device currents are always high, the drivers requiring more base current even when quiescent, and again this will be siphoned off from the VAS collector.

Speeding up this amplifier would be easier if the Miller capacitor C_{dom} was smaller. Does it really need to be that big? Well yes, because if we want the NFB factor to be reasonably low for dependable HF stability, the HF loop gain must be limited. Open-loop gain above the dominant-pole frequency $P1$ is the product of input stage g_m with the value of C_{dom}, and the g_m is already as low as it can reasonably be made by emitter degeneration. Emitter resistors R2, R3 at 100 Ω are large enough to mildly compromise the input offset voltage, because the tail current splits in two through a pair of resistors that are unlikely to be matched to better than 1%, and noise performance is also impaired by this extra resistance in the input pair emitters. Thus for a given NFB factor at 20 kHz, C_{dom} is fixed.

Despite these objections, the approach was tested by changing the distribution of open-loop gain between the input stage and the VAS. R2, R3 were increased from 100 to 220 Ω, and C_{dom} reduced to 66 pF; this does not give exactly the same NFB factor, but in essence we have halved the transconductance of the input stage, while doubling the gain of the VAS. This gain-doubling allows C_{dom} to be reduced to 66 pF without reduction of stability margins.

With R4 = 100 Ω, R13 = 68 Ω as before, the slew rate is increased to +50, −50 with C_s = 6 pF to maintain slewing symmetry. This is a 25% increase in speed rather than the 50% that might be expected from simple theory, and indicates that other restrictions on speed still exist; in fact PSPICE showed there are several.

One of these restrictions is as follows: when slewing positively, TR4 and TR12 must be turned off as quickly as possible, by pulling current out of C_{dom}. The input pair therefore causes TR10 to be turned on by an increasing voltage across TR11 and R7. As TR10 turns on, its emitter voltage rises due to R6, while at the same time the collector voltage must be pulled down to near the negative rail to turn off Q4. In the limit TR10 runs out of V_{ce}, and is unable to pull current out of C_{dom} fast enough. The simplest way to reduce this problem is to reduce the resistors R6, R7 that degenerate the current-mirror. This risks HF distortion variations due to input-pair I_c imbalance, but values down to 12 Ω have given acceptable results. Once more it is the positive rate that suffers.

Another way to reduce the value needed for C_{dom} is to lower the loop gain by increasing the feedback network attenuation or, in other words, to run the amplifier at a higher closed-loop gain. This might be no bad thing; the current 'standard' of 1 V for full output is (I suspect) due to a desire for low closed-loop gain in order to maximize the NFB factor, so reducing distortion. I recall John Linsley-Hood advocating this strategy back in 1974. However, we must take the world as we find it, and so I have left closed-loop gain alone. We could of course attenuate the input signal so it can be amplified more, though I have an uneasy feeling about this sort of thing; amplifying in a pre-amp then attenuating in the power amp implies a headroom bottleneck, if such a curdled metaphor is permissible. It might be worth exploring this approach; this amplifier has good open-loop linearity and I do not think excessive THD would be a problem.

Having previously spent some effort on minimizing distortion, we do not wish to compromise the THD of a Blameless amplifier. Mercifully, none of the modifications set out here have any significant effect on overall THD, though there may be minor variations around 10–20 kHz.

Further Improvements and Other Configurations

The results I have obtained in my attempts to improve slewing are not exactly stunning at first sight; however, they do have the merit of being as grittily realistic as I can make them. I set out in the belief that enhancing slew rate would be fairly simple; the very reverse has proved to be the case. It may well be that other VAS configurations, such as the push–pull VAS, will prove more amenable to design for rapid slew rates; however, such topologies appear to have other disadvantages to overcome.

Stochino, in a fascinating paper[25], has presented a topology that, although a good deal more complex than the conventional arrangement, claims to make slew rates up to 400 V/μs achievable.

References

[1] M. Otala, An audio power amplifier for ultimate quality requirements, IEEE Trans. Audio and Electroacoustics AU-21 (6) (December 1973).

[2] P. Baxandall, Audio power amplifier design: Part 4, Wireless World (July 1978) p. 76.

[3] S. Takahashi et al., Design and construction of high slew-rate amplifiers, AES 60th Convention, Preprint No. 1348 (A-4), 1978.

[4] D. Self, Crossover distortion and compensation (Letters), Electronics & Wireless World (August 1992) p. 657.

[5] R. Widlar, A monolithic power op-amp, IEEE J. Solid-State Circuits 23 (2) (April 1988).

[6] J. Linsley-Hood, Solid-state audio power, Electronics & Wireless World (November 1989) p. 1047.

[7] M. Leach, Feedforward compensation of the amplifier output stage for improved stability with capacitive loads, IEEE Trans. Consumer Electronics 34 (2) (May 1988).

[8] Fast compensation extends power bandwidth, Linear Brief 4, in: National Semiconductor Linear Applications Handbook, 1991.

[9] D. Feucht, Handbook of Analog Circuit Design, Academic Press, 1990, p. 264.

[10] P. Baxandall, Private communication, 1995.

[11] S.P. Pernici et al., A CMOS low-distortion amplifier with double-nested Miller compensation, IEEE J. Solid-State Circuits (July 1993) p. 758.

[12] J. Scott, G. Spears, On the advantages of nested feedback loops, JAES 39 (March 1991) p. 115.

[13] J. Atkinson, Review of Krell KSA-50S power amplifier, Stereophile (August 1995) p. 168.

[14] E. Benjamin, Audio power amplifiers for loudspeaker loads, JAES 42 (September 1994) p. 670.

[15] M. Otala et al., Input current requirements of high-quality loudspeaker systems, AES 73rd Convention, Preprint No. 1987 (D-7), March 1983.

[16] M. Otala, P. Huttunen, Peak current requirement of commercial loudspeaker systems, JAES (June 1987) 455. See Chapter 12, p. 294.

[17] R. Cordell, Interface intermodulation in amplifiers, Wireless World (February 1983) p. 32.

[18] D. Self, Distortion in power amplifiers, Part 1, Electronics & Wireless World (August 1993) p. 631.

[19] P. Baxandall, Audio power amplifier design, Wireless World (January 1978) p. 56.

[20] N. Pass, Linearity, slew rates, damping, stasis and . . ., Hi-fi News & RR (September 1983) p. 36.

[21] J. Hughes, Arcam Alpha5/Alpha6 amplifier review, Audiophile (January 1994) p. 37.

[22] D. Self, Distortion in power amplifiers, Part 7, Electronics & Wireless World (February 1994) p. 138.

[23] G. Erdi, A 300 V/μs monolithic voltage follower, IEEE J. Solid-State Circuits (December 1979) p. 1062.

[24] D. Meyer, Assembling a universal tiger, Popular Electronics (October 1970).

[25] G. Stochino, Ultra-fast amplifier, Electronics & Wireless World (October 1995) p. 835.

CHAPTER 9

Power Supplies and PSRR

'. . . my power is made perfect . . .'

2 Corinthians 12:9

Power-Supply Technologies

There are three principal ways to power an amplifier:

1. a simple unregulated power supply consisting of transformer, rectifiers, and reservoir capacitors;
2. a linear regulated power supply;
3. a switch-mode power supply.

It is immediately obvious that the first and simplest option will be the most cost-effective, but at first glance it seems likely to compromise noise and ripple performance, and possibly interchannel crosstalk. It is therefore worthwhile to examine the pros and cons of each technology in a little more detail. I am here dealing only with the main supply for the actual power amplifier rails. Many amplifiers now have some form of microcontroller to handle on/off switching by mains relays and other housekeeping functions; this is usually powered by a separate small standby transformer, which remains powered when the amplifier supply is switched off. The design of this is straightforward – or at least it was until the introduction of new initiatives to limit the amount of standby power that a piece of equipment is allowed to consume. The International Energy Agency is urging a 1 W standby power limit for all energy-using products.

Simple Unregulated Power Supplies

Advantages

- Simple, reliable, and cheap (relatively speaking – the traditional copper and iron mains transformer will probably be the most expensive component in the amplifier).
- No possibility of instability or HF interference from switch-mode frequencies.
- The amplifier can deliver higher power on transient peaks, which is just what is required.

Disadvantages

- The power into 4 Ω will not be twice that into 8 Ω, because the supply voltage will fall with increased current demand. On the other hand, the amplifier will always deliver the maximum possible power it can.

- Significant ripple is present on the DC output and so the PSRR of the amplifier will need careful attention; the problem is, however, not hard (if you read the second part of this chapter) and output hum levels below −100 dBu are easily attainable.
- The mains transformer will be relatively heavy and bulky.
- Transformer primary tappings must be changed for different countries and mains voltages.
- The absence of switch-mode technology does not mean total silence as regards RF emissions. The bridge rectifier will generate bursts of RF at a 100 Hz repetition rate as the diodes turn off. This worsens with increasing current drawn.

Linear Regulated Power Supplies

Advantages

- A regulated supply-rail voltage means that the amplifier can be made to approximate more closely to a perfect voltage source, which would give twice the power into 4 Ω than it gives into 8 Ω. This is considered to have marketing advantages in some circles, though it is not clear why you would want to operate an amplifier on the verge of clipping. There are, however, still load-dependent losses in the output stage to consider. More on this later.
- A regulated supply-rail voltage to a power amplifier gives absolutely consistent audio power output in the face of mains voltage variation.
- Clipping behavior will be cleaner, as the clipped peaks of the output waveform are not modulated by the ripple on the supply rails. Having said that, if your amplifier is clipping regularly you might consider turning it down a bit.
- Can be designed so that virtually no ripple is present on the DC output (in other words the ripple is below the white noise the regulator generates) allowing relaxation of amplifier supply-rail rejection requirements. However, you can only afford to be careless with the PSRR of the power amp if the regulators can maintain completely clean supply rails in the face of sudden current demands. If not, there will be interchannel crosstalk unless there is a separate regulator for each channel. This means four for a stereo amplifier, making the overall system very expensive.
- The possibility exists of electronic shutdown in the event of an amplifier DC fault, so that an output relay can be dispensed with. However, this adds significant circuitry, and there is no guarantee that a failed output device will not cause a collateral failure in the regulators that leaves the speakers still in jeopardy.

Disadvantages

- Complex and therefore potentially less reliable. The overall amplifier system is at least twice as complicated. The much higher component count must reduce overall reliability, and getting it working in the first place will take longer and be more difficult. For example, consider the circuit put forward by John Linsley-Hood[1]. To regulate the positive and

negative rails for the output stage, this PSU uses 16 transistors and a good number of further parts; a further six transistors are used to regulate the supplies to the small-signal stages. It is without question more complex and more expensive than most power amplifiers.

- If the power amplifier fails, due to an output device failure, then the regulator devices will probably also be destroyed, as protecting semiconductors with fuses is a very doubtful business; in fact it is virtually impossible. The old joke about the transistors protecting the fuse is not at all funny to power-amplifier designers, because this is very often precisely what happens. Electronic overload protection for the regulator sections is therefore essential to avert the possibility of a domino-effect failure, and this adds further complications as it will probably need to be some sort of foldback protection characteristic if the regulator transistors are to have a realistic prospect of survival.
- Comparatively expensive, requiring at least two more power semiconductors, with associated control circuitry and over-current protection. These power devices in turn need heat-sinks and mounting hardware, checking for shorts in production, etc.
- Transformer tappings must still be changed for different mains voltages.
- IC voltage regulators are usually ruled out by the voltage and current requirements, so it must be a discrete design, and these are not simple to make bulletproof. Cannot usually be bought in as an OEM item, except at uneconomically high cost.
- May show serious HF instability problems, either alone or in combination with the amplifiers powered. The regulator output impedance is likely to rise with frequency, and this can give rise to some really unpleasant sorts of HF instability. Some of my worst amplifier experiences have involved (very) conditional stability in such amplifiers.
- The amplifier can no longer deliver higher power on transient peaks.
- The overall power dissipation for a given output is considerably increased, due to the minimum voltage drop through the regulator system.
- The response to transient current demands is likely to be slow, affecting slewing behavior.

Switch-Mode Power Supplies

Advantages

- Has most of the advantages of linear regulated supplies, as listed above.
- Ripple can be considerably lower than for unregulated power supplies, though never as low as a good linear regulator design; 20 mV peak to peak is typical.
- There is no heavy mains transformer, giving a considerable saving in overall equipment weight. This can be important in PA equipment.
- Can be bought in as an OEM item; in fact this is virtually compulsory in most cases as switch-mode design is a specialized job for experts.

- Can be arranged to shut down if the amplifier develops a dangerous DC offset.
- Can be specified to operate properly, and give the same audio output without adjustment, over the entire possible worldwide mains-voltage range, which is normally taken as 90–260 V.

Disadvantages

- Switch-mode supplies are a prolific source of high-frequency interference. This can be extremely difficult to eradicate entirely from the audio output.
- The 100 Hz ripple output is significant, as noted above, and will require the usual PSRR precautions in the amplifiers.
- Much more complex and therefore less reliable than unregulated supplies. Dangerous if not properly cased, as high DC voltage is present.
- The response to transient current demands is likely to be relatively slow.
- Their design is very much a matter for specialists.

On perusing the above list, it seems clear to me that regulated supplies for power amplifiers are a bad thing. Not everyone agrees – see, for example, Linsley-Hood[2]. Unfortunately he did not adduce any evidence to support his case.

The usual claim – in fact it is probably the closest thing to a subjectivist consensus there is – is that linear regulated supplies give 'tighter bass' or 'firmer bass'*;* advocates of this position are always careful not to define 'tighter bass' too closely, so no one can disprove the notion. If the phrase means anything, it presumably refers to changes in the low-frequency transient response; however, since no such changes can be objectively detected, this appears to be simply untrue. If properly designed, all three approaches can give excellent sound, so it makes sense to go for the easiest solution; with the unregulated supply the main challenge is to keep the ripple out of the audio, which will be seen to be straightforward if tackled logically. The linear regulated approach presents instead the challenge of designing not one but two complex negative-feedback systems, close coupled in what can easily become a deadly embrace if one of the partners shows any HF instability. Before everyone runs off with the idea that I am irrevocably prejudiced against supply regulation, I will mention here that the first power amplifier system I ever designed did indeed have regulated power supplies, because at the time I was prepared to believe that it was the only way to achieve a good hum performance. Remarkably, considering that the only test gear I had was an old moving-coil test-meter, it all worked first time and without any misbehavior I could detect. I still have it in the cellar. However, I did take away from the experience the conviction that if the power supplies were more complex than the amplifier, something was wrong with my design philosophy.

The generic amplifier designs examined in this book have excellent supply-rail rejection, and so a simple unregulated supply is perfectly adequate. The use of regulated supplies is definitely unnecessary, and I would recommend strongly against their use. At best, you have doubled the amount of high-power circuitry to be bought, built, and tested. At worst, you could have intractable HF stability problems, peculiar slew-limiting, and some expensive device failures.

A Devious Alternative to Regulated Power Supplies

In the list of the advantages of linear regulated supplies set out above, the one that seems to have most appeal to people is the first. It allows an amplifier to approximate more closely to a perfect voltage source, which would give exactly twice the power into 4Ω than it gives into 8Ω. In the not always rational world of hi-fi, this kind of amplifier behavior is often considered a mark of solid merit, implying that there are huge output stages and heavyweight power supplies that can gracefully handle any kind of loudspeaker demand. I disagree, for the reasons set out above, but let's follow the train of thought for a bit, until it derails.

A regulated supply clearly gives a closer approach to this ideal than an unregulated supply whose voltage will droop when driving the 4Ω load. However, even if the regulated supply is as stiff as a girder of pure unbendium, there will still be load-dependent losses in the output stage that will make the 4Ω output less than twice that into 8Ω. Assume for the moment that we have an amplifier which gives 100W into 8Ω. There will be emitter resistors in the output stage, and the lowest value they are likely to have is 0.1Ω. (There are good reasons why these resistors should be as low as practicable, because this improves linearity as well as efficiency – see Chapter 6.) These resistors are in series with the output and so form a potential divider with the load. Their presence alone, without considering other losses such as increased output device V_{be} values at higher currents, and the wiring resistance, will cause the 4Ω output to be 195.1W rather than 200W. That perfect voltage source is not so easy to make after all.

However, to make a rather ambitious generalization (and all generalizations are of course dangerous) it can be said that the power deficit from this cause is rather less than that due to unregulated supply rails drooping, which can cause twice the loss in terms of watts. This factor depends very much on how big the mains transformer is, how big the reservoir capacitors are (because that affects the depth of the ripple troughs, which is where clipping occurs first) and so on – I said it was a generalization. It is therefore perhaps worthwhile to look a little closer at the regulated supply issue.

I was once faced with this situation: the managing director wanted exact power doubling in a high-power design, but I was less than enthusiastic about trying to make heavy-current regulated power supplies work dependably. Time for some thought. If you accept that there is no problem in making a hum-free amplifier that runs from unregulated and ripply rails – which is emphatically true, as demonstrated in the second half of this chapter – then the function of the regulators is simply to keep part of the supply voltage away from the amplifiers. In effect, the output stage is a giant clipping circuit. So why not do the clipping at the input of the amplifier, where it can be done with a couple of diodes, and go back to an unregulated power supply? The idea is shown in Figure 9.1. The electrical power previously wasted in the regulators is now absorbed by the output devices, perhaps necessitating a bit more heat-sinking, but all the complications of regulators disappear. As with a regulated supply, the clipping will be clean and uncontaminated by ripple – in fact probably cleaner because a small-signal clipping circuit will have no time-constants that may gather unwanted charges during overload. Now you may think that this is cheating – the managing director certainly did, but even he was forced to admit that what I proposed was functionally identical to an amplifier with regulated supplies, and *much* cheaper. However, the idea of deliberately restricting amplifier output – and this new approach simply

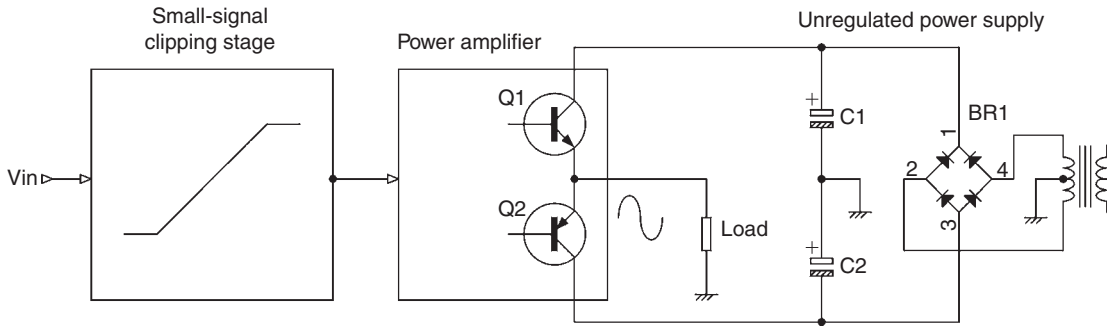


Figure 9.1: Putting a small-signal clipping circuit at the amplifier input to emulate a regulated power supply

makes it obvious that that is what regulated supplies do – did not appeal to him any more than it does to me, and the project went forward with unregulated supplies. And no hum.

In the foregoing argument there is one point that has been oversimplified a little. Making a small-signal clipping circuit is straightforward. Making a clipping circuit that is wholly distortion-free below the clipping point is anything but straightforward. As I described in Chapter 2, it can be done, with some non-obvious circuitry. You will, I hope, forgive me for not revealing it at the moment, but I rather hope that someone might buy the idea off me.

Design Considerations for Power Supplies

A typical unregulated power supply is shown in Figure 9.2. This is wholly conventional in concept, though for optimal hum performance the wiring topology and physical layout need close attention, and this point is rarely made.

In a multichannel amplifier, the power supply will fall into one of three types. In order of increasing cost, and allegedly decreased interaction between channels, these are:

1. The transformer, rectifiers, and reservoir capacitors are shared between channels.
2. Each channel has its own transformer secondary, rectifiers, and reservoirs. There is a single transformer but only the core and primary are shared.
3. Each channel has its own transformer, rectifiers, and reservoirs. Nothing except possibly the mains inlet and mains switch are shared.

In reality the only interaction experienced with (1) and (2) is a variation in maximum power output depending on how the other channels are loaded. With competent design signal crosstalk via the power supply should simply not happen.

For amplifiers of moderate power the total reservoir capacitance per rail usually ranges from 4700 to 20,000 μF, though some designs have much more. Ripple current ratings must be taken seriously, for excessive ripple current heats up the capacitors and reduces their lifetime. It is often claimed that large amounts of reservoir capacitance give ‘firmer bass’, presumably following the same sort

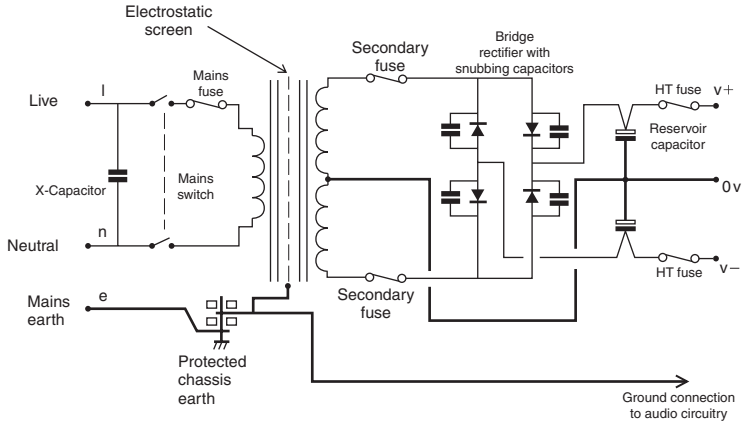


Figure 9.2: A simple unregulated power supply, including rectifier-snubbing and X-capacitor

of vague thinking that credits regulated power supplies with giving 'firmer bass', but it is untrue for all normal amplifier designs below clipping.

I do not propose to go through the details of designing a simple PSU at this point, because such information can be found in standard textbooks, but I instead offer below some hints and warnings that are either rarely published or are especially relevant to audio amplifier design.

Mains Transformers

The mains transformer will normally be either the traditional E-and-I frame type, or a toroid. The frame type is used where price is more important than compactness or external field, and vice versa. There are various other types of transformer, such as C-core or R-core, but they do not seem to be able to match the low external field of the toroid, while being significantly more expensive than the frame type.

The procurement of the mains transformer for a given voltage at a given current is simple in principle, but the field of audio power amplifiers always seems to involve a degree of trial and error. This is because when transformers are used in unregulated power supplies for audio power amplifiers, the on-load voltage has to be accurate; the power output in watts depends on the square of the rail voltage. Watts do not have a direct relation to subjective volume, but are psychologically an important part of the written spec. An amplifier that on review does not quite meet its published power output gives a poor impression. The subjective difference between 199 and 200 W is utterly negligible but the two figures look quite different laying there on the paper. It is therefore normal practice to err on the side of higher rather than lower output power; this should not be taken too far as the amplifier will be running hotter than necessary.

The main reason for output power error is that the voltage actually developed on the reservoir capacitors depends on losses that are not easily predicted, and this is inherent in any rectifier circuit where the current flows only in short sharp peaks at the crest of the AC waveform.

Firstly the voltage developed depends on the transformer regulation, i.e. the amount the voltage drops as more current is drawn. (The word 'regulation' in this context has nothing to do with negative-feedback voltage control – unfortunate and confusing, but there it is.) Transformer manufacturers are usually reluctant to predict anything more than a very approximate figure for this.

Voltage losses also depend strongly on the peak amplitude of the charging pulses from the rectifier to the reservoir; these peaks cause voltage drops in the AC wiring, transformer winding resistances, and rectifiers that are rather larger than might be expected from just considering the mean DC current. Unfortunately the magnitude of the peak current is poorly defined, being affected by wiring resistance and transformer leakage reactance (a parameter that transformer manufacturers are even more reluctant to predict), and calculations of the extra peak losses are so rough that they are of doubtful value. There may also be uncertainties in the voltage efficiency of the amplifier itself, and there are so many variables that it is only realistic to expect to try two or even three transformer designs before the exact output power required is obtained. I have run projects where the transformer was exactly right the first time, but that was maybe 10% of cases, and I might as well be honest and put them down to good luck.

The power output of an amplifier depends on when it starts clipping – a common criterion is that the rated power is given when the THD due to clipping reaches 1%. Given the usual unregulated power supply, clipping is controlled by the troughs of the ripple waveform rather than its peaks, and the depth of these troughs is a function of the size of the total reservoir capacity. Since large electrolytics have relatively wide tolerances, this introduces another uncertainty into the calculations.

Secondly, the voltage losses in the power amplifier itself are not that easy to predict, some of the clipping mechanisms being quite complicated in detail. The inevitable conclusion is that the fastest way to reach a satisfactory transformer design is to make only approximate calculations, order a prototype as soon as possible, and fine-tune the required voltage from there.

Since most amplifiers are intended to reproduce music and speech, with high peak-to-average power ratios, they will operate satisfactorily with transformers rated to supply only 70% of the current required for extended sine-wave operation, and in a competitive market the cost savings are significant. Trouble comes when the amplifiers are subjected to sine-wave testing, and a transformer so rated will probably fail from internal overheating, though it may take an hour or more for the temperatures to climb high enough. The usual symptom is breakdown of the winding insulation, the resultant shorted turns causing the primary mains fuse to blow. This process is usually undramatic, without visible transformer damage or the evolution of smoke, but it does of course ruin an expensive component.

To prevent such failures when a mains transformer is deliberately underrated, some form of thermal cut-out is essential. Self-resetting cut-outs based on snap-action bimetal disks are physically small enough to be buried in the outer winding layers and work very well. They are usually chosen to open the primary circuit at 100 or 110°C, as transformer materials are usually rated to 120°C unless special construction is required. Once-only thermal cut-outs can also be specified, but their

operation renders the transformer almost as useless as shorted turns do – it is rarely economic to rewind transformers. The point is that they are required for safety reasons; the transformer will fail in a controlled fashion rather than relying on internal shorting and consequent fuse-blowing, and they are significantly cheaper than self-resetting cut-outs.

If the primary side of the mains transformer has multiple taps for multi-country operation, remember that some of the primary wiring will carry much greater currents at low-voltage tappings; the mains current drawn on 90V input will be nearly three times that at 240V, for the same power out.

Transformer mounting

Mounting frame transformers is straightforward; bolts go through holes in the frame and into the chassis. There may be an orientation that minimizes the hum induced into the electronics, and this needs to be considered at the mechanical design stage. These transformers are usually, but not always, mounted with their sides parallel to the sides of the chassis for aesthetic reasons, and rotating them to minimize hum is not common practice.

Toroidal transformers introduce some extra considerations. It is well known that toroids can be rotated to minimize induced hum, and it is a very good idea to allow for this by making the transformer lead-out wires long enough.

Toroidal transformers are typically mounted by sandwiching them between two dished plates, or one dished plate and a dished area pressed into a chassis plate. The plates are then held in place by a single large bolt passing through the center, as shown in Figure 9.3. Neoprene washers are used at top and bottom to prevent the pressure from the plates putting undue pressure on the outer windings. In some cases a large flat washer is used underneath the chassis to spread the loading from the central bolt.

The fixing bolt must be secured with some kind of locking nut or locking washer. The toroid will be the heaviest part of the amplifier, and you really do not want it bouncing around inside the equipment

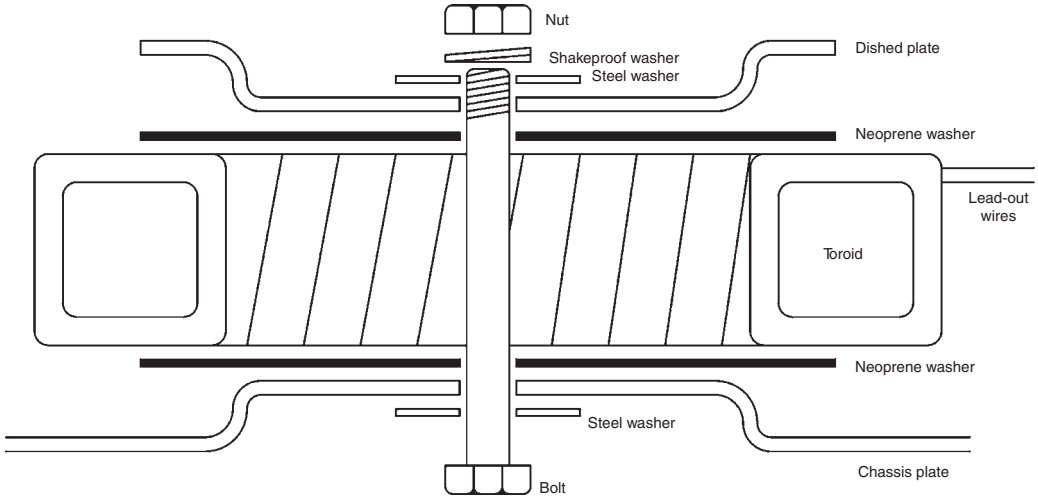


Figure 9.3: Toroidal transformer mounting. For clarity the fixing bolt is partly withdrawn

because vibration in transit has loosened the nut. It is important not to over-tighten the bolt and put undue stress on the windings; in a production situation a torque-wrench setting is usually specified.

Very small toroids can be mounted simply by putting a fixing bolt through a central filling of epoxy potting compound. This would not be safe for larger sizes as the potting compound is only adhering to the tape on the inside of the toroid, and any serious vertical force will either tear the tape or rupture the bond between tape and potting. Nevertheless, large toroids very often do have their center filled with potting compound; this is to deal with side-forces, at which it is good because one side is in compression, and *not* vertical forces. These side-forces are typically produced by the 1-meter drop-test.

It is well known that when a toroidal transformer is mounted, it is essential to avoid creating a shorted turn through the central bolt. However, the mistake shown in Figure 9.4a does still occur and the result will inevitably be blown primary fuses and profuse profanity. Slightly more subtle is the dangerous situation shown in Figure 9.4b, where a shorted turn is created when the equipment lid is slightly deformed by placing a heavy item on top of it. The clearance between the top of a toroid mounting bolt and the lid must always be checked. If you've got it wrong and you are surrounded by 1000 sets of metalwork, a thin layer of tough insulation on the inside of the lid will get you out of trouble.

Transformer specifications

A transformer specification needs to be a formal document. There are many factors to nail down and the usual result is an electrical schematic showing the primaries, secondaries, screen, etc., and a mechanical drawing showing maximum dimensions, supplemented by quite a lot of text.

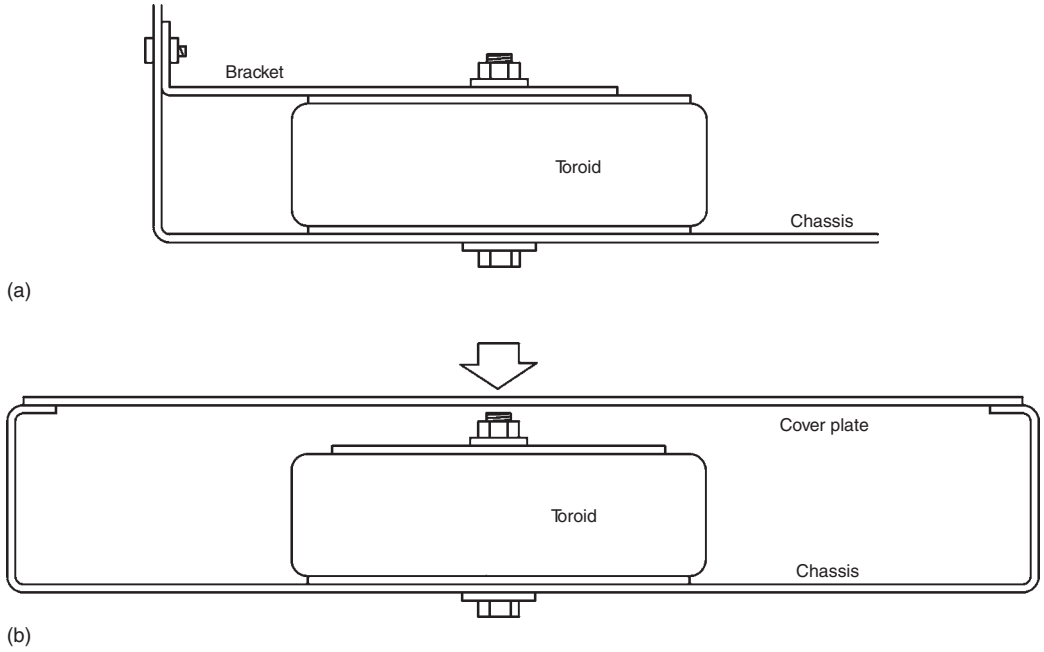


Figure 9.4: How not to mount a toroidal transformer

The specification process usually starts with an informal discussion with the manufacturer to determine the approximate physical size of the transformer for the VA required. This must be done before you freeze the mechanical size of your product.

The list below gives an idea of what you need to specify when ordering a toroidal transformer.

Electrical specifications

Having done the basic calculations, you have (you hope) a pretty good idea of what DC voltage you will need on your main supply rails to give the desired power into a given load impedance, and have done your best to translate that into a required AC voltage from the transformer secondary.

Manufacturers are not normally prepared to give exact figures how transformer regulation will affect the DC voltage available after rectification, and so when specifying the secondary AC voltage, it is realistic to aim for getting this exactly right under only one loading condition. This is usually the 'rated load', which is almost always 8 Ω. (The choice of what 'rated load' you put on the rear panel can have implications for safety testing – see Chapter 19 on testing and safety.)

1. Primary structure, i.e. will there be dual-voltage capability? If so, will there be two primary windings or a single tapped primary? The former makes better use of the copper but voltage-switching is more complicated.
2. Presence or otherwise of electrostatic screen between primary and secondary. Use one if you possibly can because it effectively stops RF getting in and out of the unit.
3. Secondary voltage with no load, i.e. with no external load on the amplifier. This will be greater than the voltage in (2) and you will need to make sure the reservoir capacitor voltage ratings are high enough.
4. Secondary voltage with rated amplifier load, such as 8 Ω (this is the 'design point', which has to be accurate).
5. Secondary voltage with heavy amplifier load, such as 2 or 4 Ω. This will obviously be less, due to resistive losses in the windings. If it comes out too low then it can be pulled up a bit by using thicker wire for windings, but this will increase the overall size.
6. Length, diameter, and color of all lead-out wires. Length stripped of insulation, and if tinned.

Mechanical matters

1. Maximum diameter (the diameter will not be constant going around the periphery – it will typically be greater near the lead-out wires).
2. Maximum height.
3. State if manufacturer is expected to supply mounting hardware, i.e. dished plate, neoprene washers, fixing bolt and nut, locking washer, etc.
4. Specify the central potting, stating the size of the hole through it for the fixing bolt. A crucial thing to specify is how close the potting comes to the top of the toroid – it must be below the dished part of the top mounting plate, or the plate will not sit properly on the top of the transformer.

Safety issues

The internal safety requirements, such as the thickness of insulation between windings, are usually left to the manufacturer. It is common, however, to specify the safety requirements for the lead-out wires, with phrases such as 'must be UL-approved'.

Transformer evaluation

When a sample transformer is ordered there are several aspects of its performance that need to be looked at. Most are straightforward, e.g. is it the right physical size, does it have the right lead-out colors and so on, but some aspects need a little more thought.

Unless your transformer manufacturer is hopelessly incompetent, the secondary voltages should be roughly what you asked for but, for the reasons detailed above, are unlikely to be exactly right. This is of no importance in secondaries for powering regulated supplies to run op-amps, etc., but because of its strong effect on the power output figure in watts, the main HT rail secondary voltage has to be correct.

When checking power output, it is obviously important to have the incoming mains at exactly the right voltage, as errors here will feed directly into erroneous measurements. Once again, this is particularly important since the output in watts varies as the square of the voltage. The usual practice is to use a variable autotransformer to fine-tune the mains voltage, its output being monitored by a DVM with the usual measure-average-but-call-it-rms calibration. Another option is to use a ferroresonant constant-voltage transformer, but these have several disadvantages: the output waveform is usually more of a square wave than a sine, and there is a fixed output voltage. They are also heavy and expensive.

The ideal solution is to use a mains synthesizer, which can output a good sine wave of variable voltage and variable frequency at a serious power level; the only downside is that it is a very expensive piece of equipment that will only be used relatively infrequently. I have only ever worked with one manufacturer that had one of these to hand (and they went bust).

Particular difficulties can arise when you are in a country with 230 V mains, and testing transformers for equipment aimed at the American and Canadian markets (115 V) and Japan (100 V). Now the variable autotransformer is required to make a major change in voltage rather than a small adjustment, and the distortion of its output waveform will usually be severe. This renders the reading of the aforementioned measure-average-but-call-it-rms voltmeter inaccurate, as the waveform distortion alters the relationship between average and peak values of the mains, and it is the peak value that determines the voltage produced by the amplifier power supply. The normal result is that the measured amplifier power output is lower than expected at 115 V and 100 V, and this can lead to baffled exchanges with your transformer manufacturer, who knows quite well that what he has supplied is correct.

If you have no plans to invest in a mains synthesizer, the second-best solution is to get a large fixed autotransformer that reduces the 230 V mains to 115 V, and use that to feed the variable autotransformer. The latter now only has to make small voltage corrections and the waveform of its output will be much less distorted than if it was doing the whole voltage step-down itself.

Evaluating a transformer sample for safety is somewhat problematical. You can do the standard insulation tests, and you can check that the lead-outs are at least labeled with the right approvals. The internal construction can only be investigated by taking a sample apart, the issue here being proper insulation between windings, especially where the lead-outs from an inner winding pass through an outer winding. If you use a reputable manufacturer you are most unlikely to have trouble in this area – if you don't, you may not find out until you submit the transformer to a test house for safety approval, by which time you're usually a long way down the road to production.

Very often the most critical part of the evaluation is the amount of hum that the transformer induces into the electronic circuitry. This has its own section just below.

Transformers and hum

All transformers, even high-quality toroidal ones, have a significant hum field, and this can present some really intractable problems if not taken very seriously from the start of the design process; the expedients available for fixing a design with excessive transformer hum are limited in number. In comparison, the fields from AC wiring are much smaller, unless the cabling arrangements are really peculiar. Here are some factors to consider:

1. Make sure that the transformer is as far as possible from any sensitive electronics. This sounds simple – you just put them on opposite sides of the box, no? Unfortunately other practical considerations may get in the way. The electronic PCBs may be so large that however they are mounted, part of their area is near the transformer. It is also not a good plan to put a heavy transformer at the extreme end of the box, as this makes it awkward to pick up and carry; when we approach a solid object like an amplifier case we expect the center of gravity to be in the middle. There may also be an aesthetic requirement that the transformer should be in the centre of the box. The visual appeal with the lid off is a significant marketing factor.
2. Use a toroidal transformer. They are more expensive for the same VA, and harder to mount, but the reduction in hum field is significant, and they are used wherever external fields are an issue.
3. If you are using a toroid, make sure it can be rotated to minimize hum. It is not usually economic to optimize the orientation for each example of a product, but toroids made by reputable manufacturers should not vary much in the shape of their hum field and the orientation can be fixed at the design stage. The limitation of this technique is that if the susceptible electronics is spread out over space, very possibly left and right channels will be on opposite sides of the enclosure, and with dreadful certainty it will be found that the hum minimum for one channel is something like the maximum for the other.

   However, with suitable layout rotation can be very effective. One prototype amplifier I have built had a sizeable toroid mounted immediately adjacent to the TO-3 end of the amplifier PCB; however, complete cancelation of magnetic hum (hum and ripple output level below −90 dBu) was possible on rotation of the transformer.

   Some toroids have single-strand secondary lead-outs, which are too stiff to allow rotation; for experimental use these can be cut short and connected to suitably large flexible wire such as 32/02, with carefully sleeved and insulated joints.

4. If you are using a frame transformer, its external field can be significantly reduced by specifying a hum strap, or 'belly-band' as it is sometimes rather indelicately called. This is a wide strip of copper that forms a closed circuit around the outside of the core and windings, so it does not form a shorted turn in the main transformer flux. Instead it intersects with the leakage flux, partially canceling it.

5. Use transformer screening. Because of its physical construction, a toroidal transformer cannot use the hum strap method to reduce the external field. The usual approach is to wrap the outside of the toroid in one or more layers of silicon steel, the intention being screening rather than the creation of a shorted turn. The success of this depends on using high-quality silicon steel, or better still GOSS (Grain-Oriented Silicon Steel), and even then the reduction in hum figures from the affected circuitry is not likely to be more than 6 dB. It may sound unlikely, but it is a fact that the method of making GOSS was discovered in 1935 – by a Mr N P Goss. Mu-metal, a nickel–iron alloy (75% nickel, 15% iron, plus copper and molybdenum), is the most effective magnetic screening material, but it is expensive and has a disconcerting habit of losing its magical properties if deformed.

6. Go to a manufacturer with a reputation for making low-field transformers. At least one toroid manufacturer specializes in low-field designs for audio applications, and their products can be 10 dB better than a standard-quality transformer; on the downside, the price will be something like twice as much. Bear in mind that low-field transformers will usually also be slightly larger than a conventional design.

7. Put the transformer in another box that can be positioned some distance away. This is obviously an expensive approach, and raises interesting questions about running high-current connections between the amplifier box and the transformer box. It is usually inappropriate. It is, however, undeniably effective. More on this approach below.

Induced hum varies proportionally with the incoming mains voltage, and this needs to be borne in mind during testing if your mains voltage varies significantly.

External Power Supplies

However much care is taken, it is very difficult to keep all traces of transformer-induced hum out of the signal circuitry. It is highly irritating to find that despite the cunning use of low-noise circuitry, the noise floor is defined by the deficiencies of a component – for the ideal transformer would obviously have no external field – rather than the laws of physics as articulated by Johnson.

The ultimate solution to the problem is to put the mains transformer in a separate box, which can be placed a meter or so away from the amplifier unit, and powering it through an umbilical lead.

Advantages

- The transformer field hum problem is authoritatively solved.
- Will appeal to some potential customers as a 'serious' approach to high-end audio.

Disadvantages

- The cost of an extra enclosure plus an extra cable and connectors, indicator lights, etc. The connectors will have to be multi-pole and capable of handling considerable voltages and currents. The transformer box must have fuses or other means of protection in case of short-circuits in the cable.
- A significant proportion of users will, exhortations to the contrary not withstanding, promptly place the amplifier box directly on top of the transformer box, immediately defeating the whole object. This is particularly likely if the two boxes have the same footprint, and so look as if they *ought* to be stacked together. However, all is not lost in this situation, as the transformer is still physically further away from the sensitive electronics (though if the transformer has a large field emerging from its ends things may actually be worse) and there are now two extra layers of steel interposed – assuming the boxes are made of steel, that is.
- The voltages involved will probably be above the limit set by the Low Voltage Directive, so it will be necessary to ensure that the connector contacts cannot be touched. If the cable has a connector at both ends then both must be checked for this. A cable that is captive at the power-supply end makes this issue simpler and will also save the cost of a mating pair of connectors, which may be considerable.

The most important design issue is the distribution of the power-supply components between the two boxes. One approach is to put just the mains transformer in the power-supply box. This has the disadvantage that the current in the umbilical cable consists of short charging pulses of large magnitude at a frequency of 100 or 120 Hz, and these will not only experience a greater voltage drop in the cable resistance than a steady current, but also give rise to much greater I^2R heating. The latter is unlikely to cause problems in the cable itself, but can easily be fatal to the contacts of connectors. Speaking from bitter experience, I can warn that connectors that appear to have a more than adequate safety margin can fail under these conditions, and it is best to keep connectors out of charging pulse circuits.

The alternative is to put not just the mains transformer but also the rectifiers and reservoir capacitors in the power-supply box. The current in the umbilical cable is now rectified and smoothed DC, and it is much easier to specify connectors to cope with it. The snag is that the reservoir capacitors have two functions; as well as smoothing the rectified DC, they also hold a store of energy that can be drawn on during output peaks. The resistance of the cable between the reservoir capacitors and the power amplifier will cause unwanted voltage drops when there are sudden demands for load current, which can significantly reduce peak power outputs during tone-burst testing. Another worry is that the extra resistance in the supply rails could imperil the stability of the amplifier, though the use of generous local decoupling capacitors should be enough to deal with this problem.

A solution to both problems is the provision of significant amounts of capacitance at both ends; the capacitors in the power-supply box deal with the smoothing, while those at the amplifier end

provide a ready reserve of electricity. In this case the current through the cable will still show some charging peaks, the size of which will depend on the proportion of the total capacitance at each end and the cable resistance. This could be artificially increased by adding series resistors of small ohmic value but high wattage, making an RC filter that will reduce the ripple seen by the amplifier. This is a bit of a doubtful remedy as it will reduce the power output on sustained signals, and it is a very poor way to reduce amplifier noise derived from the supply rails, as will be described later.

There you have some of the pros and cons of external amplifier power supplies. It is not quite the expensive but foolproof solution it first appears to be, and the design issues require careful thought.

Inrush Currents

When a transformer is abruptly connected to the mains supply it takes a large current that decays exponentially; this is called the inrush current (or sometimes the turn-on surge, or even the 'inductive surge') and it is highly inconvenient as it can be much greater than the normal current drawn, even at maximum output into the lowest rate load impedance. This inrush current is not a danger to the transformer, which has a big thermal mass, but it can and will blow primary fuses and trip house circuit-breakers. With small and medium-sized transformers the problem is not serious, but it does mean that you have to be very careful in sizing the fuse or fuses in the primary circuit, making sure that they have a high enough rating so their life is not impaired by repeated inrush currents.

With a large transformer (say bigger than 500 VA for a toroid) the inrush becomes large enough to trigger domestic overload protection. Since most houses now have magnetic circuit-breakers rather than wire fuses in the mains distribution panel, this is not as inconvenient as it used to be, but is still thoroughly annoying and will quickly earn you the enmity of your customers. The inrush issue has to be taken very seriously as it can cause problems that only show up when the unit is out in the field. There is anecdotal evidence that circuit-breakers in Germany, while nominally rated the same as those in Britain, actually respond somewhat faster, so a design that has received careful checking in one country may cause serious trouble in another.

Inrush current is most conveniently measured with purpose-built instruments such as the Voltech power analyzer range. A cheaper method is to use a current transformer (typically of the 'giant clothes-peg' type) clamped around one of the primary connections, and connected to a digital oscilloscope; this is naturally only cheaper if you already have a digital oscilloscope. It is characteristic of inrush current that its peak value varies widely from one switch-on to the next, as it depends crucially on the point of the mains cycle at which the transformer is connected. If you're unlucky the transformer core briefly saturates and a big peak current is drawn by the primary. For this reason repeated tests – possibly up to 50 – have to be done before you are confident you have experienced the worst case. This often has to be spread out over some time to avoid overtaxing inrush suppression components.

Toroidal transformers typically take greater inrush currents than frame types, due to their lower leakage reactance. There is a component of the inrush current that is due to the charging of the power-supply reservoir capacitors, but this is usually small compared with the transformer inrush. As a rough

guideline, if your transformer is bigger than 500VA you should consider using inrush suppression. If in doubt, then at least make provision for adding it to the design in the development phase.

The inrush current is controlled by making sure there is enough series resistance in the transformer primary circuit to keep the flood of amperes down to an acceptable level. The two main ways of doing this are to use an inrush suppressor component, or a relay that switches resistance into circuit for starting.

Inrush suppression by thermistor

An inrush suppressor component (sometimes called a surge limiter) is a giant thermistor whose resistance drops to a low value as it is heated by the current passing through it; they are usually of the disk type. The inrush suppressor is inserted in series with the transformer primary. The thermal inertia of its mass causes the resistance to drop relatively slowly, so the inrush current is restricted. Because of their thermistor action, these components run very hot in the low-resistance state (about 200°C) and must be mounted with caution to ensure they do not melt the plastic of adjacent components. The component leads must be left long enough to avoid thermal degradation of the solder joints with the PCB, and if these leads are insulated it must be with a high-temperature material such as fiberglass sleeving. They are also likely to burn the fingers of service personnel – it is only polite to put a HOT warning on the PCB silkscreen.

Inrush suppressors require a cool-down time after power is removed. This cool-down or 'recovery' time allows the resistance of the NTC thermistor to increase sufficiently to provide the required inrush current suppression the next time it is needed. The necessary time varies according to the particular device, the mounting method and the ambient temperature, but a typical cool-down time is about one minute. This is usually specified by the manufacturer as a thermal time-constant with values ranging from 30 to 150 seconds, the longer times being for the larger and more highly rated versions.

Inrush suppressors are available in many different sizes. The quickest design method is to select a few types that can handle the maximum current in the primary circuit, and try them out to see which is the most effective at controlling the peak inrush value.

Inrush suppression by relay

In this method there is a series resistance placed in series in the transformer primary circuit when mains is first applied. This limits the inrush current, and is then switched out by a relay after a suitable inrush control period, typically around one second. The basic circuit is shown in Figure 9.5a.

The inrush resistor has to sustain a very large short-term overload, so a chunky wire-wound type is appropriate, and it is vital to ensure that it can cope with this overload many times over the life of the amplifier.

However, resistor manufacturers are noticeably reluctant to specify how their products will cope with such conditions. It is therefore a good plan to use inrush suppression in its intended final form from the very start of the development process; by the time all other design issues have been addressed you will almost certainly have put the inrush suppression through enough operating

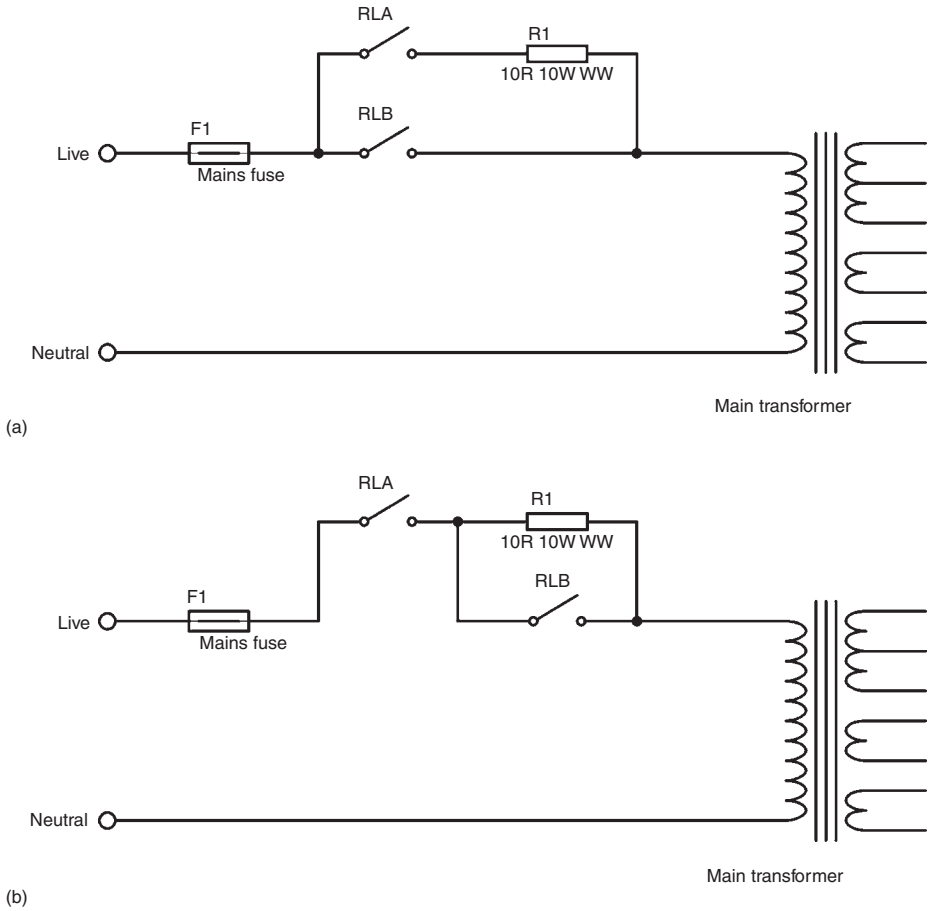


Figure 9.5: Relay-controlled inrush suppression circuits

cycles to have confidence in its durability. (Using inrush suppression from the start may well be essential anyway to prevent the workbench circuit-breakers from tripping.)

The inrush current is a complex phenomenon and the resistance value and power rating of the resistor is usually determined by experience rather than protracted mathematical analysis. Here are some typical values that I have used with success:

- 10 Ω, 10 W for an 800 VA toroid;
- 10 Ω, 20 W for a 1300 VA toroid.

Wire-wound resistors come in a limited number of types for sizes above 10 W, and it is often more convenient to use two 22 Ω 10 W resistors in parallel when a 20 W capability is required. If the resistor is correctly sized, after a single inrush event it should be warm rather than hot. Repeated and rapid cycling of the power, as may occur in testing, can cause it to get very hot and could eventually lead to failure. Fortunately this is not likely to occur in service.

The circuitry must be arranged so that if the power is turned off then immediately turned on again, inrush suppression still operates for the full period. This situation is called a 'hot restart'.

Many amplifiers are not simply switched on and off, but have an on/standby system where the mains switch initially applies power only to a small transformer that energizes a small amount of control circuitry. A low-current switch, which can be more cosmetically attractive than something hefty enough to control the full mains power, activates the control circuitry and causes it to close a relay that energizes the main supply. When this function is combined with inrush protection there are usually two identical relays in the primary circuit as shown in Figure 9.5a; at switch-on RLA closes and applies power to the transformer through the inrush resistor R1. After a second or so RLB closes and shorts out the resistor; RLA is now doing nothing so it is de-energized after a very short delay to make sure that RLB is fully closed. The alternative arrangement in Figure 9.5b should be avoided as now it is necessary to keep both relays energized all the time, which is a pointless waste of perfectly good electricity.

Fusing and Rectification

The rectifier (almost always a packaged bridge) must be generously rated to withstand the initial current surge as the reservoirs charge from empty on switch-on. Rectifier heat-sinking is definitely required for most sizes of amplifier; the voltage drop in a silicon rectifier may be low (1 V per diode is a good approximation for rough calculation) but the current pulses are large and the total dissipation is significant.

Reservoir capacitors must have the incoming wiring from the rectifier going directly to the capacitor terminals; likewise the outgoing wiring to the HT rails must leave from these terminals. In other words, do not run a tee off to the cap, because if you do its resistance combined with the high-current charging pulses adds narrow extra peaks to the ripple crests on the DC output and may worsen the hum/ripple level on the audio.

The cabling to and from the rectifiers carries charging pulses that have a considerably higher peak value than the DC output current. Conductor heating is therefore much greater due to the higher value of I^2R. Heating is likely to be especially severe if connectors are involved. Fuseholders may also heat up and consideration should be given to using heavy-duty types. Keep an eye on the fuses; if the fusewire sags at turn-on, or during transients, the fuse will fail after a few dozen hours, and the rated value needs to be increased.

When selecting the value of the mains fuse in the transformer primary circuit, remember that toroidal transformers take a large current surge at switch-on. The fuse will definitely need to be of the slow-blow type.

The bridge rectifier must be adequately rated for long-term reliability, and it needs proper heat-sinking.

RF Emissions from Bridge Rectifiers

Bridge rectifiers, even the massive ones intended solely for 100 Hz power rectification, generate surprising quantities of RF. This happens when the bridge diodes turn off; the charge carriers

are swept rapidly from the junction and the current flow stops with a sudden jolt that generates harmonics well into the RF bands. The greater the current, the more RF produced, though it is not generally possible to predict how steep this increase will be. The effect can often be heard by placing a transistor radio (long or medium wave) near the amplifier mains cable. It is the only area in a conventional power amplifier likely to give trouble in EMC emissions testing[3].

Even if the amplifier is built into a solidly grounded metal case, and the mains transformer has a grounded electrostatic screen, RF will be emitted via the live and neutral mains connections. The first line of defense against this is usually four snubbing capacitors of approximately 100 nF across each diode of the bridge, to reduce the abruptness of the turn-off. If these are to do any good, it is vital that they are all as close as possible to the bridge rectifier connections. (Never forget that such capacitors must be of a type intended to withstand continuous AC stress.)

The second line of defense against RF egress is an X-capacitor wired between Live and Neutral, as near to the mains inlet as possible (see Figure 9.1). This is usually only required on larger power amplifiers of 300 W total and above. The capacitor must be of the special type that can withstand direct mains connection; 100 nF is usually effective (some safety standards set a maximum of 470 nF). A drain resistor should be connected across the X-capacitor because if the equipment mains switch is open, and the mains lead is disconnected at the peak of the mains waveform, the X-capacitor can be left with enough charge to give a perceptible shock if the mains plug pins are touched. The resistor value should be low enough so that the X-capacitor is discharged to a safe voltage in a small fraction of a second, without being so low as to pointlessly dissipate heat. The voltage rating of the resistor should be watched; this is not usually a problem for 1/4 W sizes and above.

Relay Supplies

It is very often most economical to power relays from an unregulated supply. This is perfectly practical as most relays have a wide operating voltage range. Hum induced by electrostatic coupling from this supply rail can be sufficient to compromise the noise floor; clearly the likelihood of this depends on the physical layout, but it is inevitable that signal paths and the relay come into proximity at the relay itself. It is therefore desirable to give this line some degree of smoothing, without going to the expense of providing another regulator and heat-sink. (There should be no possibility of direct coupling between the signal ground and relay power ground; these must only join right back at the power supply.) This method of relay driving is more power efficient than a regulated supply rail as it does not require a voltage drop across a regulator that must be sufficient to prevent drop-out and consequent rail ripple in low-mains conditions.

Simple RC smoothing is quite adequate for this purpose and there is no need to consider the use of expensive chokes, which would probably cost more than a regulator, take up more space, radiate magnetic fields, and generally be a pain in the amp. Because relays draw relatively high currents, a low R and a high capacitance value for C are necessary to minimize voltage losses in R and changes in the rail voltage as different numbers of relays are energized.

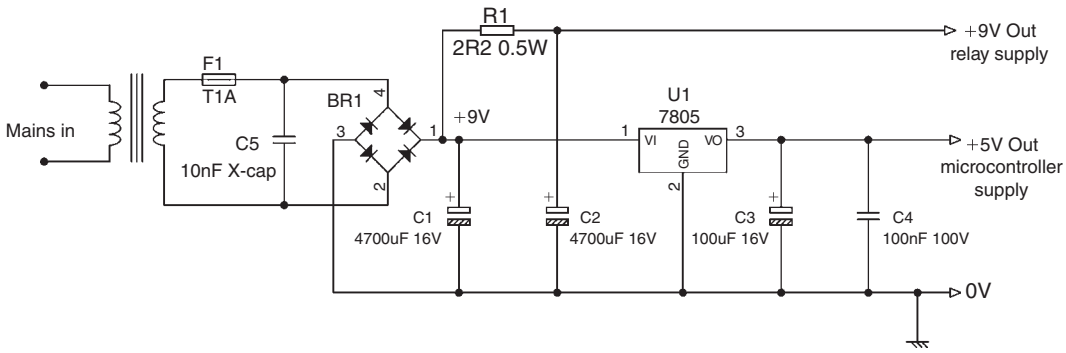


Figure 9.6: A +5 V PSU with an RC-smoothed relay supply

Figure 9.6 shows a typical power-supply circuit giving a regulated +5 V rail to power a microcontroller, with the addition of an RC-smoothed +9 V rail to power relays. The RC-smoothing values shown are typical, but are likely to need adjustment depending on how many relays are powered and how much current they draw. The R is low at 2.2 Ω and the C high at 4700 μF.

Note the 10 nF capacitor across the transformer secondary; this part must be an X-capacitor or other type rated for continuous AC stress. This is typical of the extra components required to meet modern EMC standards.

Power-Supply Rail Rejection in Amplifiers

The literature on power amplifiers frequently discusses the importance of power-supply rejection in audio amplifiers, particularly in reference to its possible effects on distortion[4]!

I have (I hope) shown in earlier chapters that regulated power supplies are just not necessary for an exemplary THD performance. I want to confirm this by examining just how supply-rail disturbances insinuate themselves into an amplifier output, and the ways in which this rail injection can be effectively eliminated. My aim is not just the production of hum-free amplifiers, but also to show that there is nothing inherently mysterious in power-supply effects, no matter what subjectivists may say on the subject.

The effects of inadequate power-supply rejection ratio (PSRR) in a typical Class-B power amplifier with a simple unregulated supply may be twofold:

1. A proportion of the 100 Hz ripple on the rails will appear at the output, degrading the noise/hum performance. Most people find this much more disturbing than the equivalent amount of distortion.

2. The rails also carry a signal-related component, due to their finite impedance. In a Class-B amplifier this will be in the form of half-wave pulses, as the output current is drawn from the two supply rails alternately; if this enters the signal path it will degrade the THD seriously.

The second possibility, the intrusion of distortion by supply-rail injection, can be eliminated in practice, at least in the conventional amplifier architecture so far examined. The most common

defect seems to be misconnected rail bypass capacitors, which add copious ripple and distortion into the signal if their return lines share the signal ground; this was denoted Distortion 5 (rail-decoupling distortion) on my list of distortion mechanisms in Chapter 3.

This must not be confused with distortion caused by *inductive* coupling of half-wave supply currents into the signal path – this effect is wholly unrelated and is completely determined by the care put into physical layout; I labeled this Distortion 6 (induction distortion).

Assuming the rail bypass capacitors are connected correctly, with a separate ground return, ripple and distortion can only enter the amplifier directly through the circuitry. It is my experience that if the amplifier is made ripple-proof under load, then it is proof against distortion components from the rails as well. This bold statement does, however, require a couple of qualifications.

Firstly, the output must be ripple-free *under load*, i.e. with a substantial ripple amplitude on the rails. If a Class-B amplifier is measured for ripple output when quiescent, there will be a very low amplitude on the supply rails and the measurement may be very good, but this gives no assurance that hum will not be added to the signal when the amplifier is operating and drawing significant current from the reservoir capacitors. Spectrum analysis could be used to sort the ripple from the signal under drive, but it is simpler to leave the amplifier undriven and artificially provoke ripple on the HT rails by loading them with a sizeable power resistor; in my work I have standardized on drawing 1 A. Thus one rail at a time can be loaded; since the rail rejection mechanisms are quite different for V+ and V−, this is a great advantage. Drawing 1 A from the V− rail of the typical power amplifier in Figure 9.7 degraded the measured ripple output from −88 dBu (mostly noise) to −80 dBu.

Secondly, I assume that any rail filtering arrangements will work with constant or increasing effectiveness as frequency increases; this is clearly true for resistor–capacitor (RC) filtering, but is by no means certain for *electronic* decoupling such as the NFB current-source biasing used in the design in Chapter 7. (These will show declining effectiveness with frequency as internal loop gains fall.) Thus, if 100 Hz components are below the noise in the THD residual, it can usually be assumed that disturbances at higher frequencies will also be invisible, and not contributing to the total distortion.

To start with some hard experimental facts, I took a power amplifier – similar to Figure 9.7 – powered by an unregulated supply on the same PCB (the significance of this proximity will become clear in a moment) driving 140 W rms into 8 Ω at 1 kHz. The PSU was a conventional bridge rectifier feeding 10,000 μF reservoir capacity per rail.

The 100 Hz rail ripple under these conditions was 1 V peak to peak. Superimposed on this were the expected half-wave pulses at signal frequency; measured at the PCB track just before the HT fuse, their amplitude was about 100 mV peak to peak. This doubled to 200 mV on the downstream side of the fuse – the small resistance of a 6.3 A slow-blow fuse is sufficient to double this aspect of the PSRR problem, and so the fine details of PCB layout and PSU wiring could well have a major effect. (The 100 Hz ripple amplitude is of course unchanged by the fuse resistance.)

It is thus clear that improving the *transmitting* end of the problem is likely to be difficult and expensive, requiring extra-heavy wire, etc., to minimize the resistance between the reservoirs and

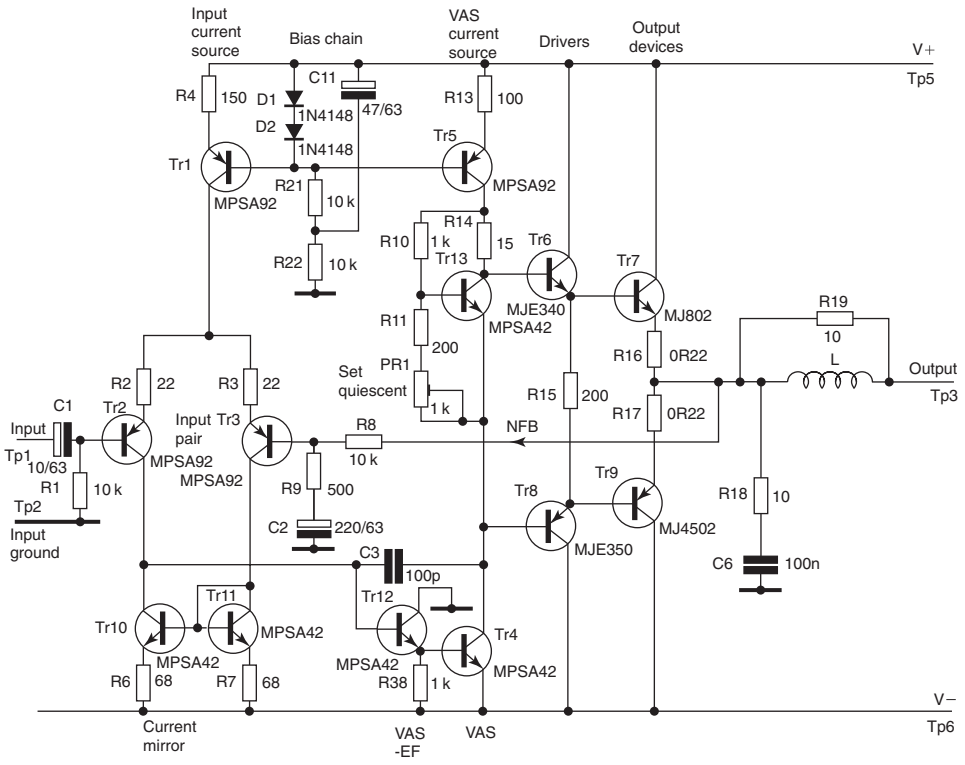


Figure 9.7: Diagram of a generic power amplifier, with diode biasing for input tail and VAS source

the amplifier. It is much cheaper and easier to attack the *receiving* end, by improving the power-amp's PSRR. The same applies to 100 Hz ripple; the only way to reduce its amplitude is to increase reservoir capacity, and this is expensive.

A Design Philosophy for Supply-Rail Rejection

First, ensure there is a negligible ripple component in the noise output of the quiescent amplifier. This should be pretty simple, as the supply ripple will be minimal; any 50 Hz components are probably due to magnetic induction from the transformer, and must be removed first by attention to physical layout.

Second, the THD residual is examined under full drive; the ripple components here are obvious as they slide evilly along the distortion waveform (assuming that the scope is synchronized to the test signal). As another general rule, if an amplifier is made visually free of ripple-synchronous artefacts on the THD residual, then it will not suffer detectable distortion from the supply rails.

PSRR is usually best dealt with by RC filtering in a discrete-component power amplifier. This will, however, be ineffective against the sub-50 Hz VLF signals that result from short-term mains voltage variations being reflected in the HT rails. A design relying wholly on RC filtering might have low AC ripple figures, but would show irregular jumps and twitches of the THD residual, hence the use of constant-current sources in the input tail and VAS to establish operating conditions more firmly.

The standard op-amp definition of PSRR is the decibel loss between each supply rail and the effective differential signal at the inputs, giving a figure independent of closed-loop gain. However, here I use the decibel loss between rail and output, in the usual non-inverting configuration with a C/L gain of 26.4 dB. This is the gain of the amplifier circuit under consideration, and allows decibel figures to be directly related to test-gear readings.

Looking at Figure 9.7, we must assume that any connection to either HT rail is a possible entry point for ripple injection. The PSRR behavior for each rail is quite different, so the two rails are examined separately.

Positive Supply-Rail Rejection

The V+ rail injection points that must be eyed warily are the input-pair tail and the VAS collector load. There is little temptation to use a simple resistor tail for the input; the cost saving is negligible and the ripple performance inadequate, even with a decoupled mid-point. A practical value for such a tail resistor would be 22 k, which in SPICE simulation gives a low-frequency PSRR of −120 dB for an undegenerated differential pair with current-mirror.

Replacing this tail resistor with the usual current source improves this to −164 dB, assuming the source has a clean bias voltage. The improvement of 44 dB is directly attributable to the greater output impedance of a current source compared with a tail resistor; with the values shown this is 4.6 M, and 4.6M/22 k is 46 dB, which is a very reasonable agreement. Since the rail signal is unlikely to exceed +10 dBu, this would result in a maximum output ripple of −154 dBu.

The measured noise floor of a real amplifier (ripple excluded) was −94.2 dBu (EIN = −121.4 dBu), which is mostly Johnson noise from the emitter degeneration resistors and the global NFB network. The tail ripple contribution would be therefore 60 dB below the noise, where I think it is safe to neglect it.

However, the tail-source bias voltage in reality will not be perfect; it will be developed from V+, with ripple hopefully excluded. The classic method is a pair of silicon diodes; LED biasing provides excellent temperature compensation, but such accuracy in setting DC conditions is probably unnecessary. It may be desirable to bias the VAS collector current source from the same voltage, which rules out anything above a volt or two. A 10 V Zener might be appropriate for biasing the input pair tail source (given suitable precautions against Zener noise) but this would seriously curtail the positive VAS voltage swing.

The negative-feedback biasing system used in the design in Chapter 7 provides a better basic PSRR than diodes, at the cost of some beta dependence. It is not quite as good as an LED, but the lower voltage generated is more suitable for biasing a VAS source. These differences become academic if the bias chain mid-point is filtered with 47 μF to V+, as Table 9.1 shows; this is C11 in Figure 9.7.

As another example, the amplifier in Figure 9.7 with diode-biasing and no bias-chain filtering gives an output ripple of −74 dBu; with 47 μF filtering this improves to −92 dBu, and 220 μF drops the reading into limbo below the noise floor.

Table 9.1: How decoupling improves hum rejection

	No decouple (dB)	Decoupled with 47 μF (dB)
Two diodes	−65	−87
LED	−77	−86
NFB low-beta	−74	−86
NFB high-beta	−77	−86

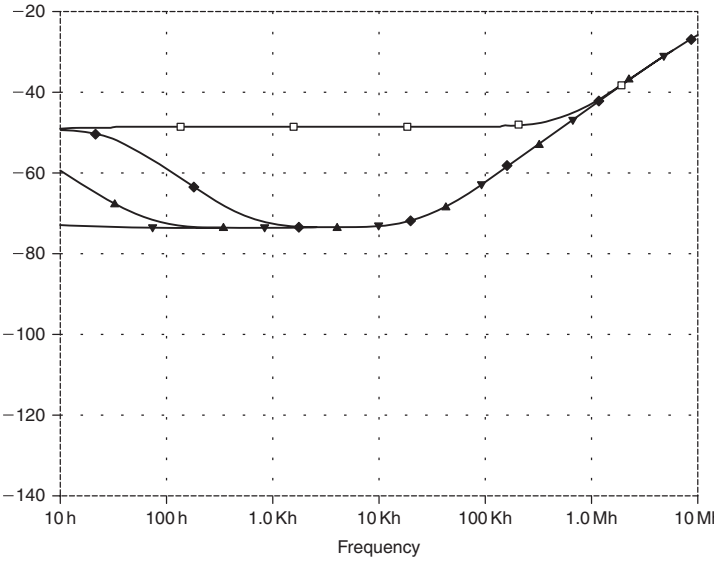


Figure 9.8: Positive-rail rejection, decoupling the tail current-source bias chain R21, R22 with 0, 1, 10, and 100 μF

Figure 9.8 shows PSPICE simulation of Figure 9.7, with a 0 dB sine wave superimposed on V+ only. A large $C_{decouple}$ (such as 100 μF) improves LF PSRR by about 20 dB, which should drop the residual ripple below the noise. However, there remains another frequency-insensitive mechanism at about −70 dB. The study of PSRR greatly resembles the peeling of onions, because there is layer after layer, and often tears. There also remains an HF injection route, starting at about 100 kHz in Figure 9.9, which is quite unaffected by the bias-chain decoupling.

Rather than digging deeper into the precise mechanisms of the next layer, it is simplest to RC filter the V+ supply to the input pair only (it makes very little difference if the VAS source is decoupled or not) as a few volts lost here are of no consequence. Figure 9.9 shows the very beneficial effect of this at middle frequencies, where the ear is most sensitive to ripple components.

Negative Supply-Rail Rejection

The V− rail is the major route for injection, and a tough nut to analyze. The well-tried wolf-fence approach is to divide the problem in half, and in this case the fence is erected by applying RC filtering to the small-signal section (i.e. input current-mirror and VAS emitter), leaving the unity-gain

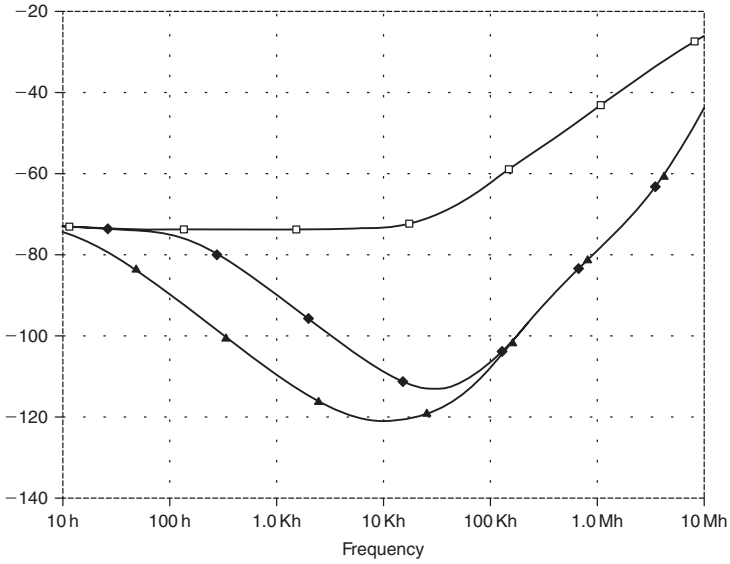


Figure 9.9: Positive-rail rejection, with input stage supply-rail RC filtered with 100 Ω and 0, 10, and 100 μF. Same scale as in Figure 9.8

output stage fully exposed to rail ripple. The output ripple promptly disappears, indicating that our wolf is getting in via the VAS or the bottom of the input pair, or both, and the output stage is effectively immune. We can do no more fencing of this kind, for the mirror has to be at the same DC potential as the VAS. SPICE simulation of the amplifier with a 1 V (0 dBV) AC signal on V− gives the PSRR curves in Figure 9.10, with C_{dom} stepped in value. As before there are two regimes, one flat at −50 dB and one rising at 6 dB/octave, implying at least two separate injection mechanisms. This suspicion is powerfully reinforced because as C_{dom} is increased, the HF PSRR around 100 kHz improves to a maximum and then degrades again, i.e. there is an optimum value for C_{dom} at about 100 pF, indicating some sort of cancelation effect. (In the V+ case, the value of C_{dom} made very little difference.)

A primary LF ripple injection mechanism is Early effect in the input-pair transistors, which determines the −50 dB LF floor of curves in Figure 9.10, for the standard input circuit (as per Figure 9.10 with $C_{dom} = 100$ pF).

To remove this effect, a cascode structure can be added to the input stage, as in Figure 9.11. This holds the V_{ce} of the input pair at a constant 5 V, and gives curve 2 in Figure 9.12. The LF floor is now 30 dB lower, although HF PSRR is slightly worse. The response to the C_{dom} value is now monotonic, simply a matter of more C_{dom}, less PSRR. This is a good indication that one of two partly canceling injection mechanisms has been deactivated.

There is a deep subtlety hidden here. It is natural to assume that Early effect in the input pair is changing the signal current fed from the input stage to the VAS, but it is not so; this current is in fact completely unaltered. What *is* changed is the integrity of the feedback subtraction performed by the input pair; modulating the V_{ce} of TR1, TR2 causes the output to alter at LF by global

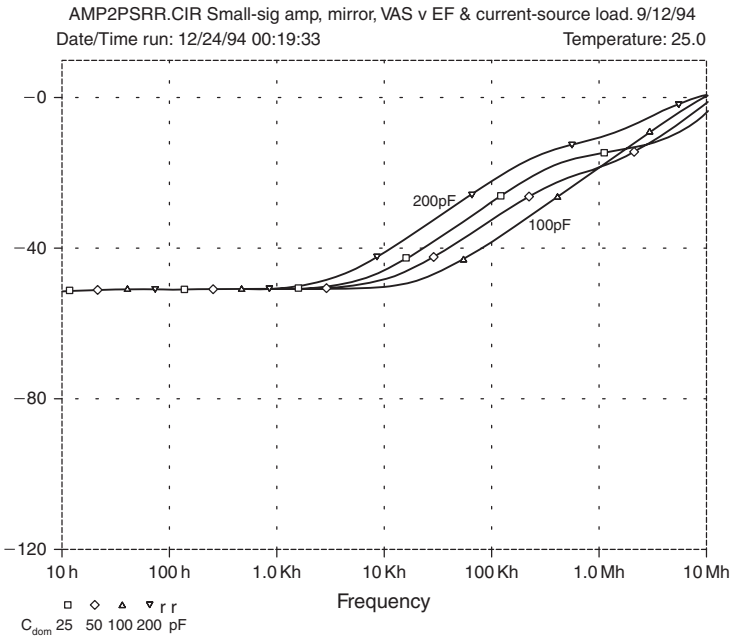


Figure 9.10: Negative-rail rejection varies with C_{dom} in a complex fashion; 100pF is the optimal value. This implies some sort of cancellation effect

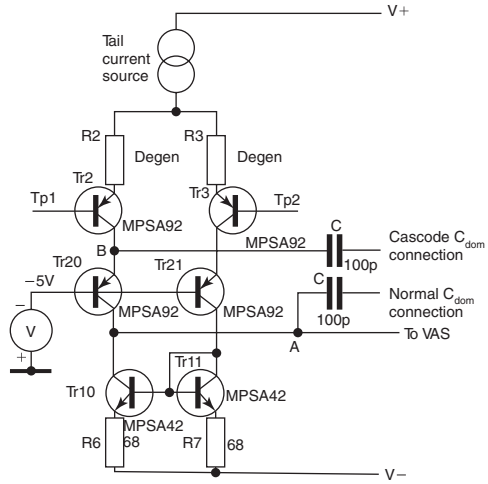


Figure 9.11: A cascoded input stage; Q21, Q22 prevent AC on V− from reaching TR2, TR3 collectors, and improve LF PSRR. B is the alternative C_{dom} connection point for cascode compensation

feedback action. Varying the amount of Early effect in TR1, TR2 by modifying VAF (Early intercept voltage) in the PSPICE transistor model alters the floor height for curve 1; the worst injection is with the lowest VAF (i.e. V_{ce} has maximum effect on I_c), which makes sense.

We still have an LF floor, though it is now at −80 rather than −50 dB. Extensive experimentation showed that this is getting in via the collector supply of TR12, the VAS beta-enhancer, modulating

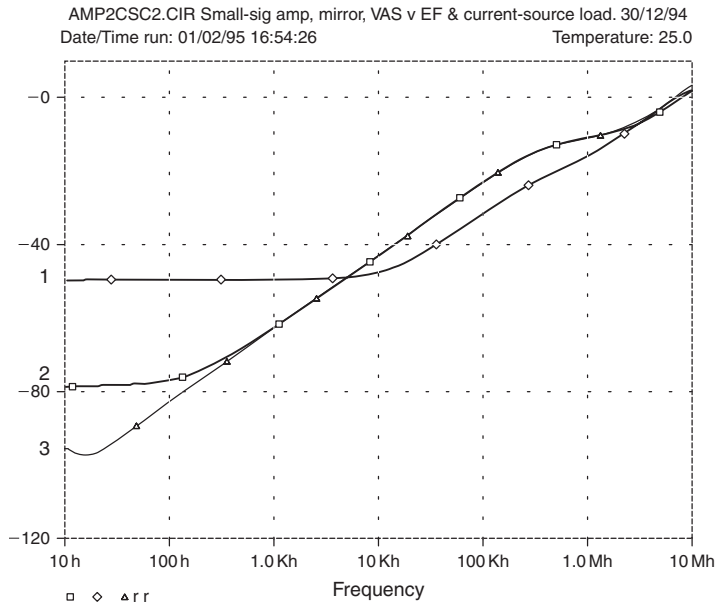


Figure 9.12: Curve 1 is negative-rail PSRR for the standard input. Curve 2 shows how cascoding the input stage improves rail rejection. Curve 3 shows further improvement by also decoupling the TR12 collector to V−

V_{ce} and adding a signal to the inner VAS loop by Early effect once more. This is easily squished by decoupling the TR12 collector to V−, and the LF floor drops to about −95 dB, where I think we can leave it for the time being (curve 3 in Figure 9.12).

Having peeled two layers from the LF PSRR onion, something needs to be done about the rising injection with frequency above 100 Hz. Looking again at the amplifier schematic in Figure 9.7, the VAS immediately attracts attention as an entry route. It is often glibly stated that such stages suffer from ripple fed in directly through C_{dom}, which certainly looks a prime suspect, connected as it is from V− to the VAS collector. However, this bald statement is untrue. In simulation it is possible to insert an ideal unity-gain buffer between the VAS collector and C_{dom}, without stability problems (A1 in Figure 9.13) and this absolutely prevents direct signal flow from V− to VAS collector through C_{dom}; the PSRR is completely unchanged.

C_{dom} has been eliminated as a direct conduit for ripple injection, but the PSRR remains very sensitive to its value. In fact the NFB factor available is the determining factor in suppressing V− ripple injection, and the two quantities are often numerically equal across the audio band.

The conventional amplifier architecture we are examining inevitably has the VAS sitting on one supply rail; full voltage swing would otherwise be impossible. Therefore the VAS input must be referenced to V−, and it is very likely that this change of reference from ground to V− is the basic source of injection. At first sight, it is hard to work out just what the VAS collector signal *is* referenced to, since this circuit node consists of two transistor collectors facing each other, with nothing to determine where it sits; the answer is that the global NFB references it to ground.

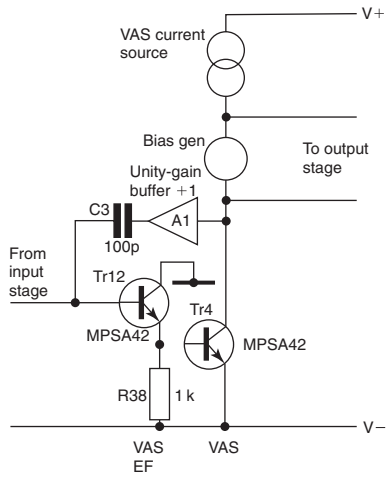


Figure 9.13: Adding a C_{dom} buffer A1 to prevent any possibility of signal entering directly from the V− rail

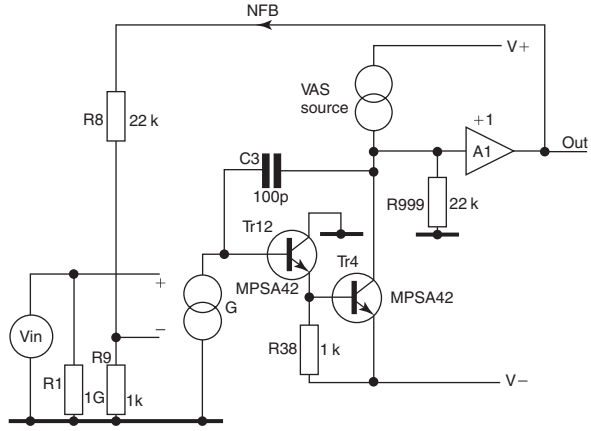


Figure 9.14: A conceptual SPICE model for V− PSRR, with only the VAS made from real components. R999 represents VAS loading

Consider an amplifier reduced to the conceptual model in Figure 9.14, with a real VAS combined with a perfect transconductance stage G and unity-gain buffer A1. The VAS beta-enhancer TR12 must be included, as it proves to have a powerful effect on LF PSRR.

To start with, the global NFB is temporarily removed, and a DC input voltage is critically set to keep the amplifier in the active region (an easy trick in simulation). As frequency increases, the local NFB through C_{dom} becomes steadily more effective and the impedance at the VAS collector falls. Therefore the VAS collector becomes more and more closely bound to the AC on V−, until at a sufficiently high frequency (typically 10 kHz) the PSRR converges on 0 dB, and everything on the V− rail couples straight through at unity gain, as shown in Figure 9.15.

There is an extra complication here; the TR12/TR4 combination actually shows *gain* from V− to the output at low frequencies; this is due to Early effect, mostly in TR12. If TR12 was omitted the LF open-loop gain drops to about −6 dB.

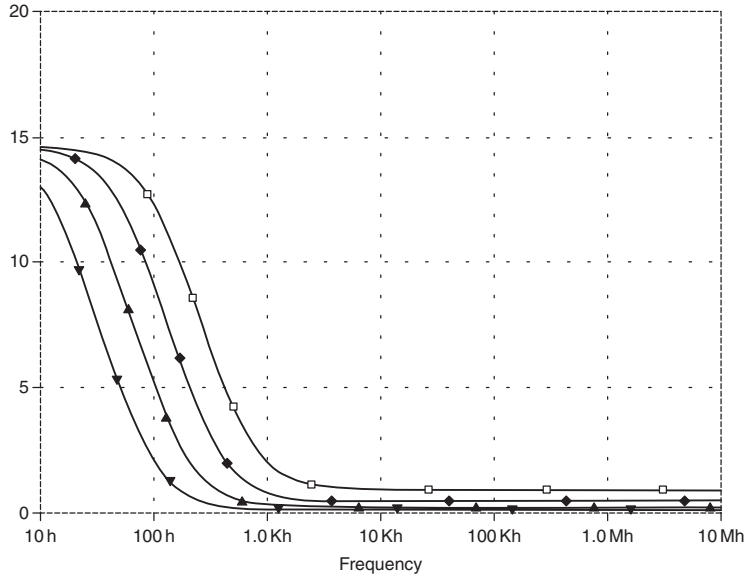


Figure 9.15: Open-loop PSRR from the model in Figure 9.14, with C_{dom} value stepped. There is actually gain below 1 kHz

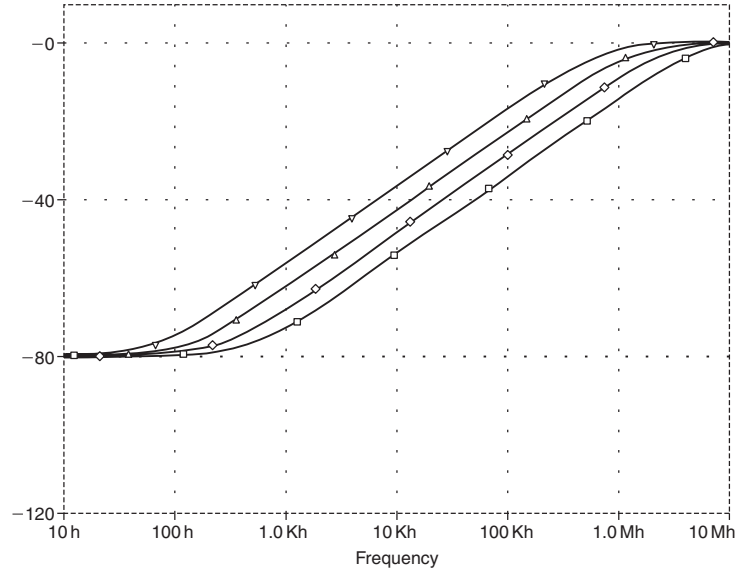


Figure 9.16: Closed-loop PSRR from Figure 9.14, with C_{dom} stepped to alter the closed-loop NFB factor

Reconnecting the global NFB, Figure 9.16 shows a good emulation of the PSRR for the complete amplifier of Figure 9.14. The 10–15 dB open-loop gain is flattened out by the global NFB, and no trace of it can be seen in Figure 9.16.

Now the NFB attempts to determine the amplifier output via the VAS collector, and if this control was perfect the PSRR would be infinite. It is not, because the NFB factor is finite, and falls with rising frequency, so PSRR deteriorates at exactly the same rate as the open-loop gain falls. This

Table 9.2: Effect of R_{nfb} value on rail ripple rejection

R_{nfb}	Ripple out (dBu)
None	83.3
470k	85.0
200k	80.1
100k	73.9

can be seen on many op-amp spec sheets, where the V− PSRR falls off from the dominant-pole frequency, assuming conventional op-amp design with a VAS on the V− rail.

Clearly a high global NFB factor at LF is vital to keep out V− disturbances. In Chapter 5, I rather tendentiously suggested that apparent open-loop bandwidth could be extended quite remarkably (without changing the amount of NFB at HF where it matters) by reducing LF loop gain; a high-value resistor R_{nfb} in parallel with C_{dom} works the trick. What I did not say was that a high global NFB factor at LF is also invaluable for keeping the hum out, a point overlooked by those advocating low NFB factors as a matter of faith rather than reason.

Table 9.2 shows how reducing global NFB by decreasing the value of R_{nfb} degraded ripple rejection in a real amplifier.

Having got to the bottom of the V− PSRR mechanism, in a just world our reward would be a new and elegant way of preventing such ripple injection. Such a method indeed exists, though I believe it has never before been applied to power amplifiers[5,6]. The trick is to change the reference, as far as C_{dom} is concerned, to ground. Figure 9.11 shows that cascode compensation can be implemented simply by connecting C_{dom} to point B rather than the usual VAS base connection at A. Figure 9.17 demonstrates that this is effective, the PSRR at 1 kHz improving by about 20 dB.

I introduced this compensation method to the hi-fi market while designing for the late lamented TAG-McLaren Audio company, and applied it to all the amplifiers I produced while I was there. It proved extremely successful and was used as one of the Unique Selling Propositions in the advertising material.

Elegant or not, the simplest way to reduce ripple below the noise floor still seems to be brute-force RC filtering of the V− supply to the input mirror and VAS, removing the disturbances before they enter. It may be crude, but it is effective, as shown in Figure 9.18. Good LF PSRR requires a large RC time-constant, and the response at DC is naturally unimproved, but the real snag is that the necessary voltage drop across R directly reduces amplifier output swing, and since the magic number of watts available depends on voltage squared, it can make a surprising difference to the raw commercial numbers (though not, of course, to perceived loudness). With the circuit values shown 10 Ω is about the maximum tolerable value; even this gives a measurable reduction in output. The accompanying C should be at least 220 μF, and a higher value is desirable if every trace of ripple is to be removed.

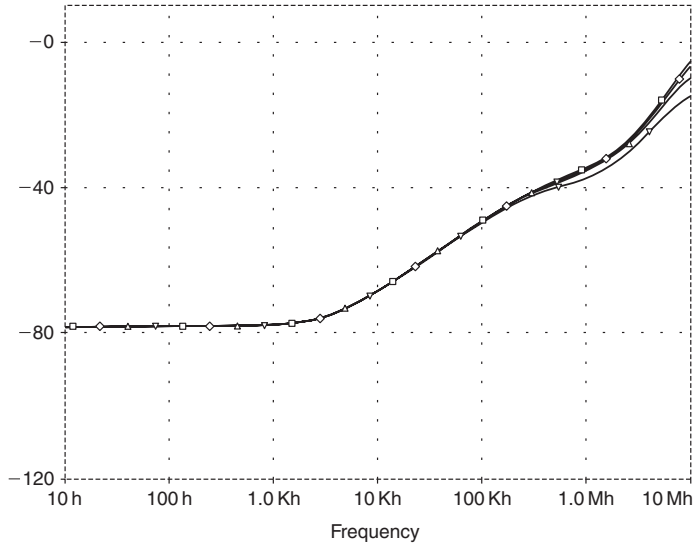


Figure 9.17: Using an input cascode to change the reference for C_{dom}. The LF PSRR is unchanged, but extends much higher in frequency (compare curve 2 in Figure 9.12). Note that the C_{dom} value now has little effect

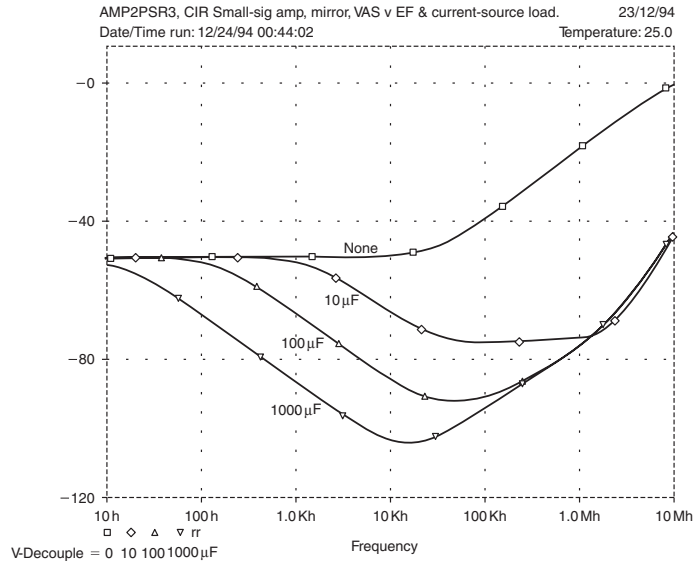


Figure 9.18: RC filtering of the V− rail is effective at medium frequencies, but less good at LF, even with 100 µF of filtering. R = 10 Ω

Negative Sub-Rails

A complete solution to this problem, which prevents ripple intruding from the negative supply rail without compromising the output voltage swing, is the use of a separate sub-rail to power the small-signal stages. This is arranged to be about 3 V below the main heavy-current V− supply rail, being supplied from an extra tap on the mains transformer secondary winding, via an extra rectifier

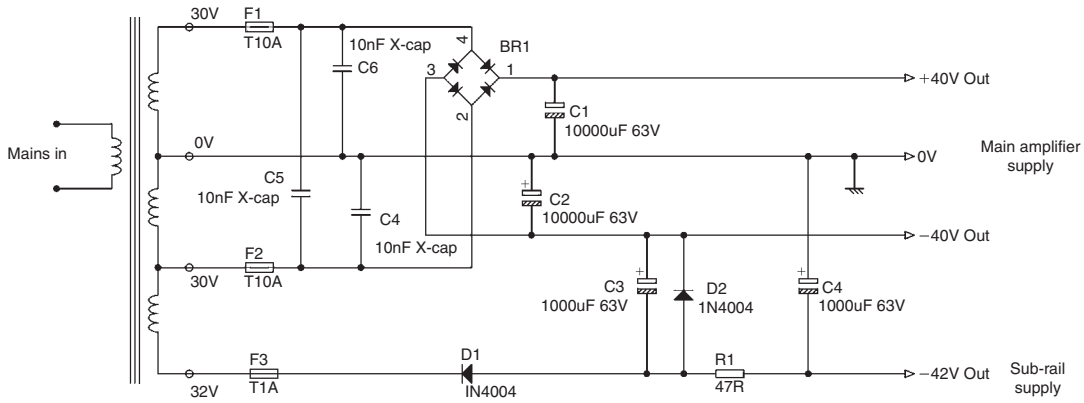


Figure 9.19: Adding a V− sub-rail supply with RC filtering to power the small-signal stages of an amplifier

and a reservoir capacitor, as in Figure 9.19, which shows a typical power supply for an amplifier or amplifiers giving about 90 W into 8 Ω.

Since the current required is only a few milliamps, half-wave rectification and a small reservoir capacitor are all that is required. The sub-rail is given some simple RC smoothing by R1 and C4, and the result is that ripple intrusion from the negative supply rail is below the noise floor. This stratagem may also improve VAS linearity, as the greatest curvature in its characteristic is likely to be at the negative end of its voltage swing.

The timing with which the rails come up and collapse is important, because if the sub-rail is lower than the main V− rail, the negative side driver may be excessively reverse-biased. Therefore the sub-rail reservoir C3 is connected to the V− rail, rather than ground; C4 has to be connected to ground if it is to perform its function of filtering the sub-rail. Note the clamping diode D2, which prevents C3 from becoming reverse-biased when the power is turned off, as C4 will probably discharge faster than the main rails.

References

[1] I. Sinclair (Ed.), Audio and Hi-fi Handbook, third ed., Newnes, 2000, p. 266.

[2] J. Linsley-Hood, Evolutionary audio. Part 3, Electronics World (January 1990) p. 18.

[3] T. Williams, EMC for Product Designers, Newnes (Butterworth-Heinemann), 1992, p. 106.

[4] G. Ball, Distorting power supplies, Electronics & Wireless World (December 1990) p. 1084.

[5] D.B. Ribner, M.A. Copeland, Design techniques for cascoded CMOS opamps, IEEE J. Solid-State Circuits (December 1984) p. 919.

[6] B.K. Abuja, Improved frequency compensation technique for CMOS opamps, IEEE J. Solid-State Circuits, (December 1983) pp. 629–633.

CHAPTER 10

Class-A Power Amplifiers

'If you can't stand the heat, get out of the kitchen.'

Harry S. Truman

An Introduction to Class-A

The two salient facts about Class-A amplifiers are that they are inefficient and that they give the best possible distortion performance. They will never supplant Class-B amplifiers, but they will always be around.

The quiescent dissipation of the classic Class-A amplifier is equal to twice the maximum output power, making massive power outputs impractical, if only because of the discomfort engendered in the summer months. However, the nature of human hearing means that the power of an amplifier must be considerably increased to sound significantly louder. Doubling the sound pressure level (SPL) is not the same as doubling subjective loudness, the latter being measured in sones rather than decibels above threshold, and it appears that doubling subjective loudness requires nearer a 10 dB rather than 6 dB rise in SPL[1]. This implies amplifier power must be increased something like 10-fold, rather than merely quadrupled, to double subjective loudness. Thus a 40 W Class-B amplifier does not sound much larger than its 20 W Class-A cousin.

There is an attractive simplicity and purity about Class-A. Most of the distortion mechanisms studied so far stem from Class-B, and we can thankfully forget crossover and switch-off phenomena (Distortions 3b and 3c), nonlinear VAS loading (Distortion 4), injection of supply-rail signals (Distortion 5), induction from supply currents (Distortion 6), and erroneous feedback connections (Distortion 7). Beta-mismatch in the output devices can also be ignored.

The only real disadvantage of Class-A is inefficiency, so inevitably efforts have been made to compromise between A and B. As compromises go, traditional Class-AB is not a happy one (see Chapters 6 and 7) because when the AB region is entered the step-change in gain generates significantly greater high-order distortion than that from optimally biased Class-B. However, a well-designed AB amplifier does give pure Class-A performance below the AB threshold, something a Class-B amp cannot do.

Another possible compromise is the so-called non-switching amplifier, with its output devices clamped to always pass a minimum current. However, it is not obvious that a sudden halt in current change as opposed to complete turn-off makes a better crossover region. Those residual oscillograms that have been published seem to show that some kind of discontinuity still exists at crossover[2].

One potential problem is the presence of maximum ripple on the supply rails at zero signal output; the PSRR must be taken seriously if good noise and ripple figures are to be obtained. This problem is simply solved by the measures proposed for Class-B designs in Chapter 9.

Class-A Configurations and Efficiency

There is a canonical sequence of efficiency in Class-A amplifiers. The simplest version is single-ended and resistively loaded, as in Figure 10.1a. When it sinks output current, there is an inevitable voltage drop across the emitter resistance, limiting the negative output capability and resulting in an efficiency of 12.5% (erroneously quoted in at least one textbook as 25%, apparently on the grounds that power not dissipated in silicon does not count). This would be of purely theoretical interest – and not much of that – except that a single-ended design by Fuller Audio has recently appeared. This reportedly produces a 10 W output for a dissipation of 120 W, with output swing predictably curtailed in one direction[3].

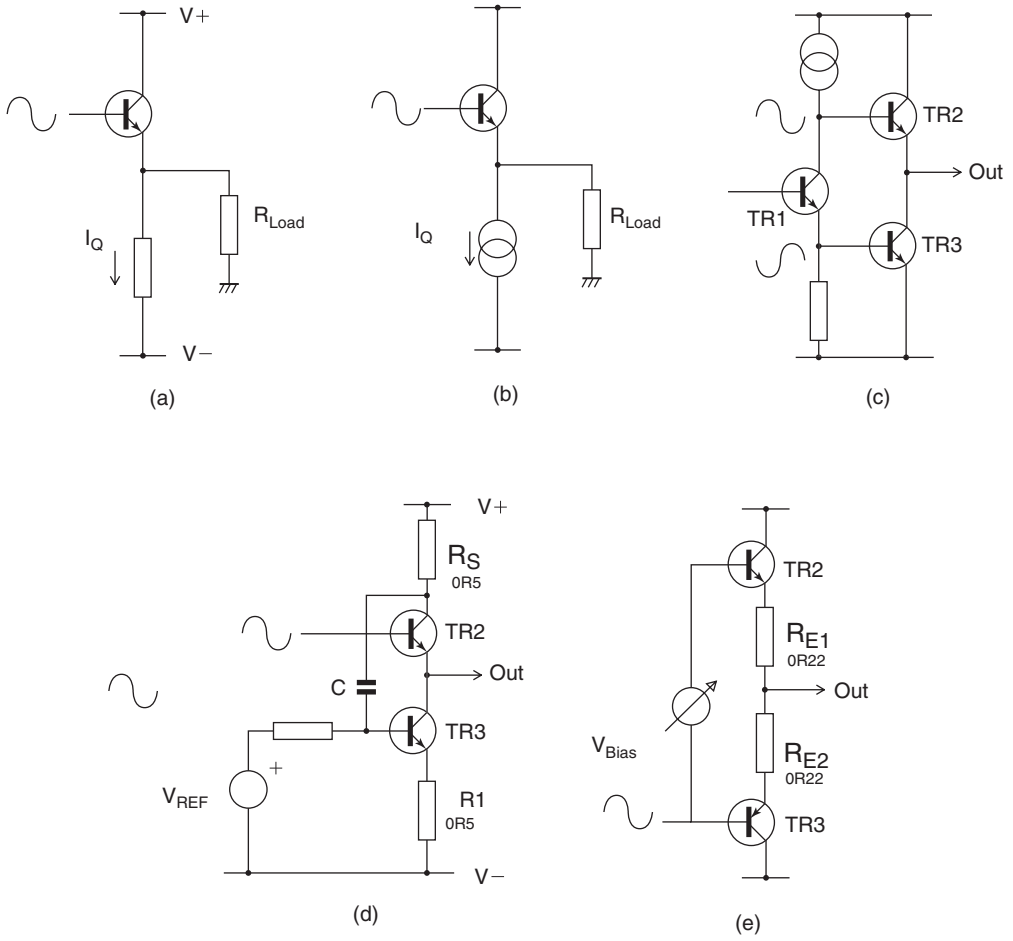


Figure 10.1: The canonical sequence of Class-A configurations. (c–e) are push–pull variants, and achieve 50% efficiency. (e) is simply a Class-B stage with higher V_{bias}

A better method – constant-current Class-A – is shown in Figure 10.1b. The current sunk by the lower constant-current source is no longer related to the voltage across it, and so the output voltage can approach the negative rail with a practicable quiescent current (hereafter shortened to I_q). Maximum efficiency is doubled to 25% at maximum output; for an example with 20 W output (and a big fan), see Pass[4]. Some versions (Krell) make the current-source value switchable, controlling it with a kind of noise gate.

Push–pull operation once more doubles full-power efficiency, getting us to a more practical 50%; most commercial Class-A amplifiers have been of this type. Both output halves now swing from zero to twice the I_q, and least voltage corresponds with maximum current, reducing dissipation. There is also the intriguing prospect of canceling the even-order harmonics generated by the output devices.

Push–pull action can be induced in several ways. Figure 10.1c, d shows the lower constant-current source replaced by a voltage-controlled current source (VCIS). This can be driven directly by the amplifier forward path, as in Figure 10.1c[5], or by a current-control negative-feedback loop, as in Figure 10.1d[6]. The first of these methods has the drawback that the stage generates gain, phase splitter TR1 doubling as the VAS; hence there is no circuit node that can be treated as the input to a unity-gain output stage, making the circuit hard to analyze, as VAS distortion cannot be separated from output stage nonlinearity. There is also no guarantee that upper and lower output devices will be driven appropriately for Class-A; in the Linsley-Hood design[5] the effective quiescent varies by more than 40% over the cycle.

The second push–pull method in Figure 10.1d is more dependable, and I have designed several versions that worked well. The disadvantage with the simple version shown is that a regulated supply is required to prevent rail ripple from disrupting the current-loop control. Designs of this type have a limited current-control range – in Figure 10.1d TR3 cannot be turned on any further once the upper device is fully off – so the lower VCIS will not be able to respond to an unforeseen increase in the output loading. In this event there is no way of resorting to Class-AB to keep the show going and the amplifier will show some form of asymmetrical hard clipping.

The best push–pull stage seems to be that in Figure 10.1e, which probably looks rather familiar. Like all the conventional Class-B stages examined in Chapters 6 and 7, this one will operate effectively in pure push–pull Class-A if the quiescent bias voltage is sufficiently increased; the increment over Class-B is typically 700 mV, depending on the value of the emitter resistors. For an example of high-biased Class-B, see Nelson-Jones[7]. This topology has the great advantage that, when confronted with an unexpectedly low load impedance, it will operate in Class-AB. The distortion performance will be inferior not only to Class-A but also to optimally biased Class-B, once above the AB transition level, but can still be made very low by proper design.

The push–pull concept has a maximum efficiency of 50%, but this is only achieved at maximum sine-wave output; due to the high peak/average ratio of music, the true average efficiency probably does not exceed 10%, even at maximum volume before obvious clipping. In my book *Self On Audio*[8], I examined the efficiency of many kinds of amplifier when handling signals with a more realistic probability density function, and it was clear that 10% was actually pretty optimistic. With

realistic listening levels, say −15 dB with respect to full output, efficiency barely reached 1%. This is not a very elegant situation.

Other possibilities are signal-controlled variation of the Class-A amplifier rail voltages, either by a separate Class-B amplifier or by a modulated switch-mode supply. Both approaches are capable of high-power output, but involve extensive extra circuitry and present some daunting design problems.

A Class-B amplifier has a limited voltage output capability, but is flexible about load impedances; more current is simply turned on when required. However, Class-A has also a current limitation, after which it enters Class-AB, and so loses its raison d'être. The choice of quiescent value has a major effect on thermal design and parts cost; so Class-A design demands a very clear idea of what load impedance is to be driven in pure A before we begin. The calculations to determine the required I_q are straightforward, though lengthy if supply ripple, $V_{ce(sat)}$ values, and Re losses, etc. are all considered, so I just give the results here. (An unregulated supply with 10,000 μF reservoirs is assumed.)

A 20 W/8 Ω amplifier will require rails of approximately ±24 V and a quiescent of 1.15 A. If this is extended to give roughly the same voltage swing into 4 Ω, then the output power becomes 37 W, and to deliver this in Class-A the quiescent must increase to 2.16 A, almost doubling dissipation. If, however, full voltage swing into 6 Ω will do (which it will for many reputable speakers), then the quiescent only needs to increase to 1.5 A; from here on I assume a quiescent of 1.6 A to give a margin of safety.

Output Stages in Class-A

I consider here only the increased-bias Class-B topology, because it is probably the best approach, effectively solving the problems presented by the other methods. Figure 10.2 shows a SPICE simulation of the collector currents in the output devices versus output voltage, and also the sum of these currents. This sum of device currents is in principle constant in Class-A, though it need not be so for low THD; the output signal is the difference of device currents and is not inherently related to the sum. However, a large deviation from this constant-sum condition means increased inefficiency, as the stage must be conducting more current than it needs to for some part of the cycle.

The constancy of this sum of currents is important because it shows that the voltage measured across Re1 and Re2 together is also effectively constant so long as the amplifier stays in Class-A. This in turn means that quiescent current can be simply set with a constant-voltage bias generator, in very much the same way as Class-B.

Figures 10.3–10.5 show SPICE gain plots for open-loop output stages, with 8 Ω loading and 1.6 A quiescent; the circuitry is exactly as for the Class-B output stages in Chapter 7. The upper traces show Class-A gain, and the lower traces optimal-bias Class-B gain for comparison. Figure 10.3 shows an emitter-follower output, Figure 10.4 a simple quasi-complementary stage, and Figure 10.5 a CFP output.

We would expect Class-A stages to be more linear than B, and they are. (Harmonic and THD figures for the three configurations, at 20 V peak, are shown in Table 10.1.) There is absolutely no

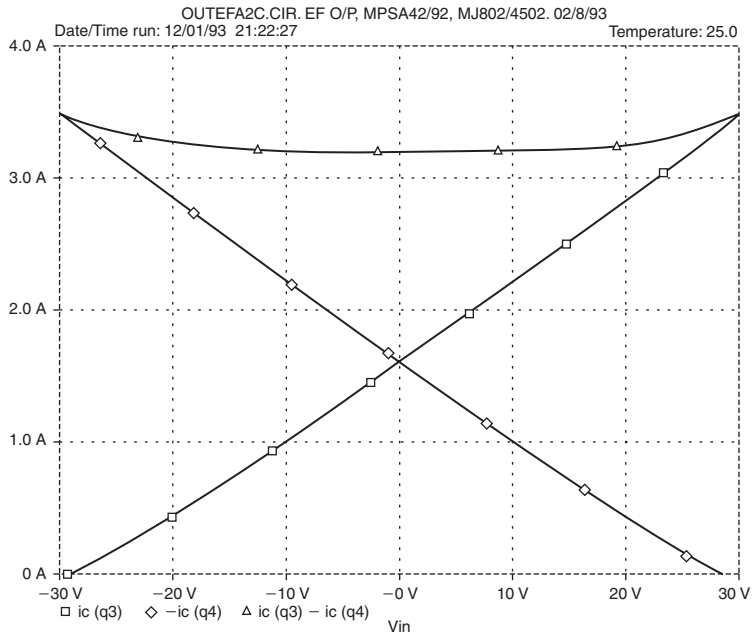


Figure 10.2: How output device current varies in push–pull Class-A. The sum of the currents is near constant, simplifying biasing

gain wobble around 0 V, as in Class-B, and push–pull Class-A really can and does cancel even-order distortion.

It is at once clear that the emitter-follower has more gain variation, and therefore worse linearity, than the CFP, while the quasi-complementary circuit shows an interesting mix of the two. The more curved side of the quasi gain plot is on the negative side, where the CFP half of the quasi circuit is passing most of the current; however, we know by comparing Figures 10.3 and 10.5 that the CFP is the more linear structure. Therefore it appears that the shape of the gain curve is determined by the output half that is turning off, presumably because this shows the biggest g_m changes. The CFP structure maintains g_m better as current decreases, and so gives a flatter gain curve with less rounding of the extremes.

The gain behavior of these stages is reflected in their harmonic generation; Table 10.1 reveals that the two symmetrical topologies give mostly odd-order harmonics, as expected. The asymmetry of the quasi-complementary version causes a large increase in even-order harmonics, and this is reflected in the higher THD figure. Nonetheless all the THD figures are still two to three times lower than for their Class-B equivalents.

This modest factor of improvement may seem a poor return for the extra dissipation of Class-A, but not so. The crucial point about the distortion from a Class-A output stage is not just that it is low in magnitude, but that it is low order, and so benefits much more from the typical NFB factor that falls with frequency than does high-order crossover distortion.

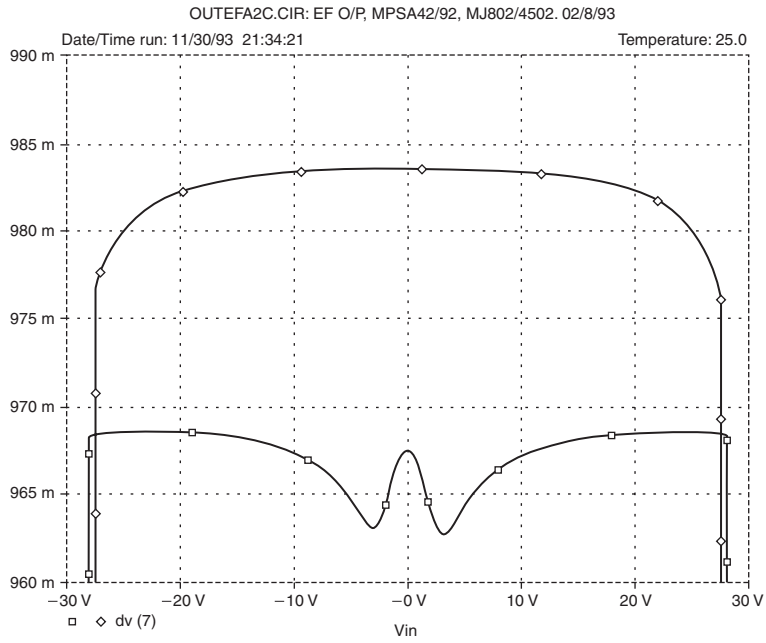


Figure 10.3: Gain linearity of the Class-A emitter-follower output stage. Load is 8 Ω and quiescent current (I_q) is 1.6 A

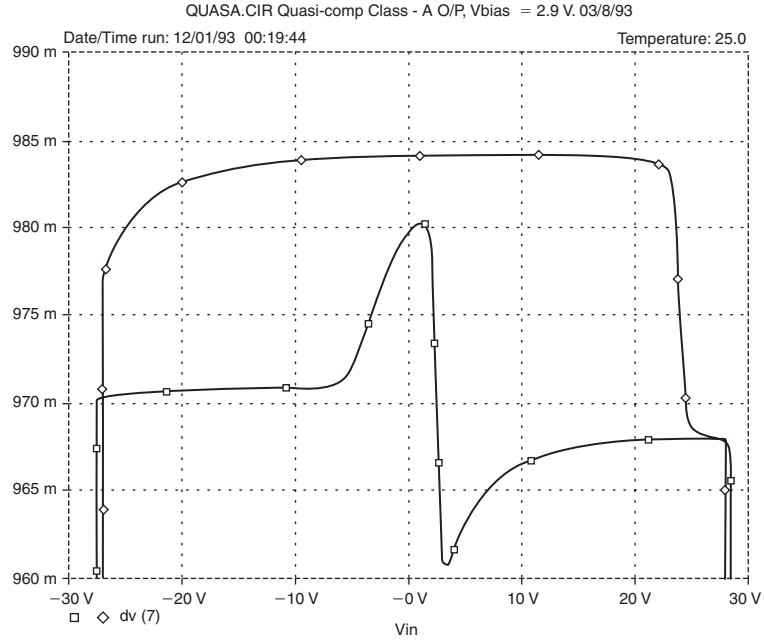


Figure 10.4: Gain linearity of the Class-A quasi-complementary output stage. Conditions as in Figure 10.3

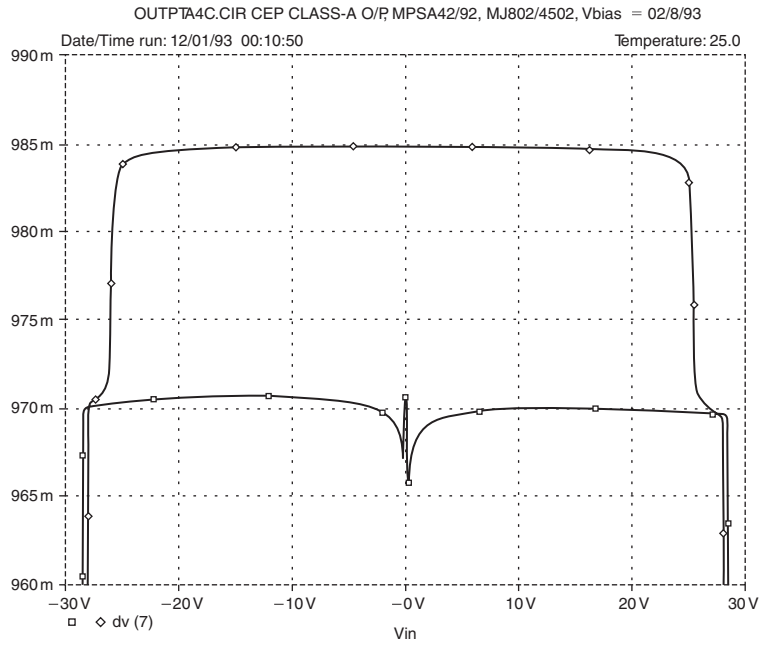


Figure 10.5: Gain linearity of the Class-A CFP output stage

Table 10.1: Harmonic levels generated by different output stages

Harmonic	Emitter-follower (%)	Quasi-comp. (%)	CFP output (%)
Second	0.00012	0.0118	0.00095
Third	0.0095	0.0064	0.0025
Fourth	0.00006	0.0011	0.00012
Fifth	0.00080	0.00058	0.00029
THD	0.0095	0.0135	0.0027

THD is calculated from the first nine harmonics, though levels above the fifth are very small.

The choice of Class-A output topology is now simple. For best performance, use the CFP; apart from greater basic linearity, the effects of output device temperature on I_q are servo-ed out by local feedback, as in Class-B. For utmost economy, use the quasi-complementary with two NPN devices; these need only a low $V_{ce(max)}$ for a typical Class-A amp, so here is an opportunity to recoup some of the money spent on heat-sinking. The rules here are somewhat different from Class-B; the simple quasi-complementary configuration gives first-class results with moderate NFB, and adding a Baxandall diode to simulate a complementary emitter-follower stage gives little improvement in linearity. See, however, Nelson-Jones[7] for an example of its use.

It is sometimes assumed that the different mode of operation of Class-A makes it inherently short-circuit proof. This may be true with some configurations, but the high-biased type studied here will continue delivering current in time-honored Class-B fashion until it bursts, and overload protection seems to be no less essential.

Quiescent Current Control Systems

Unlike Class-B, precise control of quiescent current is not required to optimize distortion; for good linearity there just has to be enough of it. However, the I_q must be under some control to prevent thermal runaway, particularly if the emitter-follower output is used. A badly designed quiescent controller can ruin the linearity, and careful design is required. There is also the point that a precisely held standing current is considered the mark of a well-bred Class-A amplifier; a quiescent that lurches around like a drunken sailor does not inspire confidence.

Straightforward thermal compensation with a V_{be}-multiplier bias generator thermally coupled to the main heat-sink can certainly be made to work[9], and will prevent thermal runaway so long as the heat-sink is of adequate size. At least one design – which in other respects appears to be a sincere homage to this chapter – has been published in which the bias generator is thermally coupled not to the main heat-sink, but to the driver heat-sinks, which are small PCB-mounting types for TO220[10]. I am not at all convinced that this will give proper control of the quiescent current.

However, this sort of approach misses a golden opportunity. Unlike Class-B, the use of Class-A offers the possibility of tightly controlling I_q by negative feedback. This is profoundly ironic because now that we can precisely control I_q, it is no longer critical. Nevertheless it seems churlish to ignore the opportunity, and so feedback quiescent control will be examined.

There are two basic methods of feedback current control. In the first, the current in one output device is monitored, either by measuring the voltage across one emitter resistor (R_s in Figure 10.6a), or by a collector sensing resistor; the second method monitors the sum of the device currents, which as described above is constant in Class-A.

The first method as implemented in Figure 10.6a[7] compares the V_{be} of TR4 with the voltage across R_s, with filtering by R_F, C_F. If quiescent is excessive, then TR4 conducts more, turning on TR5 and reducing the bias voltage between points A and B. In Figure 10.6b, which uses the VCIS approach, the voltage across collector sensing resistor R_s is compared with V_{ref} by TR4, the value of V_{ref} being chosen to allow for TR4 V_{be}[11]. Filtering is once more by R_F, C_F.

For either Figure 10.6a or b, the current being monitored contains large amounts of signal, and must be low-pass filtered before being used for control purposes. This is awkward as it adds one more time-constant to worry about if the amplifier is driven into asymmetrical clipping. In the case of collector-sensing there are unavoidable losses in the extra sense resistor. It is also my experience that imperfect filtering causes a serious rise in LF distortion.

The better way is to monitor current in both emitter resistors; as explained above, the voltage across both is very nearly constant, and in practice filtering is unnecessary. An example of this approach is shown in Figure 10.6c, based on a concept originated by Nelson Pass[12]. Here TR4 compares its own V_{be} with the voltage between X and B; excessive quiescent turns on TR4 and reduces the bias directly. Diode D is not essential to the concept, but usefully increases the current-feedback loop gain; omitting it more than doubles I_q variation with TR7 temperature in the Pass circuit.

The trouble with this method is that TR3 V_{be} directly affects the bias setting, but is outside the current-control loop. A multiple of V_{be} is established between X and B, when what we really want

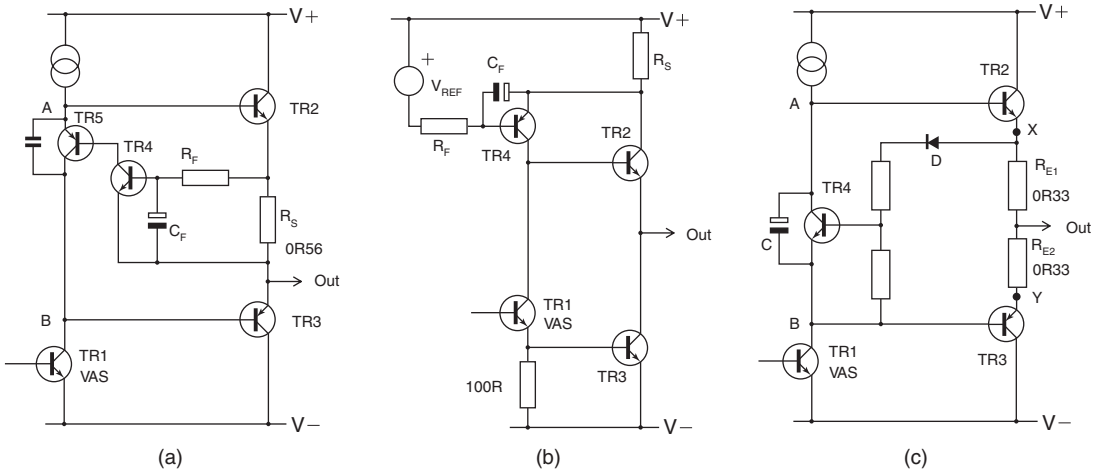


Figure 10.6: Current-control systems. Only that in (c) avoids the need to low-pass filter the control signal; (c) simply provides feedforward to speed up signal transfer to TR2

Table 10.2: I_q change per °C

	Changing TR7 temp. only (%)	Changing global temp. (%)
Quasi + V_{be}-multiplier	+0.112	−0.43
Pass: as in Figure 10.6c	+0.0257	−14.1
Pass: no diode D	+0.0675	−10.7
New system	+0.006	−0.038

Assuming OR22 emitter resistors and 1.6 A I_q.

to control is the voltage between X and Y. The temperature variations of TR4 and TR3 V_{be} partly cancel, but only partly. This method is best used with a CFP or quasi output so that the difference between Y and B depends only on the driver temperature, which can be kept low. The reference is TR4 V_{be}, which is itself temperature-dependent; even if it is kept away from the hot bits it will react to ambient temperature changes, and this explains the poor performance of the Pass method for global temperature changes (Table 10.2).

A Novel Quiescent Current Controller

To solve this problem, I would like to introduce the novel control method in Figure 10.7. We need to compare the floating voltage between X and Y with a fixed reference, which sounds like a requirement for two differential amplifiers. This can be reduced to one by sitting the reference V_{ref} on point Y; this is a very low-impedance point and can easily swallow a reference current of 1 mA or so. A simple differential pair TR15, TR16 then compares the reference voltage with that at point Y; excess quiescent turns on TR16, causing TR13 to conduct more and reducing the bias voltage.

The circuitry looks enigmatic because the high impedance of the TR13 collector would seem to prevent signal from reaching the upper half of the output stage; this is in essence true, but the vital

point is that TR13 is part of an NFB loop that establishes a voltage at A that will keep the bias voltage between A and B constant. This amounts to the same thing as maintaining a constant V_{bias} across TR5. As might be imagined, this loop does not shine at transferring signals quickly, and this duty is done by feedforward capacitor C4. Without it, the loop (rather surprisingly) works correctly, but HF oscillation at some part of the cycle is almost certain. With C4 in place the current loop does not need to move quickly, since it is not required to transfer signal but rather to maintain a DC level.

The experimental study of I_q stability is not easy because of the inaccessibility of junction temperatures. Professional SPICE implementations like PSPICE allow both the global circuit temperature and the temperature of individual devices to be manipulated; this is another aspect where simulators shine. The exact relationships of component temperatures in an amplifier are hard to predict, so I show here only the results of changing the global temperature of all devices, and changing the junction temperature of TR7 alone (Figure 10.7) with different current controllers. TR7 will be one of the hottest transistors and unlike TR9 it is not in a local NFB loop, which would greatly reduce its thermal effects.

A Class-A Design

A design example of a Blameless 20 W/8 Ω Class-A power amplifier is shown in Figure 10.7. This is as close as possible in operating parameters to the previous Class-B design, to aid comparison; in particular the NFB factor remains 30 dB at 20 kHz. The front end is as for the Class-B version, which should not be surprising as it does exactly the same job, input Distortion 1 being unaffected by output topology. As before the input pair uses a high tail current, so that R2, R3 can be

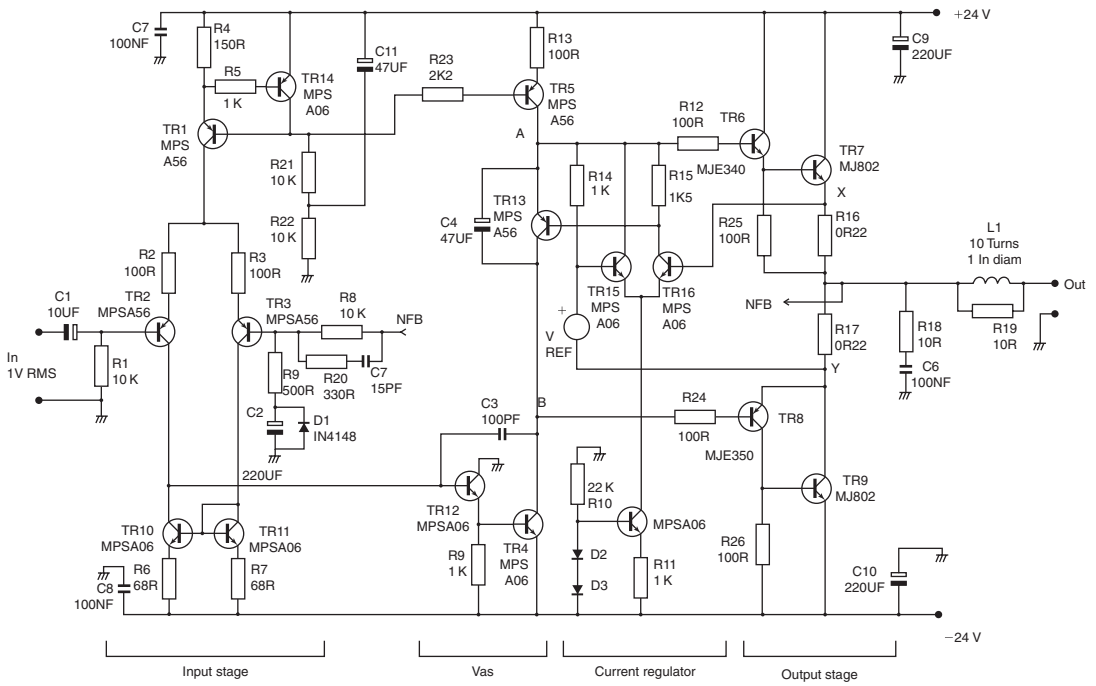


Figure 10.7: A Blameless 20 W Class-A power amplifier, using the novel current-control system

introduced to linearize the transfer characteristic and set the transconductance. Distortion 2 (VAS) is dealt with as before, the beta-enhancer TR12 increasing the local feedback through C_{dom}. There is no need to worry about Distortion 4 (nonlinear loading by output stage) as the input impedance of a Class-A output, while not constant, does not have the sharp variations shown by Class-B.

Figure 10.7 uses a standard quasi output. This may be replaced by a CFP stage without problems. In both cases the distortion is extremely low, but gratifyingly the CFP proves even better than the quasi, confirming the simulation results for output stages in isolation.

The operation of the current regulator TR13, TR15, TR16 has already been described. The reference used is a National LM385/1.2. Its output voltage is fixed at 1.223 V nominal; this is reduced to approximately 0.6 V by a 1 k–1 k divider (not shown). Using this band-gap reference, a 1.6 A I_q is held to within ±2 mA from a second or two after switch-on. Looking at Table 10.2, there seems no doubt that the new system is effective.

As before, a simple unregulated power supply with 10,000 μF reservoirs was used, and despite the higher prevailing ripple, no PSRR difficulties were encountered once the usual decoupling precautions were taken.

The closed-loop distortion performance (with conventional compensation) is shown in Figure 10.8 for the quasi-complementary output stage, and in Figure 10.9 for a CFP output version. The THD residual is pure noise for almost all of the audio spectrum, and only above 10 kHz do small amounts of third harmonic appear. The expected source is the input pair, but this so far remains unconfirmed.

The distortion generated by the Class-B and -A design examples is summarized in Table 10.3, which shows a pleasing reduction as various measures are taken to deal with it. As a final fling two-pole compensation was applied to the most linear (CFP) of the Class-A versions, reducing distortion to a rather small 0.0012% at 20 kHz, at some cost in slew rate (Figure 10.10). While this may not be the fabled Distortionless amplifier, it must be a near relation.

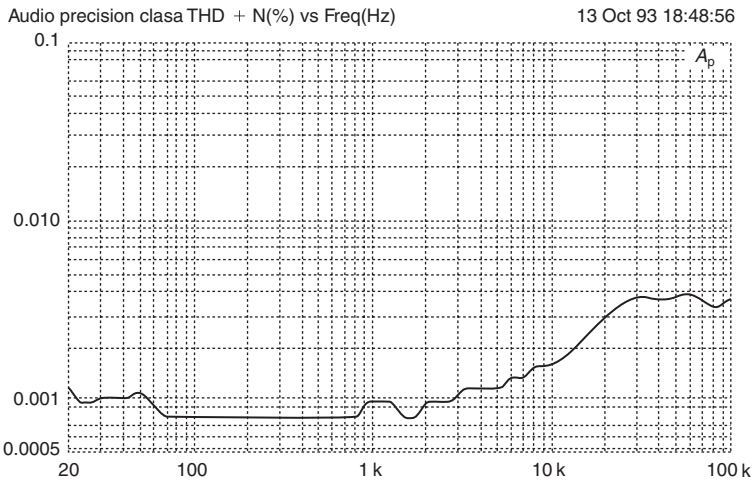


Figure 10.8: Class-A amplifier THD performance with quasi-complementary output stage. The steps in the LF portion of the trace are measurement artefacts

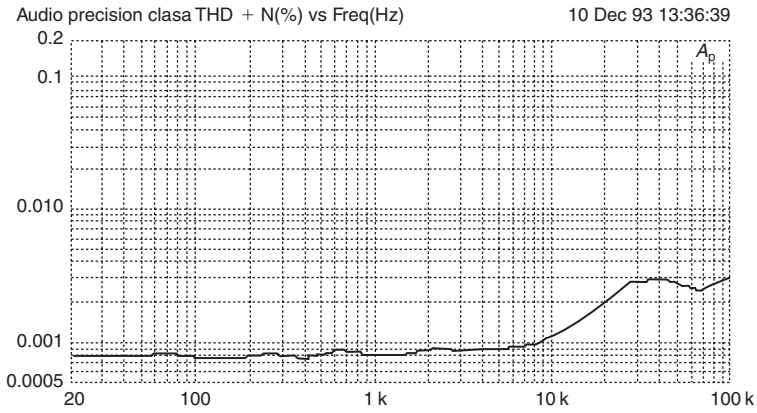


Figure 10.9: Class-A distortion performance with CFP output stage

Table 10.3: THD levels from different types of amplifier

	1 kHz (%)	10 kHz (%)	20 kHz (%)	Power (W)
Class-B EF	<0.0006	0.0060	0.012	50
Class-B CFP	<0.0006	0.0022	0.0040	50
Class-B EF two-pole	<0.0006	0.0015	0.0026	50
Class-A quasi	<0.0006	0.0017	0.0030	20
Class-A CFP	<0.0006	0.0010	0.0018	20
Class-A CFP two-pole	<0.0006	0.0010	0.0012	20

All for 8 Ω loads and 80 kHz bandwidth. Single-pole compensation unless otherwise stated.

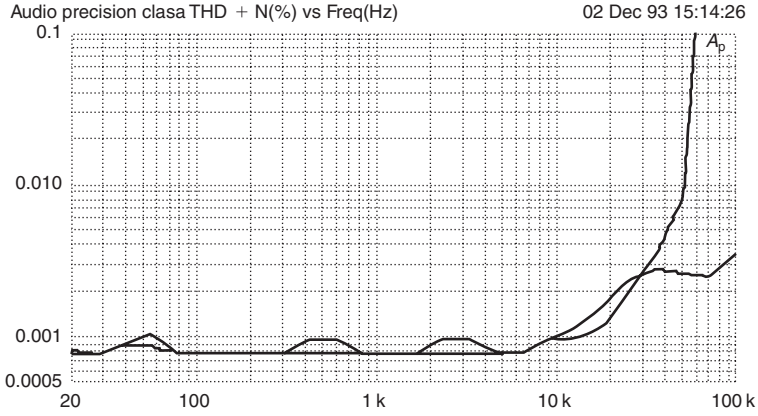


Figure 10.10: Distortion performance for CFP output stage with two-pole compensation. The THD drops to 0.0012% at 20 kHz, but the extra VAS loading has compromised the positive-going slew capability

The Trimodal Amplifier

I present here my own contribution to global warming in the shape of an improved Class-A amplifier; it is believed to be unique in that it not only copes with load impedance dips by means

of the most linear form of Class-AB possible, but will also operate as a Blameless Class-B engine. The power output in pure Class-A is 20–30 W into 8 Ω, depending on the supply rails chosen.

This amplifier uses a complementary feedback pair (CFP) output stage for best possible linearity, and some incremental improvements have been made to noise, slew rate, and maximum DC offset. The circuit naturally bears a very close resemblance to a Blameless Class-B amplifier, and so it was decided to retain the Class-B V_{be}-multiplier, and use it as a safety circuit to prevent catastrophe if the relatively complex Class-A current regulator failed. From this the idea arose of making the amplifier instantly switchable between Class-A and Class-B modes, which gives two kinds of amplifier for the price of one, and permits of some interesting listening tests. Now you really can do an A/B comparison.

In the Class-B mode the amplifier has the usual negligible quiescent dissipation. In Class-A the thermal dissipation is naturally considerable, as true Class-A operation is extended down to 6 Ω resistive loads for the full output voltage swing, by suitable choice of the quiescent current; with heavier loading the amplifier gracefully enters Class-AB, in which it will give full output down to 3 Ω before the safe operating area (SOAR) limiting begins to act. Output into 2 Ω is severely curtailed, as it must be with one output pair, and this kind of load is definitely not recommended.

In short, the amplifier allows a choice between:

1. Being very linear all the time (Blameless Class-B).
2. Being ultra-linear most of the time (Class-A) with occasional excursions into Class-AB. The AB mode is still extremely linear by current standards, though inherently it can never be quite as good as properly handled Class-B. Since there are three classes of operation I have decided to call the design a Trimodal power amplifier.

It is impossible to be sure that you have read all the literature; however, to the best of my knowledge this is the first ever Trimodal amplifier.

As previously mentioned, designing a low-distortion Class-A amplifier is in general a good deal simpler than the same exercise for Class-B, as all the difficulties of arranging the best possible crossover between the output devices disappear. Because of this it is hard to define exactly what 'Blameless' means for a Class-A amplifier. In Class-B the situation is quite different, and 'Blameless' has a very specific meaning; when each of the eight or more distortion mechanisms has been minimized in effect, there always remains the crossover distortion inherent in Class-B, and there appears to be no way to reduce it without departing radically from what might be called the generic Lin amplifier configuration. Therefore the Blameless state appears to represent some sort of theoretical limit for Class-B, but not for Class-A.

However, Class-B considerations cannot be ignored, even in a design intended to be Class-A only, because if the amplifier does find itself driving a lower load impedance than expected, it will move into Class-AB, and then all the additional Class-B requirements are just as significant as for a Class-B design proper. Class-AB can never give distortion as low as optimally biased Class-B, but it can be made to approach it reasonably closely, if the extra distortion mechanisms are correctly handled.

In a Class-A amplifier, certain sacrifices are made in the name of quality, and so it is reasonable not to be satisfied with anything less than the best possible linearity. The amplifier described here therefore uses the CFP type of output stage, which has the lowest distortion due to the local feedback loops wrapped around the output devices. It also has the advantage of better output efficiency than the emitter-follower (EF) version, and inherently superior quiescent current stability. It will shortly be seen that these are both important for this design.

Half-serious thought was given to labeling the Class-A mode 'Distortionless' as the THD is completely unmeasurable across most of the audio band. However, detectable distortion products do exist above 10 kHz, so this provocative idea was regretfully abandoned.

It seemed appropriate to take another look at the Class-A design, to see if it could be inched a few steps nearer perfection. The result is a slight improvement in efficiency and a 2 dB improvement in noise performance. In addition the expected range of output DC offset has been reduced from ±50 to ±15 mV, still without any adjustment.

Load Impedance and Operating Mode

The amplifier is 4 Ω capable in both A/AB and B operating modes, though it is the nature of things that the distortion performance is not quite so good. All solid-state amplifiers (without qualification, as far as I am aware) are much happier with an 8 Ω load, both in terms of linearity and efficiency – loudspeaker designers please note. With a 4 Ω load, Class-B operation gives better THD than Class-A/AB, because the latter will always be in AB mode, and therefore generating extra output stage distortion through g_m-doubling (which should really be called gain-deficit-halving, but somehow I do not see this term catching on). These not entirely obvious relationships are summarized in Table 10.4.

Figure 10.11 attempts to show diagrammatically just how power, load resistance, and operating mode are related. The rails have been set to ±20 V, which just allows 20 W into 8 Ω in Class-A. The curves are lines of constant power (i.e. $V \times I$ in the load), the upper horizontal line represents maximum voltage output, allowing for $V_{ce(sat)}$ values, and the sloping line on the right is the SOAR protection locus; the output can never move outside this area in either mode. The intersection between the load resistance lines sloping up from the origin and the ultimate limits of voltage-clip and SOAR protection define which of the curved constant-power lines is reached.

Table 10.4: Distortion and dissipation for different output stage classes

Load (Ω)	Mode	Distortion	Dissipation
8	A/AB	Very low	High
4	A/AB	High	High
8	B	Low	Low
4	B	Medium	Medium

Note: *High distortion* in the context of this sort of amplifier means about 0.002% THD at 1 kHz and 0.01% at 10 kHz.

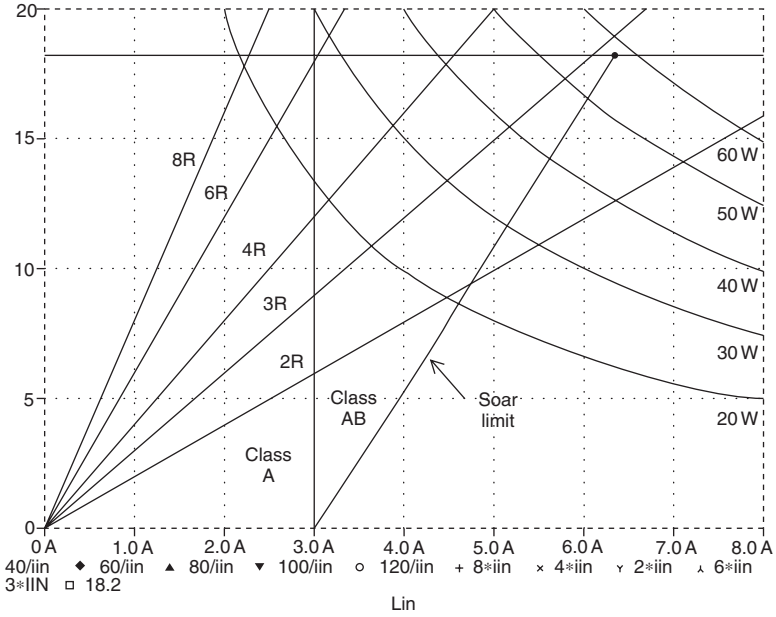


Figure 10.11: The relationships between load, mode, and power output. The intersection between the sloping load resistance lines and the ultimate limits of voltage-clipping and SOAR protection define which of the curved constant-power lines is reached. In A/AB mode, the operating point must be to the left of the vertical push–pull current-limit line for true Class-A

In A/AB mode, the operating point must be left of the vertical push–pull current-limit line (at 3A, twice the quiescent current) for Class-A. If we move to the right of this limit along one of the impedance lines, the output devices will begin turning off for part of the cycle; this is the AB operation zone. In Class-B mode, the 3A line has no significance and the amplifier remains in optimal Class-B until clipping or SOAR limiting occurs. Note that the diagram axes represent instantaneous power in the load, but the curves show sine-wave rms power, and that is the reason for the apparent factor of 2 discrepancy between them.

Efficiency

Concern for efficiency in Class-A may seem paradoxical, but one way of looking at it is that Class-A watts are precious things, wrought in great heat and dissipation, and so for a given quiescent power it makes sense to ensure that the amplifier approaches its limited theoretical efficiency as closely as possible. I was confirmed in this course by reading another recent design[13] that seems to throw efficiency to the winds by using a hybrid BJT/FET cascode output stage.
The voltage losses inherent in this arrangement demand ±50V rails and sixfold output devices for a 100W Class-A capability; such rail voltages would give 156W from a 100% efficient amplifier.

The voltage efficiency of a power amplifier is the fraction of the supply-rail voltage that can actually be delivered as peak-to-peak voltage swing into a specified load; efficiency is invariably less into 4Ω due to the greater resistive voltage drops with increased current.

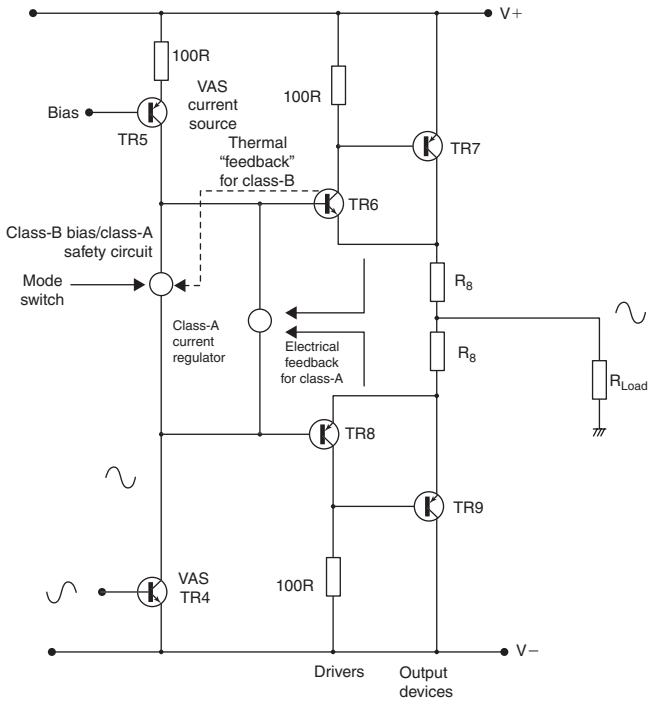


Figure 10.12: The basic CFP output stage, equally suited to operating Classes B, AB, and A, depending on the magnitude of V_{bias}. The emitter resistors Re may be from 0.1 to 0.47 Ω

The Class-B amplifier I described in Chapter 7 has a voltage efficiency of 91.7% for positive swings and 92.5% for negative, into 8 Ω. Amplifiers are not in general completely symmetrical, and so two figures need to be quoted; alternatively the lower of the two can be given as this defines the maximum undistorted sine wave. These figures above are for an emitter-follower output stage, and a CFP output does better, the positive and negative efficiencies being 94.0% and 94.7% respectively. The EF version gives a lower output swing because it has two more V_{be} drops in series to be accommodated between the supply rails; the CFP is always more voltage efficient, and so selecting it over the EF for the current Class-A design is the first step in maximizing efficiency.

Figure 10.12 shows the basic CFP output stage, together with its two biasing elements. In Class-A the quiescent current is rigidly controlled by negative feedback; this is possible because in Class-A the total voltage across both emitter resistors Re is constant throughout the cycle. In Class-B this is not the case, and we must rely on 'thermal feedback' from the output stage, though to be strictly accurate this is not feedback at all, but a kind of feedforward (see Chapter 15). Another big advantage of the CFP configuration is that I_q depends only on driver temperature, and this is important in the Class-B mode, where true feedback control of quiescent current is not possible, especially if low-value Re resistors such as 0.1 Ω are chosen, rather than the more usual 0.22 Ω; the motivation for doing this will soon become clear.

The voltage efficiency for the quasi-complementary Class-A circuit in Figure 10.7 into 8 Ω is 89.8% positive and 92.2% negative. Converting this to the CFP output stage increases this to 92.9% positive

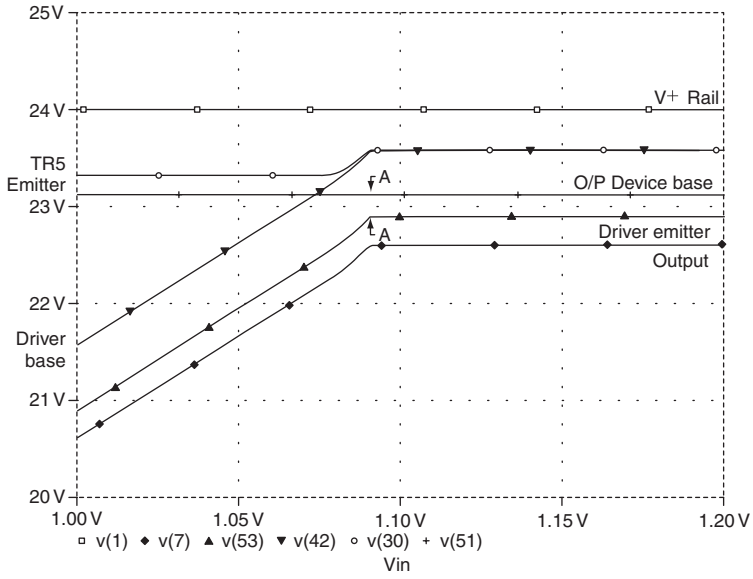


Figure 10.13: PSPICE simulation showing how positive clipping occurs in the CFP output. A higher sub-rail for the VAS cannot increase the output swing, as the limit is set by the minimum driver V_{ce} and not the VAS output swing

and 93.6% negative. Note that a Class-A quiescent current (I_q) of 1.5A is assumed throughout; this allows 31W into 8Ω in push–pull, if the supply rails are adequately high. However, the assumption that loudspeaker impedance never drops below 8Ω is distinctly doubtful, to put it mildly, and so as before this design allows for full Class-A output voltage swing into loads down to 6Ω.

So how else can we improve efficiency? The addition of extra and higher supply rails for the small-signal section of the amplifier surprisingly does not give a significant increase in output; examination of Figure 10.13 shows why. In this region, the output device TR6 base is at a virtually constant 880mV from the V+ rail, and as TR7 driver base rises it passes this level, and keeps going up; clipping has not yet occurred. The driver emitter follows the driver base up, until the voltage difference between this emitter and the output base (i.e. the driver V_{ce}) becomes too small to allow further conduction; this choke point is indicated by the arrows A–A. At this point the driver base is forced to level off, although it is still about 500mV below the level of V+. Note also how the voltage between V+ and TR5 emitter collapses. Thus a higher rail will give no extra voltage swing, which I must admit came as something of a surprise. Higher sub-rails for small-signal sections only come into their own in FET amplifiers, where the high V_{gs} for FET conduction (5V or more) makes their use almost mandatory.

The efficiency figures given so far are all greater for negative rather than positive voltage swings. The approach to the rail for negative clipping is slightly closer because there is no equivalent to the 0.6V bias established across R13; however, this advantage is absorbed by the need to lose a little voltage in the RC filtering of the V− supply to the current-mirror and VAS. This is essential if really good ripple/hum performance is to be obtained (see Chapter 9).

In the quest for efficiency, an obvious variable is the value of the output emitter resistors Re. The performance of the current regulator described, especially when combined with a CFP output stage, is more than good enough to allow these resistors to be reduced while retaining first-class I_q stability. I took 0.1 Ω as the lowest practicable value, and even this is comparable with PCB track resistance, so some care in the exact details of physical layout is essential; in particular the emitter resistors must be treated as four-terminal components to exclude unwanted voltage drops in the tracks leading to the resistor pads.

If Re is reduced from 0.22 to 0.1 Ω then voltage efficiency improves from 92.9%/93.6% to 94.2%/95.0%. Is this improvement worth having? Well, the voltage-limited power output into 8 Ω is increased from 31.2 to 32.2 W with ±24 V rails, at zero cost, but it would be idle to pretend that the resulting increase in SPL is highly significant; it does, however, provide the philosophical satisfaction that as much Class-A power as possible is being produced for a given dissipation – a delicate pleasure.

The linearity of the CFP output stage in Class-A is very slightly worse with 0.1 Ω emitter resistors, though the difference is small and only detectable open-loop; the simulated THD for 20 V peak to peak into 8 Ω is only increased from 0.0027% to 0.0029%. This is probably due simply to the slightly lower total resistance seen by the output stage.

However, at the same time, reducing the emitter resistors to 0R1 provides much lower distortion when the amplifier runs out of Class-A; it halves the size of the step-gain changes inherent in Class-AB, and so effectively reduces distortion into 4 Ω loads (see Figures 10.14 and 10.15 for output linearity simulations); the measured results from a real and Blameless Trimodal amplifier

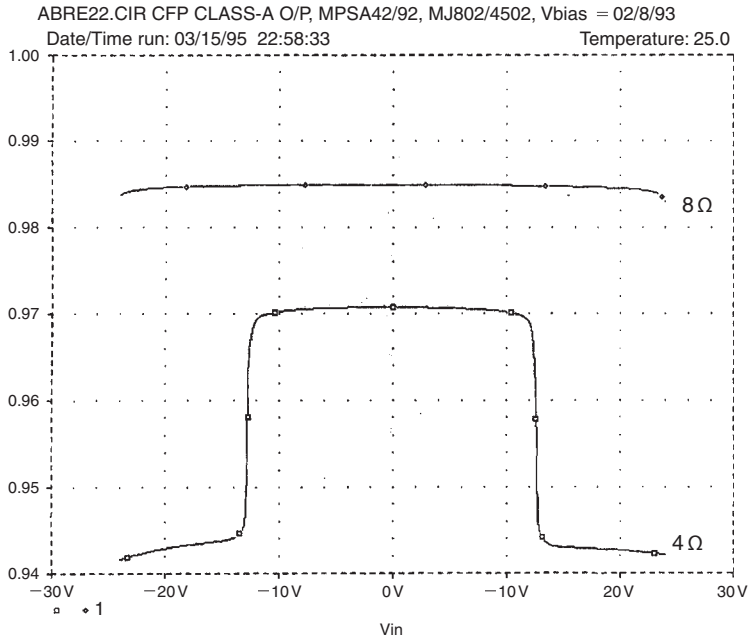


Figure 10.14: CFP output stage linearity with Re = 0R22. Upper trace is Class-A into 8 Ω, lower is Class-AB operation into 4 Ω, showing step changes in gains of 0.024 units

are shown in Figure 10.16, where it can be clearly seen that THD has been halved by this simple change. To the best of my knowledge this is a new result; if you must work in Class-AB, then keep the emitter resistors as low as possible, to minimize the gain changes.

Having considered the linearity of Classes A and AB, we must not neglect what effect this radical Re change has on Class-B linearity. The answer is not very much (see Figure 10.17, where crossover distortion seems to be slightly higher with Re = 0.2 Ω than for either 0.1 or 0.4 Ω). Whether this is a consistent effect (for CFP stages anyway) remains to be seen.

The detailed mechanisms of bias control and mode switching are described in the next section.

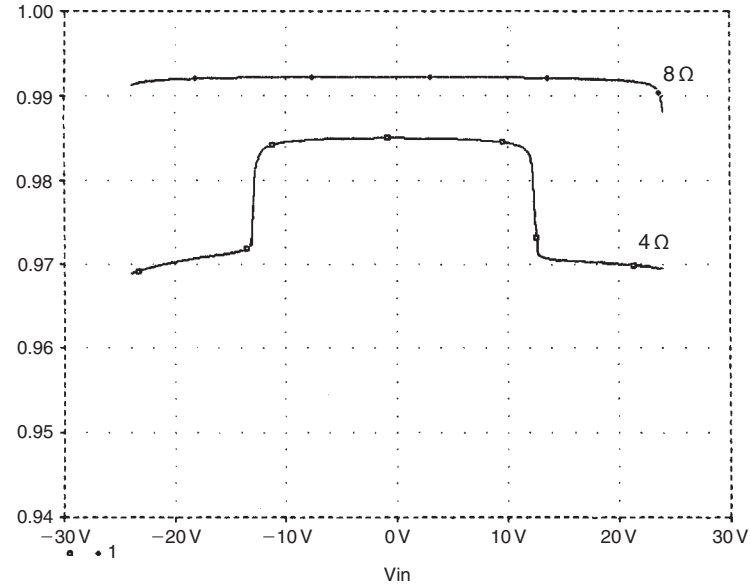


Figure 10.15: CFP output linearity with Re = 0R1, re-biased to keep I_q at 1.5 A. There is slightly poorer linearity in the flat-topped Class-A region than for Re = 0R22, but the 4 Ω AB steps are halved in size at 0.012 units. Note that both gains are now closer to unity. Same scale as in Figure 9.14

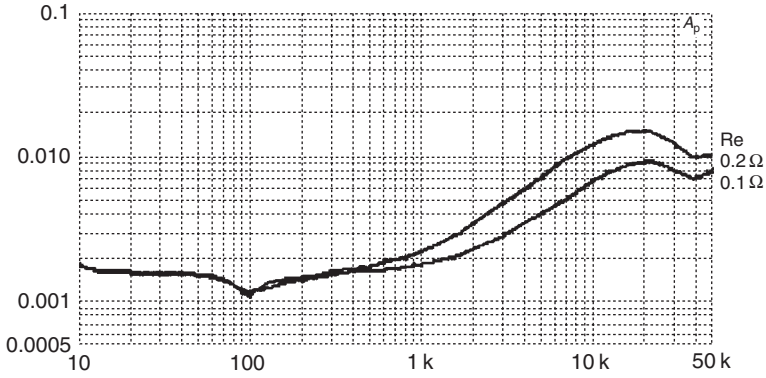


Figure 10.16: Distortion in Class-AB is reduced by lowering the value of Re

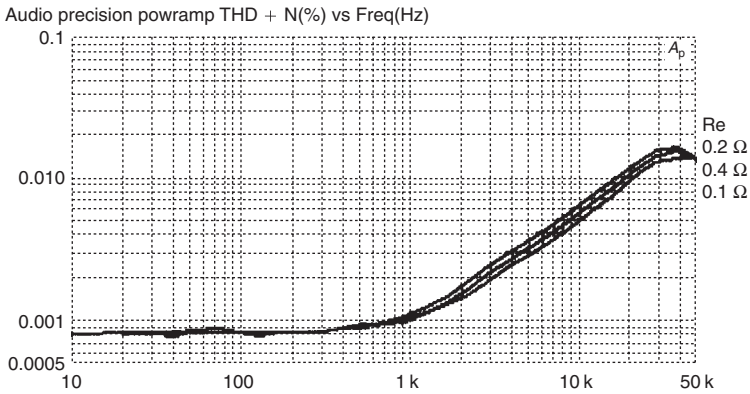


Figure 10.17: Proving that emitter resistors matter much less in Class-B. Output was 20 W in 8 Ω, with optimal bias. Interestingly, the bias does not need adjusting as the value of Re changes

Trimodal Biasing

Figure 10.18 shows a simplified rendering of the Trimodal biasing system; the full version appears in Figure 10.19. The voltage between points A and B is determined by one of two controller systems, only one of which can be in command at a time. Since both are basically shunt voltage regulators sitting between A and B, the result is that the lowest voltage wins. The novel Class-A current controller introduced on page 307 is used here adapted for 0.1 Ω emitter resistors, mainly by reducing the reference voltage to 300 mV, which gives a quiescent current (I_q) of 1.5 A when established across the total emitter resistance of 0.2 Ω.

In parallel with the current controller is the V_{be}-multiplier TR13. In Class-B mode, the current controller is disabled, and critical biasing for minimal crossover distortion is provided in the usual way by adjusting preset PR1 to set the voltage across TR13. In Class-A/AB mode, the voltage TR13 attempts to establish is increased (by shorting out PR1) to a value greater than that required for Class-A. The current controller therefore takes charge of the voltage between X and Y, and unless it fails TR13 does not conduct. Points A, B, X, and Y are the same circuit nodes as in the simple Class-A design (see Figure 10.6c).

Class-A/AB Mode

In Class-A/AB mode, the current controller (TR14, TR15, TR16 in Figure 10.18) is active and TR13 is off, as TR20 has shorted out PR1. TR15, TR16 form a simple differential amplifier that compares the reference voltage across R31 with the V_{bias} voltage across output emitter resistors R16 and R17; as explained above, in Class-A this voltage remains constant despite delivery of current into the load. If the voltage across TR16, TR17 tends to rise, then TR16 conducts more, turning TR14 more on and reducing the voltage between A and B. TR14, TR15, and TR16 all move up and down with the amplifier output, and so a tail-current source (TR17) is used.

I am very aware that the current controller is more complex than the simple V_{be}-multiplier used in most Class-B designs. There is an obvious risk that an assembly error could cause a massive current that

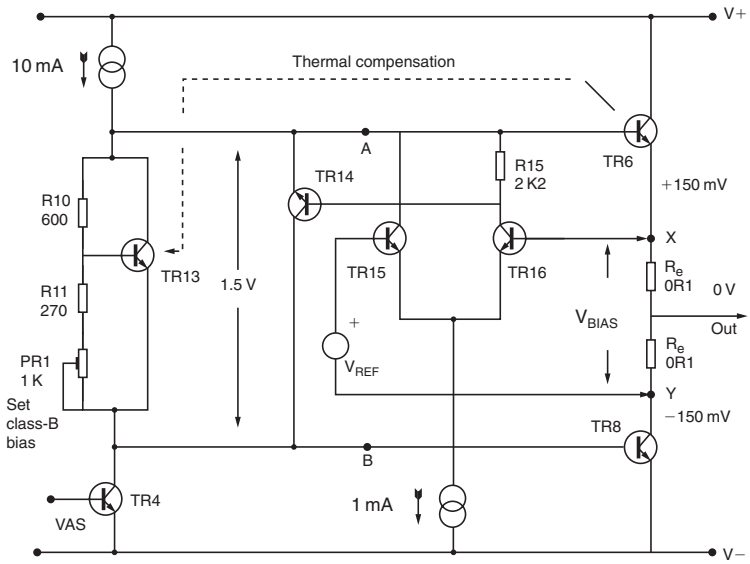


Figure 10.18: The simplified current controller in action, showing typical DC voltages in Class-A. Points A, B, X, and Y are in Figure 10.6

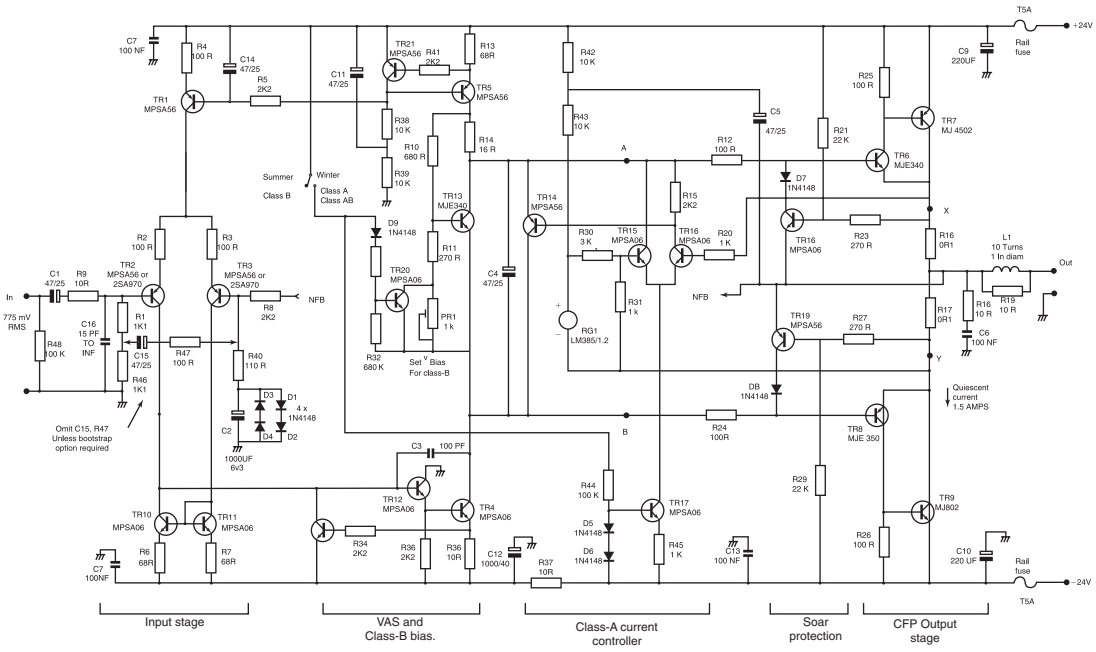


Figure 10.19: The complete circuit diagram of the Trimodal amplifier, including the optional bootstrapping components, R47 and C15

would prompt the output devices to lay down their lives to save the rail fuses. The tail-source TR17 is particularly vulnerable because any fault that extinguishes the tail current removes the drive to TR14, the controller is disabled, and the current in the output stage will be very large. In Figure 10.18 the V_{be}-multiplier TR13 acts as a safety circuit that limits V_{bias} to about 600 mV rather than the normal 300 mV, even if the current controller is completely non-functional and TR14 fully off. This gives a

quiescent of 3.0A, and I can testify this is a survivable experience for the output devices in the short term; however, they may eventually fail from overheating if the condition is allowed to persist.

There are some important points about the current controller. The entire tail current for the error amplifier, determined by TR17, is siphoned off from VAS current source TR5, and must be taken into account when ensuring that the upper output half gets enough drive current.

There must be enough tail current available to turn on TR14, remembering that most of TR16 collector current flows through R15, to keep the pair roughly balanced. If you feel moved to alter the VAS current, remember also that the base current for driver TR6 is higher in Class-A than Class-B, so the positive slew rate is slightly reduced in going from Class-A to Class-B.

The original Class-A amplifier used a National LM385/1.2, its output voltage fixed at 1.223V nominal; this was reduced to approximately 0.6V by a 1k–1k potential divider. The circuit also worked well with V_{ref} provided by a silicon diode, 0.6V being an appropriate V_{bias} drop across two 0.22Ω output emitter resistors. This is simple, and retains the immunity of I_q to heat-sink and output device temperatures, but it does sacrifice the total immunity to ambient temperature that a band-gap reference gives.

The LM385/1.2 is the lowest voltage band-gap reference commonly available; however, the voltages shown in Figure 10.18 reveal a difficulty with the new lower V_{bias} value and the CFP stage; points A and Y are now only 960mV apart, which does not give the reference room to work in if it is powered from node A, as in the original circuit. The solution is to power the reference from the V+ rail, via R42 and R43. The mid-point of these two resistors is bootstrapped from the amplifier output rail by C5, keeping the voltage across R43 effectively constant. Alternatively, a current source could be used, but this might reduce positive headroom. Since there is no longer a strict upper limit on the reference voltage, a more easily obtainable 2.56V device could be used providing R30 is suitably increased to 5k to maintain V_{ref} at 300mV across R31.

In practical use, I_q stability is very good, staying within 1% for long periods. The most obvious limitation on stability is differential heating of TR15, TR16 due to heat radiation from the main heat-sink. TR14 should also be sited with this in mind, as heating it will increase its beta and slightly imbalance TR15, TR16.

Class-B Mode

In Class-B mode, the current controller is disabled by turning off tail-source TR17 so TR14 is firmly off, and critical biasing for minimal crossover distortion is provided as usual by V_{be}-multiplier TR13. With 0.1Ω emitter resistors V_{bias} (between X and Y) is approximately 10mV. I would emphasize that in Class-B this design, if constructed correctly, will be as Blameless as a purpose-built Class-B amplifier. No compromises have been made in adding the mode-switching.

As in the previous Class-B design, the addition of R14 to the V_{be}-multiplier compensates against drift of the VAS current-source TR5. To make an old but much-neglected point, the preset should always

be in the bottom arm of the V_{be}-divider R10, R11, because when presets fail it is usually by the wiper going open; in the bottom arm this gives minimum V_{bias}, but in the upper it would give maximum.

In Class-B, temperature compensation for changes in driver dissipation remains vital. Thermal runaway with the CFP is most unlikely, but accurate quiescent setting is the only way to minimize crossover distortion. TR13 is therefore mounted on the same small heat-sink as driver TR6. This is often called thermal feedback, but it is no such thing as TR13 in no way controls the temperature of TR6; 'thermal feedforward' would be a more accurate term.

The Mode-Switching System

The dual nature of the biasing system means Class-A/Class-B switching can be implemented fairly simply. A Class-A amplifier is an uneasy companion in hot weather, and so I have been unable to resist the temptation to subtitle the mode switch summer/winter, by analogy with a car air intake.

The switchover is DC-controlled, as it is not desirable to have more signal than necessary running around inside the box, possibly compromising interchannel crosstalk. In Class-A/AB mode, SW1 is closed, so TR17 is biased normally by D5, D6, and TR20 is held on via R33, shorting out preset PR1 and setting TR13 to safety mode, maintaining a maximum V_{bias} limit of 600 mV. For Class-B, SW1 is opened, turning off TR17 and therefore TR15, TR16, and TR14. TR20 also ceases to conduct, protected against reverse-bias by D9, and reduces the voltage set by TR13 to a suitable level for Class-B. The two control pins of a stereo amplifier can be connected together, and the switching performed with a single-pole switch, without interaction or increased crosstalk.

The mode-switching affects the current flowing in the output devices, but not the output voltage, which is controlled by the global feedback loop, and so it is completely silent in operation. The mode may be freely switched while the amplifier is handling audio, which allows some interesting A/B listening tests.

It may be questioned why it is necessary to explicitly disable the current controller in Class-B; TR13 is establishing a lower voltage than the current controller, the subsystem of which will therefore turn TR14 off as it strives in a futile manner to increase V_{bias}. This is true for 8 Ω loads, but 4 Ω impedances increase the currents flowing in R16, R17 so they are transiently greater than the Class-A I_q, and the controller will therefore intermittently take control in an attempt to reduce the average current to 1.5 A. Disabling the controller by turning off TR17 via R44 prevents this.

If the Class-A controller is enabled, but the preset PR1 is left in circuit (e.g. by shorting TR20 base emitter) we have a test mode that allows suitably cautious testing; I_q is zero with the preset fully down, as TR13 overrides the current controller, but increases steadily as PR1 is advanced, until it suddenly locks at the desired quiescent current. If the current controller is faulty then I_q continues to increase to the defined maximum of 3.0 A.

Thermal Design

Class-A amplifiers are hot almost by definition, and careful thermal design is needed if they are to be reliable, and not take the varnish off the Sheraton. The designer has one good card to play; since

the internal dissipation of the amplifier is maximal with no signal, simply turning on the prototype and leaving it to idle for several hours will give an excellent idea of worst-case component temperatures. In Class-B the power dissipation is very program-dependent, and estimates of actual device temperatures in realistic use are notoriously hard to make.

Table 10.5 shows the output power available in the various modes, with typical transformer regulation, etc.; the output mode diagram in Figure 10.11 shows exactly how the amplifier changes mode from A to AB with decreasing load resistance. Remember that in this context 'high distortion' means 0.002% at 1 kHz. This diagram was produced in the analysis section of PSPICE simply by typing in equations, and without actually simulating anything at all.

The most important thermal decision is the size of the heat-sink; it is going to be expensive, so there is a powerful incentive to make it no bigger than necessary. I have ruled out fan cooling as it tends to make concern for ultra-low electrical noise look rather foolish; let us rather spend the cost of the fan on extra cooling fins and convect in ghostly silence. The exact thermal design calculations are simple but tedious, with many parameters to enter – the perfect job for a spreadsheet. The final answer is the margin between the predicted junction temperatures and the rated maximum. Once power output and impedance range are decided, the heat-sink thermal resistance to ambient is the main variable to manipulate, and this is a compromise between coolness and cost, for high junction temperatures always reduce semiconductor reliability. This is summarized very roughly in Table 10.6.

Table 10.6 shows that the transistor junctions will be 80°C above ambient, i.e. at around 100°C; the rated junction maximum is 200°C, but it really is not wise to get anywhere close to this very real limit. Note the case-sink thermal washers were high-efficiency material, and standard versions have a slightly higher thermal resistance.

Table 10.5: Power capability

	Load resistance			
	8 Ω	**6 Ω**	**4 Ω**	**Distortion**
Class-A	20 W	27 W	15 W	Low
Class-AB	n/a	n/a	39 W	High
Class-B	21 W	28 W	39 W	Medium

Table 10.6: Temperature rises resulting in a 100°C junction temperature

	Thermal resistance (°C/W)	**Heat flow (W)**	**Temp. rise (°C)**	**Temp. (°C)**
Junction to TO3 case	0.7	36	25	100 junction
Case to sink	0.23	36	8	75 TO3 case
Sink to air	0.65	72	47	67 heat-sink
Total			80	20 ambient

The heat-sinks used in the prototype had a thermal resistance of 0.65°C/W per channel. This is a substantial chunk of metal, and since aluminum is basically congealed electricity, it's bound to be expensive.

A Complete Trimodal Amplifier Circuit

The complete Class-A amplifier is shown in Figure 10.19, complete with optional input bootstrapping. It may look a little complex, but we have only added four low-cost transistors to realize a high-accuracy Class-A quiescent controller, and one more for mode-switching. Since the biasing system has been described above, only the remaining amplifier subsystems are dealt with here.

The input stage follows my design methodology by using a high tail current to maximize transconductance, and then linearizing by adding input degeneration resistors R2, R3 to reduce the final transconductance to a suitable level. Current-mirror TR10, TR11 forces the collector currents of the two input devices TR2, TR3 to be equal, balancing the input stage to prevent the generation of second-harmonic distortion. The mirror is degenerated by R6, R7 to eliminate the effects of V_{be} mismatches in TR10, TR11. With some misgivings I added the input network R9, C15, which is definitely not intended to define the system bandwidth, unless fed from a buffer stage; with practical values the HF roll-off could vary widely with the source impedance driving the amplifier. It is intended rather to give the possibility of dealing with RF interference without having to cut tracks. R9 could be increased for bandwidth definition if the source impedance is known, fixed, and taken into account when choosing R9; bear in mind that any value over 47 Ω will measurably degrade the noise performance. The values given roll-off above 150 MHz to keep out UHF.

The input stage tail current is increased from 4 to 6 mA, and the VAS standing current from 6 to 10 mA over the original Chapter 7 circuit. This increases maximum positive and negative slew rates from +21, −48 to +37, −52 V/μs; as described in Chapter 8, this amplifier architecture is bound to slew asymmetrically. One reason is feed-through in the VAS current source; in the original circuit an unexpected slew-rate limit was set by fast edges coupling through the current-source c-b capacitance to reduce the bias voltage during positive slewing. This effect is minimized here by using the negative-feedback type of current-source bias generator, with VAS collector current chosen as the controlled variable. TR21 senses the voltage across R13, and if it attempts to exceed V_{be}, turns on further to pull up the bases of TR1 and TR5. C11 filters the DC supply to this circuit and prevents ripple injection from the V+ rail. R5, C14 provide decoupling to prevent TR5 from disturbing the tail current while controlling the VAS current.

The input tail-current increase also slightly improves input stage linearity, as it raises the basic transistor g_m and allows R2, R3 to apply more local NFB.

The VAS is linearized by beta-enhancing stage TR12, which increases the amount of local NFB through Miller dominant-pole capacitor C3 (i.e. C_{dom}). R36 has been increased to 2k2 to minimize

power dissipation, as there seems no significant effect on linearity or slewing. Do not omit it altogether, or linearity will be affected and slewing much compromised.

As described in Chapter 9, the simplest way to prevent ripple from entering the VAS via the V− rail is old-fashioned RC decoupling, with a small R and a big C. We have some 200 mV in hand (see page 314) in the negative direction, compared with the positive, and expending this as the voltage-drop through the RC decoupling will give symmetrical clipping. R37 and C12 perform this function; the low rail voltages in this design allow the 1000 µF C12 to be a fairly compact component.

The output stage is of the CFP type that, as previously described, gives the best linearity and quiescent stability, due to the two local negative-feedback loops around driver and output device. Quiescent stability is particularly important with R16, R17 at 0.1 Ω, and this low value might be rather dicey in a double EF output stage. The CFP voltage efficiency is also higher than the EF version. R25, R26 define a suitable quiescent collector current for the drivers TR6, TR8, and pull charge carriers from the output device bases when they are turning off. The lower driver is now a BD136; this has a higher fT than the MJE350, and seems to be more immune to odd parasitics at negative clipping.

The new lower values for the output emitter resistors R16, R17 halve the distortion in Class-AB. This is equally effective when in Class-A with too low a load impedance, or in Class-B but with I_q maladjusted too high. It is now true in the latter case that too much I_q really is better than too little – but not much better, and AB still comes a poor third in linearity to Classes A and B.

Safe operating area (SOAR) protection is given by the networks around TR18, TR19. This is a single-slope SOAR system that is simpler than two-slope SOAR, and therefore somewhat less efficient in terms of getting the limiting characteristic close to the true SOAR of the output transistor. In this application, with low rail voltages, maximum utilization of the transistor SOAR is not really an issue; the important thing is to observe maximum junction temperatures in the A/AB mode.

The global negative-feedback factor is 32 dB at 20 kHz, and this should give a good margin of safety against Nyquist-type oscillation. Global NFB increases at 6 dB/octave with decreasing frequency to a plateau of around 64 dB, the corner being at a rather ill-defined 300 Hz; this is then maintained down to 10 Hz. It is fortunate that magnitude and frequency here are non-critical, as they depend on transistor beta and other doubtful parameters.

It is often stated in hi-fi magazines that semiconductor amplifiers sound better after hours or days of warm-up. If this is true (which it certainly is not in most cases) it represents truly spectacular design incompetence. This sort of accusation is applied with particular venom to Class-A designs, because it is obvious that the large heat-sinks required take time to reach final temperature, so I thought it important to state that in Class-A this design stabilizes its electrical operating conditions in less than a second, giving the full intended performance. No ‘warm-up time’ beyond this is required; obviously the heat-sinks take time to reach thermal equilibrium but, as described above, measures have been taken to ensure that component temperature has no significant effect on operating conditions or performance.

The Power Supply

A suitable unregulated power supply is that shown in Figure 9.2; a transformer secondary voltage of 20–0–20 V rms and reservoirs totaling 20,000 μF per rail will give approximately ±24 V. This supply must be designed for continuous operation at maximum current, so the bridge rectifier must be properly heat-sunk, and careful consideration given to the ripple-current ratings of the reservoirs. This is one reason why reservoir capacitance has been doubled to 20,000 μF per rail, over the 10,000 μF that was adequate for the Class-B design; the ripple voltage is halved, which improves voltage efficiency as it is the ripple troughs that determine clipping onset, but in addition the ripple current, although unchanged in total value, is now split between two components. (The capacitance was not increased to reduce ripple injection, which is dealt with far more efficiently and economically by making the PSRR high.) Do not omit the secondary fuses; even in these modern times rectifiers do fail, and transformers are horribly expensive.

The Performance

The performance of a properly designed Class-A amplifier challenges the ability of even the Audio Precision measurement system. To give some perspective on this, Figure 10.20 shows the distortion of the AP oscillator driving the analyzer section directly for various bandwidths. There appear to be internal mode changes at 2 and 20 kHz, causing step increases in oscillator distortion content; these are just visible in the THD plots for Class-A mode.

Figure 10.21 shows Class-B distortion for 20 W into 8 and 4 Ω, while Figure 10.22 shows the same in Class-A/AB. Figure 10.23 shows distortion in Class-A for varying measurement bandwidths. The lower bandwidths misleadingly ignore the HF distortion, but give a much clearer view of the excellent linearity below 10 kHz. Figure 10.24 gives a direct comparison of Classes A and B. The HF rise for B is due to high-order crossover distortion being poorly linearized by negative feedback that falls with frequency.

Further Possibilities

One interesting extension of the ideas presented here is the adaptive Trimodal amplifier. This would switch into Class-B on detecting device or heat-sink over-temperature, and would be a unique example of an amplifier that changed mode to suit the operating conditions. The thermal protection

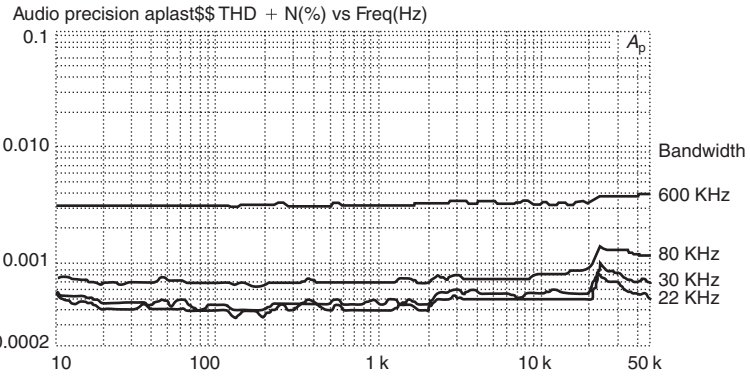


Figure 10.20: The distortion in the AP-1 system at various measurement bandwidths

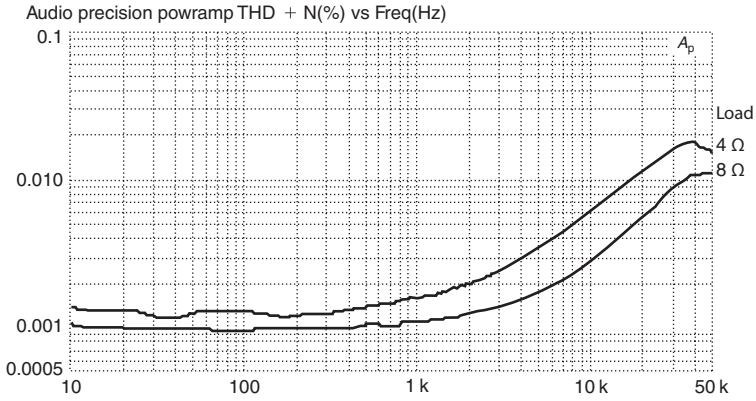


Figure 10.21: Distortion in Class-B (Summer) mode. Distortion into 4 Ω is always worse. Power was 20 W in 8 Ω and 40 W in 4 Ω, bandwidth 80 kHz

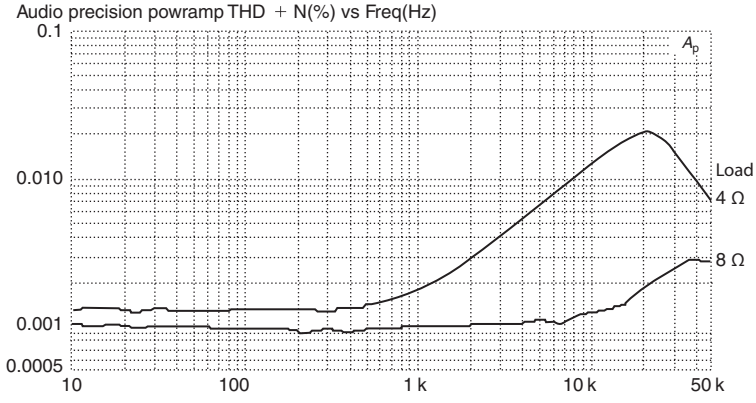


Figure 10.22: Distortion in Class-A/AB (Winter) mode, same power and bandwidth as in Figure 10.21. The amplifier is in AB mode for the 4 Ω case, and so distortion is higher than for Class-B into 4 Ω. At 80 kHz bandwidth, the Class-A plot below 10 kHz merely shows the noise floor

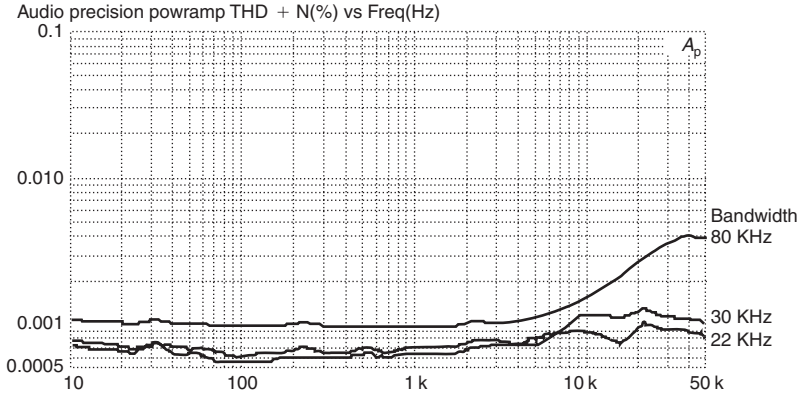


Figure 10.23: Distortion in Class-A only (20 W/8 Ω) for varying measurement bandwidths. The lower bandwidths ignore HF distortion, but give a much clearer view of the excellent linearity below 10 kHz

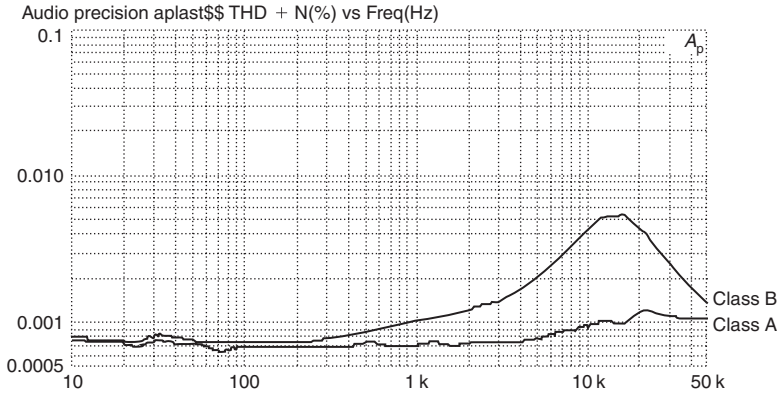


Figure 10.24: Direct comparison of Classes A and B (20 W/8 Ω) at 30 kHz bandwidth. The HF rise for B is due to the inability of negative feedback that falls with frequency to linearize the high-order crossover distortion in the output stage

would need to be latching; flipping from Class-A to Class-B every few minutes would subject the output devices to unnecessary thermal cycling.

References

[1] B.J. Moore, An Introduction to the Psychology of Hearing, Academic Press, 1982, pp. 48–50.

[2] S. Tanaka, A new biasing circuit for Class-B operation, JAES (January/February 1981) p. 27.

[3] S. Fuller, Private communication.

[4] N. Pass, Build a Class-A amplifier, Audio (February 1977) p. 28 (constant current).

[5] J. Linsley-Hood, Simple Class-A amplifier, Wireless World (April 1969) p. 148.

[6] D. Self, High-performance preamplifier, Wireless World (February 1979) p. 41.

[7] L. Nelson-Jones, Ultra-low distortion Class-A amplifier, Wireless World (March 1970) p. 98.

[8] D. Self, Self On Audio, second ed., Newnes, 2006, p. 459.

[9] T. Giffard, Class-A power amplifier, Elektor (November 1991) p. 37.

[10] L. Simpson, P. Smith, 20W Class-A amplifier module, Everyday Practical Electronics (October 2008) p. 32.

[11] J. Linsley-Hood, High-quality headphone amp, Hi-fi News & RR (January 1979) p. 81.

[12] N. Pass, The Pass/A40 power amplifier, Audio Amateur (1978) p. 4 (push–pull).

[13] N. Thagard, Build a 100 W Class-A mono amp, Audio (January 1995) p. 43.

CHAPTER 11

Class-XD™: Crossover Displacement Technology

Class-XD™ is a new output stage technology I have devised which abolishes crossover distortion up to a certain power level, without any accompanying compromises. 'XD' is derived from the phrase 'crossover displacement': the technology is covered by British patent GB2424137B and is proprietary to Cambridge Audio. At the time of writing it has so far been used in the Azur 840A and 840W power amplifiers, for both of which I did the electronic design. 'Class-XD' is a trademark of Audio Partnership PLC, and it should be pointed out that the use of the Class-XD concept and its trademark is restricted; I have permission from Cambridge Audio to use the term and describe the circuitry but no license to use the technology is implied or granted by the publication of this description.

Having held various posts in companies concerned with audio power amplifiers, I have frequently had to deal with enthusiastic inventors who feel they have come up with an output stage technology that overcomes the crossover distortion problems of conventional Class-B, and who are anxious to sell the idea to me. Two stick in the mind. There was the consortium that took out extensive worldwide patents on an idea that had been disclosed in *Wireless World* a quarter of a century before, and which did not work properly anyway. Then there was the chap who offered me an error-correcting output stage that 'only requires another 140 transistors'. I would have liked to have seen that circuit diagram, but not enough to pay money to do so.

In the light of this sort of thing, anyone is entitled to be skeptical about new and improved amplifier output stages. However, Class-XD is different; it really does work, doing what it claims with total reliability and minimal extra circuitry, as I shall now demonstrate.

One of the main themes of this book is the difficulty of dealing with crossover distortion in a Class-B output stage. I have described various methods of attack, such as the use of multiple output devices to reduce the current changes in each output transistor, and the use of two-pole compensation to increase the global negative-feedback factor. Both methods usefully reduce the amount of crossover distortion but do not eliminate it. As a result, one of the great divides in amplifier technology is still between efficient but imperfect Class-B and beautifully linear but dishearteningly inefficient Class-A. As I demonstrated in my book *Self On Audio*, a Class-A amplifier may theoretically be 50% efficient with a maximum sine-wave output, but when it reproduces a real music signal this falls to 1% or 2%[1]. For those that care at all about the economic utilization of energy, a Class-A amplifier is not an attractive proposition.

Class-B linearity can of course be very good. The Blameless amplifier design methodology, especially in its Load-Invariant form, yields less than 0.001% THD at 1 kHz. The limitation is that a Class-B amplifier inherently generates crossover distortion, and most inconveniently does so at

the zero-crossing, so it is always present no matter how low the signal amplitude. At one unique setting of quiescent conditions the distortion produced is at a minimum, and this characterizes optimal Class-B, but at no value can it be made to disappear. It is inherent in the classical Class-B operation of a pair of output transistors.

Given these two alternatives, there has always been a desire for a compromise between the efficiency of Class-B and the linearity of Class-A. The most obvious approach is to turn up the quiescent current of a Class-B stage, to create an area of Class-A operation, with both output transistors conducting, around the zero-crossing. This area widens as the quiescent current increases, until ultimately it encompasses the entire voltage output range of the amplifier, and we have created a pure Class-A design where both output transistors are conducting all the time. There is thus a range of quiescent current between Class-B and Class-A, and this mode of operation is called Class-AB. It is certainly a compromise between Class-A and Class-B, but not a good compromise, as it introduces extra distortion of its own.

This appears when the signal exceeds the limits of the Class-A region. The THD worsens abruptly due to the sudden gain changes when the output transistors turn on and off, and linearity is inferior not only to Class-A but also to optimally biased Class-B. This effect is often called 'g_m-doubling' and is dealt with in detail in Chapter 6. Class-AB distortion can be made very low by proper design, such as using the lowest practicable emitter resistors, but it remains at least twice as high as for the equivalent Class-B situation. The bias control of a Class-B amplifier does *not* give a straightforward trade-off between power dissipation and linearity at all levels, despite the constant repetition this misguided notion receives in some parts of the audio press. To demonstrate this, Figure 11.1 shows THD plotted against output level for Classes AB and B.

What we really want is an amplifier that would give Class-A performance up to the transition level, with Class-B after that, rather than the unsatisfactory Class-AB. This would abolish the abrupt AB gain changes that generate the extra distortion.

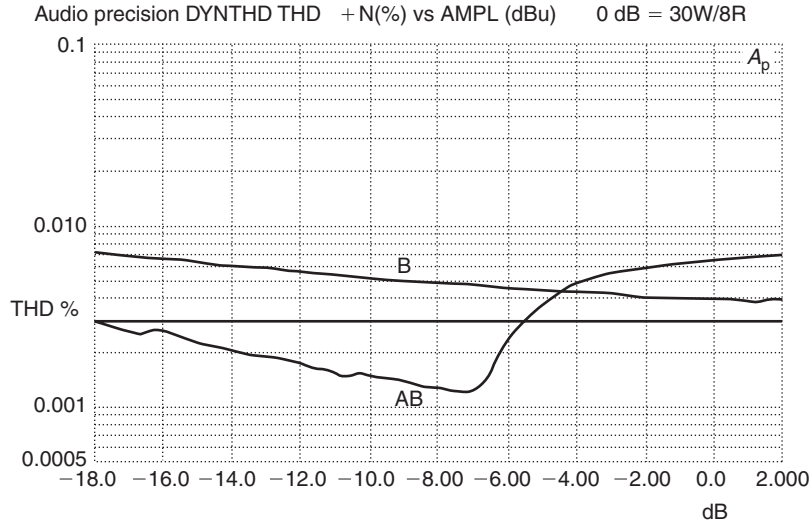


Figure 11.1: THD versus level for Class-B and Class-AB (0 dB is 30 W into 8 Ω)

The Crossover Displacement Principle

When we consider Class-B, it is clear that it would be better if the crossover region were anywhere else rather than where it is. If we can displace the crossover point away from its zero-crossing position, then the amplifier output will not traverse it until the output reaches a certain voltage level. Below this level the performance is pure Class-A; above it the performance is optimal Class-B, the only difference being that crossover discontinuities on the THD residual are no longer evenly spaced. The harmonic structure of the crossover distortion produced is not significantly changed, as explained in more detail below.

The central idea of the crossover displacement principle is the injection of an extra current, either fixed or varying with the signal, into the output point of a conventional Class-B amplifier. This is illustrated in Figure 11.2, where a black box I have called the 'displacer' draws a controlled displacement current from the output and sinks it into the negative rail; sourcing current from the positive rail and injecting it into the output would be equally valid. The displacer current may be constant, or vary with the signal.

The displacement current does not directly alter the output voltage because the output stage has an inherently low output impedance, which is further reduced by the global negative feedback. What it does do is alter the pattern of current flowing in the output devices. The displacement current in the version shown here is sunk to V− from the output, This is arbitrary as the direction of displacement makes no difference. The extra current therefore flows through Re1, and the extra voltage drop across it means the output voltage must go some way negative before the current through Re1 stops and that in Re2 starts. In other words, the crossover point when Q2 hands over to Q4 has been moved to a point negative of the 0V rail; I refer to this as the 'transition point' between Class-A and Class-B. For output levels below transition both Q2 and Q4 are conducting and no crossover distortion is generated. The resulting change in the incremental gain of the output stage is shown in Figure 11.3. Here the crossover region is moved 8V negative of ground by a 1A displacement

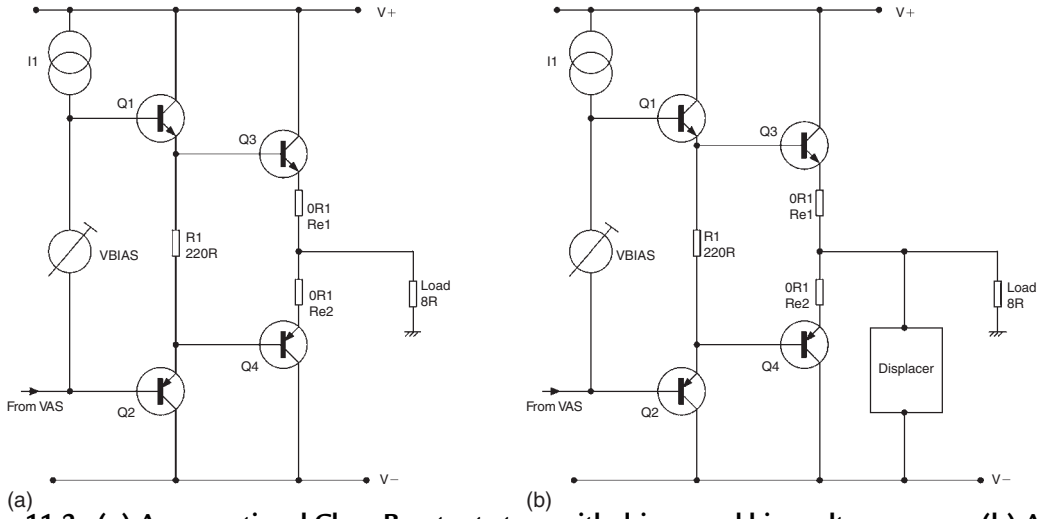


Figure 11.2: (a) A conventional Class-B output stage with drivers and bias voltage source. (b) Adding a displacer system that draws current from the output and sinks it into the negative rail

current; if the displacer had been connected to the positive rail the crossover region would have been pulled upwards. Note that the vertical scale is very much exaggerated, and that the crossover region has been moved but remains the same shape – the existing linearity has not been compromised.

I should emphasize here that crossover displacement in no way renders output stage bias adjustment unnecessary; if it is wrong the same distortion will occur, though only above a certain output level. This could be regarded as making the adjustment less critical, but getting it right costs no more than getting it roughly right, so there is really nothing to be gained by compromising on this.

We now have before us the intriguing prospect of a power amplifier with three output devices, which if nothing else is novel. The operation of the output stage is inherently asymmetrical, and indeed this is the whole point, but it should not cause alarm. Circuit symmetry is often touted as being a prerequisite for either low distortion or respectable operation in general, but this has no real foundation. A perfectly symmetrical circuit may have no even-order distortion, but it may still have frightening amounts of odd-order nonlinearity, such as a cubic characteristic. Odd-order harmonics are normally considered more dissonant than even-order, so circuit symmetry in itself is not enough.

In a conventional optimal Class-B amplifier, the crossover events are evenly spaced in time. In the crossover displacement amplifier, the crossover events are asymmetrical in time and put energy into both even and odd harmonics when operating above the transition point. However, since both even and odd exist already in conventional amplifiers, there is no cause for concern. As always, the real answer is to reduce the distortion, of whatever order, to so far below the noise floor that it could not possibly be audible and you never need to fret about it.

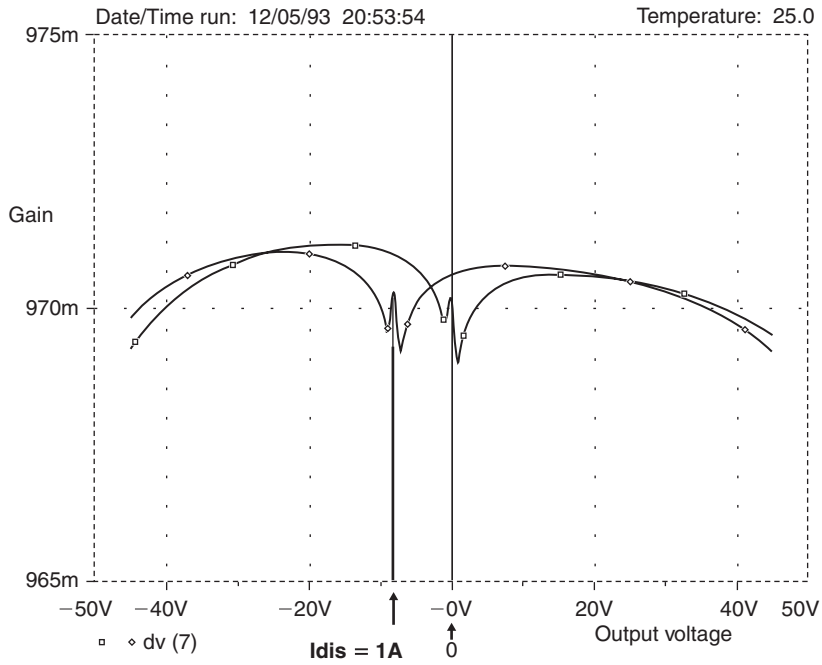


Figure 11.3: SPICE simulation of the output stage gain variation with and without a constant 1 A of displacement current. The central peak is moved left from 0 to −8 V

Crossover Displacement Realization

There are several ways in which a suitable displacement current can be drawn from the main amplifier output node. The simplest method is resistive crossover displacement. Connect a suitable power resistor between the output rail and a supply rail, as shown in Figure 11.4a, and the crossover point will be displaced. In this and all the following examples, the crossover point is displaced negatively by sinking a current into the negative rail.

The resistive method suffers from poor efficiency, as the resistance acts as another load on the amplifier output, effectively in parallel with the normal load. It also threatens ripple-rejection problems as R is connected directly to a supply rail, which in most cases is unregulated and carrying substantial 100 Hz ripple. A regulated supply to the resistor could be used, but this would be relatively expensive and even less efficient due to the voltage drop in the regulator. The resistive system is inefficient because the displacement of the crossover region occurs when the output is negative of ground, but when the output is positive the resistor is still connected and a greater current is drawn from it as the voltage across it increases. This increasing current is of no use in the displacement process and simply results in increased power dissipation in the positive output half-cycles.

This method has the other drawback that the distortion performance of the basic amplifier will be worsened because of the heavier loading it sees, the resistor being connected to ground as far as AC signals are concerned.

A superior solution is constant-current displacement, as shown in Figure 11.4b; here a constant-current source is connected between the output and negative rail. Efficiency is better as no output power is dissipated due to the high dynamic impedance of the current source. The output of the current source does not need to be controlled to very fine limits. Long-term variations in the current

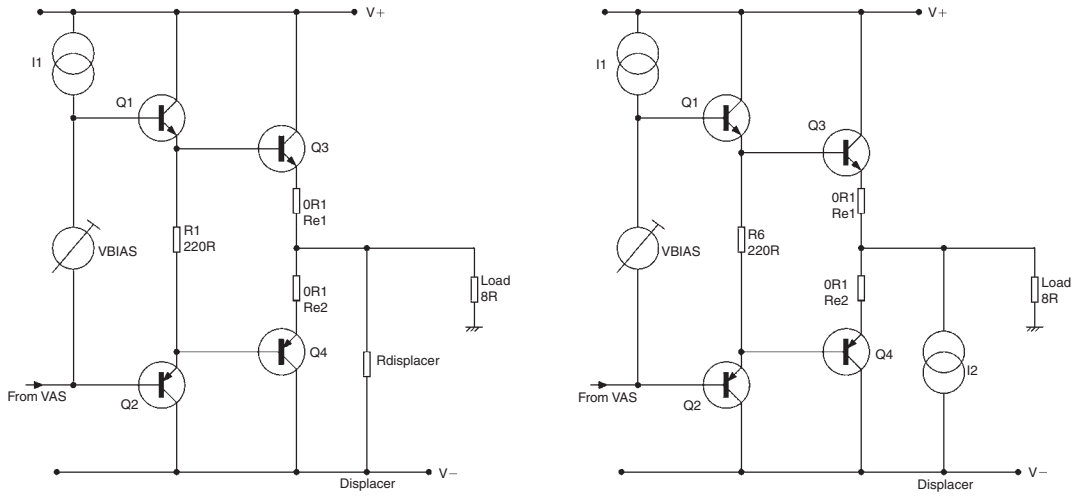


Figure 11.4: (a) The concept of resistive crossover displacement. (b) Constant-current crossover displacement

only affect the degree to which the crossover region is displaced, and this is not a critical parameter. Noise or ripple on the displacement current is greatly attenuated by the very low impedance of the basic power amplifier and its global negative feedback, so sophisticated current-control circuitry is not required. The efficiency of this configuration is greater because the output current of the displacer does not increase as the output moves more positive. The voltage across the current source increases, so its dissipation is still increased, but by a lesser amount than for the resistor. Likewise, the upper output transistor Q3 is passing less current on positive excursions so its power dissipation is less.

Having moved from a simple resistor displacer to a constant-current source, the obvious next step is to move from a constant current to a voltage-controlled current source (VCIS) whose output is modulated by the signal to further improve efficiency. The most straightforward way to do this is to make the displacement current proportional to the output voltage. Thus, if the displacement current is 1 A with the output quiescent at 0 V, it is set to increase to 2 A with the output fully negative, and to reduce to zero with the output fully positive. The displacer current is set by the equation:

$$I_d = I_q\left(1 - \frac{V_{out}}{V_{rail}}\right) \qquad \text{Equation 11.1}$$

where I_q is the quiescent displacement current (i.e. with the output at 0 V) and V_{rail} is the bottom rail voltage, which must be inserted as a negative number to make the arithmetic work. It is not essential for the displacement current to swing from zero to twice the quiescent value; it could be modulated to a lesser extent, and there is in fact a continuum of possible solutions from constant-current displacement to the full push–pull case.

Depending on the design of the VCIS, a scaling factor X is required to drive it correctly (see Figure 11.5). Since a signal polarity inversion is also necessary to get the correct mode of operation, active controlling circuitry is necessary.

The use of push–pull displacement is analogous to the use of push–pull current sources in Class-A amplifiers, where there is a well-known canonical sequence of increasing efficiency, which is fully described in Chapter 10. This begins with a resistive load giving only 12.5% efficiency at full power, moves to a constant-current source with high dynamic impedance giving 25%, and finally to a push–pull controlled current source, giving 50% efficiency. In the push–pull case the sink transistor acts in a sense as a negative resistance, though it is more usefully regarded as a driven source (VCIS) than a pure negative resistance, as the current does not depend on rail voltage. In each of these moves the efficiency doubles. These efficiency figures are ideal, ignoring circuit losses; note that Class-A efficiency is very seriously reduced at output powers less than the maximum. In the same way, there is a canonical sequence of sophistication and efficiency in crossover displacers, though the differences are smaller.

The push–pull displacement approach has another benefit; it also reduces distortion when operating above transition in the Class-B mode. This is because the push–pull system acts to reduce the current swings in the output devices, as the displacement current varies in the correct sense for this. This is equivalent to a decrease in output stage loading; this is the exact inverse of what occurs with resistive

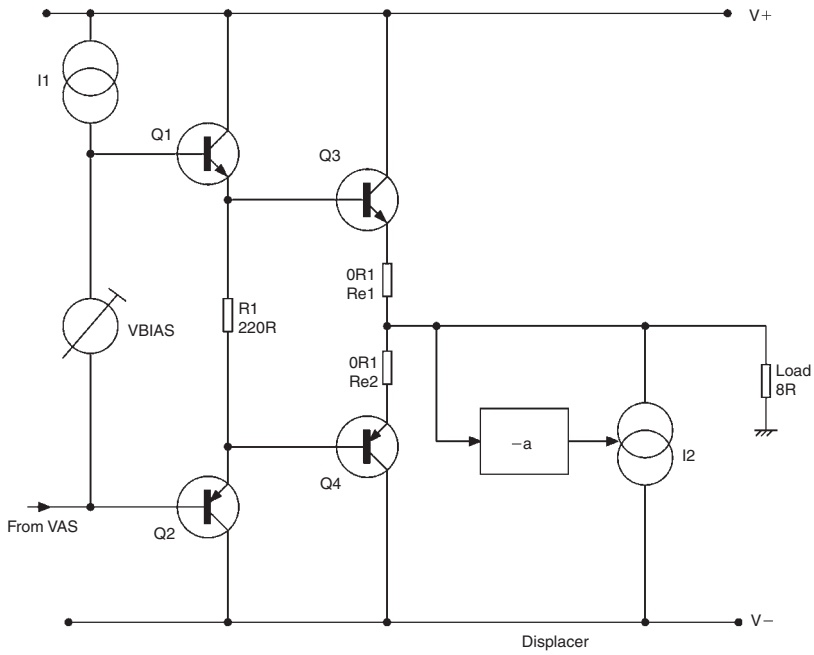


Figure 11.5: The concept of push–pull crossover displacement. The control circuitry implements a scaling factor of $-a$ in the signal to the controlled current source

displacement, which increases output loading. Lighter loading is known to make the current crossover between the output devices more gradual, and so reduces the size of the gain wobble that causes crossover distortion; this is described in Chapter 6. In addition the crossover region is spread over more of the output voltage range, so the distortion harmonics generated are lower order and receive more linearization from a negative-feedback factor that falls with frequency. Large-signal nonlinearity (typically experienced with loads of 4Ω and less) is also somewhat reduced. In push–pull displacement operation, the accuracy of the current variation does not have to be high to get the full reduction of the distortion, because of the low output impedance of the main amplifier, which maintains control of the output voltage. The global feedback around this amplifier is effective in reducing the inherently low output impedance of the output stage in the usual way, being unaffected by the addition of the displacer.

While the constant-current displacement method is simple and effective, the push–pull version of crossover displacement is to be preferred for the best linearity and efficiency; the extra control circuitry required is simple and works at low power so it adds minimally to total amplifier cost.

Circuit Techniques for Crossover Displacement

The constant-current displacer is the simplest practical displacement technique, the resistive version being discarded for the reasons given above.

A practical circuit for a constant-current displacer is shown in Figure 11.6a; for clarity the Class-B output stage is omitted. The displacement current typically chosen will be in the region of 0.5–1 A,

and therefore a driver transistor Q5 is used, exactly as drivers are used in the main amplifier, so the control circuitry can work at low power levels. The power device Q6 is going to get hot, so its V_{be} must be excluded from having a direct effect on current stability. Therefore the CFP (complementary feedback pair) structure shown is used, so the effect of V_{be} variations is reduced by the negative feedback around the local loop Q5–Q6. The bias for the constant current is shown as a Zener diode D1; if greater accuracy is required a low-voltage reference IC such as the LM385 could be used instead, but there is no real need to do so. The voltage across R1 should not be large enough to limit the output swing; but on the other hand, if it is small compared with the V_{be} of Q5, then the current value may drift excessively with temperature as Q5 warms up.

Power transistor Q6 dissipates significant heat; clearly the greater the crossover displacement required, the greater the displacement current and the greater the dissipation. Q6 is therefore normally mounted on the same heat-sink as the amplifier output devices. This provides the intriguing sight of a power amplifier with an odd number of output transistors, which might conceivably be exploited for marketing purposes.

The push–pull controller drives the displacer so that as the output rail goes positive, the displacer supplies less current. The basic problem is to apply a scaled and inverted version of the output voltage to the displacer. The signal must also have its reference transferred to the negative rail, which can be assumed to carry mains ripple and distorted signal components. Transferring the reference is done by using the high-impedance (like a current source) output from a bipolar transistor collector. As before, a driver transistor Q5 is used to drive the displacer Q6 so the control circuitry can work at low power levels. This not only minimizes total current consumption but also reduces the effect of V_{be} changes due to device heating (see Figure 11.6b).

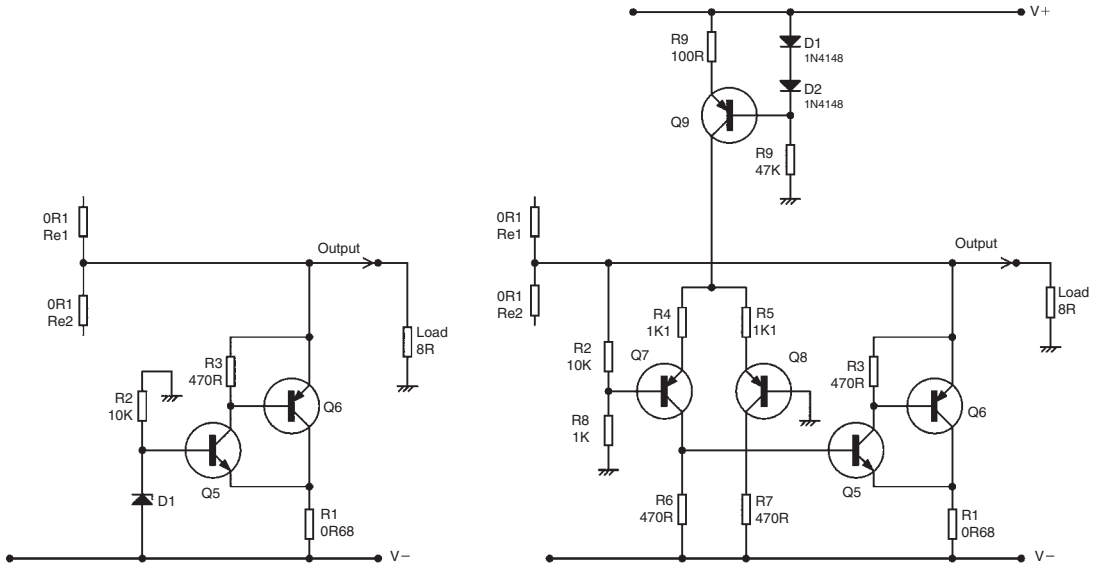


Figure 11.6: (a) Constant-current displacer with complementary feedback pair structure. (b) Push–pull displacer with differential pair controller

The controller is simply a differential pair of transistors with one input grounded and the other driven by the main amplifier output voltage, scaled down appropriately by R2, R8. The differential pair has heavy local feedback applied by the addition of the emitter resistors R4, R5, in order to minimize distortion and achieve an accurate gain. The drive to the VCIS displacer is taken from collector load R6, to give the required phase inversion. R7 is present simply to equalize the dissipation in the differential pair transistors to maintain balance.

The tail of the differential pair is fed by the 6 mA constant-current source Q9. This gives good common-mode rejection, which prevents the significant ripple voltages on the supply rails from interfering with the control signal. Since half of the standing current through the differential pair flows through R6, the value of the tail-current source sets the quiescent displacement current. The stability of the current generated by this source therefore sets the stability of the quiescent (no-signal) value of the displacement current. Figure 11.6b shows a simple current source biased by a pair of silicon diodes. This has proven to work well in practice but more sophisticated current sources using negative feedback could be used if greater stability is required. However, even if the tail-current source is perfect, the value of the displacement current still depends on the temperature of Q5. More sophisticated circuitry could be used to remove this dependency; for example, the voltage across R1 could be sensed by an op-amp instead of by Q5. The op-amp used would need to be able to work with a common-mode voltage down to the negative rail, or an extra supply rail would have to be provided.

A further possible refinement is the addition of a safety resistor in the differential pair tail to limit the amount of current flowing in the event of component failure. Such a resistor has no effect on normal operation, but it must be employed with care as its presence may mean that the circuit will not start working until the supply rails have risen to a large fraction of the working value. This is a serious drawback as it is wise to test power amplifiers by slowly raising the rail voltages from zero, and the lower the voltage at which they start working, the safer this procedure is.

A Complete Crossover Displacement Power Amplifier Circuit

Figure 11.7 shows the practical circuit of a push–pull crossover displacement amplifier. The Class-B amplifier is based on the Load-Invariant design and follows the Blameless design philosophy described elsewhere in this book. Conventional dominant-pole compensation is used. The design uses the following robust techniques described in this book to bring the distortion down to the irreducible minimum generated by a Class-B output stage.

1. The local negative feedback in the input differential pair Q1, Q2 is increased by running it at a high collector current, and then defining the stage transconductance and linearizing it by local negative feedback introduced by the emitter resistors R10, R11.

2. The crucial collector-current balance between the two halves of the input differential pair is enforced by the use of a degenerated current-mirror Q3, Q4.

3. The local negative feedback around the voltage-amplifier transistor Q10 is increased by adding the emitter-follower Q11 inside the Miller C_{dom} loop.

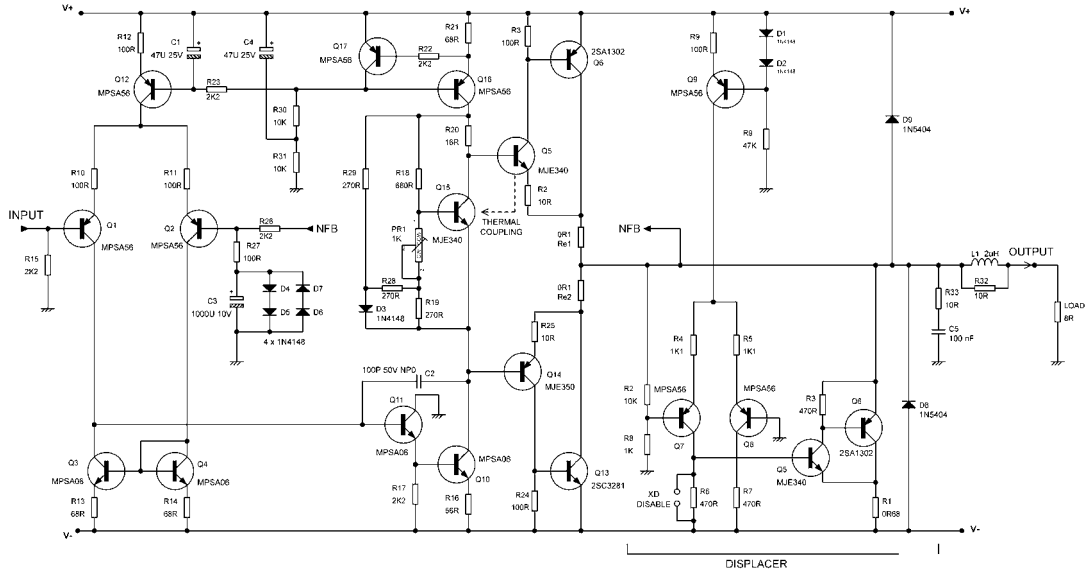


Figure 11.7: Complete circuit of an amplifier using push–pull crossover displacement

4. The output stage uses a complementary feedback pair (CFP) configuration to establish local negative feedback around the output devices. This increases linearity and also minimizes the effect of output junction temperatures on the bias conditions. The bias generator Q15 has its temperature coefficient increased by the addition of D3, R28, R29 to improve the accuracy of the thermal compensation (see Chapter 15 for more details). Q15 is thermally coupled to one of the drivers and preferably mounted on top of it; for this reason Q15 is an MJE340 simply so the packages are the same.

The circuit shown is capable of at least 50 W without modification. Powers above 100 W into 8 Ω will require two paralleled power transistors in the main amplifier output stage. The displacer transistor does not necessarily require doubling; it depends on the degree of crossover displacement desired.

The displacer control circuitry is essentially the same as in Figure 11.6b. A push-on link can be connected across R6 so that the crossover displacement action can be manually disabled to simplify testing and fault-finding.

Note that overload protection circuitry has been omitted from the diagram for simplicity.

The Measured Performance

The measurements shown here demonstrate how crossover displacement not only deals with crossover distortion, but also reduces distortion in general when the push–pull variant is employed. Tests were done with an amplifier similar to that shown in Figure 11.7.

Figure 11.8 shows THD versus frequency for a standard Blameless Class-B amplifier giving 30 W into 8 Ω. The distortion shown only emerges from the noise floor at 2 kHz, and is here wholly due to crossover artefacts; the bias is optimal and this is essentially as good as Class-B gets. The distortion

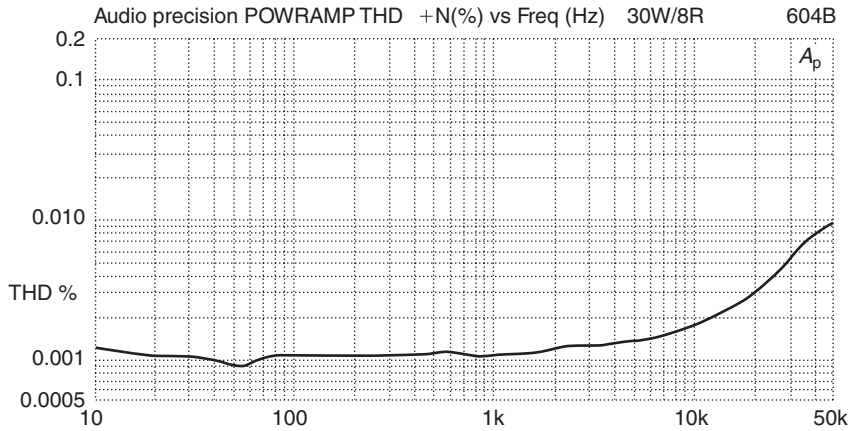


Figure 11.8: THD versus frequency for a standard Blameless Class-B amplifier at 30 W/8 Ω (604B)

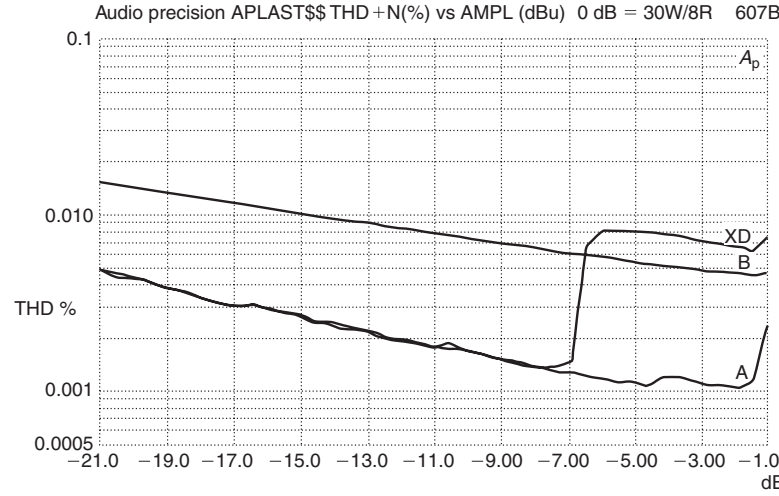


Figure 11.9: THD versus power out for Class-A, Class-B, and Class-B with constant-current crossover displacement (trace XD). Tested at 10 kHz to get enough distortion to measure; 0 dB = 30 W into 8 Ω (607B)

only gets really clear of the noise floor at 10 kHz, so this frequency was used for all the THD/amplitude tests below. This frequency provides a demanding test for an audio power amplifier. In all these tests the measurement bandwidth was 80 kHz. This may filter out some ultrasonic harmonics, but is essential to reduce the noise bandwidth; it is also a standard setting on many distortion analyzers.

Figure 11.9 plots distortion versus amplitude at 10 kHz, over the power range 200 mW–20 W; this covers the levels at which most listening is done (0 dB on the graph is 30 W into 8 Ω). Trace B is the result for the Blameless Class-B amplifier measured in Figure 11.8; the THD percentage increases as the power is reduced, partly because of the nature of crossover distortion, but more because the noise level becomes proportionally greater as level is reduced. Trace A shows the result for a Class-A amplifier of my design (see Chapter 10), which is pretty much distortion-free at 10 kHz, and once more simply shows the increasing relative noise level as power reduces; trace A is lower than trace B because of the complete

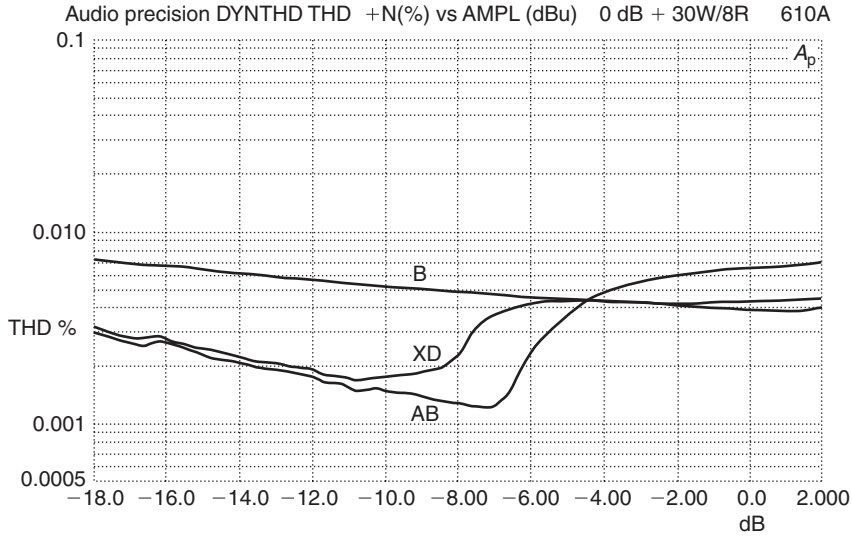


Figure 11.10: THD versus power out for Class-B, Class-AB, and Class-B with constant-current crossover displacement added (trace XD). At 10 kHz, 0 dB = 30 W/8 Ω (610A)

absence of crossover distortion. Trace XD demonstrates how a constant-current crossover displacement amplifier has the same superb linearity as Class-A up to an output of −7 dB, but distortion then rises to the Class-B level as the output swing begins to traverse the displaced crossover region. (In fact it slightly exceeds Class-B in this case, as the data was acquired before the prototype was fully optimized.)

A similar THD/amplitude plot in Figure 11.10 compares Class-B with Class-AB and constant-current crossover displacement, emphasizing that XD is superior to AB. Here the transition point from Class-A to Class-B is at −8 dB and the Class-AB case was biased so that g_m-doubling began at −7 dB. At this point the experimental amplifier was fully optimized and the constant-current crossover displacement distortion above the transition point is now the same as Class-B.

Figure 11.11 demonstrates that push–pull crossover displacement gives markedly lower distortion than constant-current crossover displacement. The transition points are not quite the same (−8 dB for push–pull versus −11 dB for constant-current) but this has no significant effect on the distortion produced. The salient point is that at −2 dB, for example, the THD is very significantly lowered from 0.0036% to 0.0022% by the use of the push–pull method, because of the way it reduces the magnitude of the current changes in the output transistors of the main amplifier.

Figure 11.12 is the same as Figure 11.11 with a THD versus level plot for Class-AB added, underlining the point that Class-AB gives significantly greater distortion above its transition point (say at −4 dB) than Class-B. As before, constant-current crossover displacement gives slightly less distortion than Class-B, and push–pull crossover displacement gives markedly less and is clearly the best mode of operation.

Figure 11.13 returns to the THD/frequency format, and is included to confirm that crossover displacement push–pull gives lower THD over the whole of the upper audio band from 1 to 20 kHz. Below 1 kHz the noise floor dominates.

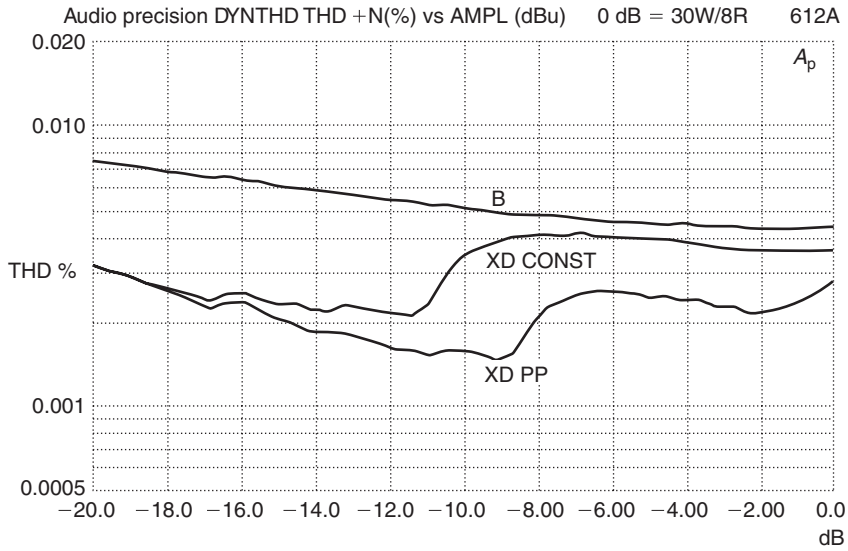


Figure 11.11: This shows that push–pull crossover displacement (XD PP) gives much lower distortion than constant-current crossover displacement (XD CONST). At 10 kHz, 0 dB = 30 W/8 Ω (612A)

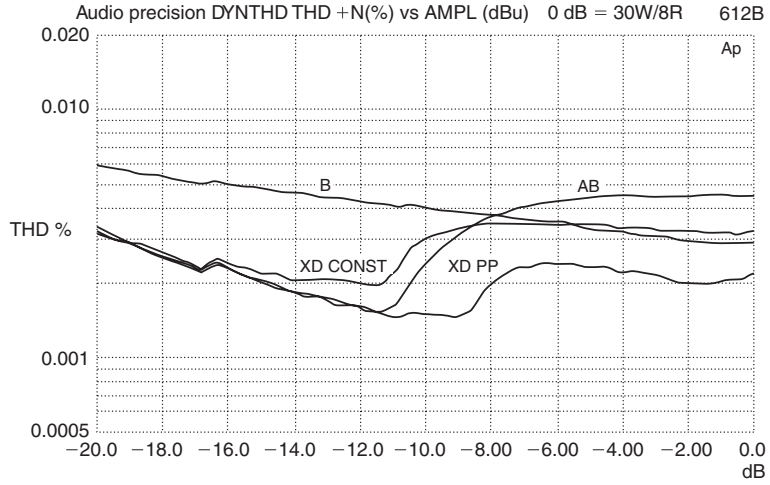


Figure 11.12: Adding the Class-AB plot shows that it is clearly the least linear mode above −8 dBu (612B)

The Effect of Loading Changes

When a new amplifier concept is considered, it is essential to consider its behavior into real loads, which deviate significantly from the classical 8 Ω resistance.

Firstly, what is the effect of changing the load resistance, for example by using a 4 Ω load? The signal currents in the output stage are doubled, so the voltage by which the crossover region is displaced is halved. Half the voltage across half the resistance means the output power is halved, so the volume at the transition point has been reduced by 3 dB. In terms of SPL and human perception, the reduction is

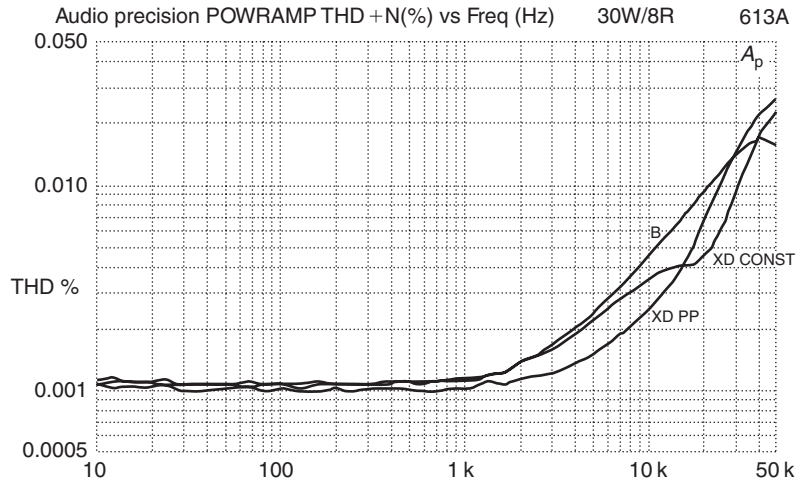


Figure 11.13: THD versus frequency for Class-B, constant-current crossover displacement (XD CONST), and push–pull crossover displacement (XD PP) at 30 W/8 Ω (613A)

not very significant. One of the concerns facing conventional Class-B amplifiers driving 4 Ω or less is the onset of large-signal nonlinearity caused by increased output device currents and consequent fall-off of beta in the driver transistors. The use of push–pull displacement reduces LSN in same way that crossover distortion is reduced – by reducing the range of current variation in the output stage.

Secondly, what about reactive loads? In particular we must scrutinize the way that the push–pull displacer is driven by output voltage rather than device currents. In a conventional Class-B amplifier, adding a reactive element to the load alters the phase relationship between the output voltage and the crossover events; this is because voltage and current are now out of phase, and crossover is a current-domain phenomenon. It has never been suggested this presents any sort of psychoacoustic problem. Putting reactive loading on a crossover displacement amplifier moves its crossover events (if the power level is above transition, otherwise they are not generated at all) in time with respect to the voltage output in exactly the same way, and there is absolutely no reason to suppose that this is any cause for concern.

Stability into reactive loads is not affected by the addition of the displacement system. As previously described, the displacer system has only a very minor effect on the output signal, its effects being confined to making the output stage operate in a more advantageous way, and so it has virtually no effect on the characteristics of the forward amplifier path, and therefore no effect on stability margins.

The Efficiency of Crossover Displacement

The crossover displacement technique obviously increases the total power dissipated in the output stage, and the efficiency is therefore somewhat worse than Class-B. The dissipation in the upper transistor is increased by the displacement current flowing through it, while that in the lower transistor is unchanged. There is also the additional dissipation in the displacer itself, which is likely to be mounted on the same heat-sink as the main output devices.

When using techniques such as Class-AB or crossover displacement that give a limited power output in Class-A, you must decide at the start just how much Class-A power you are prepared to

Table 11.1: Efficiency of amplifier types

	Full output	Half power	One-tenth power
Class-B	74%	54%	23%
Class-XD push–pull	66%	46%	14%
Class-XD constant	57%	39%	11%
Class-A	43%	23%	4%

pay for in terms of extra heat liberated, and just what load impedance you intend to drive in that mode. For example, assume that 5 W of Class-A operation into 8 Ω is required from an amplifier with a full output of 50 W. The crossover point must therefore be displaced by the peak voltage corresponding to 5 W, which is 8.9 V, and this will require a displacement current of about 1.1 A. It is well established that it takes about a 10 dB increase in sound intensity to double subjective loudness[2], and 10 dB is a power ratio of 10 times. Therefore if there is only a doubling in loudness between the transition point into Class-B and full output, the amplifier will be in the Class-A region most of the time; this seems like an eminently reasonable approach.

Table 11.1 shows the calculated efficiency for the various types of amplifier. The calculations were not based on the usual simplistic theory that ignores voltage drops in emitter resistors, transistor saturation voltages and so on, but emerged from a lengthy series of SPICE simulations of complete output stage circuits. The effects of transistor nonlinearity and so on are fully taken into account. The efficiency results are therefore slightly worse than simple theory predicts. I think they are as accurate as extensive calculation can make them. For comparison, the 'classical' calculations for Class-B give a full power efficiency of 78%, but the more detailed simulations show that it is only 74% when typical losses are included.

The output stages were simulated using 50 V rails, giving a maximum power of about 135 W into 8 Ω. Displacement currents were set to give a transition from Class-A to Class-B at 5 W. All emitter resistors were 0.1 Ω.

Here we have demonstrated that there is some penalty in efficiency when crossover displacement is used, but it is far more economical than Class-A. The push–pull XD mode is clearly better than constant-current operation. As I mentioned before, if real musical signals are used rather than sine waves, the Class-A amplifier comes off a *lot* worse with efficiency reduced to 1% or 2%.

Other Methods of Push–Pull Displacement Control

This article describes in detail push–pull crossover displacement implemented by controlling the displacer from the amplifier voltage output. This has the great merit of simplicity, but its method of operation requires design assumptions to be made about the minimum load impedances to be driven. If the load is of higher impedance than expected, which can often occur in loudspeaker loads because of voice-coil inductance, the displacer current may be increased more than is necessary for the desired amount of crossover displacement, leading to unnecessary power dissipation. This is because the voltage-control method is an open-loop or feedforward system.

This situation could be avoided by using a current-controlled system that senses the current flowing in the main amplifier output devices and turns on enough displacer current to give the amount of crossover displacement desired. This is a second negative-feedback loop operating at the full signal frequency, and experience has shown that high-frequency instability can be a serious problem with this sort of approach. Nonetheless I feel the concept is worthy of further investigation.

Summary

Crossover displacement provides a genuine way to compromise between the linearity of Class-A and the efficiency of Class-B. I think it is now firmly established that the conventional use of Class-AB, by simply turning up the bias, is not such a compromise because it introduces extra distortion.
An important merit of the technology is that it is robust and completely dependable. During the development of the Cambridge Audio Azur 840A and 840W power amplifiers, neither of the crossover displacement systems required so much as a change in a resistor value.

The pros and cons are summarized below.

Advantages

- Crossover distortion is moved away from the central point where an amplifier output spends most of its time; in normal use an amplifier can almost always run in Class-A.
- Push–pull displacement also reduces both crossover distortion and LSN when in Class-B operation.
- It is simple. Only five extra transistors are used, of which three are small-signal and of very low cost.
- There are no extra presets or adjustments.
- It does not affect HF stability
- No extra overload protection circuitry is needed, as the displacer is inherently current limited.
- The technology is versatile. It can be attached to almost any kind of Class-B amplifier.

Disadvantages

- There is some extra power dissipation, but far less than the use of Class-A.
- There is some extra cost in circuitry, but not much. Only one more power transistor is required.

References

[1] D. Self, Self On Audio, Newnes/Elsevier, 2006, p. 459.

[2] B.J. Moore, An Introduction to the Psychology of Hearing, Academic Press, 1982, pp. 48–50.

CHAPTER 12

Class-G Power Amplifiers

Most types of audio power amplifier are less efficient than Class-B; for example, Class-AB is markedly less efficient at the low end of its power capability, while it is clear from Chapter 10 that Class-A wastes virtually all the energy put into it. Building amplifiers with higher efficiency is more difficult. Class-D, using ultrasonic pulse width modulation, promises high efficiency and indeed delivers it, but it is undeniably a difficult technology, and its linearity is still a long way short of Class-B. The practical efficiency of Class-D rests on details of circuit design and device characteristics. The apparently unavoidable LC output filter – second order at least – can only give a flat response into one load impedance, and its magnetics are neither cheap nor easy to design. There are likely to be some daunting EMC difficulties with emissions. Class-D is not an attractive proposition for high-quality domestic amplifiers that must work with separate speakers of unknown impedance characteristics.

There is, however, the Class-G method. Power is drawn from either high- or low-voltage rails as the signal level demands. This technology has taken a long time to come to fruition, but is now used in very-high-power amplifiers for large PA systems, where the savings in power dissipation are important, and is also making its presence felt in home theater systems; if you have seven or eight power amplifiers instead of two their losses are rather more significant. Class-G is firmly established in powered subwoofers, and even in ADSL telephone-line drivers. Given the current concern for economy in energy consumption, Class-G may well become more popular in mainstream areas where its efficiency can be used as a marketing point. It is a technology whose time has come.

The Principles of Class-G

Music has a large peak-to-mean level ratio. For most of the time the power output is a long way below the peak levels, and this makes possible the improved efficiency of Class-G. Even rudimentary statistics for this ratio for various genres of music are surprisingly hard to find, but it is widely accepted that the range between 10 dB for compressed rock and 30 dB for classical material covers most circumstances.

If a signal spends most of its time at low power, then while this is true a low-power amplifier will be much more efficient. For most of the time lower output levels are supplied from the lowest-voltage rails, with a low voltage drop between rail and output, and correspondingly low dissipation. The most popular Class-G configurations have two or three pairs of supply rails, two being usual for hi-fi, while three is more common in high-power PA amplifiers.

When the relatively rare high-power peaks do occur they must be handled by some mechanism that can draw high power, causing high internal dissipation, but which only does so for brief periods. These infrequent peaks above the transition level are supplied from the high-voltage pair of rails. Clearly the switching between rails is the heart of the matter, and anyone who has ever done any circuit design will immediately start thinking about how easy or otherwise it will be to make this happen cleanly with a high-current 20 kHz signal.

There are two main ways to arrange the dual-rail system: series and parallel (i.e. shunt). This chapter deals only with the series configuration, as it seems to have had the greatest application to hi-fi. The parallel version is more often used in high-power PA amplifiers.

Introducing Series Class-G

A series configuration Class-G output stage using two rail voltages is shown in Figure 12.1. The so-called inner devices are those that work in Class-B; those that perform the rail-switching on signal peaks are called the outer devices – by me, anyway. In this design study the EF type of output stage is chosen because of its greater robustness against local HF instability, though the CFP configuration could be used instead for inner, outer, or both sets of output devices, given suitable care. For maximum power efficiency the inner stage normally runs in Class-B, though there is absolutely no reason why it could not be run in Class-AB or even Class-A; there will be more discussion of these intriguing possibilities later. If the inner power devices are in Class-B, and the outer ones conduct for much less than 50% of a cycle, being effectively in Class-C, then according to the classification scheme I have proposed[1], this should be denoted Class-B + C. The plus sign indicates the series rather than shunt connection of the outer and inner power devices. This basic configuration was developed by Hitachi to reduce amplifier heat dissipation[2,3]. Musical signals spend most of their time at low levels, having a high peak/mean ratio, and power dissipation is greatly reduced by drawing from the lower ±V1 supply rails at these times.

The inner stage TR3, TR4 operates in normal Class-B. TR1, TR2 are the usual drivers and R1 is their shared emitter resistor. The usual temperature-compensated V_{bias} generator is required, shown here theoretically split in half to maintain circuit symmetry when the stage is SPICE simulated; since the inner power devices work in Class-B it is their temperature that must be tracked to maintain quiescent conditions. Power from the lower supply is drawn through D3 and D4, often called the commutating diodes, to emphasize their rail-switching action. The word 'commutation' avoids confusion with the usual Class-B crossover at zero volts. I have called the level at which rail-switching occurs the transition level.

When a positive-going instantaneous signal exceeds low rail +V1, D1 conducts, TR5 and TR6 turn on and D3 turns off, so the entire output current is now taken from the high-voltage +V2 rail, with the voltage drop and hence power dissipation shared between TR4 and TR6. Negative-going signals are handled in exactly the same way. Figure 12.2 shows how the collector voltages of the inner power devices retreat away from the output rail as it approaches the lower supply level.

Class-G is commonly said to have worse linearity than Class-B, the blame usually being loaded onto the diodes and problems with their commutation. As usual, received wisdom is only half of

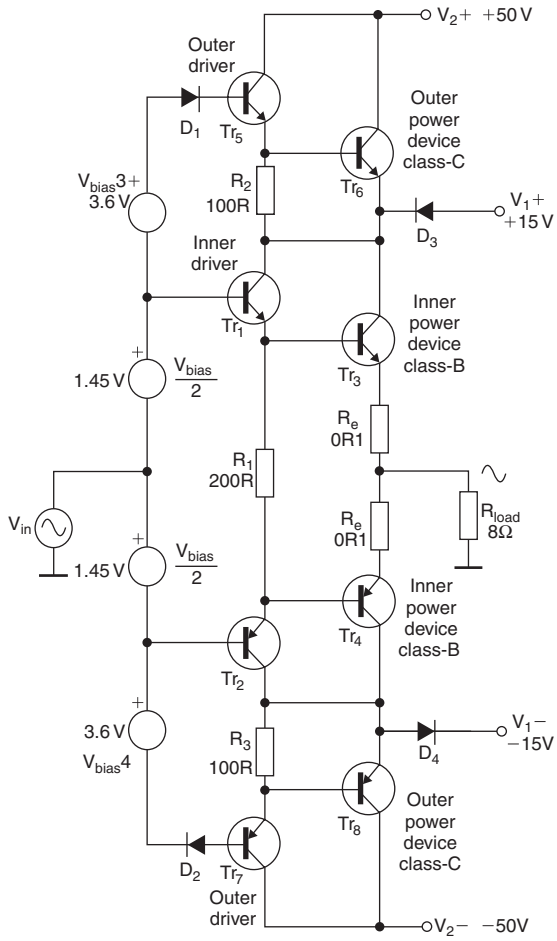


Figure 12.1: A series Class-G output stage, alternatively Class-B + C. Voltages and component values are typical. The inner stage is Class-B EF. Biasing by my method

the story, if that, and there are other linearity problems that are not due to sluggish diodes, as will be revealed shortly. It is inherent in the Class-G principle that if switching glitches do occur they only happen at moderate power or above, and are well displaced away from the critical crossover region where the amplifier spends most of its time. A Class-G amplifier has a low-power region of true Class-B linearity, just as a Class-AB amplifier has a low-power region of true Class-A performance.

Efficiency of Class-G

The standard mathematical derivation of Class-B efficiency with sine-wave drive uses straightforward integration over a half-cycle to calculate internal dissipation against voltage fraction, i.e. the fraction of possible output voltage swing. As is well known, in Class-B the maximum heat dissipation is about 40% of maximum output power, at an output voltage fraction of 63%, which also delivers 40% of the maximum output power to the load.

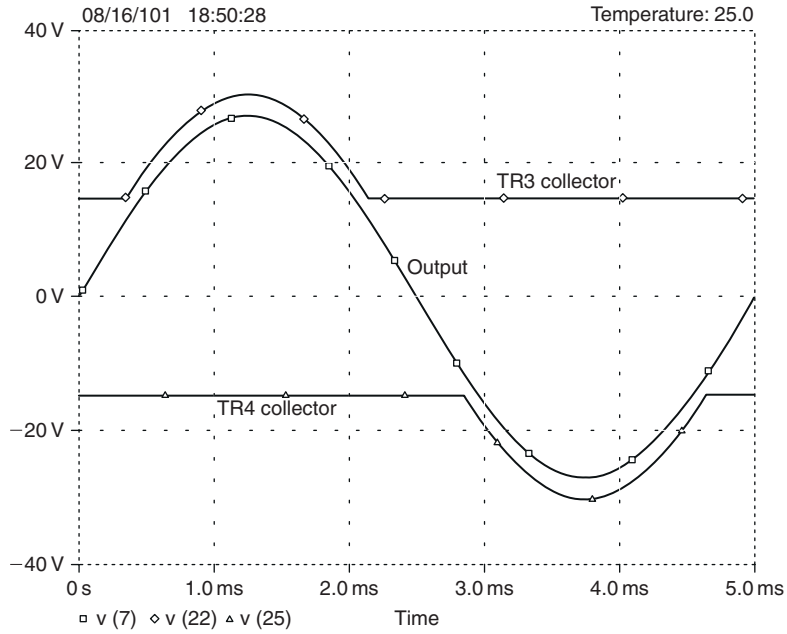


Figure 12.2: The output of a Class-G stage and the voltages on the collectors of the inner output devices

The mathematics is simple because the waveforms do not vary in shape with output level. Every possible idealization is assumed, such as zero quiescent current, no emitter resistors, no $V_{ce(sat)}$ losses and so on. In Class-G, on the other hand, the waveforms are a strong function of output level, requiring variable limits of integration and so on, and it all gets very unwieldy.

The SPICE simulation method described by Self[4] is much simpler, if somewhat laborious, and can use any input waveform, yielding a Power Partition Diagram (PPD), which shows how the power drawn from the supply is distributed between output device dissipation and useful power in the load.

No one disputes that sine waves are poor simulations of music for this purpose, and their main advantage is that they allow direct comparison with the purely mathematical approach. However, since the whole point of Class-G is power saving, and the waveform used has a strong effect on the results, I have concentrated here on the PPD of an amplifier with real musical signals or, at any rate, their statistical representation. The triangular probability density function (PDF) approach is described in Self[5].

Figure 12.3 shows the triangular PDF PPD for conventional Class-B EF, while Figure 12.4 is that for Class-G with $\pm V2 = 50V$ and $\pm V1 = 15V$, i.e. with the ratio of V1/V2 set to 30%. The PPD plots power dissipated in all four output devices, the load, and the total drawn from the supply rails. It shows how the input power is partitioned between the load and the output devices. The total sums to slightly less than the input power, the remainder being accounted for as usual by losses in the drivers and Re resistors. Note that in Class-G power dissipation is shared, though not very equally, between the inner and outer devices, and this helps with efficient utilization of the silicon.

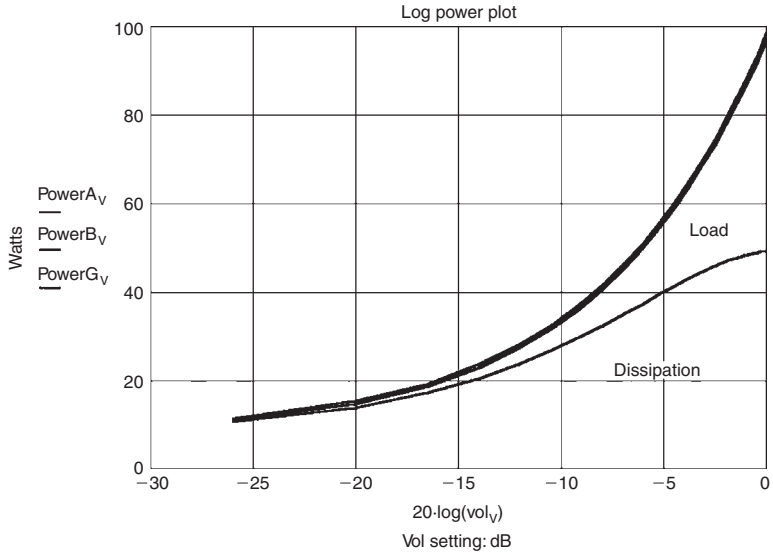


Figure 12.3: Power partition diagram for a conventional Class-B amplifier handling a typical music signal with a triangular probability density function. *X*-axis is volume

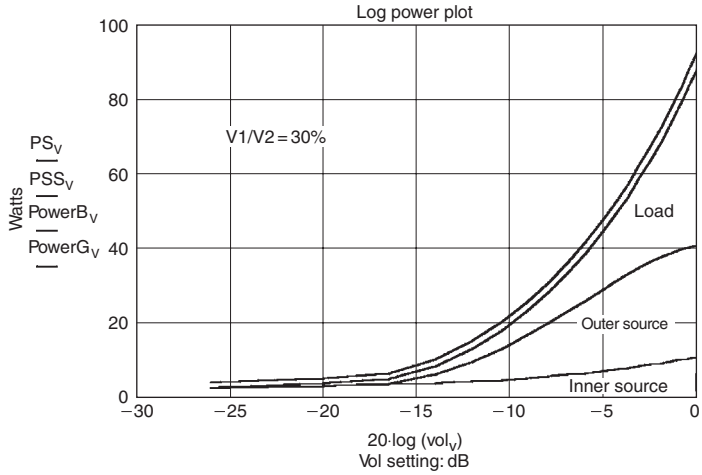


Figure 12.4: Power partition diagram for Class-G with V1/V2 = 30%. Signal has a triangular PDF. *X*-axis is volume; outer devices dissipate nothing until −15 dB is reached

In Figure 12.4 the lower area represents the power dissipated in the inner devices and the larger area just above represents that in the outer devices; there is only one area for each because in Class-B and Class-G only one side of the amplifier conducts at a time. Outer device dissipation is zero below the rail-switching threshold at −15 dB below maximum output. The total device dissipation at full output power is reduced from 48 W in Class-B to 40 W, which may not appear at first to be a very good return for doubling the power transistors and drivers.

Figure 12.5 shows the same PPD but with ±V2 = 50 V and ±V1 = 30 V, i.e. with V1/V2 set to 60%. The low-voltage region now extends up to −6 dB ref. full power, but the inner device

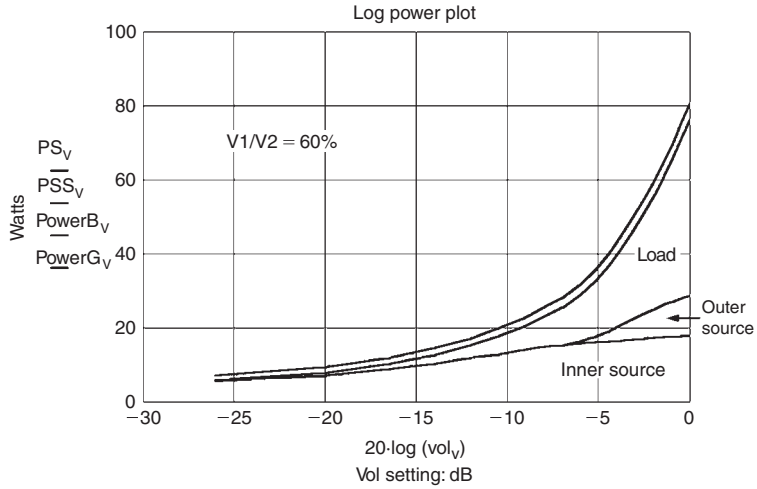


Figure 12.5: Power partition diagram for Class-G with V1/V2 = 60%. Triangular PDF. Compared with Figure 12.4, the inner devices dissipate more and the outer devices almost nothing except at maximum volume

dissipation is higher due to the higher V1 rail voltages. The result is that total device power at full output is reduced from 48 W in Class-B to 34 W, which is a definite improvement. The efficiency figure is highly sensitive to the way the ratio of rail voltages compares with the signal characteristics. Domestic hi-fi amplifiers are not operated at full volume all the time, and in real life the lower option for the V1 voltage is likely to give lower general dissipation. I do not suggest that V1/V2 = 30% is the optimum lower-rail voltage for all situations, but it looks about right for most domestic hi-fi.

Practicalities

In my time I have wrestled with many 'new and improved' output stages that proved to be anything but. When faced with a new and intriguing possibility, I believe the first thing to do is sketch out a plausible circuit such as Figure 12.1 and see if it works in SPICE simulation. It duly did.

The next stage is to build it, power it from low supply rails to minimize the size of any explosions, and see if it works for real at 1 kHz. This is a bigger step than it looks.

SPICE simulation is incredibly useful but it is not a substitute for testing a real prototype. It is easy to design clever and complex output stages that work beautifully in simulation but in reality prove impossible to stabilize at high frequencies. Some of the more interesting output-triple configurations seem to suffer from this.

The final step – and again it is a bigger one than it appears – is to prove real operation at 20 kHz and above. Again it is perfectly possible to come up with a circuit configuration that either just does not work at 20 kHz, due to limitations on power transistor speeds, or is provoked into oscillation or other misbehavior that is not set off by a 1 kHz testing.

Only when these vital questions are resolved is it time to start considering circuit details, and assessing just how good the amplifier performance is likely to be.

The Biasing Requirements

The output stage bias requirements are more complex than for Class-B. Two extra bias generators Vbias3, Vbias4 are required to make TR6 turn on before TR3 runs out of collector voltage. These extra bias voltages are not critical, but must not fall too low or become much too high. Should these bias voltages be set too low, so the outer devices turn on too late, then the V_{ce} across TR3 becomes too low, and its current sourcing capability is reduced. When evaluating this issue bear in mind the lowest impedance load the amplifier is planned to drive, and the currents this will draw from the output devices. Fixed Zener diodes of normal commercial tolerance are quite accurate and stable enough for setting Vbias3 and Vbias4.

Alternatively, if the bias voltage is set too low, then the outer transistors will turn on too early, and the heat dissipation in the inner power devices becomes greater than it need be for correct operation. The latter case is rather less of a problem so if in doubt this bias should be chosen to be on the high side rather than the low.

The original Hitachi circuit[1] put Zeners in series with the signal path to the inner drivers to set the output quiescent bias, their voltage being subtracted from the main bias generator, which was set at 10V or so, a much higher voltage than usual (see Figure 12.6). SPICE simulation showed me that the presence of Zener diodes in the forward path to the inner power devices gave poor linearity, which is not exactly a surprise. There is also the problem that the quiescent conditions will be affected by changes in the Zener voltage. The 10V bias generator, if it is the usual V_{be}-multiplier, will have much too high a temperature coefficient for proper thermal tracking.

I therefore rearrange the biasing as in Figure 12.1. The amplifier forward path now goes directly to the inner devices, and the two extra bias voltages are in the path to the outer devices; since these do not control the output directly, the linearity of this path is of lesser importance. The Zeners are out of the forward path and the bias generator can be the standard sort. It must be thermally coupled to the inner power devices; the outer ones have no effect on the quiescent conditions.

The Linearity Issues of Series Class-G

Series Class-G has often had its linearity called into question because of difficulties with supply-rail commutation. Diodes D3, D4 must be power devices capable of handling a dozen amps or more, and conventional silicon rectifier diodes that can handle such currents take a long time to turn off due to their stored charge carriers. This has the following unhappy effect: when the voltage on the cathode of D3 rises above V1, the diode tries to turn off abruptly, but its charge carriers sustain a brief but large reverse current as they are swept from its junction. This current is supplied by TR6, attempting as an emitter-follower to keep its emitter up to the right voltage. So far all is well.

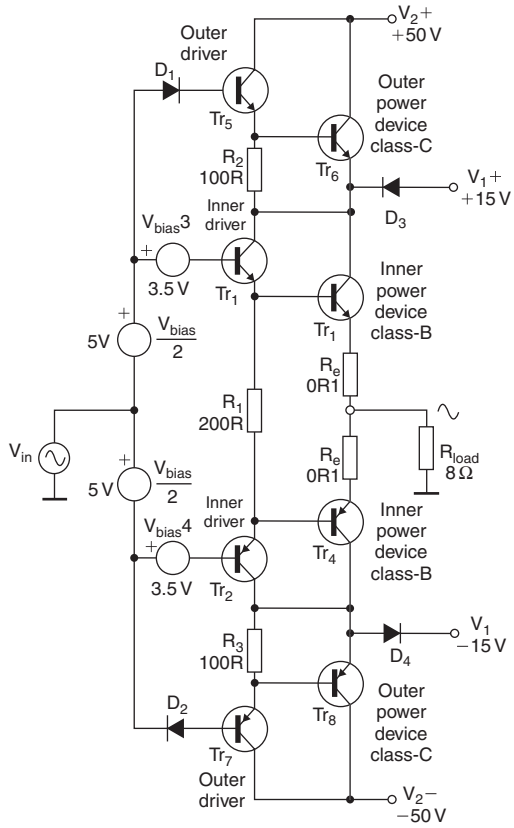


Figure 12.6: The original Hitachi Class-G biasing system, with inner device bias derived by subtracting Vbias3, Vbias4 from the main bias generator

However, when the diode current ceases, TR6 is still conducting heavily, due to its own charge-carrier storage. The extra current it turned on to feed D3 in reverse now goes through the TR3 collector, which accepts it because of TR3's low V_{ce}, and passes it onto the load via TR3 emitter and Re.

This process is readily demonstrated by a SPICE commutation transient simulation (see Figures 12.7 and 12.8). Note there are only two of these events per cycle – not four, as they only occur when the diodes turn off. In the original Hitachi design this problem was reportedly tackled by using fast transistors and relatively fast gold-doped diodes, but according to Sampei et al.[2] this was only partially successful.

It is now simple to eradicate this problem. Schottky power diodes are readily available, as they were not in 1976, and are much faster due to their lack of minority carriers and charge storage. They have the added advantage of a low forward voltage drop at large currents of 10A or more. The main snag is a relatively low reverse withstand voltage, but fortunately in Class-G usage the commutating diodes are only exposed at worst to the difference between V2 and V1, and this only when the amplifier is in its low-power domain of operation. Another good point about Schottky power diodes is that they do appear to be robust; I have subjected 50A Motorola devices to 60 A-plus repeatedly without a single failure. This is a good sign. The spikes disappear completely

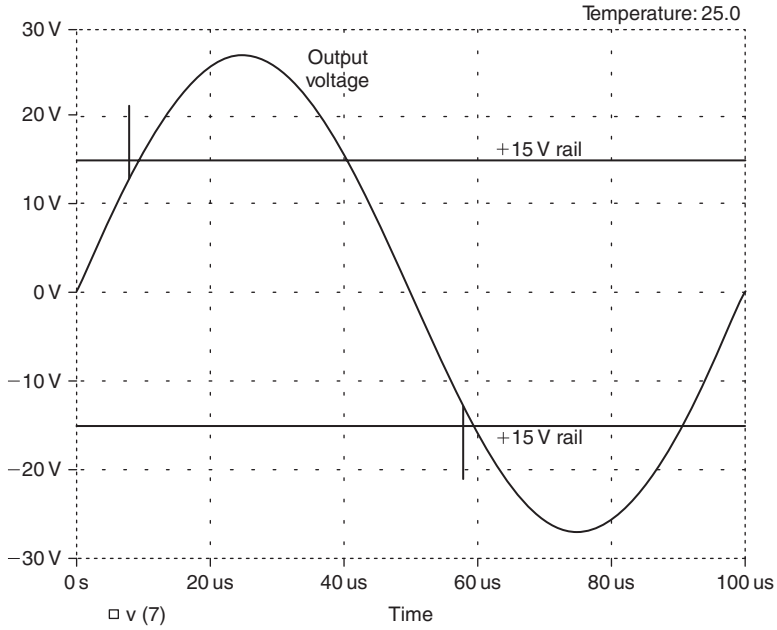


Figure 12.7: Spikes due to charge storage of conventional diodes, simulated at 10 kHz. They only occur when the diodes turn off, so there are only two per cycle. These spikes disappear completely when Schottky diodes are used in the SPICE model

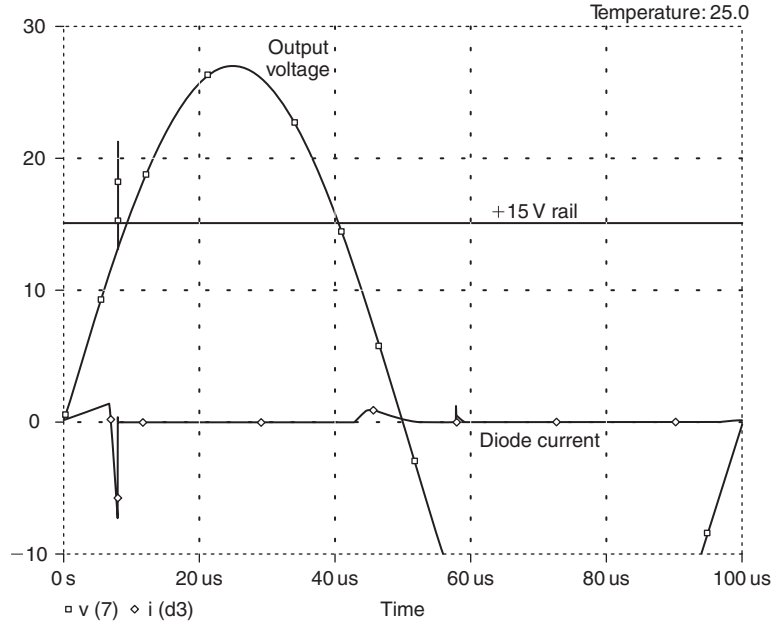


Figure 12.8: A close-up of the diode transient. Diode current rises as output moves away from zero, then reverses abruptly as charge carriers are swept out by reverse-biasing. The spike on the output voltage is aligned with the sudden stop of the diode reverse current

from the SPICE plot if the commutating diodes are Schottky rectifiers. Motorola MBR5025L diodes capable of 50 A and 25 PIV were used in simulation.

The Static Linearity

SPICE simulation shows in Figure 12.9 that the static linearity (i.e. that ignoring dynamic effects like diode charge storage) is distinctly poorer than for Class-B. There is the usual Class-B gain wobble around the crossover region, exactly the same size and shape as for conventional Class-B, but also there are now gain-steps at ±16 V. The result with the inner devices biased into push–pull Class-A is also shown, and proves that the gain-steps are not in any way connected with crossover distortion. Since this is a DC analysis the gain-steps cannot be due to diode-switching speed or other dynamic phenomena, and the Early effect was immediately suspected (the Early effect is the increase in collector current when the collector voltage increases, even though V_{be} remains constant). When unexpected distortion appears in a SPICE simulation of this kind, and effects due to finite transistor beta and associated base currents seem unlikely, a most useful diagnostic technique is to switch off the simulation of the Early effect for each transistor in turn. In SPICE transistor models the Early effect can be totally disabled by setting the parameter VAF to a much higher value than the default of 100, such as 50,000. This experiment demonstrated in short order that the gain-steps were caused wholly by the Early effect acting on both inner drivers and inner output devices. The gain-steps are completely abolished with Early effect disabled. When TR6 begins to act, TR3 V_{ce} is no longer decreasing as the output moves positive, but substantially

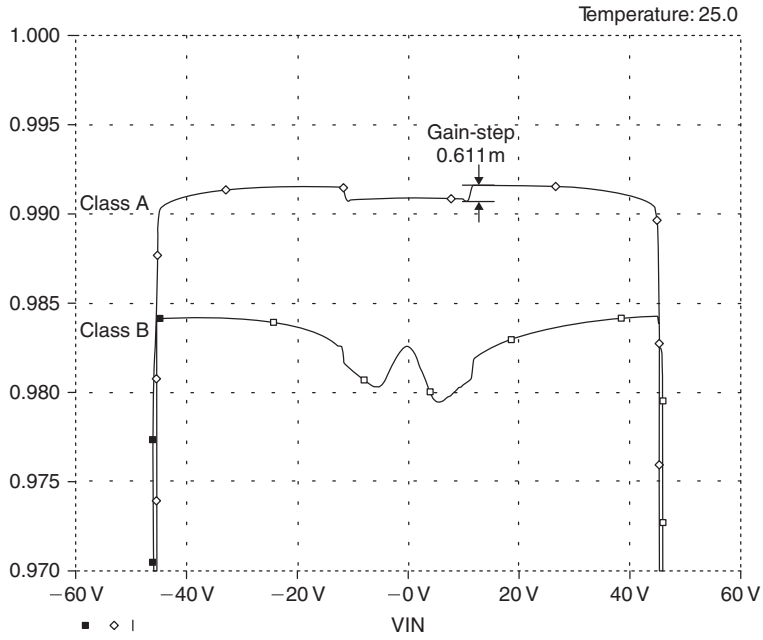


Figure 12.9: SPICE simulation shows variations in the incremental gain of an EF-type Class-G series output stage. The gain-steps at transition (at ±16 V) are due to Early effect in the transistors. The Class-A trace is the top one, with Class-B optima below. Here the inner driver collectors are connected to the switched inner rails, i.e. the inner power device collectors, as in Figure 12.1

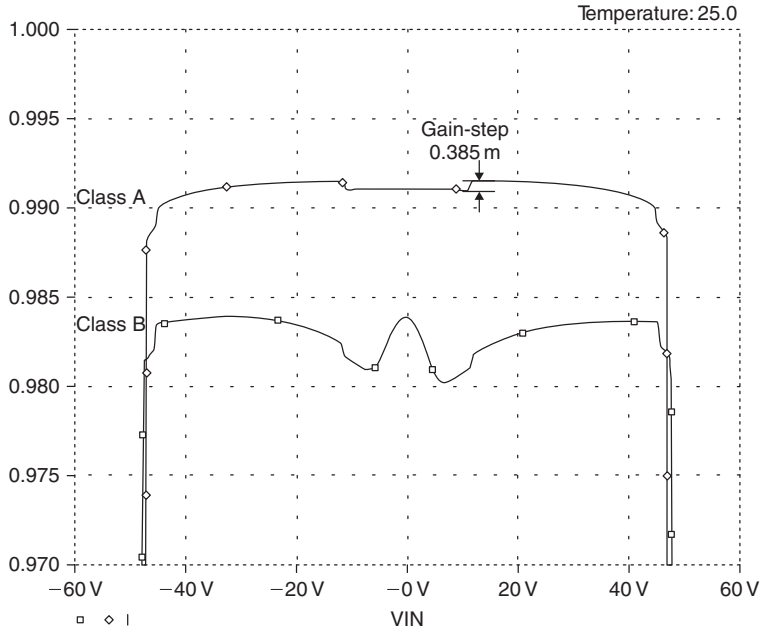


Figure 12.10: Connecting the inner driver collectors to the outer V2 rails reduces Early effect nonlinearities in them, and halves the transition gain-steps

constant as the emitter of Q6 moves upwards at the same rate as the emitter of Q3. This has the effect of a sudden change in gain, which naturally degrades the linearity.

This effect appears to occur in drivers and output devices to the same extent. It can be easily eliminated in the drivers by powering them from the outer rather than the inner supply rails. This prevents the sudden changes in the rate in which driver V_{ce} varies. The improvement in linearity is seen in Figure 12.10, where the gain-steps have been halved in size. The resulting circuit is shown in Figure 12.11. Driver power dissipation is naturally increased by the increased driver V_{ce}, but this is such a small fraction of the power consumed that the overall efficiency is not significantly reduced. It is obviously not practical to apply the same method to the output devices, because then the low-voltage rail would never be used and the amplifier is no longer working in Class-G. The small-signal stages naturally have to work from the outer rails to be able to generate the full voltage swing to drive the output stage.

We have now eliminated the commutating diode glitches and halved the size of the unwanted gain-steps in the output stage. With these improvements made it is practical to proceed with the design of a Class-G amplifier with midband THD below 0.002%.

Practical Class-G Design

The Class-G amplifier design expounded here uses very similar small-signal circuitry to the Blameless Class-B power amplifier, as it is known to generate very little distortion of its own. If the specified supply voltages of ±50 and ±15 V are used, the maximum power output is about 120 W into 8 Ω, and the rail-switching transition occurs at 28 W.

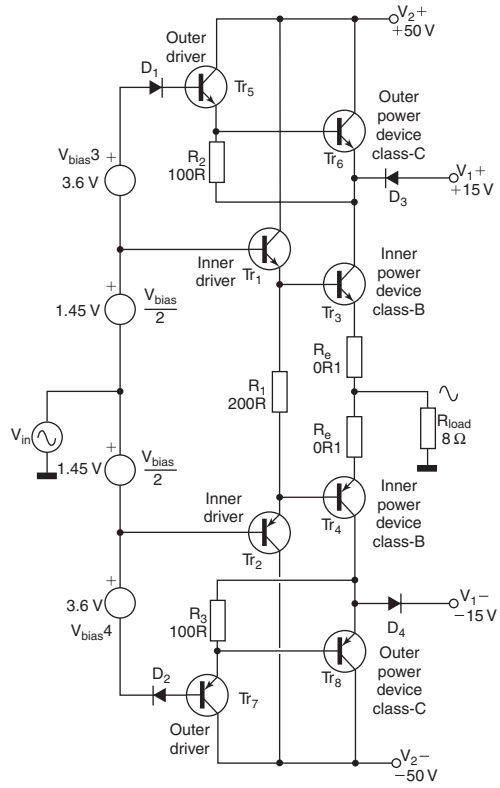


Figure 12.11: A Class-G output stage with the drivers powered from the outer supply rails

This design incorporates various techniques described in this book, and closely follows the Blameless Class-B amp described in Chapter 6, though some features derive from the Trimodal (Chapter 10) and Load-Invariant (Chapter 6) amplifiers. A notable example is the low-noise feedback network, complete with its option of input bootstrapping to give a high impedance when required. Single-slope VI limiting is incorporated for overload protection; this is implemented by Q12, Q13. Figure 12.12 shows the circuit.

As usual in my amplifiers the global NFB factor is a modest 30 dB at 20 kHz.

Controlling Small-Signal Distortion

The distortion from the small-signal stages is kept low by the same methods as for the other amplifier designs in this book, and so this is only dealt with briefly here. The input stage differential pair Q1, Q2 is given local feedback by R5 and R7 to delay the onset of third-harmonic Distortion 1. Internal r_e variations in these devices are minimized by using an unusually high tail current of 6 mA Q3, Q4 are a degenerated current-mirror that enforces accurate balance of the Q1, Q2 collector currents, preventing the production of second-harmonic distortion. The input resistance (R3 + R4) and feedback resistance R16 are made equal and made unusually low, so that base-current mismatches stemming from input device beta variations give a minimal DC offset. V_{be} mismatches in Q1 and Q2 remain, but these are much smaller than the effects of I_b. Even if Q1

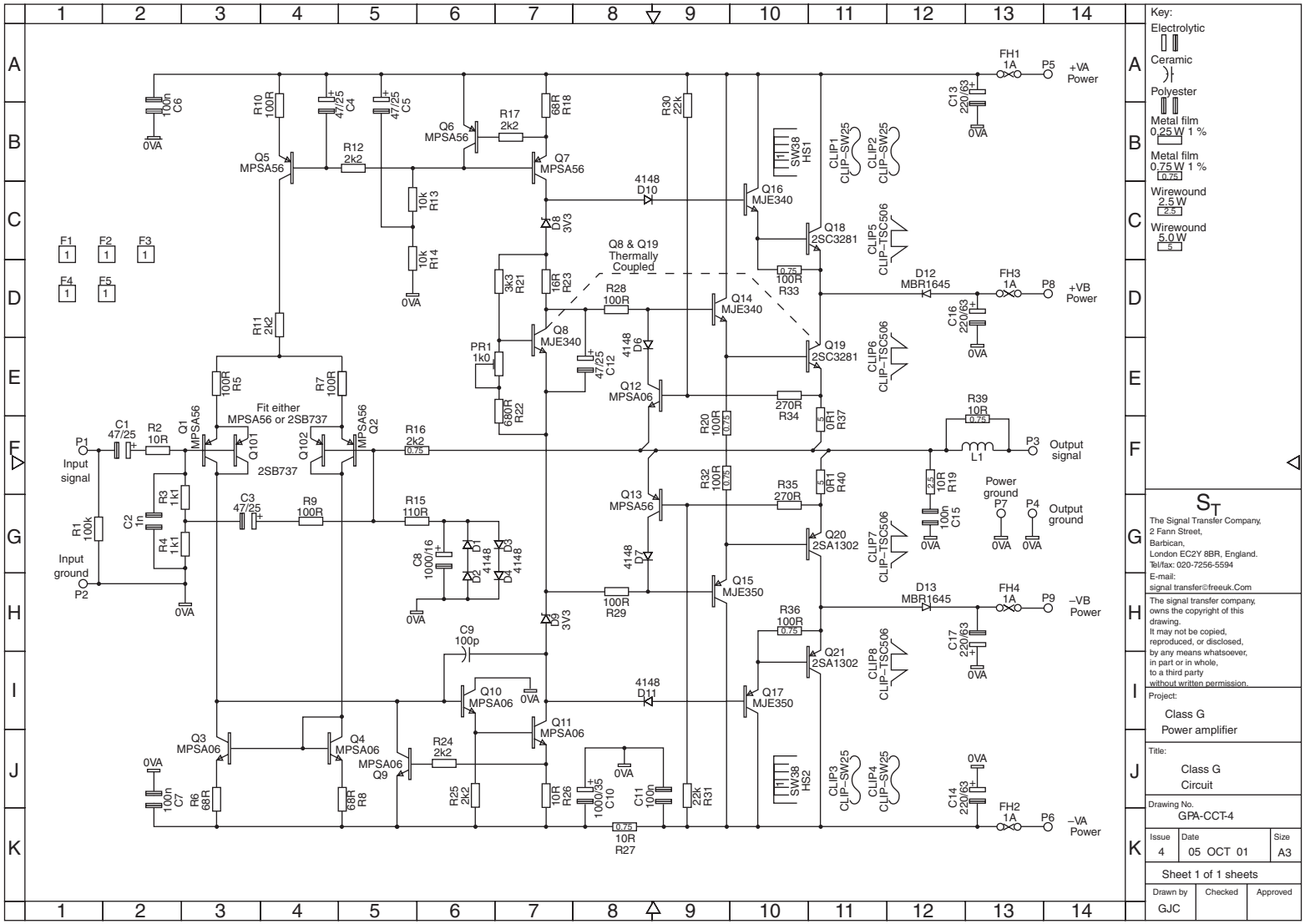


Figure 12.12: The circuit diagram of the Class-G amplifier

and Q2 are high-voltage types with relatively low beta, the DC offset voltage at the output should be kept to less than ±50 mV. This is adequate for all but the most demanding applications. This low-impedance technique eliminates the need for balance presets or DC servo systems, which is most convenient.

A lower value for R16 implies a proportionally lower value for R15 to keep the gain the same, and this reduction in the total impedance seen by Q2 improves noise performance markedly. However, the low value of R3 plus R4 at 2k2 gives an input impedance that is not high enough for many applications.

There is no problem if the amplifier is to have an additional input stage, such as a balanced line receiver. Proper choice of op-amp will allow the stage to drive a 2k2 load impedance without generating additional distortion. Be aware that adding such a stage – even if it is properly designed and the best available op-amps are used – will degrade the signal-to-noise ratio significantly. This is because the noise generated by the power amplifier itself is so very low – equivalent to the Johnson noise of a resistor of a few hundred ohms – that almost anything you do upstream will degrade it seriously.

If there is no separate input stage then other steps must be taken. What we need at the input of the power amplifier is a low DC resistance, but a high AC resistance; in other words we need either a 50 henry choke or recourse to some form of bootstrapping. There is to my mind no doubt about the way to go here, so bootstrapping it is. The signal at Q2 base is almost exactly the same as the input, so if the mid-point of R3 and R4 is driven by C3, so far as input signals are concerned R3 has a high AC impedance. When I first used this arrangement I had doubts about its high-frequency stability, and so added resistor R9 to give some isolation between the bases of Q1 and Q2. In the event I have had no trouble with instability, and no reports of any from the many constructors of the Trimodal and Load-Invariant designs, which incorporate this option.

The presence of R9 limits the bootstrapping factor, as the signal at the R3–R4 junction is thereby a little smaller than at Q2 base, but it is adequate. With R9 set to 100R, the AC input impedance is raised to 13 k, which should be high enough for almost all purposes. Higher values than this mean that an input buffer stage is required.

The value of C8 shown (1000 μF) gives an LF roll-off in conjunction with R15 that is −3 dB at 1.4 Hz. The purpose is not impossibly extended sub-bass, but the avoidance of a low-frequency rise in distortion due to nonlinearity effects in C8. If a 100 μF capacitor is used here the THD at 10 Hz worsens from <0.0006% to 0.0011%, and I regard this as unacceptable esthetically – if not perhaps audibly. This is not the place to define the low-frequency bandwidth of the system – this must be done earlier in the signal chain, where it can be properly implemented with more accurate non-electrolytic capacitors. The protection diodes D1–D4 prevent damage to C2 if the amplifier suffers a fault that makes it saturate in either direction; it looks like an extremely dubious place to put diodes but since they normally have no AC or DC voltage across them no measurable or detectable distortion is generated.

The voltage-amplifier stage (VAS) Q11 is enhanced by emitter-follower Q10 inside the Miller compensation loop, so that the local negative feedback that linearizes the VAS is increased. This

effectively eliminates VAS nonlinearity. Thus increasing the local feedback also reduces the VAS collector impedance, so a VAS buffer to prevent Distortion 4 (loading of VAS collector by the nonlinear input impedance of the output stage) is not required. Miller capacitor C_{dom} is relatively big at 100 pF, to swamp transistor internal capacitances and circuit strays, and make the design predictable. The slew rate calculates as 40 V/μs use in each direction. VAS collector load Q7 is a standard current source.

Almost all the THD from a Blameless amplifier derives from crossover distortion, so keeping the quiescent conditions optimal to minimize this is essential. The bias generator for an EF output stage, whether in Class-B or Class-G, is required to cancel out the V_{be} variations of four junctions in series; those of the two drivers and the two output devices. This sounds difficult, because the dissipation in the two types of devices is quite different, but the problem is easier than it looks. In the EF type of output stage the driver dissipation is almost constant as power output varies, and so the problem is reduced to tracking the two output device junctions. The bias generator Q8 is a standard V_{be}-multiplier, with R23 chosen to minimize variations in the quiescent conditions when the supply rails change. The bias generator should be in contact with the top of one of the inner output devices, and not the heat-sink itself. This position gives much faster and less attenuated thermal feedback to Q8. The VAS collector circuit incorporates not only bias generator Q8 but also the two Zeners D8, D9, which determine how early rail-switching occurs as the inner device emitters approach the inner (lower) voltage rails.

The output stage was selected as an emitter-follower (EF) type as this is known to be less prone to parasitic or local oscillations than the CFP configuration, and since this design was to some extent heading into the unknown it seemed wise to be cautious where possible. R32 is the usual shared emitter resistor for the inner drivers. The outer drivers Q16 and Q17 have their own emitter resistors R33 and R36, which have their usual role of establishing a reasonable current in the drivers as they turn on, to increase driver transconductance, and also in speeding up turn-off of the outer output devices by providing a route for charge carriers to leave the output device bases.

As explained above, the inner driver collectors are connected to the outer rails to minimize the gain-steps caused by the abrupt change in collector voltage when rail transition occurs.

Deciding the size of heat-sink required for this amplifier is not easy, mainly because the heat dissipated by a Class-G amplifier depends very much on the rail voltages chosen and the signal statistics. A Class-B design giving 120 W into 8 Ω would need a heat-sink with thermal resistance of the order of 1°C/W (per channel); a good starting point for a Class-G version giving the same power would be about half the size, i.e. 2°C/W. The Schottky commutating diodes do not require much heat-sinking, as they conduct only intermittently and have a low forward voltage drop. It is usually convenient to mount them on the main heat-sink, even if this does mean that most of the time they are being heated rather than cooled.

C15 and R38 make up the usual Zobel network. The coil L1, damped by R39, isolates the amplifier from load capacitance. A component with 15–20 turns at 1 inch diameter should work well; the value of inductance for stability is not all that critical.

The Performance

Figure 12.13 shows the THD at 20 W and 50 W (into 8 Ω) and I think this demonstrates at once that the design is a practical competitor for Class-B amplifiers. Compare these results with the upper trace of Figure 12.14, taken from a Blameless Class-B amplifier at 50 W, 8 Ω. Note the lower trace of Figure 12.14 is for 30 kHz bandwidth, used to demonstrate the lack of distortion below 1 kHz; the THD data above 30 kHz is in this case meaningless as all the harmonics are filtered out. All the Class-G plots here are taken at 80 kHz to make sure any high-order glitching is properly measured.

Figure 12.15 shows the actual THD residual at 50 W output power. The glitches from the gain-steps are more jagged than the crossover disturbances, as would be expected from the output stage gain

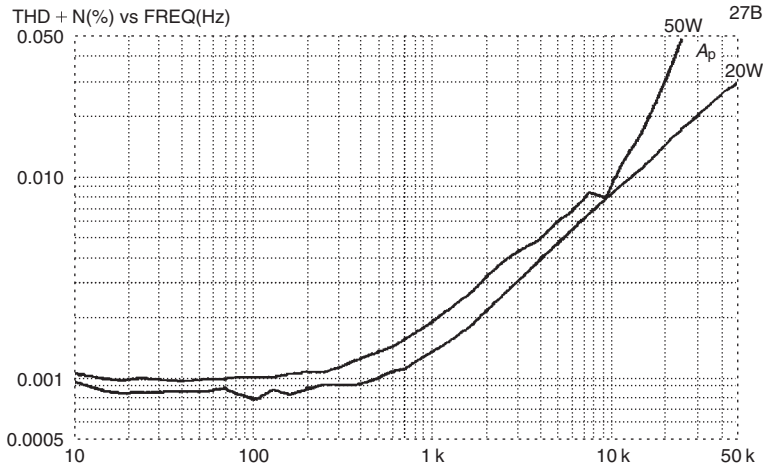


Figure 12.13: THD versus frequency, at 20 W (below transition) and 50 W into an 8 Ω load. The joggle around 8 kHz is due to a cancelation of harmonics from crossover and transition. Bandwidth 80 kHz

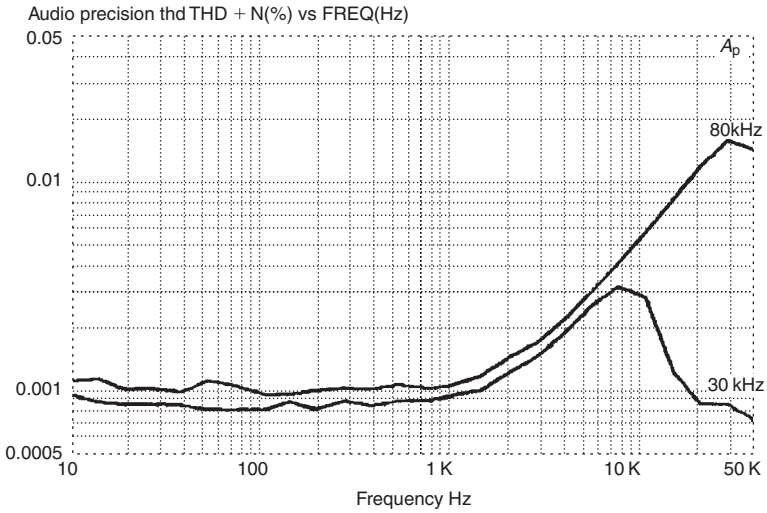


Figure 12.14: THD versus frequency for a Blameless Class-B amplifier at 50 W, 8 Ω

plot in Figures 12.9 and 12.11. Figure 12.16 confirms that at 20 W, below transition, the residual is indistinguishable from that of a Blameless Class-B amplifier; in this region, where the amplifier is likely to spend most of its time, there are no compromises on quality.

Figure 12.17 shows THD versus level, demonstrating how THD increases around 28 W as transition begins. The steps at about 10 W are nothing to do with the amplifier – they are artefacts due to internal range-switching in the measuring system.

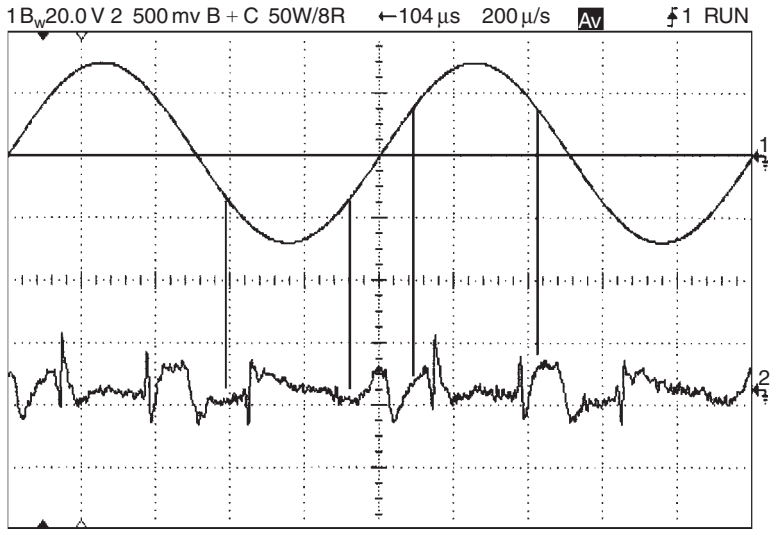


Figure 12.15: The THD residual waveform at 50 W into 8 Ω. This residual may look rough, but in fact it had to be averaged eight times to dig the glitches and crossover out of the noise; THD is only 0.0012%. The vertical lines show where transition occurs

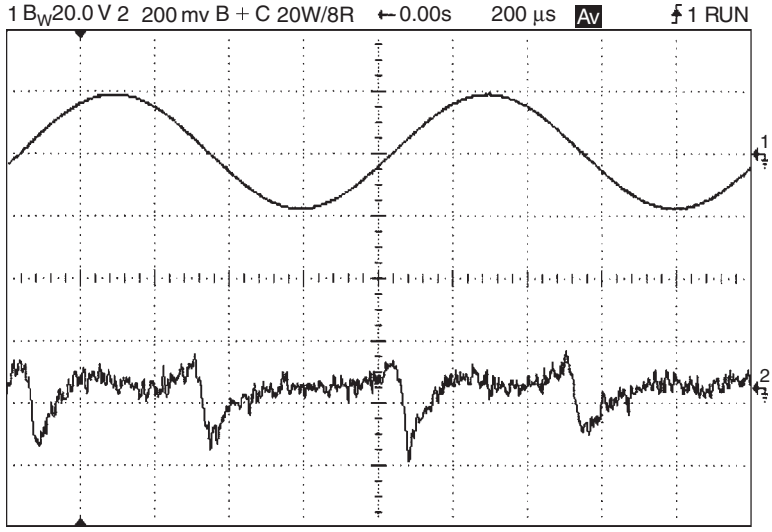


Figure 12.16: The THD residual waveform at 20 W into 8 Ω, below transition. Only crossover artefacts are visible as there is no rail-switching

Figure 12.18 shows for real the benefits of powering the inner drivers from the outer supply rails. In SPICE simulation (see above) the gain-steps were roughly halved in size by this modification, and Figure 12.18 duly confirms that the THD is halved in the HF region, the only area where it is sufficiently clear of the noise floor to be measured with any confidence.

Deriving a New Kind of Amplifier: Class-A + C

A conventional Class-B power amplifier can be almost instantly converted to push–pull Class-A simply by increasing the bias voltage to make the required quiescent current flow. This is the only real circuit change, though naturally major increases in heat-sinking and power-supply capability are required for practical use. Exactly the same principle applies to the Class-G amplifier. In the book *Self On Audio*[6] I suggested a new and much more flexible system for classifying amplifier types and here it comes in very handy. Describing Class-G operation as Class-B + C immediately

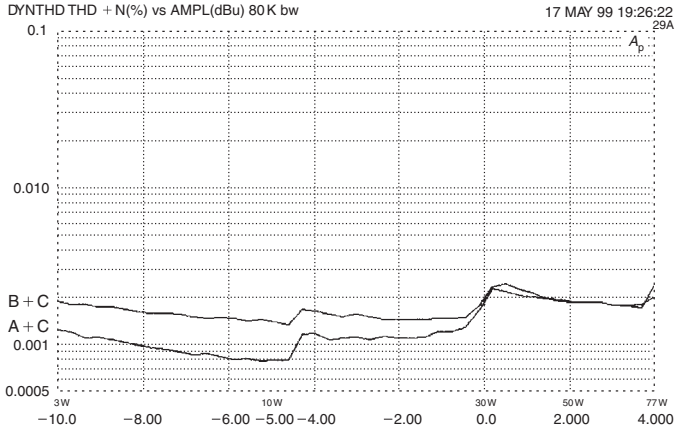


Figure 12.17: THD versus level, showing how THD increases around 28 W as transition begins. Class-A + C is the lower and Class-B + C the upper trace

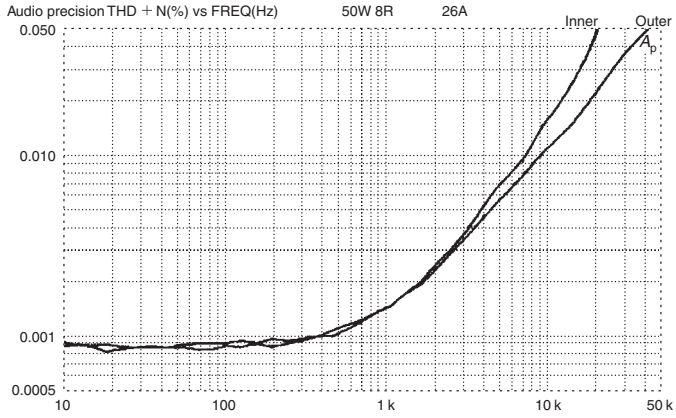


Figure 12.18: THD plot of real amplifier driving 50 W into 8 Ω. Rails were ±40 and ±25 V. Distortion at HF is halved by connecting the inner drivers to the outer supply rails rather than the inner rails

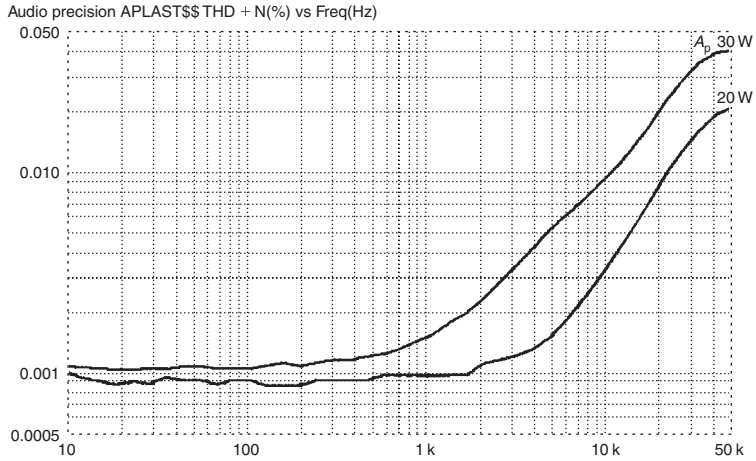


Figure 12.19: The THD plot of the Class-A + C amplifier (30 and 20 W into 8 Ω). Inner drivers powered from outer rails

indicates that only a bias increase is required to transform it into Class-A + C, and a new type of amplifier is born. This amplifier configuration combines the superb linearity of classic Class-A up to the transition level, with only minor distortion artefacts occurring at higher levels, as demonstrated for Class-B + C above. Using Class-A means that the simple V_{be}-multiplier bias generator can be replaced with precise negative-feedback control of the quiescent current, as implemented in the Trimodal amplifier in Chapter 10. There is no reason why an amplifier could not be configured as a Class-G Trimodal, i.e. manually switchable between Classes A and B. That would indeed be an interesting machine.

In Figure 12.19 is shown the THD plot for such an A + C amplifier working at 20 and 30 W into 8 Ω. At 20 W the distortion is very low indeed, no higher than a pure Class-A amplifier. At 30 W the transition gain-steps appear, but the THD remains very well controlled, and no higher than a Blameless Class-B design. Note that as in Class-B, when the THD does start to rise it only does so at 6 dB/octave. The quiescent current was set to 1.5 A.

Figure 12.20 reveals the THD residual during A + C operation. There are absolutely no crossover artefacts, and the small disturbances that do occur happen at such a high signal level that I really do think it is safe to assume they could never be audible. Figure 12.21 shows the complete absence of artefacts on the residual when this new type of amplifier is working below transition; it gives pure Class-A linearity. Finally, Figure 12.22 gives the THD when the amplifier is driving the full 50 W into 8 Ω; as before the A + C THD plot is hard to distinguish from Class-B, but there is the immense advantage that there is no crossover distortion at low levels, and no critical bias settings.

Adding Two-Pole Compensation

I have previously shown elsewhere in this book that amplifier distortion can be very simply reduced by changes to the compensation, which means a scheme more sophisticated than the near-universal dominant-pole method. It must be borne in mind that any departure from the conventional 6 dB/

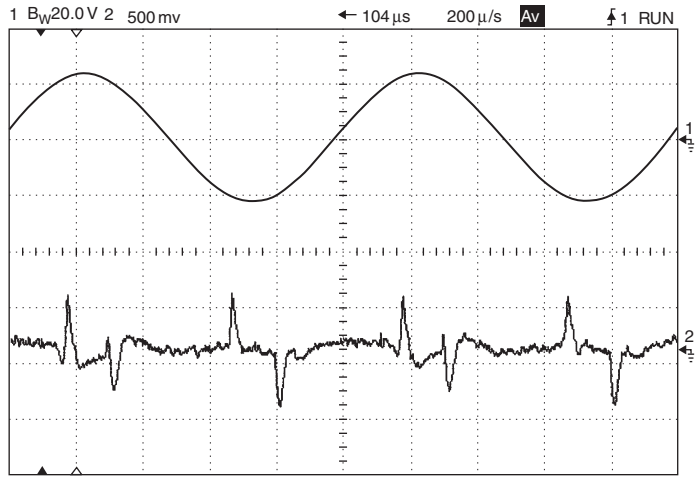


Figure 12.20: The THD residual waveform of the Class-A + C amplifier above transition, at 30 W into 8 Ω. Switching artefacts are visible but not crossover distortion

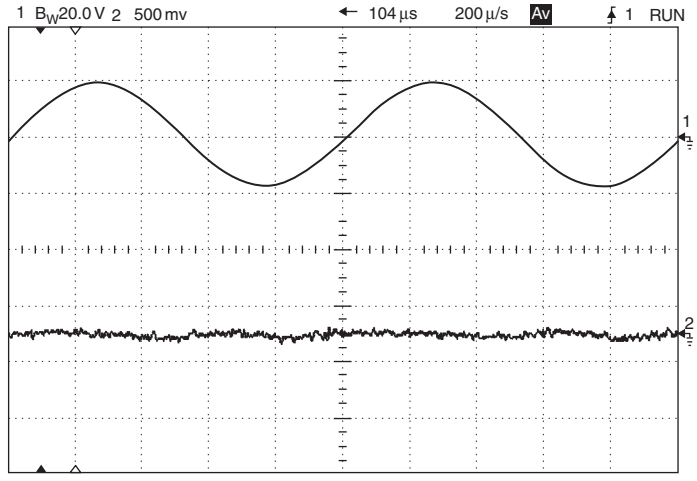


Figure 12.21: The THD residual waveform plot of the Class-A + C amplifier (20 W into 8 Ω)

octave all-the-way compensation scheme is likely to be a move away from unconditional stability. (I am using this phrase in its proper meaning; in Control Theory unconditional stability means that increasing open-loop gain above a threshold causes instability, but the system is stable for all lower values. Conditional stability means that lower open-loop gains can also be unstable.)

A conditionally stable amplifier may well be docile and stable into any conceivable reactive load when in normal operation, but shows the cloven hoof of oscillation at power-up and power-down, or when clipping. This is because under these conditions the effective open-loop gain is reduced.

Class-G distortion artefacts are reduced by normal dominant-pole feedback in much the same way as crossover nonlinearities, i.e. not all that effectively, because the artefacts take up a very small part of the cycle and are therefore composed of high-order harmonics. Therefore a compensation

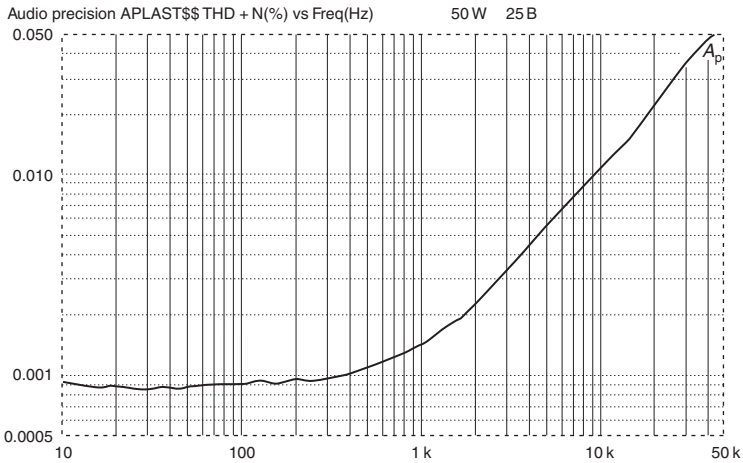


Figure 12.22: The THD plot of the Class-A + C amplifier (50 W into 8 Ω). Inner drivers powered from outer rails

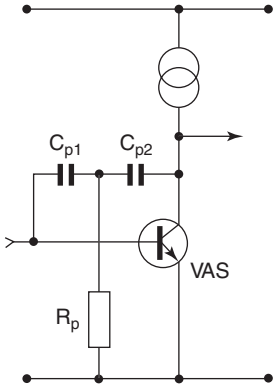


Figure 12.23: The circuit modification for two-pole compensation

system that increases the feedback factor at high audio frequencies will be effective on switching artefacts, in the same way that it is for crossover distortion. The simplest way to implement two-pole circuit compensation is shown in Figure 12.23. Further details are given in Chapter 8.

The results of two-pole compensation for B + C are shown in Figure 12.24; comparing it with Figure 12.13 (the normally compensated B + C amplifier) the above-transition (30 W) THD at 10 kHz has dropped from 0.008% to 0.005%; the sub-transition (20 W) THD at 10 kHz has fallen from 0.007% to 0.003%. Comparisons have to be done at 10 kHz or thereabouts to ensure there is enough to measure.

Now comparing the two-pole B + C amplifier with Figure 12.19 (the A + C amplifier) the above-transition (30 W) THD at 10 kHz of the former is lower at 0.005% compared with 0.008%. As I have demonstrated before, proper use of two-pole compensation can give you a Class-B amplifier that is hard to distinguish from Class-A – at least until you put your hand on the heat-sink.

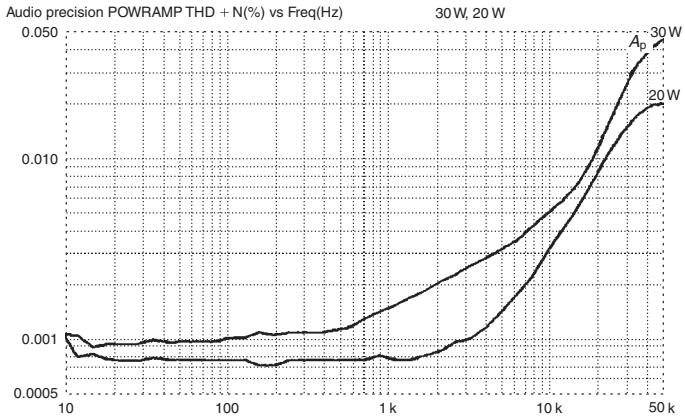


Figure 12.24: The THD plot for B + C operation with two-pole compensation (20 and 30 W into 8 Ω). Compare with Figures 12.13 (B + C) and 12.19 (A + C)

Further Variations on Class-G

This by no means exhausts the possible variations that can be played on Class-G. For example, it is not necessary for the outer devices to operate synchronously with the inner devices. So long as they turn on in time, they can turn off much later without penalty except in terms of increased dissipation. In so-called syllabic Class-G, the outer devices turn on quickly but then typically remain on for 100 ms or so to prevent glitching (see Funada and Akiya[7] for one version). Given the good results obtained with straight Class-G, this no longer seems a promising route to explore.

With the unstoppable advance of multichannel amplifier and powered subwoofers, Class-G is at last coming into its own. It has recently even appeared in a Texas ADSL driver IC. I hope I have shown how to make it work, and then how to make it work better. I modestly suggest that this might be the lowest distortion Class-G amplifier so far.

References

[1] D. Self, Self On Audio, Newnes, 2006, p. 347.

[2] T. Sampei et al., Highest efficiency and super quality audio amplifier using MOS power FETs in Class-G operation, IEEE Trans. Consumer Electronics CE-24 (3) (August 1978) p. 300.

[3] L. Feldman, Class-G high efficiency hi-fi amplifier, Radio Electronics (August 1976) p. 47.

[4] D. Self, Self On Audio, Newnes, 2006, p. 369.

[5] D. Self, Self On Audio, Newnes, 2006, p. 386.

[6] D. Self, Self On Audio, Newnes, 2006, p. 293.

[7] S. Funada, H. Akiya, A study of high-efficiency audio power amplifiers using a voltage switching method, JAES 32 (10) (October 1984) p. 755.

CHAPTER 13

Class-D Amplifiers

Since the first edition of this book, Class-D amplifiers have increased enormously in popularity. This is because Class-D gives the highest efficiency of any of the amplifier classes, although the performance, particularly in terms of linearity, is not so good. The rapid rate of innovation means that this section of the book is much more of a snapshot of a fast-moving scene than the rest of the material. I do not want to keep repeating 'at the time of writing' as each example is introduced, so I hope you will take that as read.

The fields of application for Class-D amplifiers can be broadly divided into two areas: low- and high-power outputs. The low-power field reaches from a few milliwatts (for digital hearing aids) to around 5 W, while the high-power applications go from 80 to 1400 W. At present there seems to be something of a gap in the middle, for reasons that will emerge.

The low-power area includes applications such as mobile phones, personal stereos, and laptop computer audio. These products are portable, and battery driven, so power economy is very important. A major application of Class-D is the production of useful amounts of audio power from a single low-voltage supply rail. A good example is the National Semiconductor LM4671, a single-channel amplifier IC that gives 2.1 W into a 4 Ω speaker from a 5 V supply rail, using a 300 kHz switching frequency. This is a very low voltage by conventional power amplifier standards, and requires an H-bridge output structure, of which more later.

The high-power applications include PA amplifiers, home theater systems, and big subwoofers. These are all energized from the mains supply, so power economy is not such a high priority. Here Class-D is used because it keeps dissipation and therefore power supply and heat-sink size to a minimum, leading to a smaller and neater product. High-power Class-D amplifiers are also used in car audio systems, with power capabilities of 1000 W or more into 2 Ω; here minimizing the power drain is of rather greater importance, as the capabilities of the engine-driven alternator that provides the 12 V supply are finite.

There is a middle ground between these two areas, where an amplifier is powered from the mains but of no great output power – say a stereo unit with an output of 30 W into 8 Ω per channel. The heat-sinks will be small, and eliminating them altogether will not be a great cost saving. The power supply will almost certainly be a conventional toroid-and-bridge-rectifier arrangement, and the cost savings on reducing the size of this component by using Class-D will not be large. In this area the advantage gained by accepting the limitations of Class-D are not at present enough to justify it.

Class-D amplifiers normally come as single ICs or as chip sets with separate output stages. Since the circuitry inside these ICs is complex, and not disclosed in detail, they are not very instructive to those planning to design their own discrete Class-D amplifier.

History

The history of the Class-D amplifier goes back, as is so often the case with technology, further than you might think. The principle is generally regarded as having surfaced in the 1950s, but the combination of high switching frequencies and valve output transformers probably did not appear enticing. The first public appearance of Class-D in the UK was the Sinclair X10, which claimed an output of 10 W. This was followed by the X20, alleging a more ambitious 20 W. I resurrected one of the latter in 1976, when my example proved to yield about 3 W into 8 Ω. The THD was about 5% and the rudimentary output filter did very little to keep the low switching frequency out of the load. The biggest problem of the technology at that time was that bipolar transistors of suitable power-handling capacity were too slow for the switching frequencies required; this caused serious losses that undermined the whole point of Class-D, and also produced unappealingly high levels of distortion. It was not until power FETs, with their very fast switching times, appeared that Class-D began to become a really practical proposition.

Basic Principles

Amplifiers working in Class-D differ radically from the more familiar classes of A, B, and G. In Class-D there are no output devices operating in the linear mode. Instead they are switched on and off at an ultrasonic frequency, the output being connected alternately to each supply rail. When the mark/space ratio of the input signal is varied, the average output voltage varies with it, the averaging being done by a low-pass output filter, or by the loudspeaker inductance alone. Note that the output is also directly proportional to the supply voltage; there is no inherent supply rejection at all with this sort of output stage, unlike the Class-B output stage. The use of negative feedback helps with this. The switching frequencies used range from 50 kHz to 1 MHz. A higher frequency makes the output filter simpler and smaller, but tends to increase switching losses and distortion.

The classic method of generating the drive signal is to use a differential comparator. One input is driven by the incoming audio signal and the other by a sawtooth waveform at the required switching frequency. A basic Class-D amplifier is shown in Figure 13.1, and the PWM process is illustrated in Figure 13.2.

Clearly the sawtooth needs to be linear (i.e. with constant slope) to prevent distortion being introduced at this stage. There are other ways to create the required waveform, such as a sigma-delta modulator.

When the aim is to produce as much audio power as possible from a low-voltage supply such as 5 V, the H-bridge configuration is employed, as shown in Figure 13.3. It allows twice the voltage swing across the load, and therefore theoretically four times the output power, and also permits the amplifier to run from one supply rail without the need for bulky output capacitors of doubtful linearity. This method is also called the bridge-tied load, or BTL.

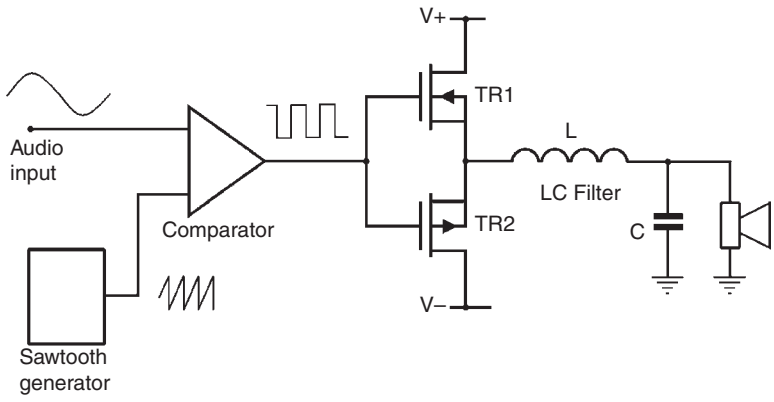


Figure 13.1: A basic Class-D amplifier with PWM comparator, FET output stage, and second-order LC output filter

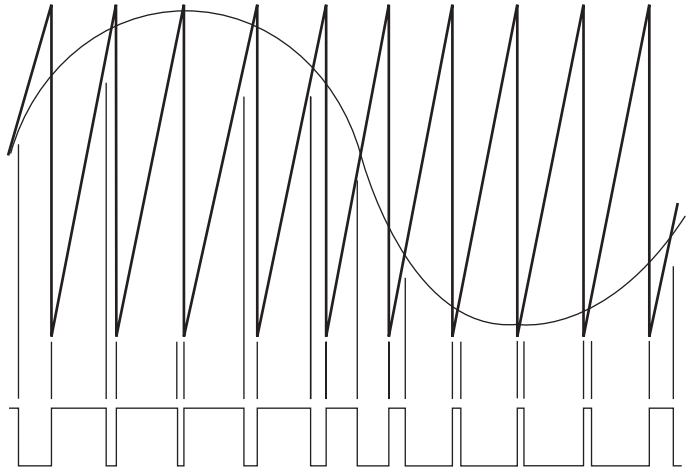


Figure 13.2: The PWM process as performed by a differential comparator

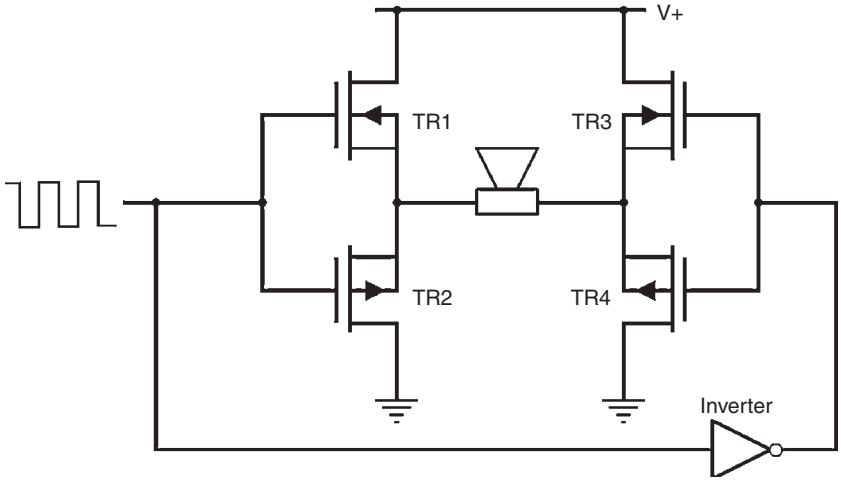


Figure 13.3: The H-bridge output configuration. The output filter is not shown

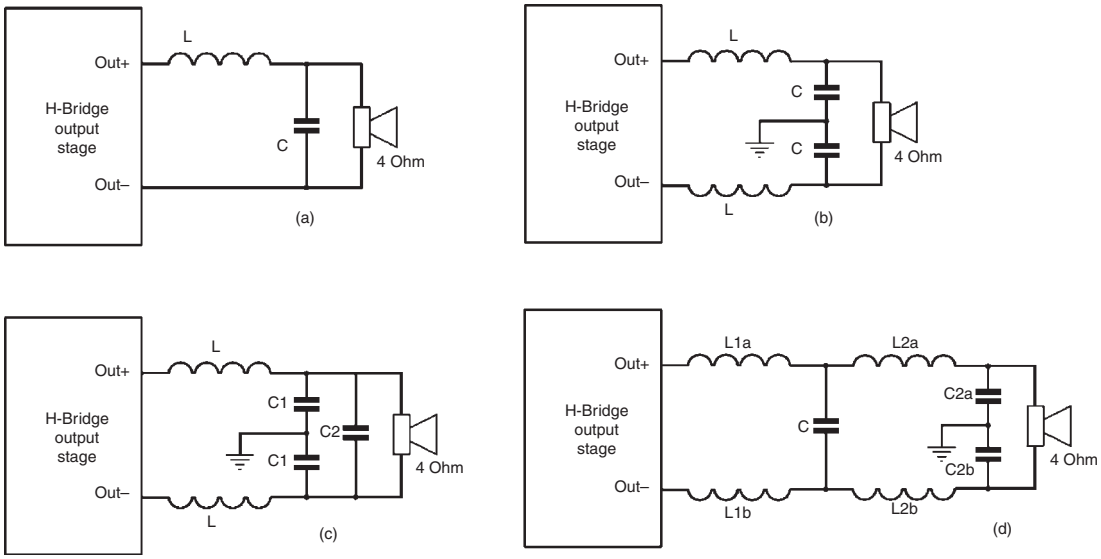


Figure 13.4: Filter arrangements for the H-bridge output. (a) is the simplest but allows a common-mode signal on the speaker cabling. (b) and (c) are the most usual versions. (d) is a four-pole filter

The use of two amplifier outputs requires a somewhat more complex output filter. If the simple two-pole filter of Figure 13.4a is used, the switching frequency is kept out of the loudspeaker, but the wiring to it will carry a large common-mode signal from Out. A balanced filter is therefore commonly used, in either the Figure 13.4b or Figure 13.4c versions. Figure 13.4d illustrates a four-pole output filter – note that you can save a capacitor. This is only used in quality applications because inductors are never cheap.

Technology

The theory of Class-D has an elegant simplicity about it, but in real life complications quickly begin to intrude.

While power FETs have a near-infinite input resistance at the gate, they require substantial current to drive them at high frequencies, because of the large device capacitances, and the gate drive circuitry is a non-trivial part of the amplifier. Power FETs, unlike bipolars, require several volts on the gate to turn them on. This means that the gate-drive voltage needed for the high-side FET TR1 in Figure 13.1 is actually above the positive voltage rail. In many designs a bootstrap supply driven from the output is used to power the gate-drive circuitry. Since this supply will not be available until the high-side FET is working, special arrangements are needed at start-up.

The more powerful amplifiers usually have external Schottky diodes connected from output to the supply rails for clamping flyback pulses generated by the inductive load. These are not merely to protect the output stage from damage, but to improve efficiency, as described in the section below.

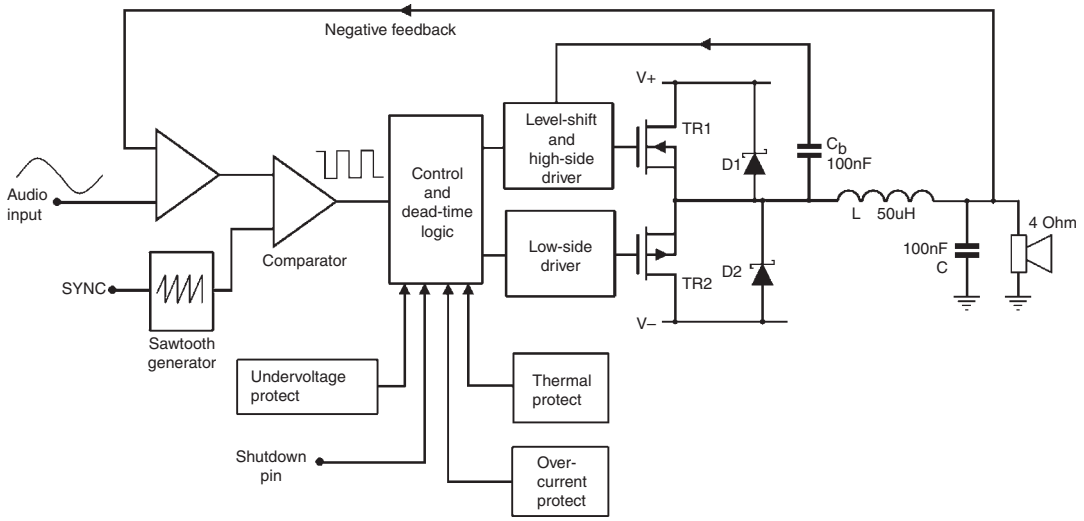


Figure 13.5: The main features of a practical Class-D amplifier, including Schottky clamp diodes, bootstrap supply, and one form of negative feedback

The application of negative feedback to reduce distortion and improve supply-rail rejection is complicated by the switched nature of the output waveform. Feedback can be taken from after the output filter, or alternatively taken from before it and passed through an op-amp active filter to remove the switching frequency. In either case the filtering adds phase shift and limits the amount of negative feedback that can be applied while still retaining Nyquist stability.

Other enhancements that are common are selectable input gain and facilities for synchronizing the switching clocks of multiple amplifiers to avoid audible heterodyne tones. Figure 13.5 shows a Class-D amplifier including these features.

Perhaps the gravest problem with Class-D is that you either have to use proprietary and therefore single-sourced parts, or design and build it yourself from standard components. The latter is a very serious undertaking; do not underestimate the amount of background research, the length of the design investigations, and the protracted periods of optimization that will be required before you have a reliable product with reasonable performance. In most cases the only realistic option is to use proprietary parts, which as always carry with them the risk that the manufacturer will suddenly disappear, leaving you well and truly in the lurch. If you are lucky you may be able to do a last-time buy that will give you enough time to do some high-speed redesign, but you may not be lucky, and this is the sort of thing that sinks companies.

A recent example is Tripath Technologies, who called their approach to Class-D by the name ‘Class-T’, though this was just a trademark rather than an actual class of operation. Financial difficulties caused Tripath to file for Chapter 11 bankruptcy protection in February 2007.

Protection

All the implementations of Class-D on the market have internal protection systems to prevent excessive output currents and device temperatures.

In the published circuitry DC offset protection is conspicuous by its absence. It is understandable that there is little enthusiasm for adding output relays to personal stereos – they might consume more power than the amplifier. However, it is surprising that they also appear to be absent from 500 W designs where relay size and power consumption are minor issues. Are such amplifiers really that reliable?

Most Class-D systems also have undervoltage protection. If the supply voltage falls too low then there may not be enough gate-drive voltage to turn the output FETs fully on, and they will dissipate excessive power. A lockout circuit prevents operation below a certain voltage. A shutdown facility is almost always provided; this inhibits any switching in the output stage and allows power consumption to be very low indeed in the standby mode.

Output Filters

The purpose of the output filter is to prevent radiation of switching frequencies for amplifiers that have external speaker cables, and also to improve efficiency. The inductance of a loudspeaker coil alone will in general be low enough to allow some of the switching-frequency energy to pass through it to ground, causing significant losses. While some low-power integrated applications have no output filter at all, most Class-D amplifiers have a second-order LC filter between the amplifier output and the loudspeaker. In some cases a fourth-order filter is used, as in Figure 13.4d. The Butterworth alignment is usually chosen to give maximal flatness of frequency response.

As described in the chapter on real speaker loads, a loudspeaker, even a single-element one, is a long way from being a resistive load. It is therefore rather surprising that at least one manufacturer provides filter design equations that assume just that. When a Class-D amplifier is to be used with separate loudspeakers of unknown impedance characteristic, the filter design can only proceed on the basis of plausible assumptions, and there are bound to be some variations in frequency response.

The inductor values required are typically in the region 10–50 μH, which is much larger than the 1–2 μH air-cored coils used to ensure stability with capacitive loads in Class-B amplifiers. It is therefore necessary to use ferrite-cored inductors, and care must be taken that they do not saturate at maximum output.

Efficiency

At the most elementary level of theory, the efficiency of a Class-D amplifier is always 100%, at all output levels. In practice, of course, the mathematical idealizations do not hold, and the real-life efficiency of most implementations is between 80% and 90% over most of the power output range. At very low powers the efficiency falls off steeply, as there are fixed losses that continue to dissipate power in the amplifier when there is no audio output at all (see Figure 13.6).

The losses in the output stage are due to several mechanisms. The most important are outlined below.

Firstly, the output FETs have a non-zero resistance even when they are turned hard on. This is typically in the range 100–200 mΩ, and can double as the device temperature increases from 0 to 150°C, the latter being the usual maximum operating temperature. This resistance causes I^2R losses.

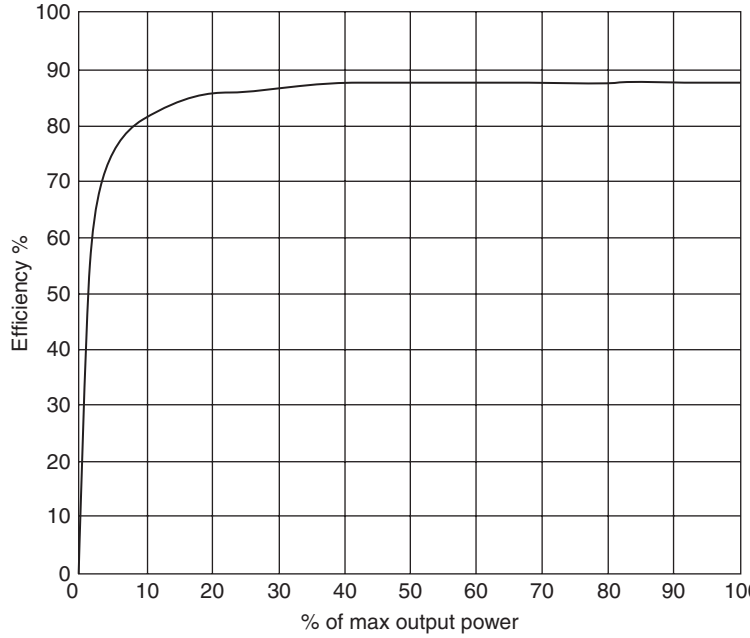


Figure 13.6: A typical efficiency curve for a Class-D amplifier driving a 4 Ω load at 1 kHz

Secondly, the output devices have non-zero times for switching on and off. In the period when the FET is turning on or turning off, it has an intermediate value of resistance that again causes I^2R losses. It is essential to minimize the stray inductance in the drain and source circuits as this not only extends the switching times but also causes voltage transients at turn-off that can overstress the FETs.

Thirdly, flyback pulses generated by an inductive load can cause conduction of the parasitic diodes that are part of the FET construction. These diodes have relatively long reverse recovery times and more current will flow than is necessary. To prevent this many Class-D designs have Schottky clamp diodes connected between the output line and supply rails as in Figure 13.5. These turn on at a lower voltage than the parasitic FET diodes and deal with the flyback pulses. They also have much faster reverse recovery times.

Last, and perhaps most dangerous, is the phenomenon known as 'shoot-through'. This somewhat opaque term refers to the situation when one FET has not stopped conducting before the other starts. This gives rise to an almost direct short between the supply rails, although very briefly, and large amounts of unwanted heating can occur. To prevent this the gate drive to the FET that is about to be turned on is slightly delayed, by a 'dead-time' circuit. The introduction of dead time increases distortion, so only the minimum is applied; a 40 ns delay is sufficient to create more than 2% THD in a 1 kHz sine wave.

CHAPTER 14

FET Output Stages

The Characteristics of Power FETs

A field-effect transistor (FET) is essentially a voltage-controlled device. So are bipolar junction transistors (BJTs), despite the conventional wisdom that persists in regarding them as current controlled. They are not, even if BJT base currents are non-negligible.

The power FETs normally used are enhancement devices – in other words, with no voltage between gate and source they remain off. In contrast, the junction FETs found in small-signal circuitry are depletion devices, requiring the gate to be taken negative of the source (for the most common N-channel devices) to reduce the drain current to usable proportions. (Please note that the standard information on FET operation is in many textbooks and will not be repeated here.)

Power FETs have large internal capacitances, both from gate to drain and from gate to source. The gate-source capacitance is effectively bootstrapped by the source-follower configuration, but the gate-drain capacitance, which can easily total 2000 pF, remains to be driven by the previous stage. There is an obvious danger that this will compromise the amplifier slew rate if the VAS is not designed to cope.

FETs tend to have much larger bandwidths than BJT output devices. My own experience is that this tends to manifest itself as a greater propensity for parasitic oscillation rather than anything useful, but the tempting prospect of higher global NFB factors due to a higher output stage pole remains. The current state of knowledge does not yet permit a definitive judgment on this.

A great deal has been said on the thermal coefficients of the V_{bias} voltage. It is certainly true that the temperature coefficient at high drain currents is negative – in other words drain current falls with increasing temperature – but on the other hand the coefficient reverses sign at low drain currents, and this implies that precise quiescent-current setting will be very difficult. A negative-temperature coefficient provides good protection against thermal runaway, but this should never be a problem anyway.

FET versus BJT Output Stages

On beginning any power amplifier design, one of the first decisions that must be made is whether to use BJTs or FETs in the output stage. This decision may of course already have been taken for you by the marketing department, as the general mood of the marketplace is that if FETs are more

expensive, they must be better. If, however, you are lucky enough to have this crucial decision left to you, then FETs normally disqualify themselves on the same grounds of price. If the extra cost is not translated into either better performance and/or a higher sustainable price for the product, then it appears to be foolish to choose anything other than BJTs.

Power MOSFETS are often hailed as the solution to all amplifier problems, but they have their own drawbacks, not the least being low transconductance, poor linearity, and a high on-resistance that makes output efficiency mediocre. The high-frequency response may be better, implying that the second pole $P2$ of the amplifier response will be higher, allowing the dominant pole $P1$ to be raised with the same stability margin, and so in turn giving more NFB to reduce distortion. However, we would need this extra feedback (if it proves available in practice) to correct the worse open-loop distortion, and even then the overall linearity would almost certainly be worse. To complicate matters, the compensation cannot necessarily be lighter because the higher output resistance makes more likely the lowering of the output pole by capacitive loading.

The extended FET frequency response is, like so many electronic swords, two-edged if not worse, and the HF capabilities mean that rigorous care must be taken to prevent parasitic oscillation, as this is often promptly followed by an explosion of disconcerting violence. FETs should at least give freedom from switch-off troubles (Distortion 3c) as they do not suffer from BJT charge-storage effects.

Advantages of FETs

1. For a simple complementary FET output stage, drivers are not required. This is somewhat negated by the need for gate-protection Zener diodes.
2. There is no second-breakdown failure mechanism. This may simplify the design of overload protection systems, especially when arranging for them to cope with highly reactive loads.
3. There are no charge-storage effects to cause switch-off distortion.

Disadvantages of FETs

1. Linearity is very poor by comparison with a BJT degenerated to give the same transconductance. The Class-B conduction characteristics do not cross over smoothly, and there is no equivalent to the optimal Class-B bias condition that is very obvious with a BJT output stage.
2. The V_{gs} required for conduction is usually of the order of 4–6 V, which is much greater than the 0.6–0.8 V required by a BJT for base drive. This greatly reduces the voltage efficiency of the output stage unless the preceding small-signal stages are run from separate and higher-voltage supply rails.
3. Power FETs draw negligible DC current through their gate connections, but they have high internal capacitances, which must be charged and discharged rapidly for high-frequency operation. This often requires extra complications in the driver circuitry to provide these large currents at low distortion.

4. The minimum channel resistance of the FET, known as $R_{ds(on)}$, is high and gives a further reduction in efficiency compared with BJT outputs.
5. Power FETs are liable to parasitic oscillation. In severe cases a plastic-package device will literally explode. This is normally controllable in the simple complementary FET output stage by adding gate-stopper resistors, but is a serious disincentive to trying radical experiments in output stage circuit design.
6. Some commentators claim that FET parameters are predictable; I find this hard to understand as they are notorious for being anything but. From one manufacturer's data (Harris), the V_{gs} for the IRF240 FET varies between 2.0 and 4.0 V for an I_d of 250 μA; this is a range of two to one. In contrast the V_{be}/I_c relation in bipolars is fixed by a mathematical equation for a given transistor type, and is much more reliable. Nobody uses FETs in log converters.
7. Since the V_{gs} spreads are high, this will complicate placing devices in parallel for greater power capability. Paralleled BJT stages rarely require current-sharing resistors of greater than 0.1 Ω, but for the FET case they may need to be a good deal larger, reducing efficiency further.
8. At the time of writing, there is a significant economic penalty in using FETs. Taking an amplifier of given power output, the cost of the output semiconductors is increased by between 1.5 and 2 times with FETs.

IGBTs

Insulated-gate bipolar transistors (IGBTs) represent a relatively new option for the amplifier designer. They have been held up as combining the best features of FETs and BJTs. In my view, this is a dubious proposition as I find the advantages of FETs for audio to be heavily outweighed by the drawbacks, and if IGBTs have any special advantages they have not so far emerged. According to the Toshiba application notes[1], IGBTs consist of an FET controlling a bipolar power transistor; I have no information on the linearity of these devices, but the combination does not sound promising.

The most discouraging aspect of IGBTs is the presence of a parasitic BJT that turns the device hard on above a critical current threshold. This inbuilt self-destruct mechanism will at the very least make overload protection an extremely critical matter; it seems unlikely that IGBTs will prove popular for audio amplification.

Not withstanding this, at least one IGBT design has been put forward for amateur construction[2]. The output stage of this 90 W/8 Ω design is a hybrid (described below) with BJT drivers and IGBT output devices, arranged to give gain in the output stage, which is unusual. Interestingly, there are very few circuit changes from a 60 W/8 Ω version using HEXFETs for output devices, and also having gain in the output stage.

Power FET Output Stages

Three types of FET output stage are shown in Figure 14.1, and Figures 14.2–14.5 show SPICE gain plots, using 2SK135/2SJ50 devices. Most FET amplifiers use the simple source-follower

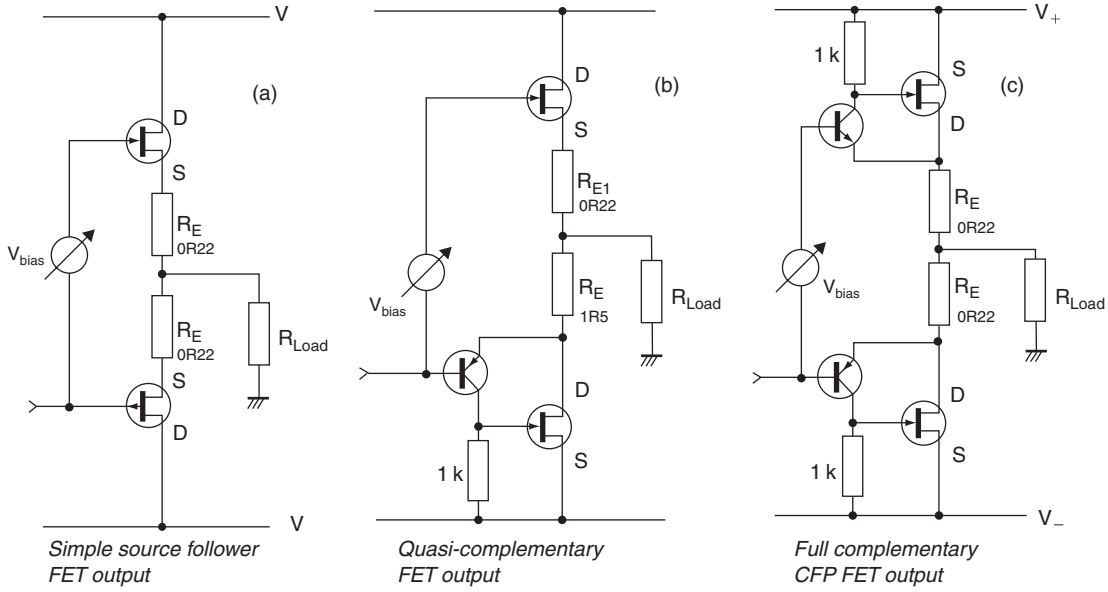


Figure 14.1: Three MOSFET output architectures

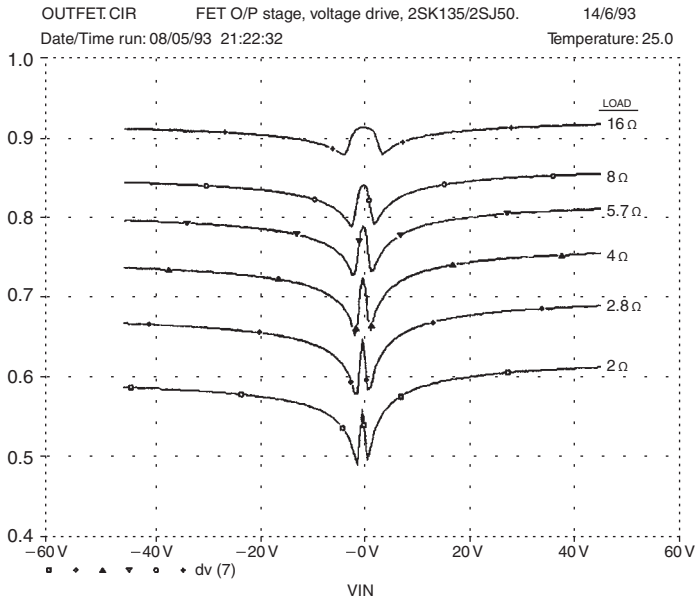


Figure 14.2: Source-follower FET large-signal gain versus output

configuration in Figure 14.1a; the large-signal gain plot in Figure 14.2 shows that the gain for a given load is lower (0.83 rather than 0.97 for bipolar, at 8Ω) because of low g_m, and this, with the high on-resistance, reduces output efficiency seriously. Open-loop distortion is markedly higher; however, LSN does not increase with heavier loading, there being no equivalent of 'bipolar gain-droop'.

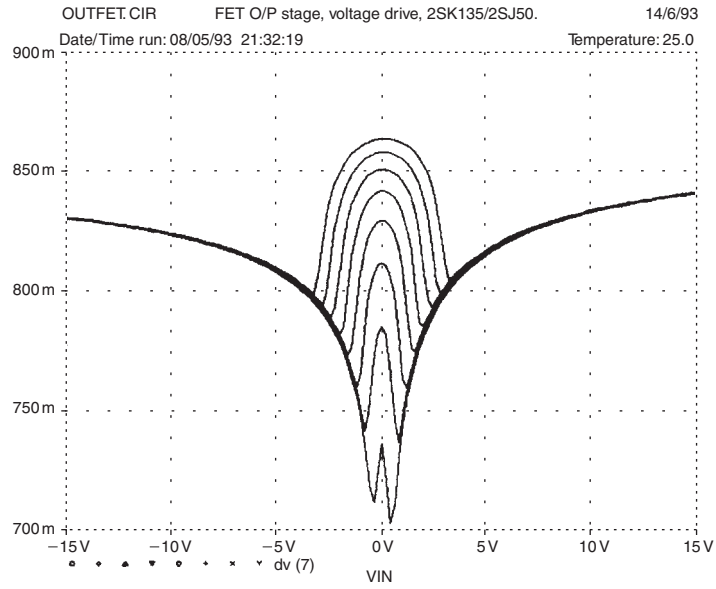


Figure 14.3: Source-follower FET crossover region, ±15 V range

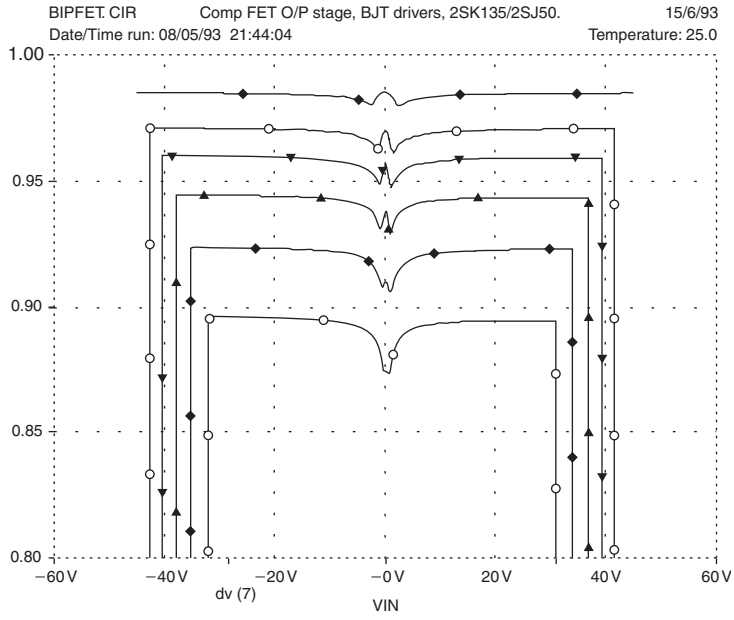


Figure 14.4: Complementary bipolar-FET gain versus output

The crossover region has sharper and larger gain deviations than a bipolar stage, and generally looks pretty nasty; Figure 14.3 shows the impossibility of finding a correct V_{bias} setting.

Figure 14.1b shows a hybrid (i.e. bipolar/FET) quasi-complementary output stage, first described by me[3]. This topology is intended to maximize economy rather than performance, once the decision

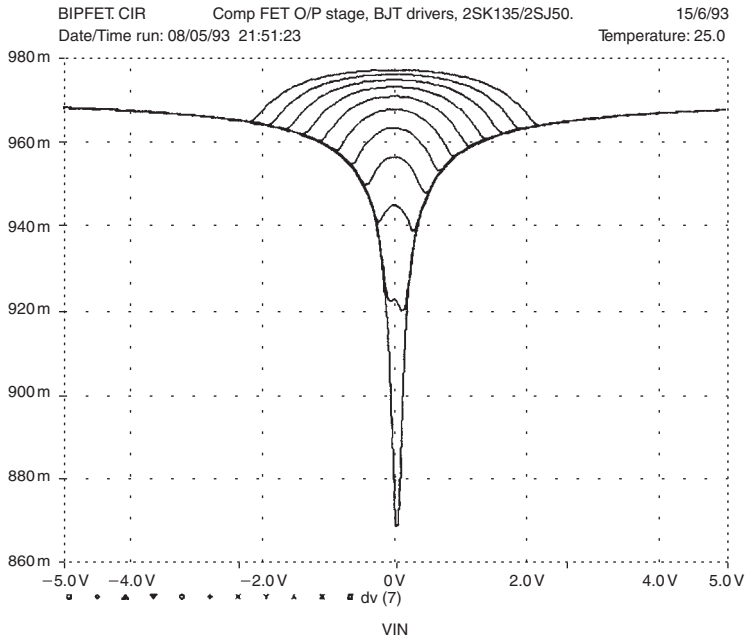


Figure 14.5: Complementary BJT–FET crossover region, ±15 V range

has been made (presumably for marketing reasons) to use FETs, by making both output devices cheap N-channel devices; complementary MOSFET pairs remain relatively rare and expensive. The basic configuration is badly asymmetrical, the hybrid lower half having a higher and more constant gain than the source-follower upper half. Increasing the value of Re2 gives a reasonable match between the gains of the two halves, but leaves a daunting crossover discontinuity. To the best of my knowledge this idea has not caught on, which is probably a good thing. No gain plot is given.

The hybrid full-complementary stage in Figure 14.1c was conceived[4] to maximize FET performance by linearizing the output devices with local feedback and reducing I_q variations due to the low power dissipation of the bipolar drivers. It is very linear, with no gain-droop at heavier loadings (Figure 14.4), and promises freedom from switch-off distortions; however, as shown, it is rather inefficient in voltage swing. The crossover region in Figure 14.5 still has some unpleasant sharp corners, but the total crossover gain deviation (0.96–0.97 at 8 Ω) is much smaller than for the quasi-hybrid (0.78–0.90) and so less high-order harmonic energy is generated.

Table 14.1 summarizes the SPICE curves for 4 and 8 Ω loadings. Each was subjected to Fourier analysis to calculate THD percentage results for a ±40 V input. The BJT results from Chapter 6 are included for comparison.

Power FETs and Bipolars: The Linearity Competition

There has been much debate as to whether power FETs or BJTs are superior in power amplifier output stages, e.g. Hawtin[5]. As the debate rages, or at any rate flickers, it has often been flatly

Table 14.1: THD percentages and average gains for various types of output stage, for 8 and 4 Ω loading

	Emitter-follower	CFP	Quasi simple	Quasi Bax	Triple Type 1	Simple MOSFET	Quasi MOSFET	Hybrid MOSFET
8 Ω THD (%)	0.031	0.014	0.069	0.050	0.13	0.47	0.44	0.052
Gain	0.97	0.97	0.97	0.96	0.97	0.83	0.84	0.97
4 Ω THD (%)	0.042	0.030	0.079	0.083	0.60	0.84	0.072	0.072
Gain	0.94	0.94	0.94	0.94	0.92	0.72	0.73	0.94

stated that power FETs are more linear than BJTs, usually in tones that suggest that only the truly benighted are unaware of this.

In audio electronics it is a good rule of thumb that if an apparent fact is repeated times without number, but also without any supporting data, it needs to be looked at very carefully indeed. I therefore present my own view of the situation here.

I suggest that it is now well established that power FETs when used in conventional Class-B output stages are a good deal less linear than BJTs. The gain deviations around the crossover region are far more severe for FETs than the relatively modest wobbles of correctly biased BJTs, and the shape of the FET gain plot is inherently jagged, due to the way in which two square-law devices overlap. The incremental gain range of a simple FET output stage is 0.84–0.79 (range 0.05) and this is actually much greater than for the bipolar stages examined in Chapter 6; the EF stage gives 0.965–0.972 into 8 Ω (range 0.007) and the CFP gives 0.967–0.970 (range 0.003). The smaller ranges of gain variation are reflected in the much lower THD figures when PSPICE data is subjected to Fourier analysis.

However, the most important difference may be that the bipolar gain variations are gentle wobbles, while all FET plots seem to have abrupt changes that are much harder to linearize with NFB that must decline with rising frequency. The basically exponential I_c/V_{be} characteristics of two BJTs approach much more closely the ideal of conjugate (i.e. always adding up to 1) mathematical functions, and this is the root cause of the much lower crossover distortion.

A close-up examination of the way in which the two types of device begin conducting as their input voltages increase shows that FETs move abruptly into the square-law part of their characteristic, while the exponential behavior of bipolars actually gives a much slower and smoother start to conduction.

Similarly, recent work[6] shows that less conventional approaches, such as the CC-CE configuration of Mr Bengt Olsson[7], also suffer from the non-conjugate nature of FETs, and show sharp changes in gain. Gevel[8] shows that this holds for both versions of the stage proposed by Olsson, using both N- and P-channel drivers. There are always sharp gain changes.

FETs in Class-A Stages

It occurred to me that the idea that FETs are more linear was based not on Class-B power-amplifier applications, but on the behavior of a single device in Class-A. It might be argued that the roughly

square-law nature of an FET's I_d/V_{gs} law is intuitively more 'linear' than the exponential I_c/V_{be} law of a BJT, but it is a bit difficult to know quite how to define 'linear' in this context. Certainly a square-law device will generate predominantly low-order harmonics, but this says nothing about the relative amounts produced.

In truth the BJT/FET contest is a comparison between apples and aardvarks, the main problem being that the raw transconductance (g_m) of a BJT is far higher than for any power FET. Figure 14.6 illustrates the conceptual test circuit; both a TO-3 BJT (MJ802) and a power-FET (IRF240) have an increasing DC voltage V_{in} applied to their base/gate, and the resulting collector and drain currents from PSPICE simulation are plotted in Figure 14.7. V_{offset} is used to increase the voltage

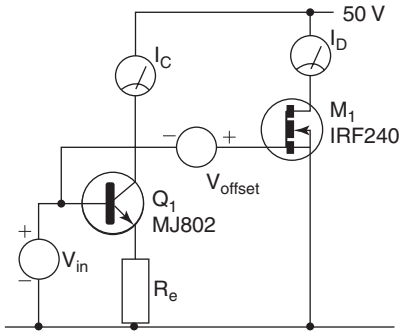


Figure 14.6: The linearity test circuit. V_{offset} adds 3 V to the DC level applied to the FET gate, purely to keep the current curves helpfully adjacent on a graph

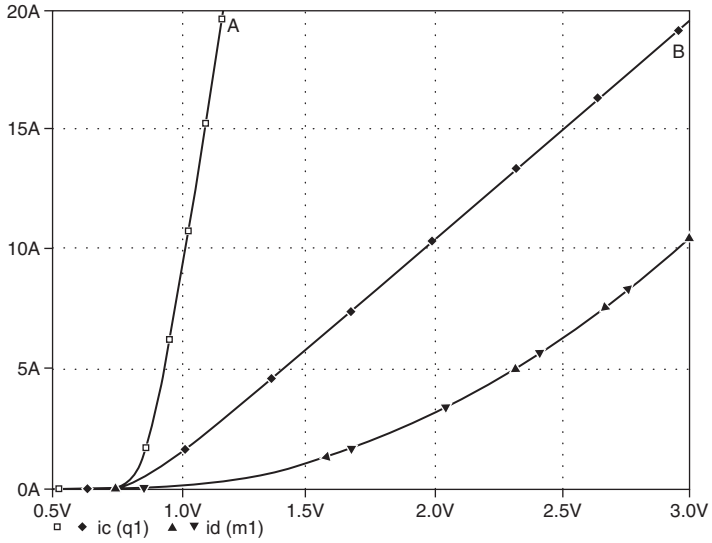


Figure 14.7: Graph of I_c and I_d for the BJT and the FET. Curve A shows I_c for the BJT alone, while curve B shows the result for Re = 0.1 Ω. The curved line is the I_d result for a power FET without any degeneration

applied to FET M1 by 3.0V because nothing much happens below $V_{gs} = 4$V, and it is helpful to have the curves on roughly the same axis. Curve A, for the BJT, goes almost vertically skywards, as a result of its far higher g_m. To make the comparison meaningful, a small amount of local negative feedback is added to Q1 by Re, and as this emitter degeneration is increased from 0.01 to 0.1 Ω, the I_c curves become closer in slope to the I_d curve.

Because of the curved nature of the FET I_d plot, it is not possible to pick an Re value that allows very close g_m equivalence; Re = 0.1 Ω was chosen as a reasonable approximation (see curve B). However, the important point is that I think no one could argue that the FET I_d characteristic is more linear than curve B.

This is made clearer by Figure 14.8, which directly plots transconductance against input voltage. There is no question that FET transconductance increases in a beautifully linear manner – but this 'linearity' is what results in a square-law I_d increase. The near-constant g_m lines for the BJT are a much more promising basis for the design of a linear amplifier.

To forestall any objections that this comparison is all nonsense because a BJT is a current-operated device, I add here a reminder that this is completely untrue. The BJT is a voltage-operated device, and the base current that flows is merely an inconvenient side-effect of the collector current induced by the said base voltage. This is why beta varies more than most BJT parameters; the base current is an unavoidable error rather than the basis of transistor operation.

The PSPICE simulation shown here was checked against manufacturers' curves for the devices, and the agreement was very good – almost unnervingly so. It therefore seems reasonable to rely

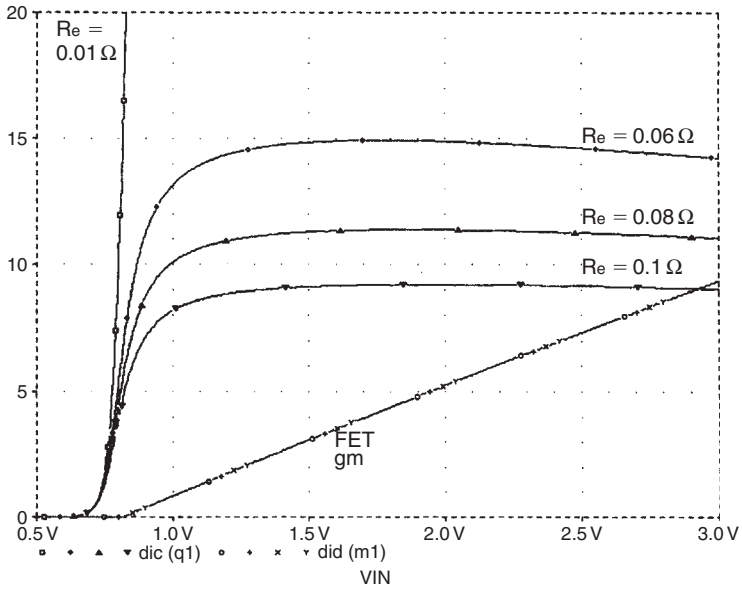


Figure 14.8: Graph of transconductance versus input voltage for BJT and FET. The near-horizontal lines are BJT g_m for various Re values

on simulator output for these kind of studies – it is certainly infinitely quicker than doing the real measurements, and the comprehensive power FET component libraries that are part of PSPICE allow the testing to be generalized over a huge number of component types without actually buying them.

To conclude, I think it is probably irrelevant to simply compare a naked BJT with a naked FET. Perhaps the vital point is that a bipolar device has much more raw transconductance gain to begin with, and this can be handily converted into better linearity by local feedback, i.e. adding a little emitter degeneration. If the transconductance is thus brought down roughly to FET levels, the bipolar has far superior large-signal linearity. I must admit to a sneaking feeling that if practical power BJTs had come along after FETs, they would have been seized upon with glee as a major step forward in power amplification.

References

[1] S. Langdon, Audio amplifier designs using IGBTs, MOSFETs, and BJTs, Toshiba Application Note X3504, vol. 1, March 1991.

[2] Build Your Own High-end Audio Equipment, Elektor Electronics (1995), p. 57.

[3] D. Self, Sound MOSFET design, Electronics & Wireless World (September 1990) p. 760.

[4] D. Self, MOSFET audio output, Letter, Electronics & Wireless World (May 1989) p. 524 (see also Ref. [2]).

[5] V. Hawtin, Letters, Electronics World (December 1994) p. 1037.

[6] D. Self, Two-stage amplifiers and the Olsson output stage, Electronics World (September 1995) p. 762.

[7] B. Olsson, Better audio from non-complements? Electronics World (December 1994) p. 988.

[8] M. Gevel, Private communication, January 1995.

CHAPTER 15

Thermal Compensation and Thermal Dynamics

Why Quiescent Conditions are Critical

In earlier sections of this book we looked closely at the distortion produced by amplifier output stages, and it emerged that a well-designed Class-B amplifier with proper precautions taken against the easily fixed sources of nonlinearity, but using basically conventional circuitry, can produce startlingly low levels of THD. The distortion that actually is generated is mainly due to the difficulty of reducing high-order crossover nonlinearities with a global negative-feedback factor that declines with frequency; for 8 Ω loads this is the major source of distortion, and unfortunately crossover distortion is generally regarded as the most pernicious of nonlinearities. For convenience, I have chosen to call such an amplifier, with its small-signal stages freed from unnecessary distortions, but still producing the crossover distortion inherent in Class-B, a Blameless amplifier (see Chapter 3).

Chapter 6 suggests that the amount of crossover distortion produced by the output stage is largely fixed for a given configuration and devices, so the best we can do is ensure the output stage runs at optimal quiescent conditions to minimize distortion.

Since it is our only option, it is therefore particularly important to minimize the output stage gain irregularities around the crossover point by holding the quiescent conditions at their optimal value. This conclusion is reinforced by the finding that for a Blameless amplifier increasing quiescent current to move into Class-AB makes the distortion worse, not better, as g_m-doubling artefacts are generated. In other words the quiescent setting will only be correct over a relatively narrow band, and THD measurements show that too much quiescent current is as bad (or at any rate very little better) than too little.

The initial quiescent setting is simple, given a THD analyzer to get a good view of the residual distortion; simply increase the bias setting from minimum until the sharp crossover spikes on the residual merge into the noise. Advancing the preset further produces edges on the residual that move apart from the crossover point as bias increases; this is g_m-doubling at work, and is a sign that the bias must be reduced again.

It is easy to attain this optimal setting, but keeping it under varying operating conditions is a much greater problem because quiescent current (I_q) depends on the maintenance of an accurate voltage drop V_q across emitter resistors Re of tiny value, by means of hot transistors with varying V_{be} drops. It's surprising it works as well as it does.

Some kinds of amplifier (e.g. Class-A or current-dumping types) manage to evade the problem altogether, but in general the solution is some form of thermal compensation, the output stage bias voltage being set by a temperature sensor (usually a V_{be}-multiplier transistor) coupled as closely as possible to the power devices.

There are inherent inaccuracies and thermal lags in this sort of arrangement, leading to program dependency of I_q. A sudden period of high-power dissipation will begin with the I_q increasing above the optimum, as the junctions will heat up very quickly. Eventually the thermal mass of the heat-sink will respond, and the bias voltage will be reduced. When the power dissipation falls again, the bias voltage will now be too low to match the cooling junctions and the amplifier will be underbiased, producing crossover spikes that may persist for some minutes. This is very well illustrated in an important paper by Sato et al.[1].

Accuracy Required of Thermal Compensation

Quiescent stability depends on two main factors. The first is the stability of the V_{bias} generator in the face of external perturbations, such as supply voltage variations. The second and more important is the effect of temperature changes in the drivers and output devices, and the accuracy with which V_{bias} can cancel them out.

V_{bias} must cancel out temperature-induced changes in the voltage across the transistor base–emitter junctions, so that V_q remains constant. From the limited viewpoint of thermal compensation (and given a fixed Re) this is very much the same as the traditional criterion that the quiescent current must remain constant, and no relaxation in exactitude of setting is permissible.

I have reached some conclusions on how accurate the V_{bias} setting must be to attain minimal distortion. The two major types of output stage, the emitter-follower (EF) and the complementary feedback pair (CFP), are quite different in their behavior and bias requirements, and this complicates matters considerably. The results are approximate, depending partly on visual assessment of a noisy residual signal, and may change slightly with transistor type, etc. Nonetheless, Table 15.1 gives a much-needed starting point for the study of thermal compensation.

From these results, we can take the permissible error band for the EF stage as about $\pm 100\,\text{mV}$ and for the CFP as about $\pm 10\,\text{mV}$. This goes some way to explaining why the EF stage can give satisfactory quiescent stability despite its dependence on the V_{be} of hot power transistors.

Returning to the PSPICE simulator, and taking Re = 0R1, a quick check on how the various transistor junction temperatures affect V_q yields:

- The EF output stage has a V_q of 42 mV, with a V_q sensitivity of $-2\,\text{mV/°C}$ to driver temperature, and $-2\,\text{mV/°C}$ to output junction temperature. No surprises here.
- The CFP stage has a much smaller V_q (3.1 mV). V_q sensitivity is $-2\,\text{mV/°C}$ to driver temperature, and only $-0.1\,\text{mV/°C}$ to output device temperature. This confirms that local NFB in the stage makes V_q relatively independent of output device temperature, which is just as well, as Table 15.1 shows it needs to be about 10 times more accurate.

Table 15.1: V_{bias} tolerance for 8 Ω

		EF output	CFP output
Crossover spikes obvious	Underbias	2.25 V	1.242 V
Spikes just visible	Underbias	2.29	1.258
Optimal residual	Optimal	2.38	1.283
g_m-doubling just visible	Overbias	2.50	1.291
g_m-doubling obvious	Overbias	2.76	1.330

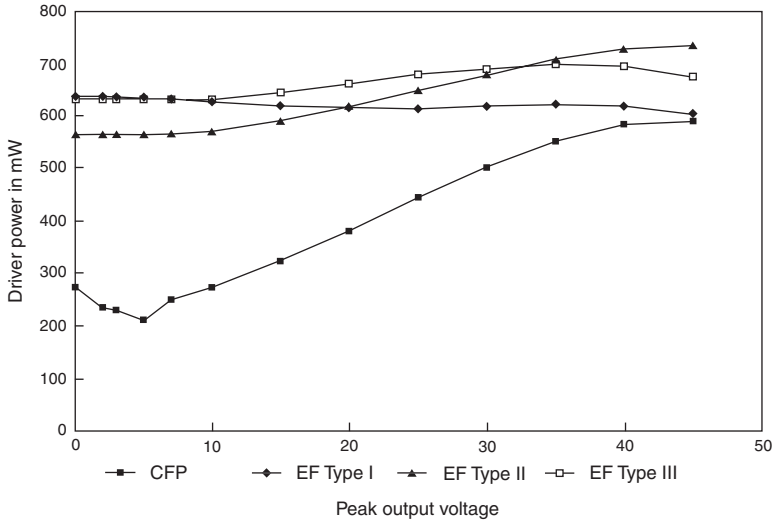


Figure 15.1: Driver dissipation versus output level. In all variations on the EF configuration, power dissipation varies little with output; CFP driver power, however, varies by a factor of 2 or more

The CFP output devices are about 20 times less sensitive to junction temperature, but the V_q across Re is something like 10 times less; hence the actual relationship between output junction temperature and crossover distortion is not so very different for the two configurations, indicating that as regards temperature stability the CFP may only be twice as good as the EF, and not vastly better, which is perhaps the common assumption. In fact, as will be described, the CFP may show poorer thermal performance in practice.

In real life, with a continuously varying power output, the situation is complicated by the different dissipation characteristics of the drivers as output varies (see Figure 15.1, which shows that the CFP driver dissipation is more variable with output, but on average runs cooler). For both configurations driver temperature is equally important, but the EF driver dissipation does not vary much with output power, though the initial drift at switch-on is greater as the standing dissipation is higher. This, combined with the twofold greater sensitivity to output device temperature and the greater self-heating of the EF output devices, may be the real reason why most designers have a general feeling that the EF version has inferior quiescent stability. The truth as to which type of stage is more thermally stable is much more complex, and depends on several design choices and assumptions.

Having assimilated this, we can speculate on the ideal thermal compensation system for the two output configurations. The EF stage has V_q set by the subtraction of four dissimilar base–emitter junctions from V_{bias}, all having an equal say, and so all four junction temperatures ought to be factored into the final result. This would certainly be comprehensive, but four temperature sensors per channel is perhaps overdoing it. For the CFP stage, we can ignore the output device temperatures and only sense the drivers, which simplifies things and works well in practice.

If we can assume that the drivers and outputs come in complementary pairs with similar V_{be} behavior, then symmetry prevails and we need only consider one half of the output stage, so long as V_{bias} is halved to suit. This assumes the audio signal is symmetrical over timescales of seconds to minutes, so that equal dissipations and temperature rises occur in the top and bottom halves of the output stage. This seems a pretty safe bet, but as one example the unaccompanied human voice has positive and negative peak values that may differ by up to 8 dB, so prolonged a cappella performances have at least the potential to mislead any compensator that assumes symmetry. One amplifier that does use separate sensors for the upper and lower output sections is the Adcom GFA-565.

For the EF configuration, both drivers and outputs have an equal influence on the quiescent V_q, but the output devices normally get much hotter than the drivers, and their dissipation varies much more with output level. In this case the sensor goes on or near one of the output devices, thermally close to the output junction. It has been shown experimentally that the top of the TO-3 can is the best place to put it (see page 396). Recent experiments have confirmed that this holds true also for the TO-3P package (a large flat plastic package like an overgrown TO-220, and nothing at all like TO-3), which can easily get 20° hotter on its upper plastic surface than does the underlying heat-sink.

Since the first edition of this book, the TO-3 has all but disappeared; being replaced in high-power applications by the MT200 package. This is a large, flat, plastic format like a wider version of the TO-3P package. Once again, the top of the package gets hotter than the adjacent heat-sink.

In the CFP the drivers have most effect and the output devices, although still hot, have only one-twentieth the influence on the quiescent conditions. Driver dissipation is also much more variable, so now the correct place to put the thermal sensor is as near to the driver junction as you can get it.

Schemes for the direct servo control of quiescent current have been mooted[2], but all suffer from the difficulty that the quantity we wish to control is not directly available for measurement, as except in the complete absence of signal it is swamped by Class-B output currents. Nevertheless, several designs have been put forward in which the quiescent current is controlled by measuring the average current in the output stage. While this could be made to apparently work with steady test signals, it seems inevitable that there would be serious bias errors with a dynamic music signal.

In contrast the quiescent current of a Class-A amplifier is easily measured, allowing very precise feedback control; ironically its value is not critical for distortion performance.

So, how accurately must quiescent current be held? This is not easy to answer, not least because it is the wrong question. Chapter 6 established that the crucial parameter is not quiescent current (hereafter I_q) as such, but rather the quiescent voltage drop V_q across the two emitter resistors Re.

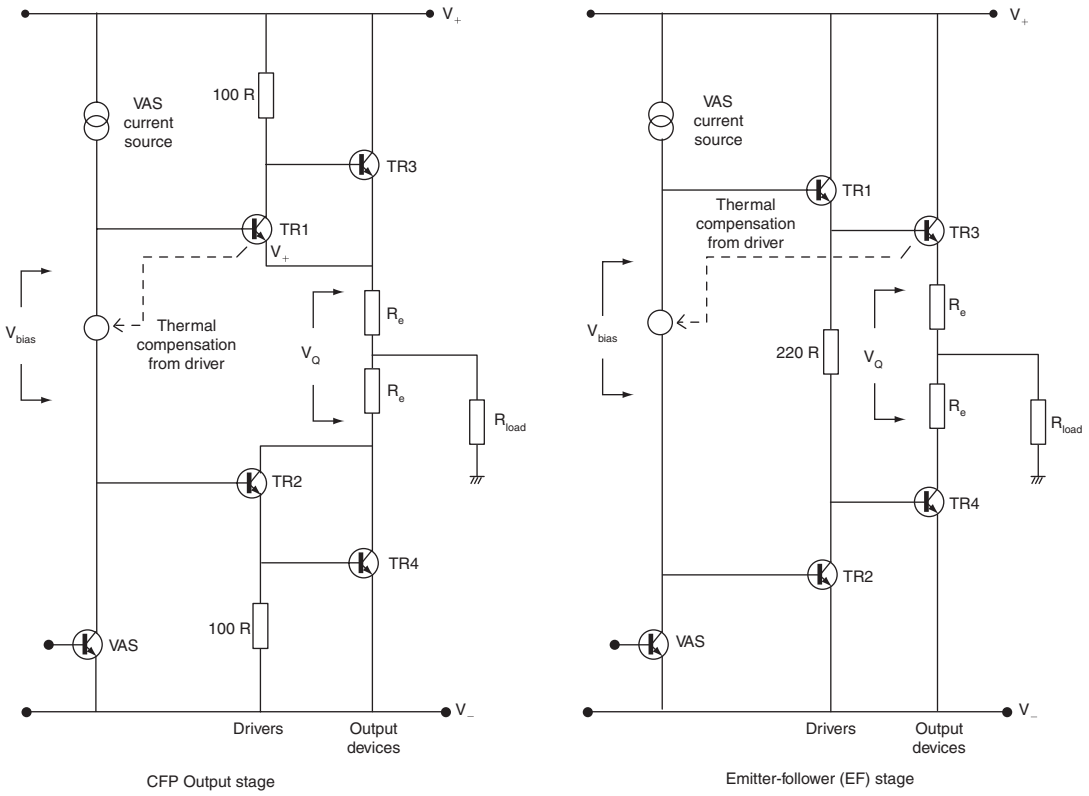


Figure 15.2: The emitter-follower (EF) and complementary feedback pair (CFP) output configurations, showing V_{bias} and V_q

This takes a little swallowing – after all, people have been worrying about quiescent current for 30 years or more – but it is actually good news, as the value of Re does not complicate the picture. The voltage across the output stage inputs (V_{bias}) is no less critical, for once Re is chosen V_q and I_q vary proportionally. The two main types of output stage, the emitter-follower (EF) and the complementary feedback pair (CFP), are shown in Figure 15.2. Their V_q tolerances are quite different.

From the measurements shown above the permissible error band for V_q in the EF stage is $\pm 100\,mV$ and for the CFP is $\pm 10\,mV$. These tolerances are not defined for all time; I only claim that they are realistic and reasonable. In terms of total V_{bias}, the EF needs $2.93\,V \pm 100\,mV$, and the CFP $1.30\,V \pm 10\,mV$. V_{bias} must be higher in the EF as four V_{be} values are subtracted from it to get V_q, while in the CFP only two driver V_{be} values are subtracted.

The CFP stage appears to be more demanding of V_{bias} compensation than EF, needing 1% rather than 3.5% accuracy, but things are not so simple. V_q stability in the EF stage depends primarily on the hot output devices, as EF driver dissipation varies only slightly with power output. V_q in the CFP depends almost entirely on driver junction temperature, as the effect of output device temperature is reduced by the local negative feedback; however, CFP driver dissipation varies strongly with power output so the superiority of this configuration cannot be taken for granted.

Driver heat-sinks are much smaller than those for output devices, so the CFP V_q time-constants promise to be some 10 times shorter.

From the statements made above, it would appear that with an EF-type output stage, it is essential to put the sensor on an output device, or as a second best onto the main heat-sink, particularly as EF driver dissipation varies only slightly with power output, and therefore apparently gives little indication of what the output device dissipation is. In fact, this is not the complete story. An EF-type amplifier with satisfactory bias stability *can* be made by putting the sensor on one of the driver heat-sinks. I used such a driver sensor in a recent project, as a result of some mechanical constraints that made putting a sensor on the main heat-sink very difficult, and the amplifier is still on the market and selling well. The approach can be made to work satisfactorily; unfortunately I have no space to explore it further here.

Basic Thermal Compensation

In Class-B, the usual method for reducing quiescent variations is so-called 'thermal feedback'. V_{bias} is generated by a thermal sensor with a negative temperature coefficient, usually a V_{be}-multiplier transistor mounted on the main heat-sink. This system has proved entirely workable over the last 30-odd years, and usually prevents any possibility of thermal runaway. However, it suffers from thermal losses and delays between output devices and temperature sensor that make maintenance of optimal bias rather questionable, and in practice quiescent conditions are a function of recent signal and thermal history. Thus the crossover linearity of most power amplifiers is intimately bound up with their thermal dynamics, and it is surprising this area has not been examined more closely; Sato et al.[1] produced one of the few serious papers on the subject, though the conclusions it reaches appear to be unworkable, depending on calculating power dissipation from amplifier output voltage without considering load impedance.

As is almost routine in audio design, things are not as they appear. So-called 'thermal feedback' is not feedback at all – this implies the thermal sensor is in some way controlling the output stage temperature; it is not. It is really a form of approximate *feedforward* compensation, as shown in Figure 15.3. The quiescent current (I_q) of a Class-B design causes a very small dissipation compared with the signal, and so there is no meaningful feedback path returning from I_q to the left of the diagram. (This might be less true of Class-AB, where quiescent dissipation may be significant.) Instead, this system aspires to make the sensor junction temperature mimic the driver or output junction temperature, though it can never do this promptly or exactly because of the thermal resistances and thermal capacities that lie between driver and sensor temperatures in Figure 15.3. It does not place either junction temperature or quiescent current under direct feedback control, but merely aims to cancel out the errors. Hereafter I simply call this *thermal compensation*.

Assessing the Bias Errors

The temperature error must be converted to mV error in V_q, for comparison with the tolerance bands suggested above. In the CFP stage this is straightforward; both driver V_{be} and the halved

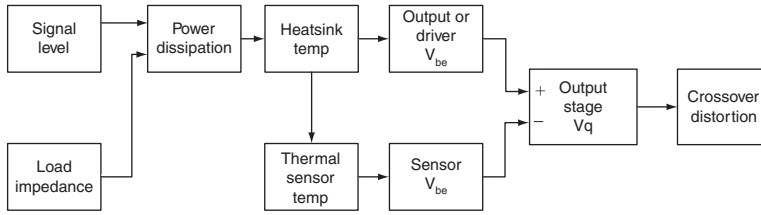


Figure 15.3: Thermal signal flow of a typical power amplifier, showing that there is no thermal feedback to the bias generator. There is instead feedforward of driver junction temperature, so that the sensor V_{be} will hopefully match the driver V_{be}

V_{bias} voltage decrease by 2 mV/°C, so temperature error converts to voltage error by multiplying by 0.002. Only half of each output stage will be modeled, exploiting symmetry, so most of this chapter deals in half-V_q errors, etc. To minimize confusion this use of half-amplifiers is adhered to throughout, except at the final stage when the calculated V_q error is doubled before comparison with the tolerance bands quoted above.

The EF error conversion is more subtle. The EF V_{bias} generator must establish $4 \times V_{be}$ plus V_q, so the V_{be} of the temperature-sensing transistor is multiplied by about 4.5 times, and so decreases at 9 mV/°C. The CFP V_{bias} generator only multiplies 2.1 times, decreasing at 4 mV/°C. The corresponding values for a half-amplifier are 4.5 and 2 mV/°C.

However, the EF drivers are at near-constant temperature, so after two driver V_{be} values have been subtracted from V_{bias}, the remaining voltage decreases faster with temperature than does output device V_{be}. This runs counter to the tendency to under-compensation caused by thermal attenuation between output junctions and thermal sensor; in effect the compensator has thermal *gain,* and this has the potential to reduce long-term V_q errors. I suspect this is the real reason why the EF stage, despite looking unpromising, can in practice give acceptable quiescent stability.

Thermal Simulation

Designing an output stage requires some appreciation of how effective the thermal compensation will be, in terms of how much delay and attenuation the 'thermal signal' suffers between the critical junctions and the V_{bias} generator.

We need to predict the thermal behavior of a heat-sink assembly over time, allowing for things like metals of dissimilar thermal conductivity and the very slow propagation of heat through a mass compared with near-instant changes in electrical dissipation. Practical measurements are very time-consuming, requiring special equipment such as multi-point thermocouple recorders. A theoretical approach would be very useful.

For very simple models, such as heat flow down a uniform rod, we can derive analytical solutions to the partial differential equations that describe the situation; the answer is an equation directly relating temperature to position along the rod and time. However, even slight complications (such as a non-uniform rod) involve rapidly increasing mathematical complexities, and anyone who is not already deterred should consult Carslaw and Jaeger[3]; this will deter them.

Table 15.2: The relation between real thermal units and the electrical units used in simulation

	Reality	Simulation
Temperature	°C	Volts
Heat quantity	Joules (watt-seconds)	Coulombs (amp-seconds)
Heat flow rate	Watts	Amps
Thermal resistance	°C/W	Ohms
Thermal capacity	°C/J	Farads
Heat source	Dissipative element, e.g. transistor	Current source
Ambient	Medium-sized planet	Voltage source

To avoid direct confrontation with higher mathematics, finite element and relaxation methods were developed; the snag is that Finite Eement Analysis (FEA) is a rather specialized taste, and so commercial FEA software is expensive.

I therefore cast about for another method, and found I already had the wherewithal to solve problems of thermal dynamics; the use of electrical analogs is the key. If the thermal problem can be stated in terms of lumped electrical elements, then a circuit simulator of the SPICE type can handle it, and as a bonus has extensive capabilities for graphical display of the output. The work here was done with PSPICE. A more common use of electrical analogs is in the electromechanical domain of loudspeakers (see Murphy[4] for a virtuoso example).

The simulation approach treats temperature as voltage, and thermal energy as electric charge, making thermal resistance analogous to electrical resistance, and thermal capacity to electrical capacitance. Thermal capacity is a measure of how much heat is required to raise the temperature of a mass by 1°C. (And if anyone can work out what the thermal equivalent of an inductor is, I would be interested to know.) With the right choice of units the simulator output will be in volts, with a one-to-one correspondence with degrees Celsius, and amps similarly representing watts of heat flow (see Table 15.2). It is then simple to produce graphs of temperature against time.

Since heat flow is represented by current, the inputs to the simulated system are current sources. A voltage source would force large chunks of metal to change temperature instantly, which is clearly wrong. The ambient is modeled by a voltage source, as it can absorb any amount of heat without changing temperature.

Modeling the EF Output Stage

The major characteristic of EF output stages is that the output device junction temperatures are directly involved in setting I_q. This junction temperature is not accessible to a thermal compensation system, and measuring the heat-sink temperature instead provides a poor approximation, attenuated by the thermal resistance from junction to heat-sink mass, and heavily time-averaged by heat-sink thermal inertia. This can cause serious production problems in initial setting up; any drift of I_q will be very slow as a lot of metal must warm up.

For EF outputs, the bias generator must attempt to establish an output bias voltage that is a summation of four driver and output V_{be} values. These do not vary in the same way. It seems at first a bit of a mystery how the EF stage, which still seems to be the most popular output topology, works as well as it does. The probable answer is in Figure 15.1, which shows how driver dissipation (averaged over a cycle) varies with peak output level for the three kinds of EF output described in Chapter 6, and for the CFP configuration. The SPICE simulations used to generate this graph used a triangle waveform, to give a slightly closer approximation to the peak-average ratio of real waveforms. The rails were ±50V and the load 8Ω.

It is clear that the driver dissipation for the EF types is relatively constant with power output, while the CFP driver dissipation, although generally lower, varies strongly. This is a consequence of the different operation of these two kinds of output. In general, the drivers of an EF output remain conducting to some degree for most or all of a cycle, although the output devices are certainly off half the time. In the CFP, however, the drivers turn off almost in synchrony with the outputs, dissipating an amount of power that varies much more with output. This implies that EF drivers will work at roughly the same temperature and can be neglected in arranging thermal compensation; the temperature-dependent element is usually attached to the main heat-sink, in an attempt to compensate for the junction temperature of the outputs alone. The Type I EF output keeps its drivers at the most constant temperature; this may (or may not) have something to do with why it is the most popular of the EF types.

(The above does not apply to integrated Darlington outputs, with drivers and assorted emitter resistors combined in one ill-conceived package, as the driver sections are directly heated by the output junctions. This would seem to work directly against quiescent stability, and why these compound devices are ever used in audio amplifiers remains a mystery to me.)

The drawback with most EF thermal compensation schemes is the slow response of the heat-sink mass to thermal transients, and the obvious solution is to find some way of getting the sensor closer to one of the output junctions (symmetry of dissipation is assumed). If TO-3 devices are used, then the flange on which the actual transistor is mounted is as close as we can get without a hacksaw. This is, however, clamped to the heat-sink, and almost inaccessible, though it might be possible to hold a sensor under one of the mounting bolts. A simpler solution is to mount the sensor on the top of the TO-3 can. This is probably not as accurate an estimate of junction temperature as the flange would give, but measurement shows the top gets much hotter much faster than the heat-sink mass, so while it may appear unconventional, it is probably the best sensor position for an EF output stage. Figure 15.4 shows the results of an experiment designed to test this. A TO-3 device was mounted on a thick aluminum L-section thermal coupler in turn clamped to a heat-sink; this construction is representative of many designs. Dissipation equivalent to 100 W/8Ω was suddenly initiated, and the temperature of the various parts monitored with thermocouples. The graph clearly shows that the top of the TO-3 responds much faster and with a larger temperature change, though after the first two minutes the temperatures are all increasing at the same rate. The whole assembly took more than an hour to asymptote to thermal equilibrium.

Figure 15.5 shows a TO-3 output device mounted on a thermal coupling bar, with a silicone thermal washer giving electrical isolation. The coupler is linked to the heat-sink proper via a second

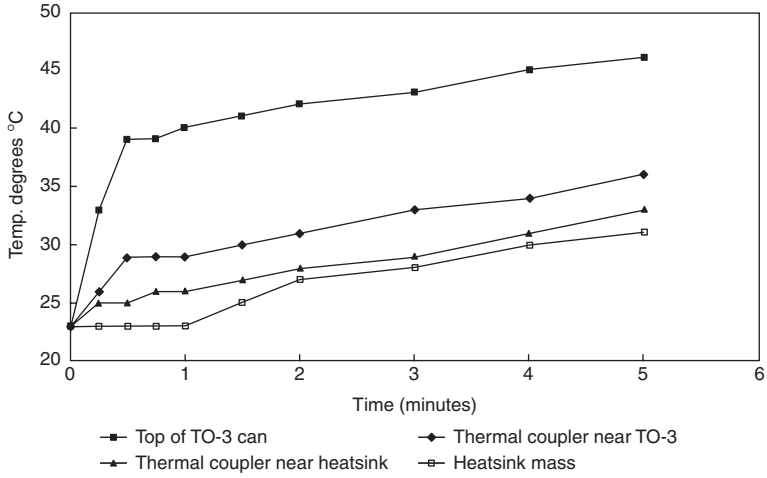


Figure 15.4: Thermal response of a TO-3 device on a large heat-sink when power is suddenly applied. The top of the TO-3 can responds most rapidly

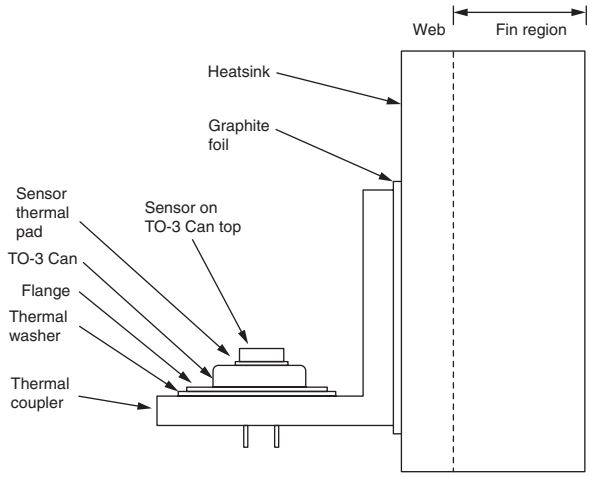


Figure 15.5: A TO-3 power transistor attached to a heat-sink by a thermal coupler. The thermal sensor is shown on the can top; the more usual position would be on the thermal coupler

conformal material; this need not be electrically insulating so highly efficient materials like graphite foil can be used. This is representative of many amplifier designs, though a good number have the power devices mounted directly on the heat-sink; the results hardly differ. A simple thermal-analogy model of Figure 15.5 is shown in Figure 15.6; the situation is radically simplified by treating each mass in the system as being at a uniform temperature, i.e. isothermal, and therefore representable by one capacity each. The boundaries between parts of the system are modeled, but the thermal capacity of each mass is concentrated at a notional point. In assuming this we give capacity elements zero thermal resistance, e.g. both sides of the thermal coupler will always be at the same temperature. Similarly, elements such as the thermal washer are assumed to have zero heat capacity, because they are very thin and have negligible mass compared with other elements in the system.

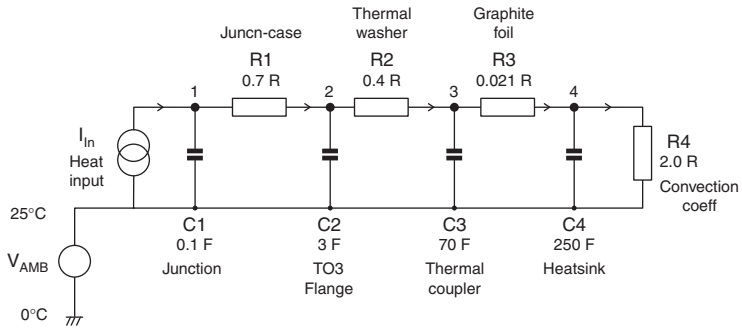


Figure 15.6: A thermal/electrical model of Figure 15.5, for half of one channel only. Node 1 is junction temperature, node 2 flange temperature, and so on. V_{amb} sets the baseline to 25°C. Arrows show heat flow

Thus the parts of the thermal system can be conveniently divided into two categories: pure thermal resistances and pure thermal capacities. Often this gives adequate results; if not, more subdivision will be needed. Heat losses from parts other than the heat-sink are neglected.

Real output stages have at least two power transistors; the simplifying assumption is made that power dissipation will be symmetrical over anything but the extreme short term, and so one device can be studied by slicing the output stage, heat-sink, etc., in half.

It is convenient to read off the results directly in °C, rather than temperature rise above ambient, so Figure 15.6 represents ambient temperature with a voltage source V_{amb} that offsets the baseline (node 10) 25°C from simulator ground, which is inherently at 0°C (0V).

Values of the notional components in Figure 15.6 have to be filled in with a mixture of calculation and manufacturer's data. The thermal resistance R1 from junction to case comes straight from the data book, as does the resistance R2 of the TO-3 thermal washer, as well as R4, the convection coefficient of the heat-sink itself, otherwise known as its thermal resistance to ambient. This is always assumed to be constant with temperature, which it very nearly is. Here R4 is 1°C/W, so this is doubled to 2 as we cut the stage in half to exploit symmetry.

R3 is the thermal resistance of the graphite foil; this is cut to size from a sheet and the only data is the bulk thermal resistance of 3.85 W/mK, so R3 must be calculated. Thickness is 0.2 mm, and the rectangle area in this example was 38 mm × 65 mm. We must be careful to convert all lengths to meters:

$$\begin{aligned}\text{Heat flow/°C} &= \frac{3.85 \times \text{Area}}{\text{Thickness}} \\ &= \frac{3.83 \times (0.038 \times 0.065)}{0.0002} \\ &= 47.3 \text{ W/°C}\end{aligned} \qquad \text{Equation 15.1}$$

$$\begin{aligned}\text{So thermal resistance} &= \frac{1}{47.3} \\ &= 0.021\text{°C/W}\end{aligned}$$

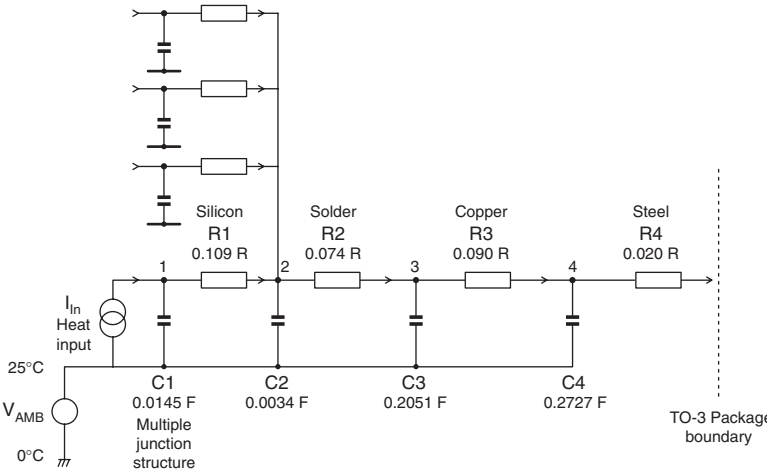


Figure 15.7: Internal thermal model for a TO-3 transistor. All the heat is liberated in the junction structure, shown as *N* multiples of C1 to represent a typical interdigitated power transistor structure

Thermal resistance is the reciprocal of heat flow per degree, so R3 is 0.021°C/W, which just goes to show how efficient thermal washers can be if they do not have to be electrical insulators as well.

In general all the thermal capacities will have to be calculated, sometimes from rather inadequate data, thus:

$$\text{Thermal capacity} = \text{Density} \times \text{Volume} \times \text{Specific heat}$$

A power transistor has its own internal structure and its own internal thermal model (Figure 15.7). This represents the silicon die itself, the solder that fixes it to the copper header, and part of the steel flange the header is welded to. I am indebted to Motorola for the parameters, from an MJ15023 TO-3 device[5]. The time-constants are all extremely short compared with heat-sinks, and it is unnecessary to simulate in such detail here.

The thermal model of the TO-3 junction is therefore reduced to lumped component C1, estimated at 0.1 J/°C; with a heat input of 1 W and no losses its temperature would increase linearly by 10°C/s. The capacity C2 for the transistor package was calculated from the volume of the TO-3 flange (representing most of the mass) using the specific heat of mild steel. The thermal coupler is known to be aluminum alloy (not pure aluminum, which is too soft to be useful) and the calculated capacity of 70 J/°C should be reliable. A similar calculation gives 250 J/°C for the larger mass of the aluminum heat-sink. Our simplifying assumptions are rather sweeping here because we are dealing with a substantial chunk of finned metal, which will never be truly isothermal.

The derived parameters for both output TO-3s and TO-225AA drivers are summarized in Table 15.3. The drivers are assumed to be mounted onto small individual heat-sinks with an isolating thermal washer; the data is for the popular Redpoint SW38-1 vertical heat-sink.

Figures 15.8 and 15.9 show the result of a step-function in heat generation in the output transistor; 20 W dissipation is initiated, corresponding approximately to a sudden demand for full sine-wave

Table 15.3: The parameters for an output stage with TO-3 outputs and TO-225AA drivers

			Output device	Driver
C1	Junction capacity	J/°C	0.1	0.05
R1	Junction-case resistance	°C/W	0.7	6.25
C2	Transistor package capacity	J/°C	3.0	0.077
R2	Thermal washer resitance	°C/W	0.4	6.9
C3	Coupler capacity	J/°C	70	–
R3	Coupler heat-sink resistance	°C/W	0.021	–
C4	Heat-sink capacity	J/°C	250	20.6
R4	Heat-sink convective resistance	°C/W	2.0	10.0

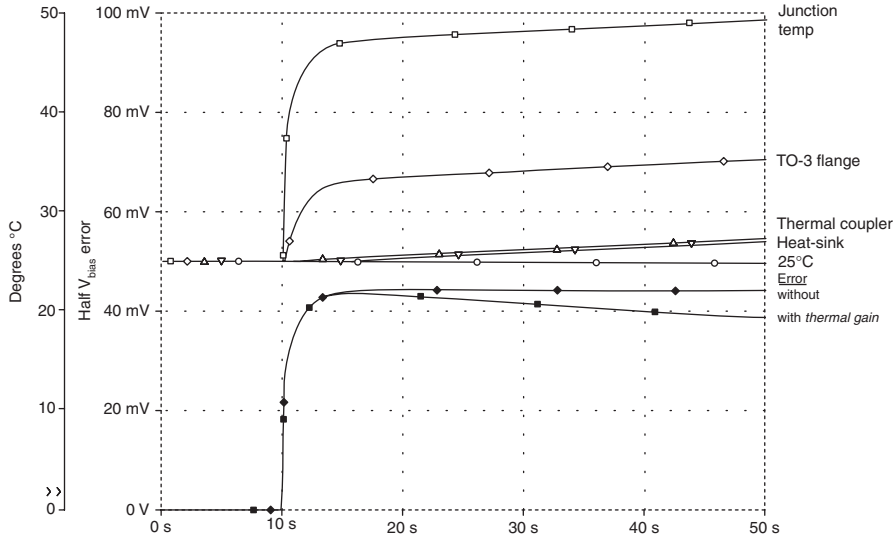


Figure 15.8: Results for Figure 15.6, with step heat input of 20W to junction initiated at time=10s. Upper plot shows temperatures, lower the V_{bias} error for half of output stage

power from a quiescent 100W amplifier. The junction temperature V(1) takes off near-vertically, due to its small mass and the substantial thermal resistance between it and the TO-3 flange; the flange temperature V(2) shows a similar but smaller step as R2 is also significant. In contrast the thermal coupler, which is so efficiently bonded to the heat-sink by graphite foil that they might almost be one piece of metal, begins a slow exponential rise that will take a very long time to asymptote. Since after the effect of C1 and C2 have died away the junction temperature is offset by a constant amount from the temperature of C3 and C4, V(1) also shows a slow rise. Note the *X*-axis of Figure 15.9 must be in kiloseconds, because of the relatively enormous thermal capacity of the heat-sink.

This shows that a temperature sensor mounted on the main heat-sink can never give accurate bias compensation for junction temperature, even if it is assumed to be isothermal with the heat-sink; in practice there will be some sensor cooling that will make the sensor temperature slightly under-read the heat-sink temperature V(4). Initially the temperature error V(1)–V(4) increases rapidly as

the TO-3 junction heats, reaching 13° in about 200 ms. The error then increases much more slowly, taking 6 seconds to reach the effective final value of 22°. If we ignore the thermal-gain effect mentioned above, the long-term V_q error is +44 mV, i.e. V_q is too high. When this is doubled to allow for both halves of the output stage we get +88 mV, which uses up nearly all of the ±100 mV error band, without any other inaccuracies. (Hereafter all V_{bias}/V_q error figures quoted have been doubled and so apply to a complete output stage.) Including the thermal gain actually makes little difference over a 10-second timescale; the lower V_q error trace in Figure 15.8 slowly decays as the main heat-sink warms up, but the effect is too slow to be useful.

The amplifier V_q and I_q will therefore rise under power, as the hot output device V_{be} voltages fall, but the cooler bias generator on the main heat-sink reduces its voltage by an insufficient amount to compensate.

Figure 15.9 shows the long-term response of the system. At least 2500 s pass before the heat-sink is within a degree of final temperature.

In the past I have recommended that EF output stages should have the thermal sensor mounted on the top of the TO-3 can, despite the mechanical difficulties. This is not easy to simulate as no data is available for the thermal resistance between junction and can top. There must be an additional thermal path from junction to can, as the top very definitely gets *hotter* than the flange measured at the very base of the can. In view of the relatively low temperatures, this path is probably due to internal convection rather than radiation.

A similar situation arises with TO-3P packages, for the top plastic surface can get at least 20° hotter than the heat-sink just under the device. Recent work has shown that this also applies to the MT200 and the TO-264 plastic packages.

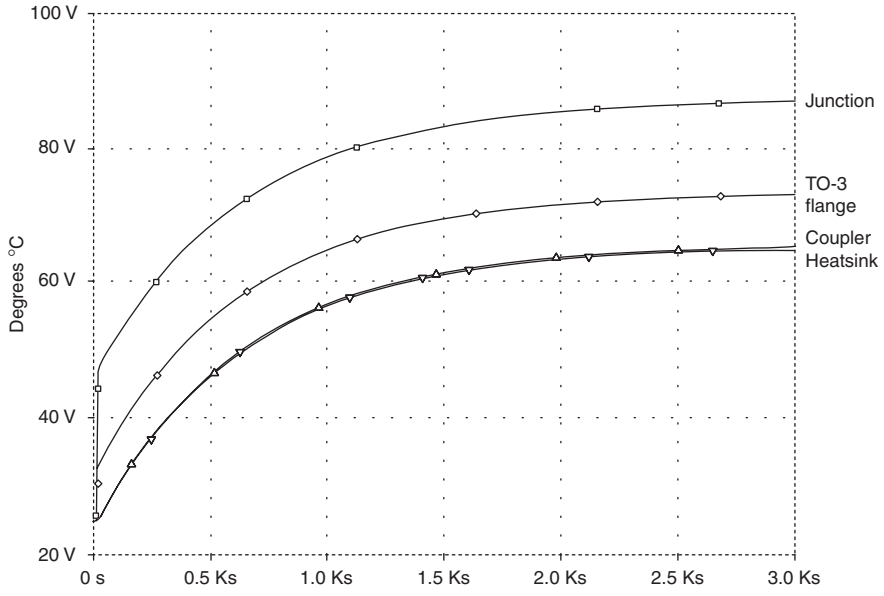


Figure 15.9: The long-term version of Figure 15.8, showing that it takes over 40 minutes for the heat-sink to get within 1° of final temperature

Using the real thermocouple data from page 392, I have estimated the parameters of the thermal paths to the TO-3 top. This gives Figure 15.10, where the values of elements R20, R21, C5 should be treated with considerable caution, though the temperature results in Figure 15.11 match reality fairly well; the can top (V20) gets hotter faster than any other accessible point. R20 simulates the heating path from the junction to the TO-3 can and R21 the can-to-flange cooling path, C5 being can thermal capacity.

Figure 15.10 includes approximate representation of the cooling of the sensor transistor, which now matters. R22 is the thermal pad between the TO-3 top and the sensor, C6 the sensor thermal capacity, and R23 is the convective cooling of the sensor, its value being taken as twice the datasheet free-air thermal resistance as only one face is exposed. The sensor transistor is assumed to be isothermal, and not significantly heated by its own standing current.

Placing the sensor on top of the TO-3 would be expected to reduce the steady-state bias error dramatically. In fact, it overdoes it, as after factoring in the thermal gain of a V_{be}-multiplier in an EF stage, the bottom-most trace of Figure 15.11 shows that the bias is overcompensated; after the initial positive transient error, V_{bias} falls too low giving an error of −30 mV, slowly worsening as the main heat-sink warms up. If thermal gain had been ignored, the simulated error would have apparently fallen from +44 (Figure 15.8) to +27 mV, apparently a useful improvement, but actually illusory.

Since the new sensor position overcompensates for thermal errors, there should be an intermediate arrangement giving near-zero long-term error. I found this condition occurs if R22 is increased to 80°C/W, requiring some sort of semi-insulating material rather than a thermal pad, and gives the upper error trace in the lower half of Figure 15.11. This peaks at +30 mV after 2 s, and then decays to nothing over the next 20. This is much superior to the persistent error in Figure 15.8, so this new technique may be useful, but bear in mind that it slows the sensor response.

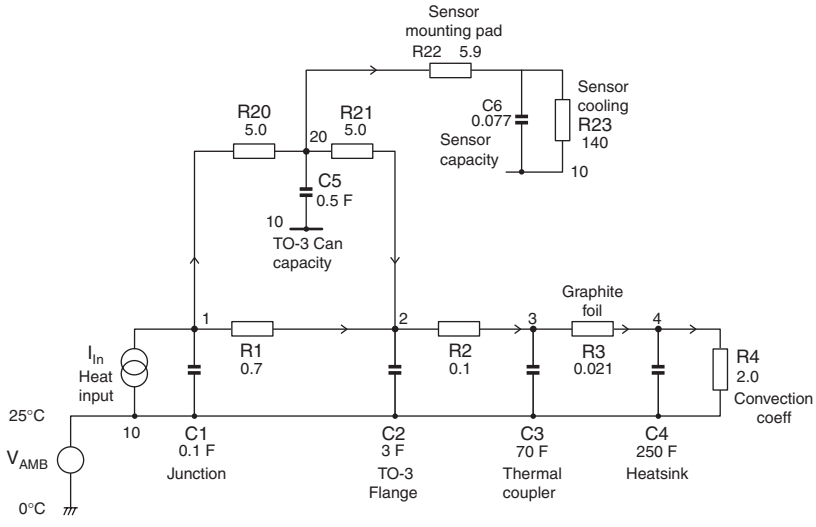


Figure 15.10: Model of EF output stage with thermal paths to TO-3 can top modeled by R20, R21. C5 simulates can capacity. R23 models sensor convection cooling; node 21 is sensor temperature

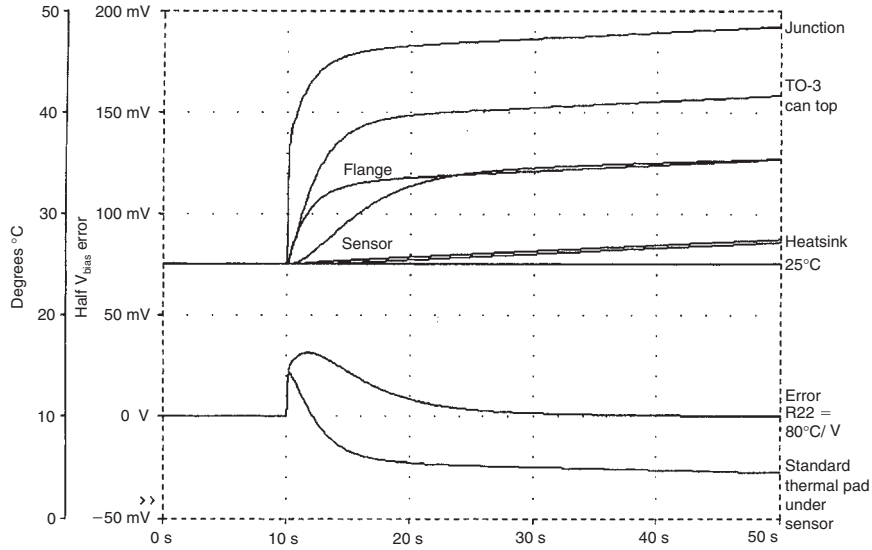


Figure 15.11: The simulation results for Figure 15.10; lower plot shows V_{bias} errors for normal thermal pad under sensor, and 80°C/W semi-insulator. The latter has near-zero long-term error

It has been suggested that a sensor position that needs long wires to connect to the rest of the circuitry could make HF stability uncertain. I have had no trouble with wires up to 20 cm in length, and I think this is not surprising because the result is presumably an increase in the capacitance to ground at the VAS collector of a few pF. Since the effect of such capacitance is, perhaps counter-intuitively but quite definitely, to increase HF stability (see Chapter 8 for more detail), this seems to be something of a non-problem.

Modeling the CFP Output Stage

In the CFP configuration, the output devices are inside a local feedback loop and play no significant part in setting V_q, which is dominated by thermal changes in the driver V_{be} values. Such stages are virtually immune to thermal runaway; I have found that assaulting the output devices with a powerful heat gun induces only very small I_q changes. Thermal compensation is mechanically simpler as the V_{be}-multiplier transistor is usually mounted on one of the driver heat-sinks, where it aspires to mimic the driver junction temperature.

It is now practical to make the bias transistor of the same type as the drivers, which may help to give the best matching of V_{be} [6], though given the differences in $I_{c,}$ how important this is in practice is uncertain. It definitely avoids the difficulty of trying to attach a small-signal (probably TO92) transistor package to a heat-sink.

Since it is the driver junctions that count, output device temperatures are here neglected. The thermal parameters for a TO-225AA driver (e.g. MJE340/350) on the SW38-1 vertical heat-sink are shown in Table 15.3; the drivers are on individual heat-sinks so their thermal resistance is used directly, without doubling.

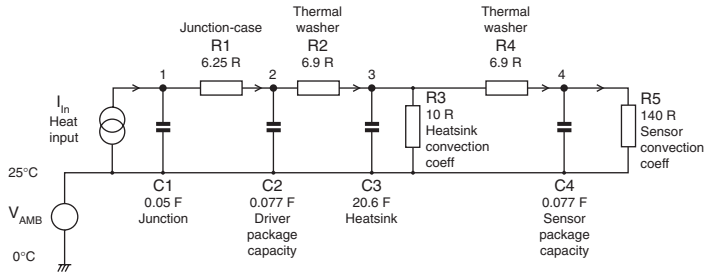


Figure 15.12: Model of a CFP stage. Driver transistor is mounted on a small heat-sink, with sensor transistor on the other side. Sensor dynamics and cooling are modelled by R4, C4, and R5

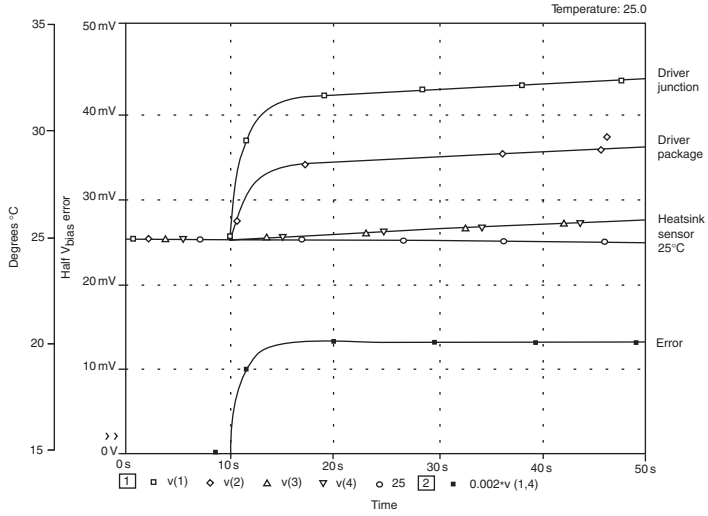


Figure 15.13: Simulation results for CFP stage, with step heat input of 0.5 W. Heat-sink and sensor are virtually isothermal, but there is a persistent error as driver is always hotter than heat-sink due to R1, R2

In the simulation circuit (Figure 15.12) V(3) is the heat-sink temperature; the sensor transistor (also MJE340) is mounted on this sink with thermal washer R4, and has thermal capacity C4. R5 is convective cooling of the sensor. In this case the resulting differences in Figure 15.13 between sink V(3) and sensor V(4) are very small.

We might expect the CFP delay errors to be much shorter than in the EF; however, simulation with a step heat input suitably scaled down to 0.5 W (Figure 15.13) shows changes in temperature error V(1)–V(4) that appear rather paradoxical; the error reaches 5° in 1.8 s, leveling out at 6.5° after about 6 s. This is markedly slower than the EF case, and gives a total bias error of +13 mV, which after doubling to +26 mV is well outside the CFP error band of ±10 mV.

The initial transients are slowed down by the much smaller step heat input, which takes longer to warm things up. The *final* temperature, however, is reached in 500 rather than 3000 s, and the timescale is now in hundreds rather than thousands of seconds. The heat input is smaller, but the driver heat-sink capacity is also smaller, and the overall time-constant is less.

It is notable that both timescales are much longer than musical dynamics.

The Integrated Absolute Error Criterion

Since the thermal sensor is more or less remote from the junction whose gyrations in temperature will hopefully be canceled out, heat losses and thermal resistances cause the temperature change reaching the sensor to be generally too little and too late for complete compensation.

In this section, all the voltages and errors here are for one-half of an output stage, using symmetry to reduce the work involved. These 'half-amplifiers' are used throughout this chapter, for consistency, and the error voltages are only doubled to represent reality (a complete output stage) when they are compared against the tolerance bands previously quoted.

We are faced with errors that vary not only in magnitude, but also in their persistence over time; judgment is required as to whether a prolonged small error is better than a large error that quickly fades away.

The same issue faces most servomechanisms, and I borrow from Control Theory the concept of an *error criterion* that combines magnitude and time into one number[7,8]. The most popular criterion is the Integrated Absolute Error (IAE), which is computed by integrating the absolute value of the error over a specified period after giving the system a suitably provocative stimulus; the absolute value prevents positive and negative errors canceling over time. Another common criterion is the Integrated Square Error (ISE), which solves the polarity problem by squaring the error before integration – this also penalizes large errors much more than small ones. It is not immediately obvious which of these is most applicable to bias control and the psychoacoustics of crossover distortion that changes with time, so I have chosen the popular IAE.

One difficulty is that the IAE criterion for bias voltage tends to accumulate over time, due to the integration process, so any constant bias error quickly comes to dominate the IAE result. In this case, the IAE is little more than a counter-intuitive way of stating the constant error, and must be quoted over a specified integration time to mean anything at all. This is why the IAE concept was not introduced earlier in this chapter.

Much more useful results are obtained when the IAE is applied to a situation where the error decays to a very small value after the initial transient and stays there. This can sometimes be arranged in amplifiers, as I hope to show. In an ideal system where the error decayed to zero without overshoot, the IAE would asymptote to a constant value after the initial transient. In real life, residual errors make the IAE vary slightly with time, so for consistency all the IAE values given here are for 30 s after the step-input.

Improved Thermal Compensation for the EF Stage

It was shown above that the basic EF stage with the sensor on the main heat-sink has significant thermal attenuation error and therefore undercompensates temperature changes. (The V_q error is +44 mV, the positive sign showing it is too high. If the sensor is on the TO-3 can top it overcompensates instead; V_q error is then −30 mV.)

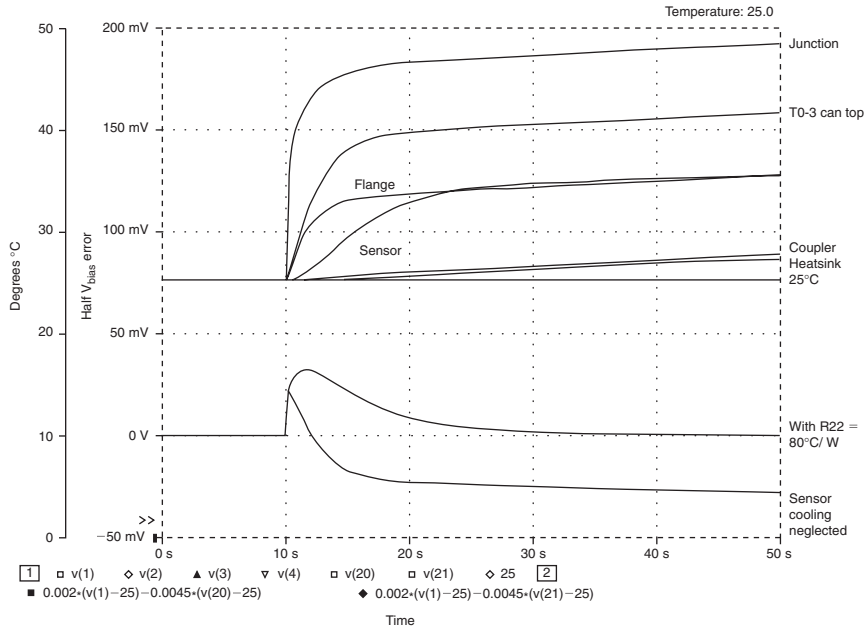


Figure 15.14: EF behavior with semi-insulating pad under sensor on TO-3 can top. The sensor in the upper temperature plot rises more slowly than the flange, but much faster than the main heat-sink or coupler. In the lower V_q error section, the upper trace is for an 80°C/W thermal resistance under the sensor, giving near-zero error. The bottom trace shows the serious effect of ignoring sensor cooling in the TO-3-top version

If an intermediate configuration is contrived by putting a layer of controlled thermal resistance (80°C/W) between the TO-3 top and the sensor, then the 50 seconds timescale component of the error can be reduced to near zero. This is the top error trace in the bottom half of Figure 15.14; the lower trace shows the wholly misleading result if sensor heat losses are neglected in this configuration.

Despite this medium-term accuracy, if the heat input stimulus remains constant over the very long term (several kiloseconds) there still remains a very slow drift towards overcompensation due to the slow heating of the main heat-sink (Figure 15.15).

This long-term drift is a result of the large thermal inertia of the main heat-sink and, since it takes 1500 s (25 minutes) to go from zero to −32 mV, is of doubtful relevance to the timescales of music and signal level changes. On doubling to −64 mV, it remains within the EF V_q tolerance of ±100 mV. On the shorter 50 seconds timescale, the half-amplifier error remains within a ±1 mV window from 5 to 60 seconds after the step-input.

For the EF stage, a very-long-term drift component will always exist so long as the output device junction temperature is kept down by means of a main heat-sink that is essentially a weighty chunk of finned metal.

The EF system stimulus is a 20 W step as before, being roughly worst-case for a 100 W amplifier. Using the 80°C/W thermal semi-insulator described above gives the upper error trace in Figure 15.16,

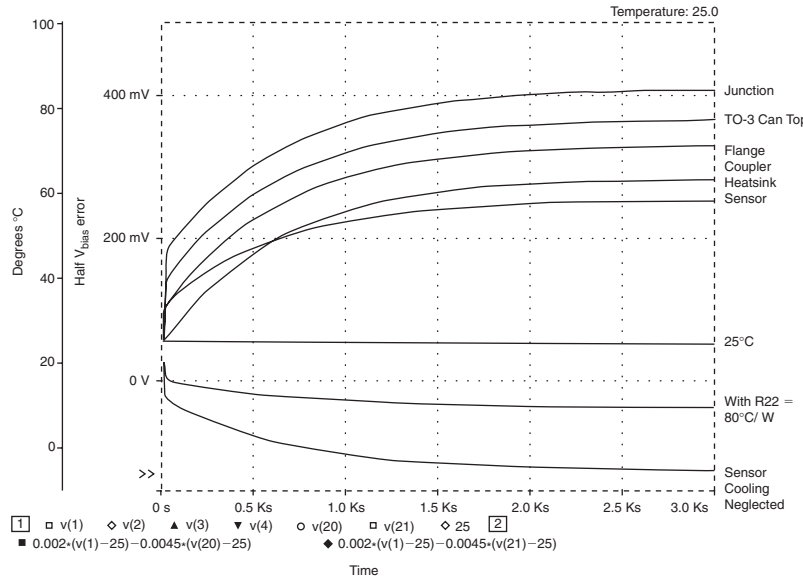


Figure 15.15: Over a long timescale, the lower plot shows that the V_q error, although almost zero in Figure 15.14, slowly drifts into overcompensation as the heat-sink temperature (upper plot) reaches asymptote

and an IAE of 254 mV-s after 30 seconds. This is relatively large because of the extra time delay caused by the combination of an increased R22 with the unchanged sensor thermal capacity C6. Once more, this figure is for a half-amplifier, as are all IAEs in this chapter.

Up to now I have assumed that the temperature coefficient of a V_{be}-multiplier bias generator is rigidly fixed at -2 mV/°C times the V_{be}-multiplication factor, which is about 4.5× for EF and 2× for CFP. The reason for the extra thermal gain displayed by the EF was set out on page 389.

The above figures are for both halves of the output stage, so the half-amplifier value for EF is -4.5 mV/°C and for CFP -2 mV/°C. However, if we boldly assume that the V_{bias} generator can have its thermal coefficient varied at will, the insulator and its aggravated time-lag can be eliminated.

If a thermal pad of standard material is once more used between the sensor and the TO-3 top, the optimal V_{bias} coefficient for minimum error over the first 40 seconds proves to be -2.8 mV/°C, which is usefully less than -4.5. The resulting 30-s IAE is 102 mV-s, more than a two times improvement (see the lower trace in Figure 15.16 for comparison with the semi-insulator method described above).

From here on I am assuming that a variable temperature coefficient (tempco) bias generator can be made when required; the details of how to do it are not given here. It is an extremely useful device, as thermal attenuation can then be countered by increasing the thermal gain*;* it does not, however, help with the problem of thermal delay.

In the second EF example above, the desired tempco is -2.8 mV/°C, while an EF output stage plus has an actual tempco of -4.5 mV/°C. (This inherent thermal gain in the EF was explained on page 389.)

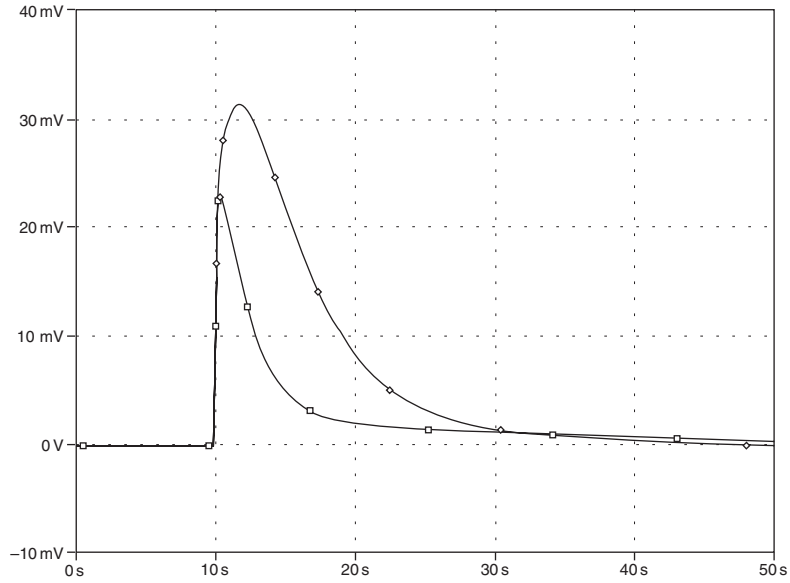


Figure 15.16: The transient error for the semi-insulating pad and the low-tempco version. The latter responds much faster, with a lower peak error, and gives less than half the integrated absolute error (IAE)

In this case we need a bias generator that has a *smaller* tempco than the standard circuit. The conventional EF with its temperature sensor on the relatively cool main heat-sink would require a *larger* tempco than standard.

A potential complication is that amplifiers should also be reasonably immune to changes in ambient temperature, quite apart from changes due to dissipation in the power devices. The standard tempco gives a close approach to this automatically, as the V_{be}-multiplication factor is naturally almost the same as the number of junctions being biased. However, this will no longer be true if the tempco is significantly different from standard, so it is necessary to think about a bias generator that has one tempco for power-device temperature changes and another for ambient changes. This sounds rather daunting, but is actually fairly simple.

Improved Compensation for the CFP Output Stage

As revealed earlier, the CFP output stage has a much smaller bias tolerance of ± 10 mV for a whole amplifier, and surprisingly long time-constants. A standard CFP stage therefore has larger relative errors than the conventional EF stage with thermal sensor on the main heat-sink; this is the opposite of conventional wisdom. Moving the sensor to the top of the TO-3 can was shown to improve the EF performance markedly, so we shall attempt an analogous improvement with driver compensation.

The standard CFP thermal compensation arrangements have the sensor mounted on the driver heat-sink, so that it senses the heat-sink temperature rather than that of the driver itself (see Figure 15.17a for mechanical arrangement and Figure 15.18 for the thermal model). As in the

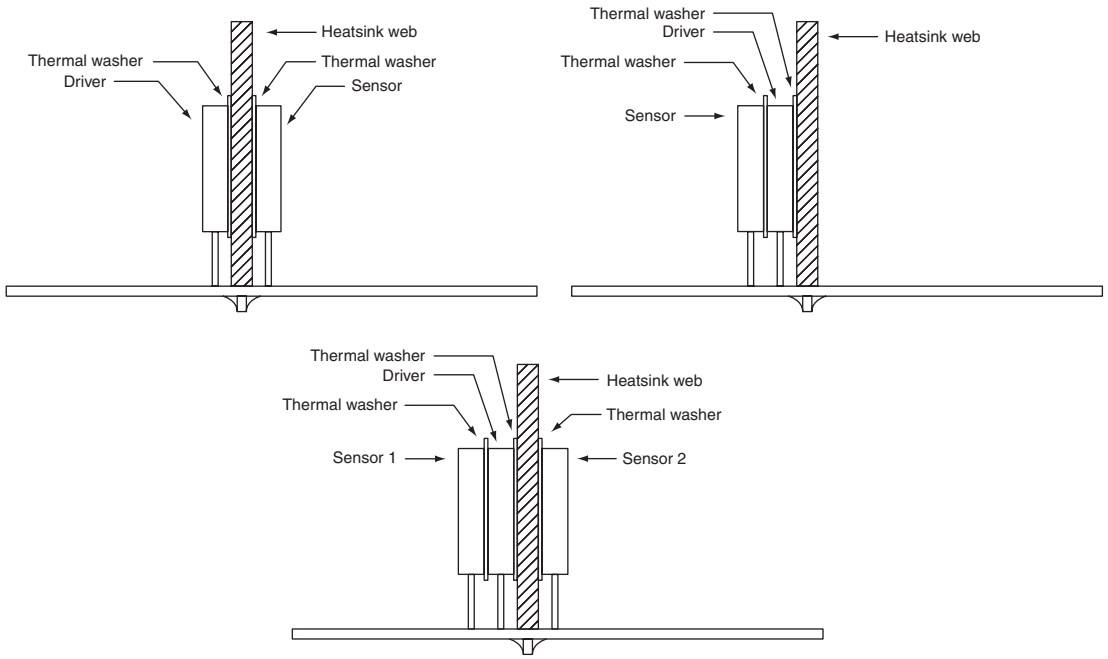


Figure 15.17: (a) The sensor transistor on the driver heat-sink. (b) An improved version, with the sensor mounted on top of the driver itself, is more accurate. (c) Using two sensors to construct a junction estimator

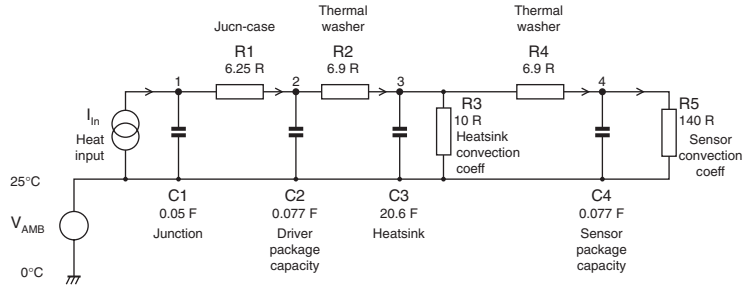


Figure 15.18: Thermal circuit of normal CFP sensor mounting on a heat-sink. R3 is the convective cooling of the heat-sink, while R5 models heat losses from the sensor body itself

EF, this gives a constant long-term error due to the sustained temperature difference between the driver junction and heat-sink mass (see the upper traces in Figure 15.20, plotted for different bias tempcos). The CFP stimulus is a 0.5 W step, as before. This constant error cannot be properly dealt with by choosing a tempco that gives a bias error passing through a zero in the first 50 seconds, as was done for the EF case with a TO-3-top sensor, as the heat-sink thermal inertia causes it to pass through zero very quickly and head rapidly south in the direction of ever-increasing negative error. This is because it has allowed for thermal attenuation but has not decreased thermal delay. It is therefore pointless to compute an IAE for this configuration.

A Better Sensor Position

By analogy with the TO-3 and TO-3P transistor packages examined earlier, it will be found that driver packages such as TO-225AA on a heat-sink get hotter faster on their exposed plastic face than any other accessible point. It looks as if a faster response will result from putting the sensor on top of the driver rather than on the other side of the sink as usual. With the Redpoint SW38-1 heat-sink this is fairly easy as the spring-clips used to secure one plastic package will hold a stack of two TO-225AAs with only a little physical persuasion. A standard thermal pad is used between the top of the driver and the metal face of the sensor, giving the sandwich shown in Figure 15.17b. The thermal model is shown in Figure 15.19. This scheme greatly reduces both thermal attenuation and thermal delay (lower traces in Figure 15.20), giving an error that falls within a ±1 mV window after about 15.5 seconds, when the tempco is set to −3.8 mV/°C. The IAE computes to 52 mV, as shown in Figure 15.21, which demonstrates how the IAE criterion tends to grow without limit

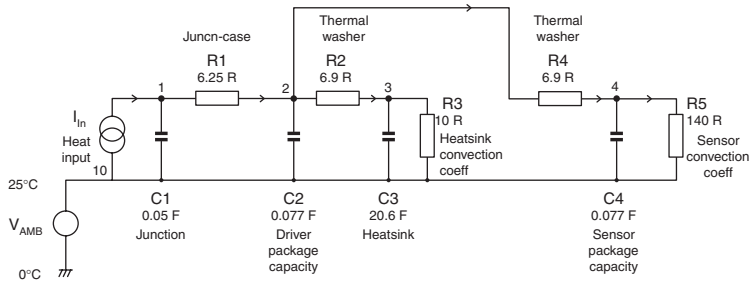


Figure 15.19: Thermal circuit of driver-back mounting of sensor. The large heat-sink time-constant R2–C2 is no longer in the direct thermal path to the sensor, so the compensation is faster and more accurate

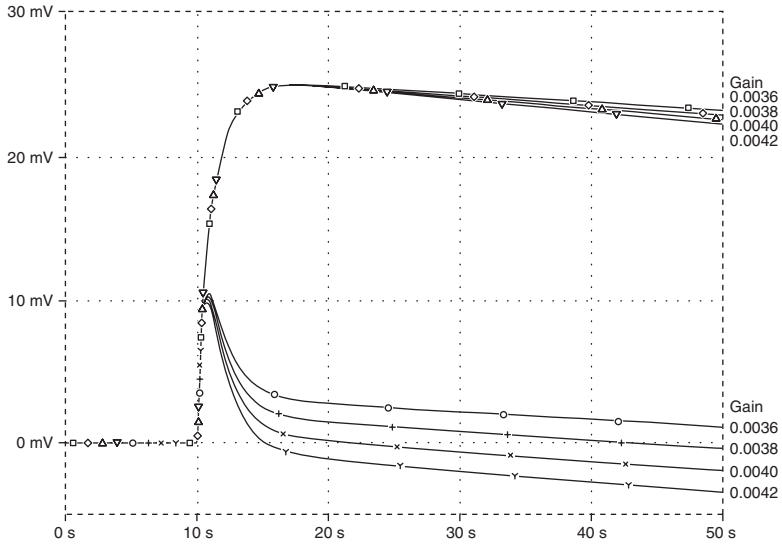


Figure 15.20: The V_q errors for normal and improved sensor mounting, with various tempcos. The improved method can have its tempco adjusted to give near-zero error over this timescale – not so for the usual method

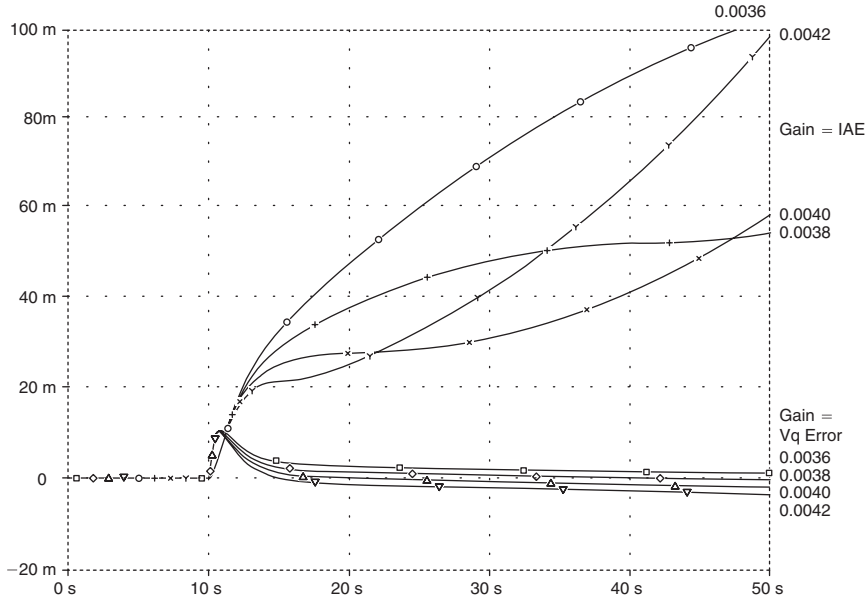


Figure 15.21: The V_q error and IAE for the improved sensor mounting method on driver back. Error is much smaller, due both to lower thermal attenuation and to less delay. The best IAE is 52 mV-s (with gain = 0.0038), twice as good as the best EF version

unless the error subsides to zero. This value is a distinct improvement on the 112 mV IAE, which is the best that could be got from the EF output.

The effective delay is much less because the long heat-sink time-constant is now partly decoupled from the bias compensation system.

A Junction-Temperature Estimator

It appears that we have reached the limit in what can be done, as it is hard to get one transistor closer to another than they are in Figure 15.17b. It is, however, possible to get better performance, not by moving the sensor position, but by using more of the available information to make a better estimate of the true driver junction temperature. Such 'estimator' subsystems are widely used in servo control systems where some vital variable is inaccessible, or only knowable after such a time delay as to render the data useless[9]. It is often almost as useful to have a model system, usually just an abstract set of gains and time-constants, which gives an estimate of what the current value of the unknown variable must be, or at any rate, *ought* to be.

The situation here is similar, and the first approach makes a better guess at the junction temperature V(1) by using the known temperature drop between the package and the heat-sink. The simplifying assumption is made that the driver package (not including the junction) is isothermal, so it is modeled by one temperature value V(2).

If two sensors are used, one placed on the heat-sink as usual and the other on top of the driver package, as described above (Figure 15.17c), then things get interesting. Looking at Figure 15.19,

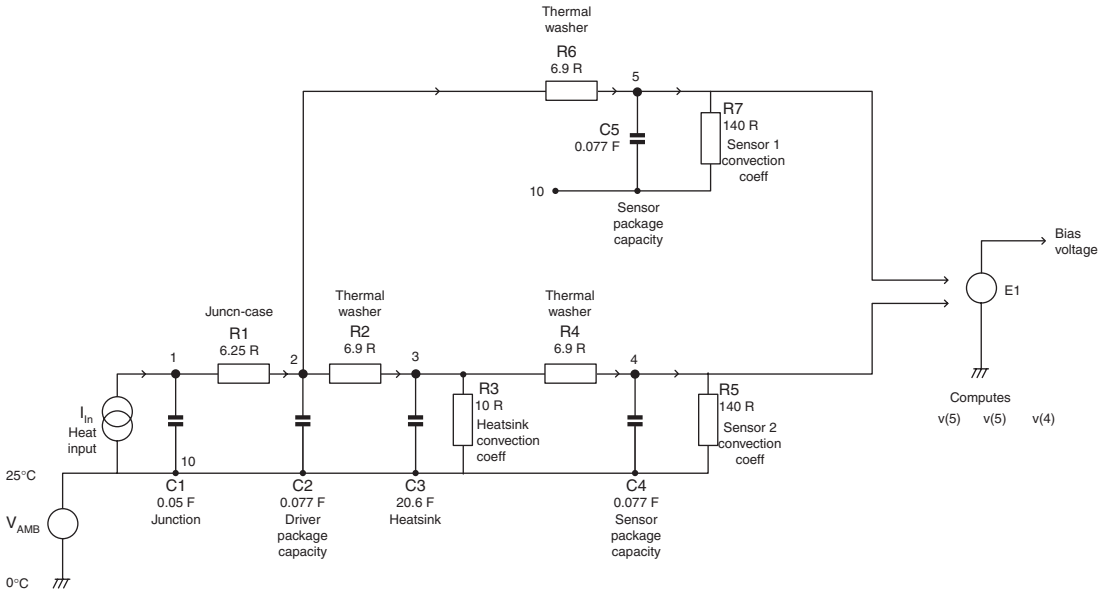


Figure 15.22: Conceptual diagram of the junction estimator. Controlled voltage source E1 acts as an analog computer performing the scaling and subtraction of the two sensor temperatures V(4) and V(5), to derive the bias voltage

it can be seen that the difference between the driver junction temperature and the heat-sink is due to R1 and R2; the value of R1 is known, but not the heat flow through it. Neglecting small incidental losses, the temperature drop through R1 is proportional to the drop through R2. Since C2 is much smaller than C3, this should remain reasonably true even if there are large thermal transients. Thus measuring the difference between V(2) and V(3) allows a reasonable estimate of the difference between V(1) and V(2); when this difference is added to the known V(2), we get a rather good estimation of the inaccessible V(1). This system is shown conceptually in Figure 15.22, which gives only the basic method of operation; the details of the real circuitry must wait until we have decided exactly what we want it to do.

We can only measure V(2) and V(3) by applying thermal sensors to them, as in Figure 15.17c, so we actually have as data the sensor temperatures V(4) and V(5). These are converted to bias voltage and subtracted, thus estimating the temperature drop across R1. The computation is done by voltage-controlled voltage source (VCVS) E1, which in PSPICE can have any equation assigned to define its behavior. Such definable VCVSs are very handy as little 'analog computers' that do calculations as part of the simulation model. The result is then multiplied by a scaling factor called *estgain*, which is incorporated into the defining equation for E1, and is adjusted to give the minimum error; in other words the variable-tempco bias approach is used to allow for the difference in resistance between R1 and R2.

The results are shown in Figure 15.23, where an *estgain* of 1.10 gives the minimum IAE of 25 mV-s. The transient error falls within a ± 1 mV window after about 5 seconds. This is a major improvement, at what promises to be little cost.

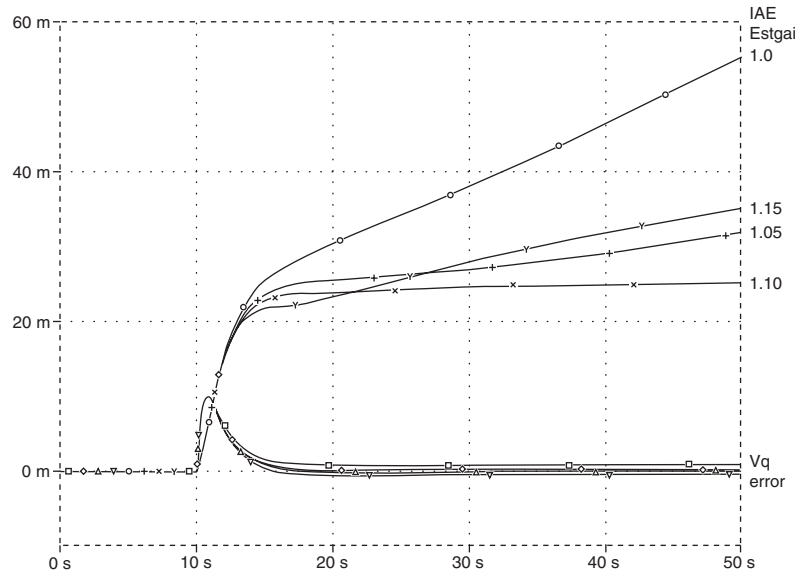


Figure 15.23: Simulation results for the junction estimator, for various values of *estgain*. The optimal IAE is halved to 25 mV-s (compare with Figure 15.21)

A Junction Estimator with Dynamics

The remaining problem with the junction-estimator scheme is still its relatively slow initial response; nothing can happen before heat flows through R6 into C5 in Figure 15.22. It will take even longer for C4 to respond, due to the inertia of C3, so we must find a way to speed up the dynamics of the junction estimator.

The first obvious possibility is the addition of phase advance to the forward bias-compensation path. This effectively gives a high gain initially, to get things moving, which decays back over a carefully set time to the original gain value that gave near-zero error over the 50 seconds timescale. The conceptual circuit in Figure 15.24 shows the phase-advance circuitry added to the compensation path; the signal is attenuated 100× by R50 and R51, and then scaled back up to the same level by VCVS E2, which is defined to give a gain of 110× incorporating estimated gain = 1.10. C causes fast changes to bypass the attenuation, and its value in conjunction with R50, R51 sets the degree of phase advance or lead. The slow behavior of the circuit is thus unchanged, but transients pass through C and are greatly amplified by comparison with steady-state signals.

The result on the initial error transient of varying C around its optimal value can be seen in the expanded view of Figure 15.25. The initial rise in V_q error is pulled down to less than a third of its value if C is made 10 μF; with a lower C value the initial peak is still larger than it need be, while a higher value introduces some serious undershoot that causes the IAE to rise again, as seen in the upper traces in Figure 15.26. The big difference between no phase advance and a situation where it is even approximately correct is very clear.

With C set to 10 μF, the transient error falls a ±1 mV window after only 0.6 s, which is more than 20 times faster than the first improved CFP version (sensor put on driver) and gives a nicely reduced IAE of 7.3 mV-s at 50 s. The real-life circuitry to do this has not been designed in detail,

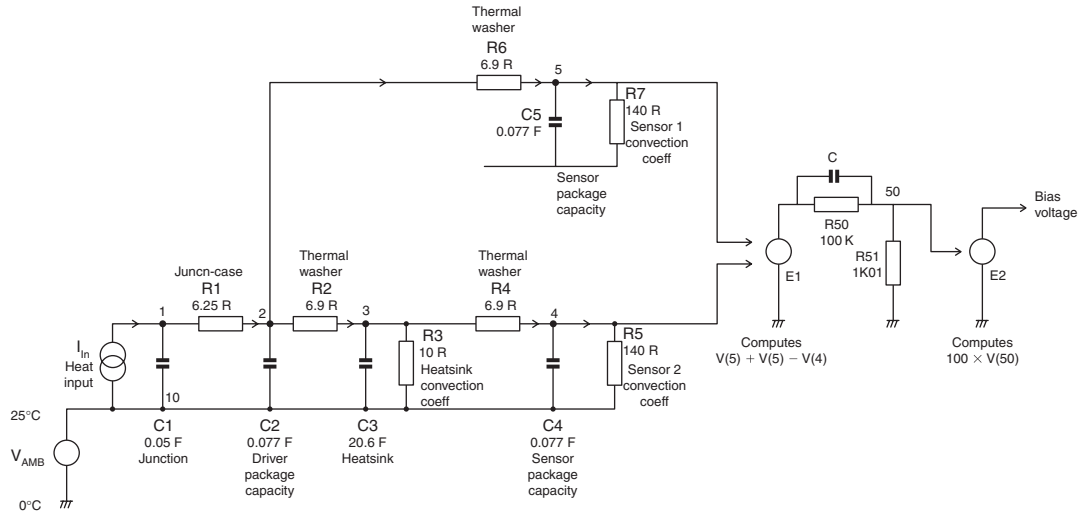


Figure 15.24: The conceptual circuit of a junction estimator with dynamics. C gives higher gain for fast thermal transients and greatly reduces the effects of delay

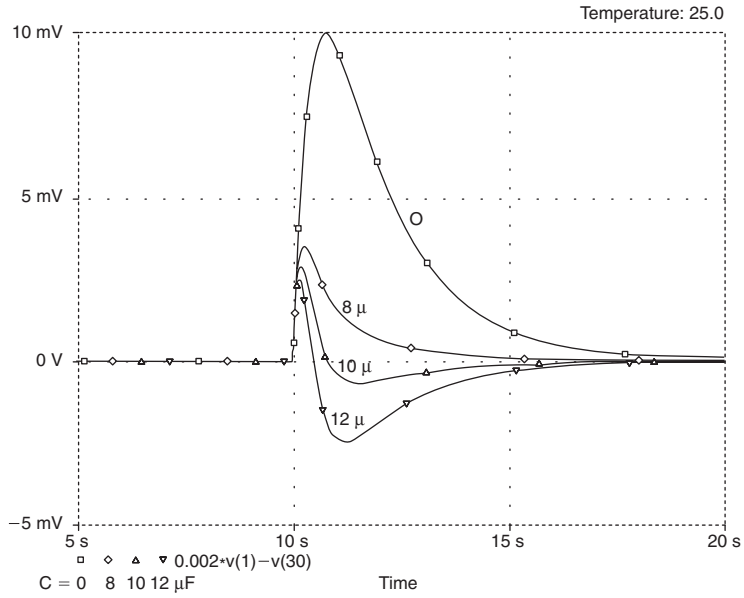


Figure 15.25: The initial transient errors for different values of C. Too high a value causes undershoot

but presents no obvious difficulties. The result should be the most accurately bias-compensated Class-B amplifier ever conceived.

Conclusions About the Simulations

Some of the results of these simulations and tests were rather unexpected. I thought that the CFP would show relatively smaller bias errors than the EF, but it is the EF that stays within its much

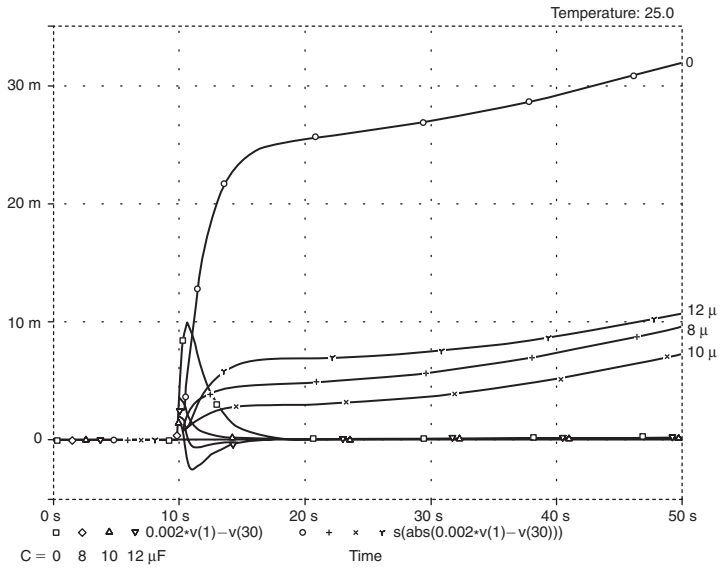


Figure 15.26: The IAE for different values of C; 10 µF is clearly best for minimum integrated error (IAE = 7.3 mV-s) but even a rough value is a great improvement

wider tolerance bands, with either heat-sink or TO-3-top mounted sensors. The thermal-gain effect in the EF stage seems to be the root cause of this, and this in turn is a consequence of the near-constant driver dissipation in the EF configuration.

However, the cumulative bias errors of the EF stage can only be reduced to a certain extent, as the system is never free from the influence of the main heat-sink with its substantial thermal inertia. In contrast the CFP stage gives much more freedom for sensor placement and gives scope for more sophisticated approaches that reduce the errors considerably.

Hopefully it is clear that it is no longer necessary to accept 'V_{be}-multiplier on the heat-sink' as the only option for the crucial task of V_{bias} compensation. The alternatives presented promise greatly superior compensation accuracy.

Power Transistors with Integral Temperature Sensors

For a very long time it was obvious that all attempts at estimating device junction temperature would come a poor second to having a sensor built right into the junction structure. At last such power transistors appeared when Sanken introduced the SAP series of transistors with integral sensing diodes. These were Darlington devices; these are usually not good for bias stability as the driver transistor is heated up by the adjacent output device, but in this case the integral diodes were intended to compensate both driver and output V_{be} changes. The SAP transistors had one diode built into the NPN part and five diodes built into the PNP part, designed so that 2.5 mA through the diodes gave good matching with a transistor quiescent I_c of 40 mA. They also had an integral 0R22 emitter resistor, which proved not to be a good idea as it was more electrically fragile than the transistors themselves. The SAP series has now been replaced by the STD03N and the STD03P, which have

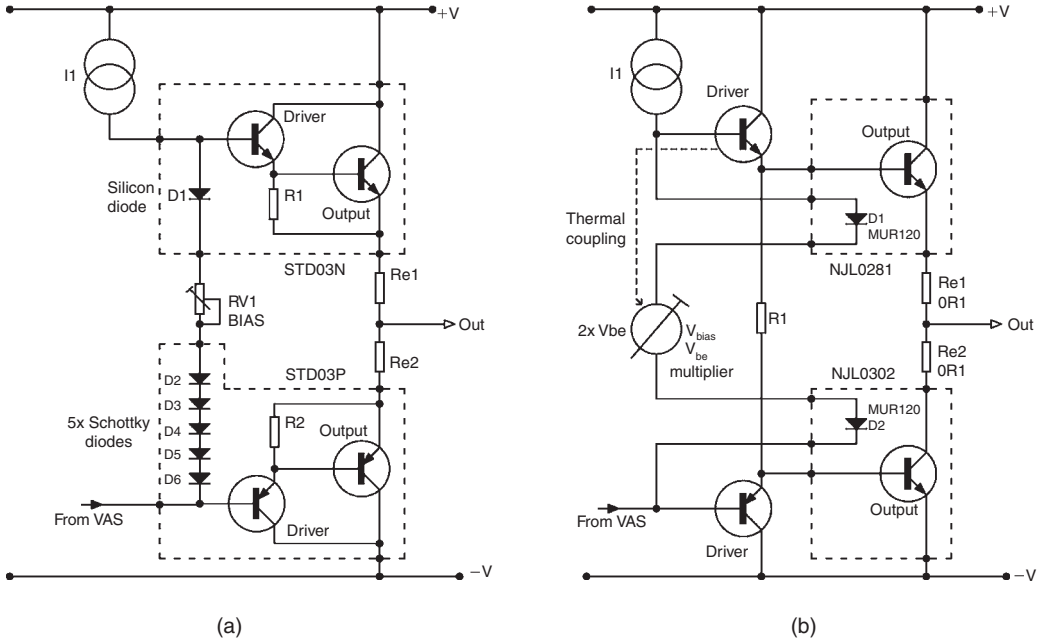


Figure 15.27: The internal biasing diodes of the Sanken STD03 (a) and the Motorola NJL (b) transistors

the same diode structure but no emitter resistor (see Figure 15.27a). A Darlington output stage needs four V_{be} drops, plus a few millivolts across the emitter resistors, so six diodes worth of bias may appear excessive; the answer is that the five diodes in the PNP device are Schottkys with a lower voltage drop. The diodes are part of the main transistor die and so have the fastest response possible. The main drawback of this approach is that it permits little flexibility in circuit design.

Quite recently ON Semiconductor (Motorola that was) introduced some very interesting output pairs with integral diodes, under the name ThermalTrak. They are the NJL4281 (NPN) and NJL4302 (PNP) pair, with a V_{ceo} of 350 V, and the NJL0281 (NPN) and NJL0302 (PNP) pair, with a V_{ceo} of 260 V. They are different from the Sanken devices in that a single silicon diode of the MUR120 type is mounted on the copper lead frame and is electrically isolated from the transistors. This gives much greater freedom of design, at the cost of a thermal response that is slower, but presumably still a lot faster than a sensor mounted on the top of the package. It has been pointed out that if thermal compensation is *too* fast, thermal distortion might be introduced as the bias changed during a cycle.

Bob Cordell has made some measurements[10] that indicate that the transistor V_{be} temperature coefficient is -2.14 mV/°C while the diode has -1.7 mV/°C, and so the best voltage matching occurs when the diode current is a quarter of the transistor I_c; so it would appear that applying these devices to get the best compensation is going to take a bit of thought. The drivers are now separate devices and will have their own temperature, so the internal diodes can only compensate the output transistors. One possible configuration is shown in Figure 15.27b, where a conventional V_{be}-multiplier

generating two V_{be} drops is thermally coupled to the drivers only; current source I1 will need to be set for the best diode-output-device matching. The Motorola devices have beautifully flat β/I_c curves and so should give low large-signal distortion (see Chapter 6).

Just as this book was going to press I was able to procure some NJL0281 and NJL0302, and I have done some tests, which are described at the end of this chapter.

Variable-Tempco Bias Generators

The standard V_{be}-multiplier bias generator has a temperature coefficient that is fixed by the multiplication factor used, and so ultimately by the value of V_{bias} required. At many points in this chapter it has been assumed that it is possible to make a bias generator with an arbitrary temperature coefficient. This section shows how to do it.

Figure 15.28 shows two versions of the usual V_{be}-multiplier bias generator. Here the lower rails are shown as grounded to simplify the results. The first version in Figure 15.28a is designed for an EF output stage, where the voltage V_{bias} to be generated is $(4 \times V_{be}) + V_q$, which totals +2.93 V. Recall that V_q is the small quiescent voltage across the emitter resistors Re; it is *this* quantity we are aiming to keep constant, rather than the quiescent current, as is usually assumed. The optimal V_q for an EF stage is in the region of 50 mV.

The second bias generator in Figure 15.28b is intended for a CFP output stage, for which the required V_{bias} is less at $(2 \times V_{be}) + V_q$, or approximately 1.30 V in total. Note that the optimal V_q is also much smaller for the CFP type of output stage, being about 5 mV.

It is assumed that V_{bias} is trimmed by varying R2, which will in practice be a preset with a series end-stop resistor to limit the maximum V_{bias} setting. It is important that this is the case, because a preset normally fails by the wiper becoming disconnected, and if it is in the R2 position the bias will default to minimum. In the R1 position an open-circuit preset will give maximum bias, which may blow fuses or damage the output stage. The adjustment range provided should be no greater than that required to

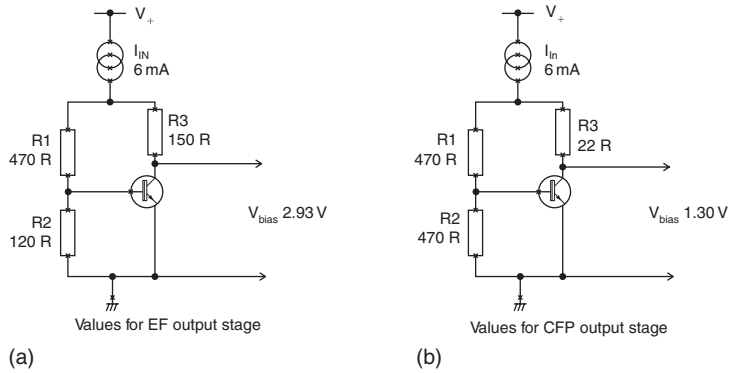


Figure 15.28: The classical V_{be}-multiplier bias generator. Two versions are shown: for biasing EF (a) and CFP (b) output stages. The EF requires more than twice the bias voltage for optimal crossover performance

take up production tolerances; it is, however, hard to predict just how big that will be, so the range is normally made wide for pre-production manufacture and then tightened in the light of experience.

The EF version of the bias generator has a higher V_{bias}, so there is a larger V_{be}-multiplication factor to generate it. This is reflected in the higher temperature coefficient (hereafter shortened to 'tempco') – see Table 15.4.

Creating a Higher Tempco

A higher (i.e. more negative) tempco than normal may be useful to compensate for the inability to sense the actual output junction temperatures. Often the thermal losses to the temperature sensor are the major source of steady-state V_{bias} error, and to reduce this a tempco is required that is larger than the standard value given by 'V_{be}-multiplication factor times −2 mV/°C'. Many approaches are possible, but the problem is complicated because in the CFP case the bias generator has to work within two rails only 1.3 V apart. Additional circuitry outside this voltage band can be accommodated by bootstrapping, as in the Trimodal amplifier biasing system in Chapter 10, but this does add to the component count.

A simple new idea is shown in Figure 15.29. The aim is to increase the multiplication factor (and hence the negative tempco) required to give the same V_{bias}. The diagram shows a voltage source V1 inserted in the R2 arm. To keep V_{bias} the same, R2 is reduced. Since the multiplication factor

Table 15.4: Temperature coefficients for different V_{bias} voltages

	V_{bias} (V)	R1 (Ω)	R2 (Ω)	R3 (Ω)	Tempco (mV/°C)
EF	2.93	120R	470R	22R	−9.3
CFP	1.30	470R	470R	150R	−3.6

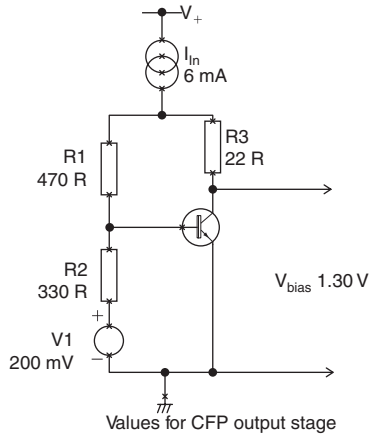


Figure 15.29: Principle of a V_{be}-multiplier with increased tempco. Adding voltage source V1 means the voltage multiplication factor must be increased to get the same V_{bias}. The tempco is therefore also increased, here to −4.4 mV/°C

Table 15.5: Creating a higher tempco by varying V1

V1 (mV)	V_{bias} (V)	R2 (Ω)	Tempco (mV/°C)
0	1.287	470	−3.6
100	1.304	390	−4.0
200	1.287	330	−4.4
300	1.286	260	−5.0
400	1.285	190	−6.9

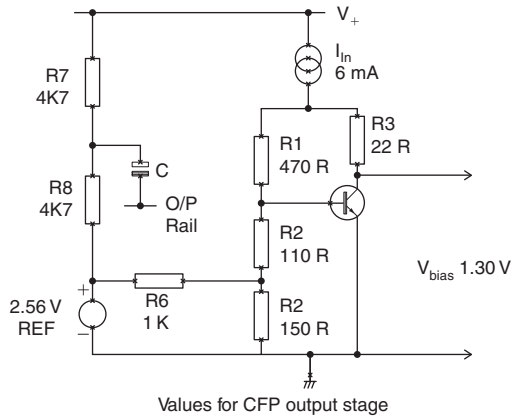


Figure 15.30: A practical version of a V_{be}-multiplier with increased tempco. The extra voltage source is derived from the band-gap reference by R6, R4. Tempco is increased to −5.3 mV/°C

(R1 + R2)/R2 is increased, the tempco is similarly increased. In Table 15.5, a CFP bias circuit has its tempco varied by increasing V1 in 100 mV steps; in each case the value of R2 is then reduced to bring V_{bias} back to the desired value, and the tempco is increased.

A practical circuit is shown in Figure 15.30, using a 2.56 V band-gap reference to generate the extra voltage across R4. This reference has to work outside the bias generator rails, so its power-feed resistors R7, R8 are bootstrapped by C from the amplifier output, as in the Trimodal amplifier design.

Ambient Temperature Changes

Power amplifiers must be reasonably immune to ambient temperature changes, as well as changes due to dissipation in power devices. The standard compensation system deals with this pretty well, as the V_{be}-multiplication factor is inherently almost the same as the number of junctions being biased. This is no longer true if the tempco is significantly modified. Ideally we require a bias generator that has one increased tempco for power-device temperature changes only, and another standard tempco for ambient changes affecting all components. One approach to this is given in Figure 15.31, where V1 is derived via R6, R4 from a silicon diode rather than a band-gap reference, giving a voltage reducing with temperature. The tempco for temperature changes to Q1 only is −4.0 mV/°C, while the tempco for global temperature changes to *both* Q1 and D1 is lower at −3.3 mV/°C. Ambient temperatures vary much less than output device junction temperatures, which may easily range over 100°C.

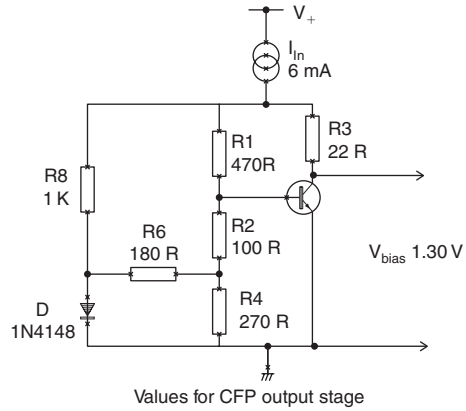


Figure 15.31: Practical V_{be}-multiplier with increased tempco, and also improved correction for ambient temperature changes, by using diode D to derive the extra voltage

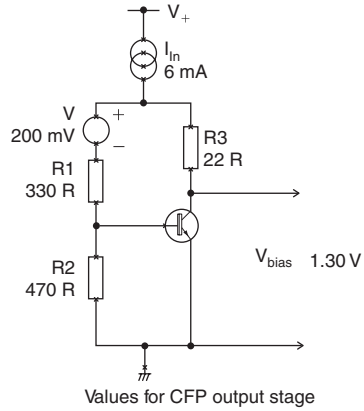


Figure 15.32: The principle of a V_{be}-multiplier with reduced tempco. The values shown give −3.1 mV/°C

Creating a Lower Tempco

Earlier in this chapter I showed that an EF output stage has 'thermal gain' in that the thermal changes in V_q make it appear that the tempco of the V_{bias} generator is higher than it really is. This is because the bias generator is set up to compensate for four base–emitter junctions, but in the EF output configuration the drivers have a roughly constant power dissipation with changing output power, and therefore do not change much in junction temperature. The full effect of the higher tempco is thus felt by the output junctions, and if the sensor is placed on the power device itself rather than the main heat-sink, to reduce thermal delay, then the amplifier can be seriously overcompensated for temperature. In other words, after a burst of power V_q will become too low rather than too high, and crossover distortion will appear. We now need a V_{bias} generator with a *lower* tempco than the standard circuit.

The principle is exactly analogous to the method of increasing the tempco. In Figure 15.32, a voltage source is inserted in the upper leg of potential divider R1, R2; the required V_{be}-multiplication factor for the same V_{bias} is reduced, and so therefore is the tempco.

Table 15.6: Creating a lower tempco by varying V1

V1 (mV)	V_{bias} (V)	R1 (Ω)	Tempco (mV/°C)
0	1.287	470	−3.6
100	1.304	390	−3.3
200	1.287	330	−3.1
300	1.286	260	−2.8
400	1.285	190	−2.5

Table 15.6 shows how this works as V1 is increased in 100 mV steps. R1 has been varied to keep V_{bias} constant, in order to demonstrate the symmetry of resistor values with Table 15.5; in reality R2 would be the variable element, for the safety reasons described above.

Current Compensation

Both bias generators in Figure 15.28 are fitted with a current-compensation resistor R3. The V_{be}-multiplier is a very simple shunt regulator, with low loop gain, and hence shows a significant series resistance. In other words, the V_{bias} generated shows unwanted variations in voltage with changes in the standing current through it. R3 is added to give first-order cancelation of V_{bias} variations caused by these current changes. It subtracts a correction voltage proportional to this current. Rather than complete cancelation, this gives a peaking of the output voltage at a specified current, so that current changes around this peak value cause only minor voltage variations. This peaking philosophy is widely used in IC bias circuitry.

R3 should never be omitted, as without it mains voltage fluctuations can seriously affect V_q. Table 15.4 shows that the optimal value for peaking at 6 mA depends strongly on the V_{be}-multiplication factor.

Figure 7.18 in Chapter 7 demonstrates the application of this method to the Class-B amplifier. The graph in Figure 7.19 shows the variation of V_{bias} with current for different values of R3. The slope of the uncompensated (R3 = 0) curve at 6 mA is approximately 20 Ω, and this linear term is canceled by setting R14 to 18 Ω in Figure 7.18.

The current through the bias generator will vary because the VAS current source is not a perfect circuit element. Biasing this current source with the usual pair of silicon diodes does not make it wholly immune to supply-rail variations. I measured a generic amplifier (essentially the original Class-B Blameless design) and varied the incoming mains from 212 to 263 V, a range of 20%. This, in these uncertain times, is perfectly plausible for a power amplifier traveling around Europe. The VAS current-source output varied from 9.38 to 10.12 mA, which is a 7.3% range. Thanks to the current-compensating resistor in the bias generator, the resulting change in quiescent voltage V_q across the two Re resistors is only from 1.1 mV (264 V mains) to 1.5 mV (212 V mains). This is a very small absolute change of 0.4 mV, and within the V_q tolerance bands. The ratio of change is greater, because V_{bias} has had a large fixed quantity (the device V_{be} values) subtracted from it, so the residue varies much more. V_q variation could be further suppressed by making the VAS current source more stable against supply variations.

The finite ability of even the current-compensated bias generator to cope with changing standing current makes a bootstrapped VAS collector load much less attractive than the current-source version; from the above data, it appears that V_q variations will be at least three times greater.

A quite different approach reduces V_{bias} variations by increasing the loop gain in the V_{be}-multiplier. Figure 15.33 shows the circuit of a two-transistor version that reduces the basic resistance slope from 20 to 1.7 Ω. The first transistor is the sensor. An advantage is that V_{bias} variations will be smaller for all values of VAS current, and no optimization of a resistor value is required.
A drawback is slightly greater complexity in an area where reliability is vital. Figure 15.34

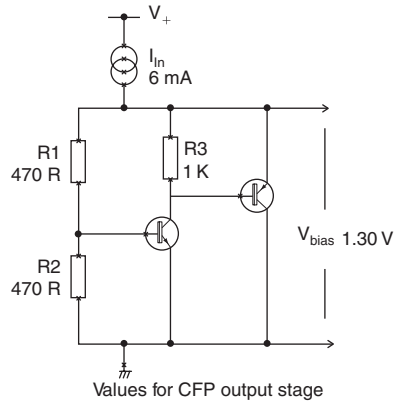


Figure 15.33: Circuit of a two-transistor V_{be}-multiplier. The increased loop gain holds V_{bias} more constant against current changes

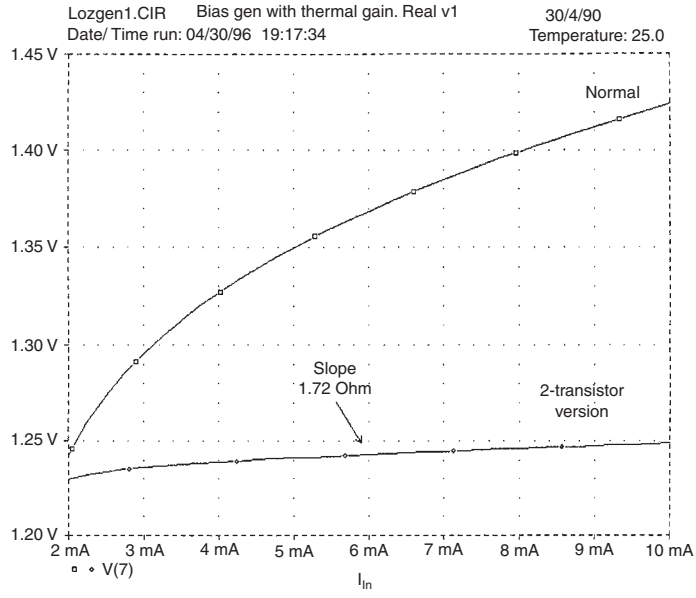


Figure 15.34: The two-transistor configuration gives a consistently lower series resistance, and hence V_{bias} variation with current, compared with the standard version without R3

compares the two-transistor configuration with the standard version (without R3). Multi-transistor feedback loops raise the possibility of instability and oscillation, and this must be carefully guarded against, as it is unlikely to improve amplifier reliability.

This section of the thermal dynamics chapter describes simple V_{bias} generators with tempcos ranging from −2.5 to −6.9 mV/°C. It is hoped that this, in combination with the techniques described earlier, will enable the design of Class-B amplifiers with greater bias accuracy, and therefore less afflicted by crossover distortion. However, there is another factor that causes quiescent conditions to vary, and which must be considered when setting the current compensation. This is dealt with in the next section.

Early Effect in Output Stages

There is another factor that affects the accuracy with which quiescent conditions can be maintained. If you take a typical power amplifier (with an unregulated power supply) and power it from a variable-voltage transformer, you are very likely to find that V_q varies with the mains voltage applied. This at first seems to indicate that the apparently straightforward business of compensating the bias generator for changes in standing current has fallen somewhat short of success. However, even if this compensation appears to be correct, and the constant-current source feeding the bias generator and VAS is made absolutely stable, the quiescent conditions are still likely to vary. At first this seems utterly mysterious, but the true reason is that the transistors in the output stage are reacting directly to the change in their collector-emitter voltage (V_{ce}). As V_{ce} increases, so does the V_q and the quiescent current. This is called the Early effect. It is a narrowing of the base-collector region as V_{ce} increases, which will cause an increase in the collector current I_c even if V_{be} and I_b are held constant. In a practical EF output stage the result is a significant variation in quiescent conditions when the supply voltage is varied over a range such as ±10%.

Table 15.7 shows the effect as demonstrated by SPICE simulation, using MJE340/50 for drivers and MJ15022/23 as output devices, with fixed bias voltage of 2.550 V, which gave optimal crossover in this case. It is immediately obvious that (as usual) things are more complicated than they at first appear. The V_q increases with rail voltage, which matches reality. However, the way in which this occurs is rather unexpected. The V_{be} values of the drivers Q1 and Q2 reduce with increasing V_{ce} as expected. However, the output devices Q3 and Q4 show a V_{be} that increases – but

Table 15.7: SPICE V_{be} changes with supply-rail voltage (MJE340/50 and MJ15022/3) – all devices held at 25°C

±Rail (V)	V_q (mV)	Q1 V_{be} (mV)	Q3 V_{be} (mV)	Q2 V_{be} (mV)	Q4 V_{be} (mV)	Sum (V)
10	7.8	609	633	654	646	2.550
20	13	602	640	647	648	2.550
30	18	597	643	641	649	2.550
40	23	593	647	637	650	2.550
50	28	589	649	634	650	2.550

Table 15.8: Real V_{be} changes with supply-rail voltage (2SC4382, 2SA1668 drivers and 2SC2922, 2SA1216 output)

±Rail (V)	V_q (mV)	Q1 V_{be} (mV)	Q3 V_{be} (mV)	Q2 V_{be} (mV)	Q4 V_{be} (mV)	Sum (V)
40	1.0	554	568	541	537	2.201
45	1.0	544	556	533	542	2.176
50	1.0	534	563	538	536	2.172
55	1.0	533	549	538	540	2.161
60	1.0	527	552	536	535	2.151
65	1.0	525	540	536	539	2.141
70	1.0	517	539	537	539	2.133

by a lesser amount, so that after subtracting all the V_{be} drops from the fixed bias voltage the aggregate effect is that V_q, and hence quiescent current I_q, both increase. Note that the various voltages have been summed as a check that they really do add up to 2.550 V in each case.

Table 15.8 has the results of real V_{be} measurements. These are not easy to do, because any increase in I_q increases the heating in the various transistors, which will cause their V_{be} values to drift. This happens to such an extent that sensible measurements are impossible. The measurement technique was therefore slightly altered. The amplifier was powered up on the minimum rail voltage, with its V_q set to 1.0 mV only. This is far too low for good linearity, but minimizes heating while at the same time ensuring that the output devices are actually conducting. The various voltages were measured, the rail voltage increased by 5 V, and then the bias control turned down as quickly as possible to get V_q back to 1.0 mV, and the process is repeated. The results are inevitably less tidy as the real V_{be} values are prone to wander around by a millivolt or so, but it is clear that in reality, as in SPICE, most of the Early effect is in the drivers, and there is a general reduction in aggregate V_{be} as rail voltage increases. The sum of V_{be} values is no longer constant as V_q has been constrained to be constant instead.

It may seem at this point as if the whole business of quiescent control is just too hopelessly complicated – not so. The cure for the Early effect problem is to overcompensate for VAS standing-current changes, by making the value of resistor R3 described above larger than usual. The best and probably the only practical way to find the right value is the empirical method. Wind the HT up and down on the prototype design with a variable transformer and adjust the value of R3 until the V_q change is at a minimum. (Unfortunately this interacts with the bias setting, so there is a bit of twiddling to do – however, for a given design you only need to find the optimal value for R3 once.) The resistance value will be a good deal larger than that required to merely compensate for changes in the output of the VAS current source; in one design, with a negative feedback biased current source, R3 = 16 Ω gave optimal rejection of current-source changes, but 100 Ω was the value required to give minimal change in V_q as the variable transformer was wound up and down over a mains range from 80% to 110%. The results (for a different amplifier, but with the same output stage configuration) are shown in Table 15.9. It is a good question as to how much this effect will vary between different specimens of the same output transistor type – right now I have no answers on that.

Table 15.9: Changes in *Vq* with mains voltage

Mains voltage (%)	V_q (mV) with R3 = 16 Ω	V_q (mV) with R3 = 100 Ω
110	18.2	14.6
100	14.4	14.4
90	10.6	14.3
80	8.0	14.2

If R3 is as high as 100 Ω, the extra voltage drop across R3 will be between 600 mV and 1 V, depending on the VAS standing current, and this may reduce the positive output swing slightly. This simple method assumes that the supply-rail rejection of the VAS current source and its biasing circuitry is predictable and stable; with the circuits normally used this seems to be the case, but some further study in this area is required. A potential problem is that the current-source biasing circuitry is likely to include RC filtering to prevent rail ripple getting in, and this could introduce a delay so that rapid mains variations are not properly compensated.

Thermal Dynamics by Experiment

One of the main difficulties in the study of amplifier thermal dynamics is that some of the crucial quantities, such as transistor junction temperatures, are not directly measurable; this is why simulation is so important in this field. However, some insight into the way that bias conditions are altering can be obtained by observing changes in the THD residual as viewed on a scope or recorded against time. This does not of course tell you anything about how the various contributions to the bias state are varying – you just get the single result as a THD figure.

At the end of the day, what really matters is the crossover distortion produced by the output stage, and measuring this gets to the heart of the matter. One method I have used with success works as follows. The amplifier under study is deliberately underbiased by a modest amount. I choose a bias setting that gives about 0.02% THD with a peak responding measurement mode. This creates crossover spikes that are clear of the rest of the THD residual, to ensure the analyzer is reading these spikes and ignoring noise and other distortions at a lower level. The AP System-1 has a mode that plots a quantity against time (it has to be said that the way to do this is not at all obvious from the AP screen menus – essentially 'time' is treated as an external stimulus – but it *is* in the manual) and this effectively gives that most desirable of plots: crossover conditions against time. In both cases below the amplifier was turned on with the input signal already present, so that the power dissipation stabilized within a second or so.

One limitation is that it appears to take the AP System-1 about a second to make a THD reading, and this limits the time resolution.

Crossover Distortion Against Time – Some Results

The first test amplifier examined has a standard EF output stage. The drivers have their own small heat-sinks and have no thermal coupling with the main output device heat-sink. The most important

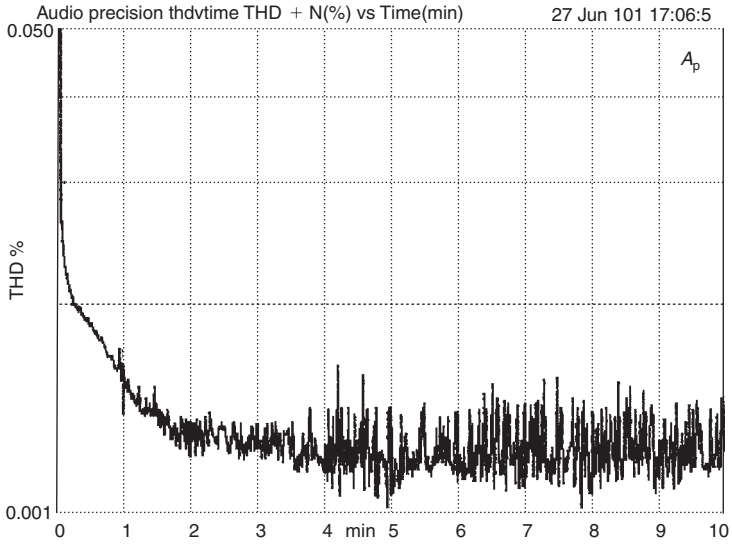


Figure 15.35: Peak THD versus time over 10 minutes

feature is that the bias sensor transistor is not mounted on the main heat-sink, as is usual, but on the back of one of the output devices, as I recommended above. This puts the bias sensor much closer thermally to the output device junction. A significant feature of this test amplifier is its relatively high supply rails. This means that even under no load, there is a drift in the bias conditions due to the drivers heating up to their working temperature. This drift can be reduced by increasing the size of the driver heat-sinks, but not eliminated. Figure 15.35 shows the THD plot taken over 10 minutes, starting from cold and initiating some serious power dissipation at $t = 0$. The crossover distortion drops at once; Figure 15.1 at the start of this chapter shows that driver dissipation is not much affected by output level, so this appears to be due to the output device junctions heating up and increasing V_q. There is then a slower reduction until the THD reading stabilizes at about 3 minutes.

The second amplifier structure examined is more complex. It is a triple-EF design with drivers and output devices mounted on a large heat-sink with considerable thermal inertia. The pre-drivers are TO220 devices mounted separately without heat-sinks. It may seem perverse to mount the drivers on the same heat-sink as the outputs, because some of the time they are being heated up rather than cooled down, which is exactly the opposite of what is required to minimize V_{be} changes. However, they need a heat-sink of some sort, and given the mechanical complications of providing a separate thermally isolated heat-sink just for the drivers, they usually end up on the main heat-sink. All that can be done (as in this case) is to mount them in the heat-sink in the area that stays coolest in operation. Once more the bias sensor transistor is not mounted on the main heat-sink, but on the back of one of the output devices (see Figure 15.36 for the electrical circuit and thermal coupling paths).

The results are quite different. Figure 15.37 shows at A the THD plot taken over 10 minutes, again starting from cold and initiating dissipation at $t = 0$. Initially THD falls rapidly, as before, as the output device junctions heat. It then commences a slow rise over 2 minutes, indicative of falling bias, and this represents the time lag in heating the sensor transistor. After this there is a much

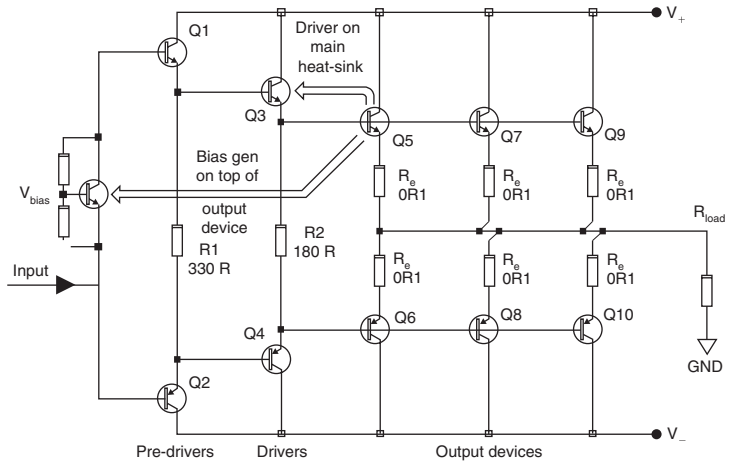


Figure 15.36: Circuit and thermal paths of the triple-EF output stage

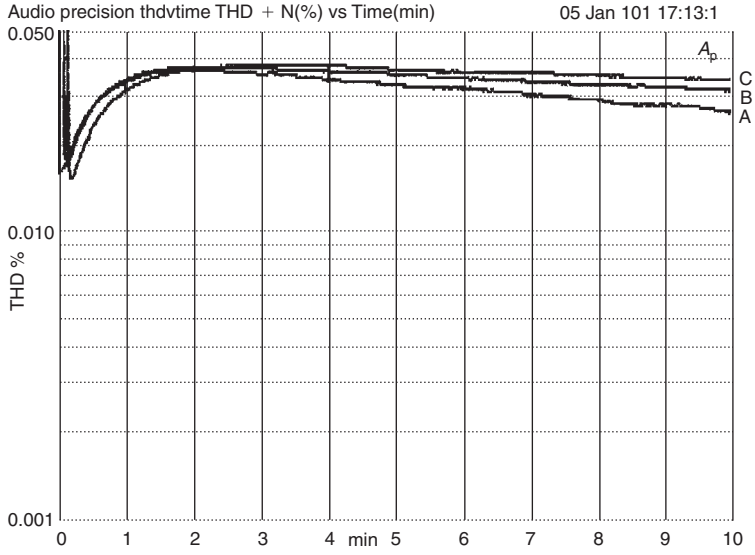


Figure 15.37: Peak THD versus time over 10 seconds

slower drift downwards, at about the same rate as the main heat-sink is warming up. There are clearly at least three mechanisms operating with very different time-constants. The final time-constant is very long, and the immediate suspicion is that it must be related to the slow warming of the main heat-sink. Nothing else appears to be changing over this sort of timescale. In fact this long-term increase in bias is caused by cooling of the bias sensor compared with the output device it is mounted on. This effect was theoretically predicted above, and it is pleasing to see that it really exists, although it does nothing but further complicate the quest for optimal Class-B operation. As the main heat-sink gets hotter, the heat losses from the sensor become more significant, and its temperature is lower than it should be. Therefore the bias voltage generated is too high, and this effect grows over time as the heat-sink warms up.

Knowledge of how the long-term drift occurs leads at once to a strategy for reducing it. Adding thermal insulation to cover the sensor transistor, in the form of a simple pad of plastic foam, gives plot B, with the long-term variation reduced. Plot C reduces it still further by more elaborate insulation: a rectangular block of foam with a cut-out for the sensor transistor. This is about as far as it is possible to go with sensor insulation; the long-term variation is reduced to about 40% of what it was. While this technique certainly appears to improve bias control, bear in mind that it is being tested with a steady sine wave. Music is noted for not being at the same level all the time, and its variations are much faster than the slow effect we are examining. It is very doubtful if elaborate efforts to reduce sensor cooling are worthwhile. I must admit this is the first time I have applied thermal lagging to an amplifier output stage.

More Measurements – Conventional and ThermalTrak

I recently revisited this technique to investigate the ThermalTrak transistors described earlier in this chapter. I used the NJL0281 (NPN) and NJL0302 (PNP) pair. The test amplifier was essentially the Load-Invariant design with the output stage converted from CFP to EF Type II. The maximum power was 20 W/8R but the amplifier was run at 8.3 W output to increase the dissipation; it is always a good idea to keep the power level of experimental amplifiers low if possible as they are much more tolerant of misplaced probes and slipping screwdrivers.

The main heat-sink was deliberately small to speed up heating and cooling; it got warm but not hot in 200 seconds – measuring it with the Mk1 fingertip, I would say about 40°C. The driver heat-sinks were made large compared with the power level, to minimize driver heating and keep the thermal situation simpler. They stayed at ambient temperature. The first three tests used conventional compensation with a simple V_{be}-multiplier to establish some basis for comparisons. The ThermalTrak diodes were not used.

The first test, shown in Figure 15.38, had the V_{be}-multiplier sensor transistor mounted on top of one of the drivers. Because these stayed cold the amplifier was definitely undercompensated, with the bias level steadily increasing, and therefore the crossover distortion steadily decreasing. Note the initial transient lasting about 20 seconds. I assume – but I do *not* know for certain at this point – that this is due to output device junction heating, while a temperature gradient junction-case heat-sink is being established.

Next the sensor was moved to a conventional position on the main heat-sink, about 2 cm from one of the output devices (see Figure 15.39). The long-term result is much better, but somewhat overcompensated as the sensor V_{be} is being multiplied 4 times but the drivers are staying cold. The bias therefore slowly decreases over 200 seconds. The initial transient appears unchanged.

In the third test the sensor was mounted on top of one of the output devices, with a silicone washer between them (see Figure 15.40). This is the sensor position that has been recommended earlier in this chapter. The initial transient is now faster (10 rather than 20 seconds) but no smaller, which is what might be expected. The long-term compensation is also good, though that is more luck than judgment – the amount of thermal coupling to the sensor happens to be about right. This is intended to represent the best that conventional compensation can do.

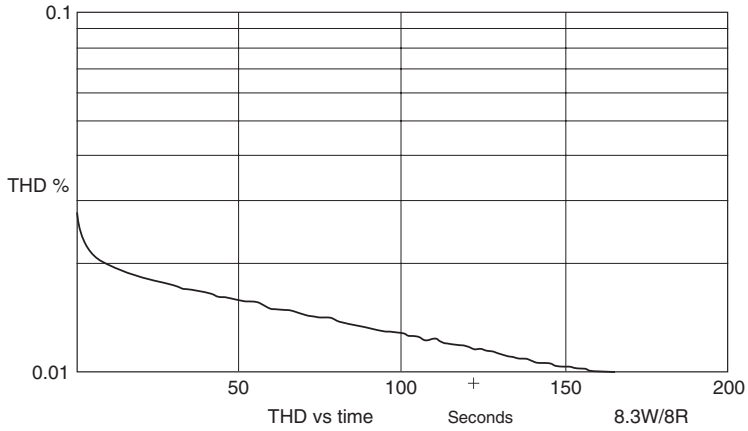


Figure 15.38: Sensor transistor mounted on the driver heat-sink – undercompensated so crossover distortion decreases as bias increases with time

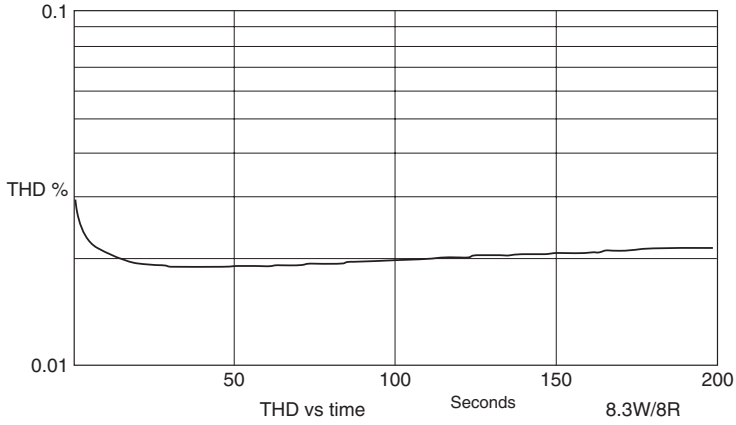


Figure 15.39: Sensor transistor mounted on the main heat-sink, somewhat overcompensated, so bias decreases with time

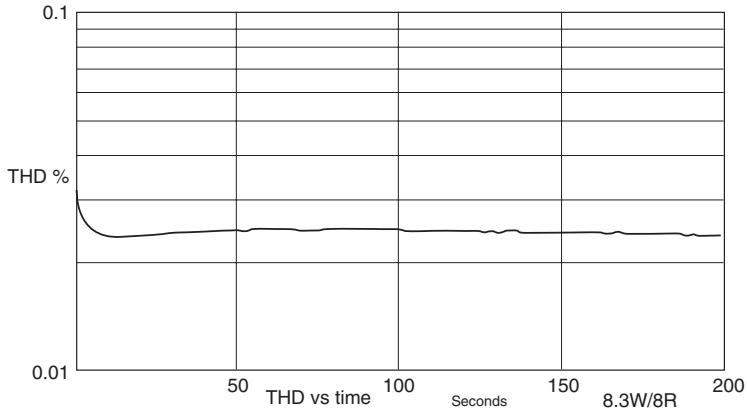


Figure 15.40: Sensor transistor mounted on top of one of the output devices – good compensation over 200 seconds

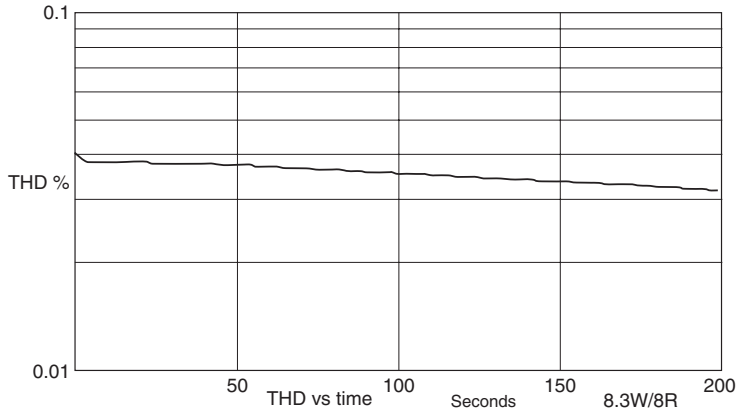


Figure 15.41: ThermalTrak diodes in series with the sensor V_{be}-multiplier – clearly somewhat undercompensated

Now we make use of the ThermalTrak diodes for the first time (Figure 15.41). They were simply put in series with the sensor V_{be}-multiplier, as shown in Figure 15.27b above. The V_{be}-multiplier was turned down in voltage as it is now only compensating for the driver V_{be} values, and the sensor was moved back to one of the driver heat-sinks. Everything there remained cold, so we should be able to see clearly how the ThermalTrak diodes compensate their associated transistors. The VAS current was 9.2 mA.

The initial transient is now both smaller and faster, but over 200 seconds the amplifier is somewhat undercompensated. This was expected, because as stated earlier in this chapter, it appears that the transistor V_{be} temperature coefficient is -2.14 mV/°C while the diode has -1.7 mV/°C, at a current of 25 mA[10]. The initial transient is a lot faster.

In the ThermalTrak discussions on the diyAudio Forums, several people suggested that the discrepancy in temperature coefficients could be corrected by running the diodes at a much lower current; diode tempcos increase slowly as current is reduced (by approximately 0.2 mV/°C per decade decrease of current). The diode tempco needs to be increased by 0.44 mV/°C, so the diode current must be reduced by 2.2 decades, or 158 times. That gives us a diode current of only 158 μA (no, not a typo – that's just how the numbers work out). This is much too low for a VAS operating current so the diodes were moved to a circuit that replicated the diode voltage at much greater current in the VAS collector. The ThermalTrak diodes were actually fed with 800 μA, rather than 158 μA, for practical reasons, so we do not expect perfect compensation, but it should be much better. This is where theory and practice diverge, because the results seen in Figure 15.42 are better, but only slightly so; the THD change over 200 seconds is 0.006% rather than the 0.008% in the previous test.

For the next test I decided to radically increase the ThermalTrak diode tempco rather than just tweak it. The diodes are once more passing about 9.2 mA but their V_f is now multiplied by a factor of 2. The long-term compensation as seen in Figure 15.43 is now quite good, though our theory says it should be seriously overcompensated as 2×1.7 mV/°C $= 3.4$ mV/°C, much more

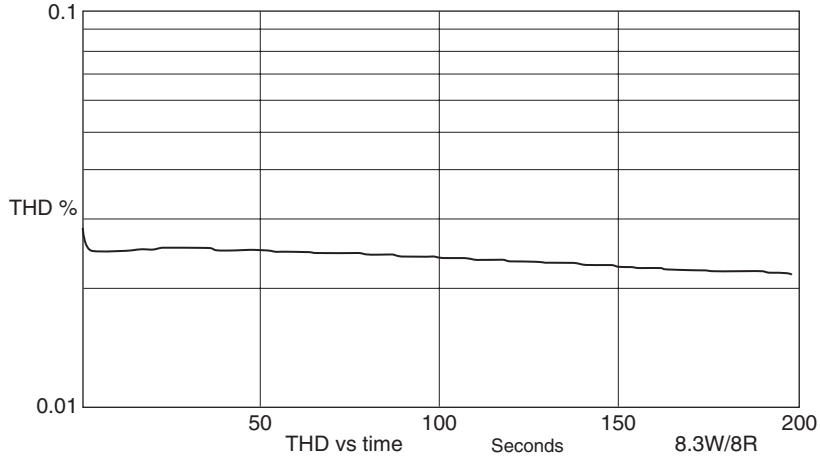


Figure 15.42: ThermalTrak diodes running at 800 μA to increase their tempco – a puzzlingly small improvement, and still definitely undercompensated

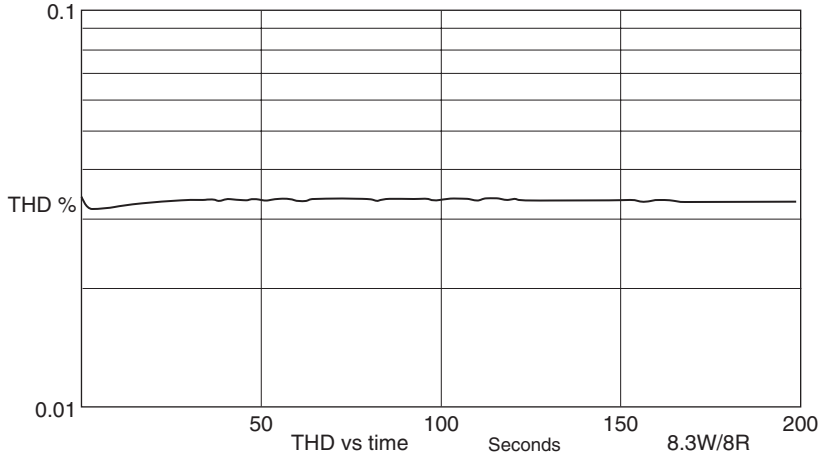


Figure 15.43: ThermalTrak sensor transistor mounted on top of one of the output devices

than 2.14 mV/°C. There is also a rather worrying dip in the first 20 seconds – if the diodes are rapidly following the transistor temperatures, then where does that come from? Clearly we are some way from having all the answers.

The final two tests were intended to investigate the greater speed of compensation response that the internal diodes should give. Figure 15.44 zooms in on the first 20 seconds of the response for conventional compensation with the sensor on top of the output device as shown in Figure 15.40 above. Figure 15.45 similarly shows the first 20 seconds of the first ThermalTrak experiment as in Figure 15.41. It is clear that the latter has a faster initial transient. For conventional compensation the time-constant (time to reach 63% of final value) is about 2.5 seconds, while for the ThermalTrak case it is about 0.5 seconds. I must admit I was expecting the ThermalTrak response to be faster than that, and I suspect there is more here than meets the eye.

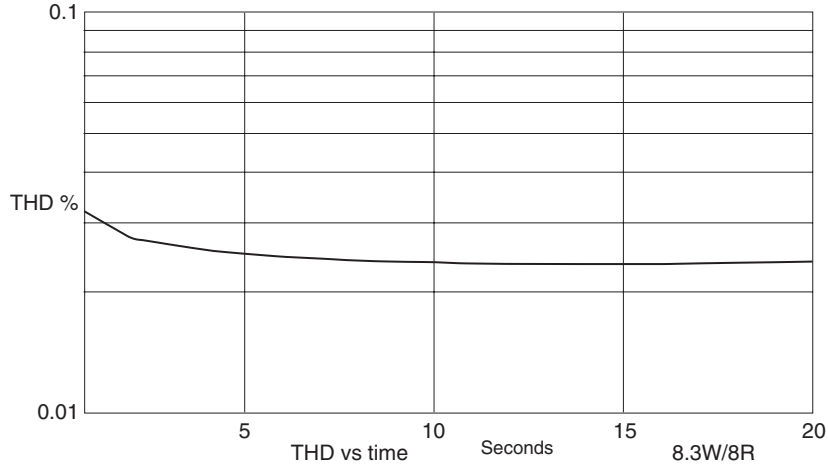


Figure 15.44: Conventional compensation with the sensor on top of the output device – the first 20 seconds

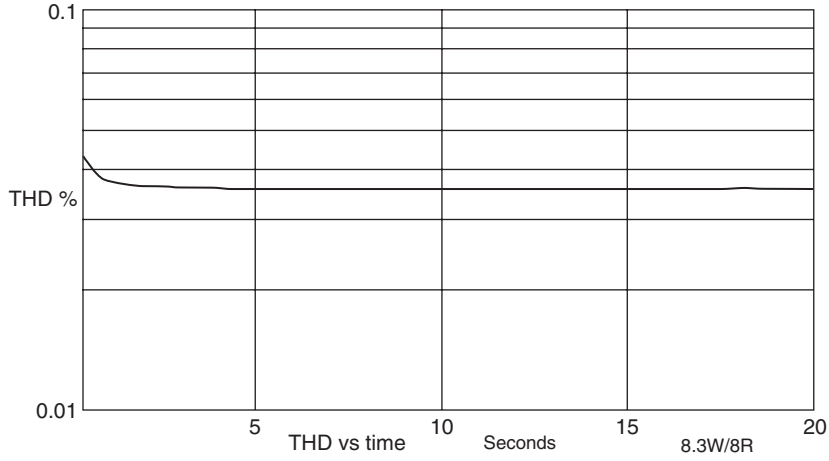


Figure 15.45: ThermalTrak compensation – the first 20 seconds

You will have gathered that this is work in progress, but it clearly shows now that the bias problem is more complex than it looks. I cannot at present demonstrate the perfect biasing system, any more than I can demonstrate the perfect amplifier. These tests were done only a few days before this book went to press, and are included because I think these new devices are potentially very important for amplifier design, and what information there is should be published as soon as possible.

References

[1] T. Sato et al., Amplifier transient crossover distortion resulting from temperature change of output power transistors, AES Preprint 1896 for 72nd Convention, October 1982.

[2] I. Brown, Opto-bias basis for better power amps, Electronics World (February 1992) p. 107.

[3] H. Carslaw, J. Jaeger, Conduction of Heat in Solids, Oxford University Press, 1959.

[4] D. Murphy, Axisymmetric model of a moving-coil loudspeaker, JAES (September 1993) p. 679.

[5] Motorola, Toulouse, Private communication, 1995.

[6] J. Evans, Audio amplifier bias current (Letters), Electronics & Wireless World (January 1991) p. 53.

[7] C.-T. Chen, Analog and Digital Control System Design, Saunders-HBJ, 1993 p. 346.

[8] P. Harriot, Process Control, McGraw-Hill, 1964 pp. 100–102.

[9] B. Liptak (Ed.), Instrument Engineer's Handbook – Process Control, Butterworth-Heinemann, 1995, p. 66.

[10] B. Cordell, diyAudio Forums – Biasing and thermal compensation of ThermalTrak transistors, p. 5. htm. Go to http://www.diyaudio.com. September 2006.

CHAPTER 16

The Design of DC Servos

In the section of this book dealing with input stages I have gone to some lengths to demonstrate that a plain unassisted amplifier – if designed with care – can provide DC offset voltages at an output that is are low enough for most practical purposes, without needing either an offset-nulling preset or a DC servo system. For example, the Trimodal amplifier can be expected not to exceed ±15 mV at the output. However, there may be premium applications where this is not good enough. In this case the choice is between manual adjustment and DC servo technology. As precision op-amps have got cheaper, the use of DC servos has increased.

DC Offset Trimming

Preset adjustment to null the offset voltage has the advantage that it is simple in principle and most unlikely to cause any degradation of audio performance. In servicing the offset should not need renulling unless one of relatively few components is changed; the input devices have the most effect, because the new parts are unlikely to have exactly the same beta, but the feedback resistors also have some influence as the input stage base currents flow through them.

The disadvantages are that an extra adjustment is required in production, and since this is a set-and-forget preset, it can have no effect on DC offsets that may accumulate due to input stage thermal drift or component ageing.

Figure 16.1 shows one simple way to add a DC trim control to an amplifier, by injecting a small current of whatever polarity is required into the feedback point. Since the trim circuit is powered from the main HT rails, which are assumed to be unregulated, careful precautions against the injection of noise, ripple, and DC fluctuations must be taken. The diodes D1, D2 set up a stable voltage across the potentiometer. They do of course have a thermal coefficient, but this is not likely to be significant over the normal temperature range. R3 and C1 form a low-pass filter to reduce noise and ripple, and the trimming current is injected through R4. This resistor has a relatively high value to minimize its effect on the closed-loop gain, and to give a powerful filtering action in conjunction with the large value of C101, to remove any remaining noise and ripple. Note that the trim current is injected at the bottom of R103 and not into the actual feedback point at B, as this would feed any disturbances on C1 directly into the amplifier path. From the point of view of the amplifier, R4 is simply a resistance to ground in parallel with R101, so its effect on the gain can be easily taken into account if required. This DC trim circuit should not degrade the noise performance of the amplifier when it is added, even though the amplifier itself is unusually quiet due to the low impedance of the feedback network.

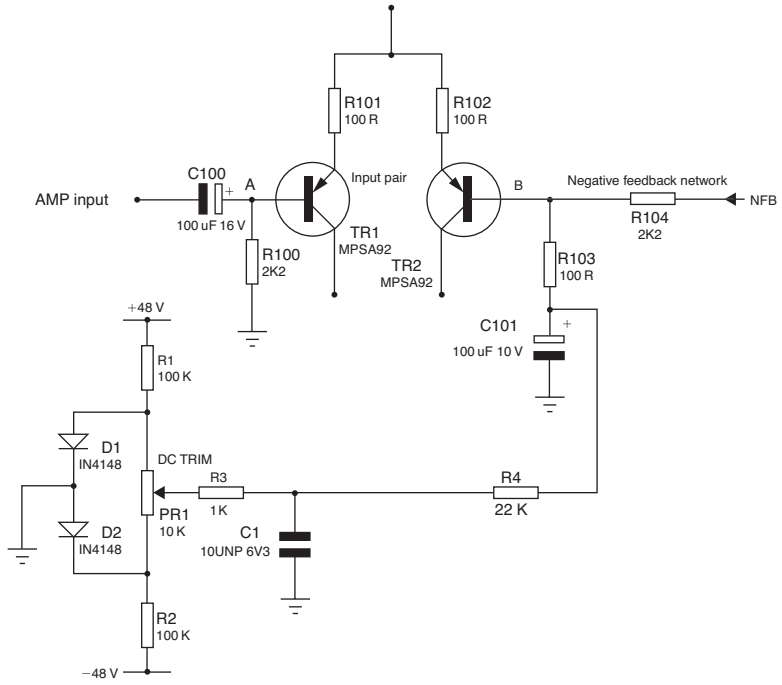


Figure 16.1: DC offset trim with injection into the negative-feedback network

So long as the input is properly AC-coupled (DC-blocked) the trim current can also be fed into the input at point A, but the possible effect on the noise and hum performance is less predictable as the impedance feeding the amplifier input is not known.

DC Offset Control by Servo-Loop

A DC servo system (presumably so-called to emphasize that it does not get directly involved in the main feedback loop) provides continuous active nulling of the amplifier offset by creating another feedback path that has a high gain at DC and very low frequencies, but limited control of the output DC level. This second path uses an op-amp, usually configured as an integrator, to perform the feedback subtraction in which the output DC level is compared with ground. It is straightforward to select an op-amp whose input offset specification is much better than the discrete input stage, because DC precision is where op-amp technology can really excel. For example, both the Analog Devices AD711JN and OPA134 offer a maximum offset of ±2 mV at 25°C, rising to 3 mV over the full commercial temperature range. Performance an order of magnitude better than this is available, e.g. the OPA627, but the price goes up by an order of magnitude too. FET input op-amps are normally used to avoid bias-current offsets with high-value resistors.

An unwelcome complication is the need to provide ±15 V (or thereabouts) supply rails for the op-amp, if it does not already exist. It is absolutely essential that this supply is not liable to drop- out if the main amplifier reproduces a huge transient that pulls down the main supply rails. If it does drop out sufficiently to disrupt the operation of the servo, disturbances will be fed into the main

amplifier, possibly causing VLF oscillation. This may not damage the amplifier, but is likely to have devastating results for the loudspeakers connected to it.

The Advantages of DC Servos

1. The output op-amp DC offset of the amplifier can be made almost as low as desired. The technology of DC precision is mature and well understood.
2. The correction process is continuous and automatic, unlike the DC trimming approach. Thermal drift and component ageing are dealt with, and there is only one part on which the accuracy of offset nulling depends – the servo op-amp, which should not significantly change its characteristics over time.
3. The low-frequency roll-off of the amplifier can be made very low without using huge capacitors. It can also be made more accurate, as the frequency is now set by a non-electrolytic capacitor.
4. The use of electrolytics in the signal path can be avoided, and this will impress some people.
5. The noise performance of the power amplifier can be improved because lower value resistances can be used in the feedback network, yielding a very quiet amplifier indeed.

Points 3, 4, and 5 are all closely related, so they are dealt with at greater length below.

Basic Servo Configurations

Figure 16.2a shows a conventional feedback network, as used in the Load-Invariant amplifier in this book. The usual large capacitor C is present at the bottom of the feedback network; its function is to improve offset accuracy by reducing the closed-loop gain to unity at DC. Figure 16.2b shows a power amplifier with a DC servo added, in the form of a long-time-constant integrator feeding into the feedback point. C is no longer required, as the servo can do all the work of maintaining the DC conditions, though sometimes it might be a good idea to retain it to keep the DC loop gain of the servo system high, and so improve its accuracy; if you do, check carefully for LF stability, as you have introduced another time-constant. Note that the output of the integrator is at ground as far as audio frequencies are concerned, and so the addition of R3 puts it effectively in parallel with R2 and causes a small increase in closed-loop gain that must be taken into account.

It had better be said at once that if the integrator constant is suitably long, a negligible amount of the audio signal passes through it, and the noise and distortion of the main amplifier should not be degraded in any way (more on this later).

As with manual trimming, there are many ways to implement a DC servo. This method works very well, and I have used it many times. One important point is that the integrator block must be non-inverting for the servo feedback to be in the correct phase. The standard shunt-feedback integrator is of course inverting, so something needs to be done about that. Several non-inverting integrators are examined below.

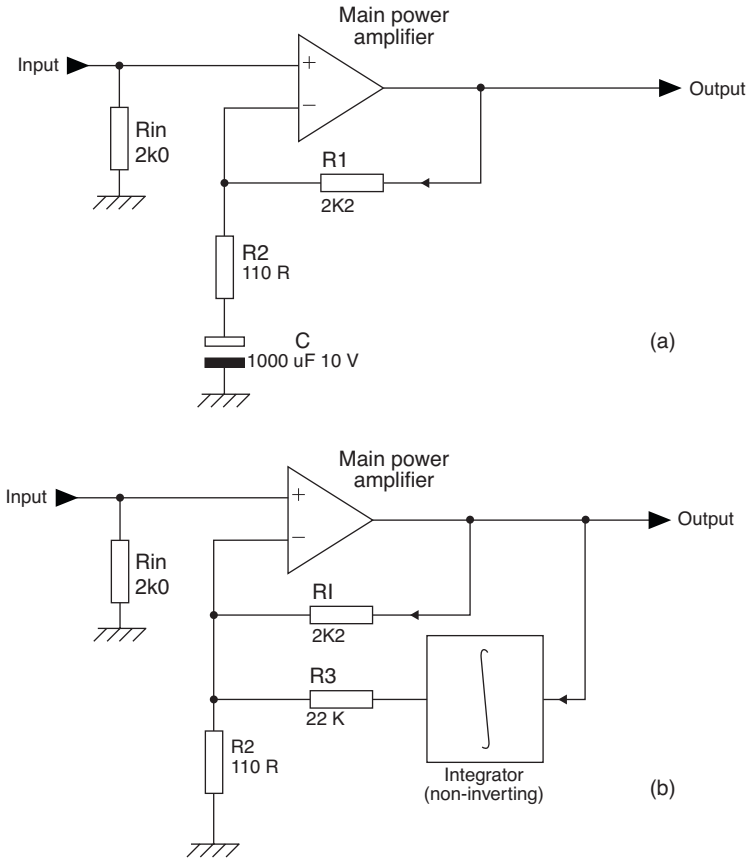


Figure 16.2: Power amplifiers without and with a DC servo in the feedback path

Injecting the servo signal into the input is possible, and in this case a standard inverting integrator can be used. However, as for manual trimming, using the input gives a greater degree of uncertainty in the operating conditions as the source impedance is unknown. If there is no DC blocking on the input, the DC servo will probably not work correctly as the input voltage will be controlled by the low impedance of a pre-amp output. If there is DC blocking then the blocking capacitor may introduce an extra pole into the servo response, which if nothing else complicates things considerably.

Injection of the servo correction into the amplifier forward path is not a good idea as the amplifier has its own priorities – in particular keeping the input pair exactly balanced. If, for example, you feed the servo output into the current-mirror at the bottom of the input pair, the main amplifier can only accommodate its control demands by unbalancing the input pair collector currents, and this will have dire effects on the high-frequency distortion performance.

Noise, Component Values, and the Roll-Off

When you design an amplifier feedback network, there is a big incentive to keep the Johnson noise down by making the resistor values as low as possible. In the simple feedback network shown

in Figure 16.2a, the source impedance seen by the input stage of the amplifier is effectively that of R2; if the rest of the amplifier has been thoughtfully designed then this will be a significant contributor to the overall noise level. Since the Johnson noise voltage varies as the square root of the resistance, minor changes (such as allowing for the fact that R1 is effectively in parallel in R2) are irrelevant. Because of the low value of R2, the feedback capacitor C tends to be large as its RC time-constant with R2 (not R1 + R2) is what sets the LF roll-off. If R2 is low then C is big, and practical values of C put a limit on how far R2 can be reduced. Hence there is a trade-off between low-frequency response and noise performance, controlled by the physical size of C.

When a DC servo is fitted, it is usual to let it do all the work, by removing capacitor C from the bottom arm of the negative-feedback network. The components defining the LF roll-off are now transferred to the servo, which will use high-value resistors and small non-electrolytic capacitors. The value of R2 is no longer directly involved in setting the LF roll-off and there is the possibility that its resistance can be further reduced to minimize its noise contribution, while at the same time the LF response is extended to whatever frequency is thought desirable. The limit of this approach to noise reduction is set by how much power it is desirable to dissipate in R1.

There is a temptation to fall for the techno-fallacy that if it can be done, it should be done. A greatly extended LF range (say below 0.5 Hz) exposes the amplifier to some interesting new problems of DC drift. A design with its lower point set at 0.1 Hz is likely to have its output wavering up and down by tens of millivolts, as a result of air currents differentially cooling the input pair, introducing variations that are slow but still too fast for the servo to correct. Whether these perturbations are likely to cause subtle intermodulations in speaker units is a moot point; it is certain that it does not look good on an oscilloscope, and could cause reviewers to raise their eyebrows. Note that unsteady air currents can exist even in a closed box due to convection from internal heating.

A cascode input stage reduces this problem by greatly lowering the voltage drop across the input transistors, and hence their dissipation, package temperature, and vulnerability to air currents. While it has been speculated that an enormously extended LF range benefits reproduction by reducing phase distortion at the bottom of the audio spectrum, there seems to be no hard evidence for this, and in practical terms there is no real incentive to extend the LF bandwidth greatly beyond what is actually necessary.

Non-Inverting Integrators

The obvious way to build a non-inverting integrator is to use a standard inverting integrator followed by an inverter. The first op-amp must have good DC accuracy as it is here the amplifier DC level is compared with 0 V. The second op-amp is wholly inside the servo loop so its DC accuracy is not important. This arrangement is shown in Figure 16.3. It is not a popular approach because it is perfectly possible to make a non-inverting integrator with one op-amp. It does, however, have the advantage of being conceptually simple; it is very easy to calculate. The frequency response of the integrator is needed to calculate the low-frequency response of the whole system.

The component values shown in Figure 16.3 give unity gain at 1 Hz.

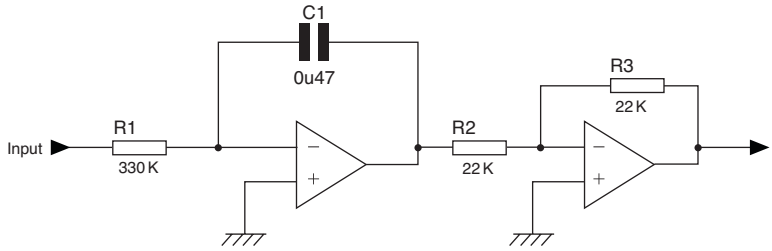


Figure 16.3: A conventional inverting integrator followed by an inverter

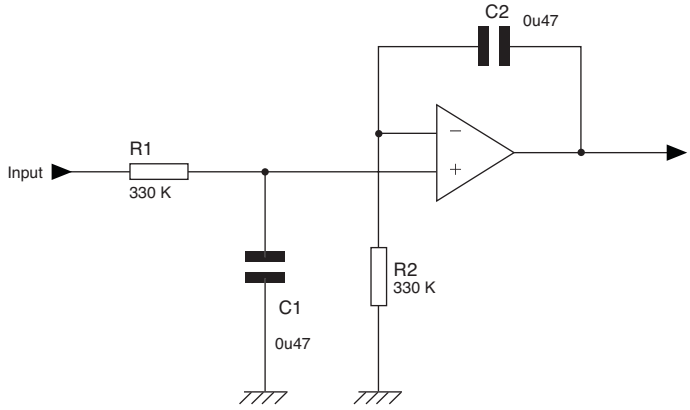


Figure 16.4: A non-inverting integrator that requires only one op-amp

The 2C Integrator

Figure 16.4 shows a non-inverting integrator that has often been used in DC servo applications, having the great advantage of requiring one op-amp. It does, however, use two capacitors; if you are aiming for a really low roll-off these can become quite large for non-electrolytics and will be correspondingly expensive. Despite the presence of two RC time-constants, this circuit is still a simple integrator with a standard −6 dB/octave frequency response.

At the input is a simple RC lag, with the usual exponential time response to step changes; its deviation from being an integrator is compensated for by the RC lead network in the feedback network. A good question is what happens if the two RC time-constants are not identical – does the circuit go haywire? Fortunately not. A mismatch only causes gain errors at very low frequencies, and these are unlikely to be large enough to be a problem. An RC mismatch of ±10% leads to an error of ±0.3 dB at 1.0 Hz, and this error has almost reached its asymptote of ±0.8 dB at 0.1 Hz.

The frequency domain response of Figure 16.4 is:

$$A = \frac{1}{j\omega RC} \qquad \text{Equation 16.1}$$

where $\omega = 2\pi f$, exactly as for the simple integrator of Figure 16.3. The values shown give unity gain at 1 Hz.

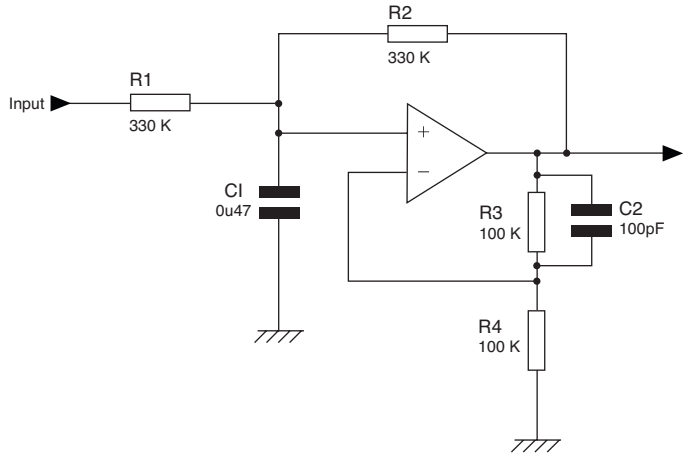


Figure 16.5: A non-inverting integrator that requires only one op-amp and one capacitor

The 1C Integrator

Figure 16.5 displays an apparently superior non-inverting integrator circuit that requires only one op-amp and one capacitor. How it works is by no means immediately obvious, but work it does. R1 and C1 form a simple lag circuit at the input. By itself, this naturally does not give the desired integrator response of a steadily rising or falling capacitor voltage as a result of a step input; instead it gives the familiar exponential response, because as the capacitor voltage rises the voltage across R1 falls, and the rate of capacitor charging is reduced. In this circuit, however, as the capacitor voltage rises the output of the op-amp rises at twice the rate, due to the gain set up by R3 and R4, and so the increasing current flowing into C1 through R2 exactly compensates for the decreasing current flowing through R1, and the voltage on C1 rises linearly, as though it were being charged from a constant-current source. This is in fact the case, because the circuit can be viewed as equivalent to a Howland current source driving into a capacitor.

As for the previous circuit, doubts may be entertained as to what happens when the compensation is less than perfect. For example, here it depends on R1 and R2 being the same value, and also the equality of R3 and R4, to set a gain of exactly 2. Note that R3 and R4 can be high-value resistors. Stray capacitances are dealt with by the addition of C2, which in most cases will be found to be essential for the HF stability of this configuration; this extra capacitor somewhat detracts from the economy of the circuit, but it will be a small ceramic type and of much less cost than the non-electrolytic capacitor used to set the integrator time-constant.

The frequency domain response is now different:

$$A = \frac{1}{j\omega\frac{R}{2}C} \qquad \text{Equation 16.2}$$

where R = R1 = R2.

The *R*/2 term appears because C1 is now being charged through two equal resistors R1 and R2. The values shown therefore give unity gain at 2 Hz.

Choice of Integrator Type

The 1C integrator is clearly the most economical, because big non-electrolytic capacitors are relatively expensive, and I have used it successfully in several applications where it was appropriate. However, it does have some non-obvious disadvantages. If there is a significant power amplifier offset to be servoed out, the accuracy with which it is done depends rather critically on the matching of the two resistor pairs R1, R2 and R3, R4 in Figure 16.5. This holds even if a perfect servo op-amp with zero input offset voltage of its own is assumed.

A significant offset typically occurs when the bases of the two transistors in the power amplifier input differential pair are fed from resistances of very different values. Looking at Figure 16.2b, R2 connects the inverting input to ground with a low resistance of 110 Ω, the value being kept as low as possible to minimize Johnson noise. A resistor R_{in} is connected from the non-inverting input to ground, to define the DC conditions, but this is typically much larger, so as not to load unduly the signal source. It is usually of the order of 2 kΩ if there is some op-amp circuitry (such as a balanced input amplifier) upstream, as this is high enough to avoid excessively loading an op-amp and so introducing distortion. However, it could be a good deal higher at 10 kΩ or more if the amplifier is intended to be driven directly from the outside world. Even if we assume exactly equal base currents, the much higher value of R_{in} will give a positive offset of tens of millivolts at the non-inverting input. This would not be the case in Figure 16.2a, which has a capacitor at the bottom of the NFB network, as it often possible to make R_{in} = R1 and so aim for offsets that are equal at each input and so cancel out.

In a respectable power amplifier the collector currents of the input pair should be almost exactly equal to minimize distortion (see Chapter 4 on input stages) but this does not mean that the base currents, or what would in an op-amp spec sheet be called the input bias currents, are equal, as the input devices will have differing betas.

To take a real example, an amplifier as in Figure 16.2b with R_{in} = 2k2 and R2 = 110 Ω gave an offset of +26 mV on the non-inverting input; if the input transistors had had the minimum beta on their spec sheet it could have been several times greater. Using this value in a SPICE simulation using a 1C servo circuit as per Figure 16.5, with R3 = 2k2 and zero op-amp offset, gave a highly satisfactory offset of +37 μV at the power amp output. But this simulation had both resistor pairs R1, R2 and R3, R4 set to be exactly correct. If R3 was set just 1% high, the power amp output offset leapt up to +29 mV; when it was 1% low, the output offset was −31 mV. Deviations of 1% in the values of R1, R2 gave similar errors.

This is a very good illustration of the caution you need to apply to simulator results; it is not obvious on inspecting the circuit that its operation depends crucially on perfectly matched resistors. The simulator answer is absolutely correct, but not applicable to the real world of imperfect components.

Another disadvantage of the 1C circuit is that when you use a real op-amp, as opposed to a simulated perfect one, its own input offset appears doubled at the power amplifier output, due to

the gain of 2 set up by R3 and R4. If you do not have access to the op-amp offset-null pins there is no easy way to add a DC trimming network as connecting even high-value resistors such as 10M disturbs the balance of this circuit and stops it working properly.

Having examined the quite serious limitations of the 1C non-inverting integrator, let's go back to the 2C version and see if that is more tractable.

Firstly, the 2C circuit does not require accurately matched resistors to work properly. In Figure 16.4, R1 – R2 mismatch has a negligible effect on the DC accuracy and only a microscopic effect on the AC response below 1 Hz. Secondly, the op-amp offset is not multiplied by 2.

Another important point is that it is now possible to add a DC trimming network without disrupting the integrator's operation. This can be used to null to zero the small offset (typically 1–2 mV) that remains when a servo is added. If the op-amp used has offset-null pins then these should be used with whatever nulling circuitry the manufacturer recommends, but for economy it is often the case that the servo is one-half of a dual op-amp with no offset-null pins, the other half typically being used for over-temperature detection. If this is so, a DC trimming network can be added to the 2C circuit – unlike the 1C version. Figure 16.6 shows a network that can be added to allow nulling to less than a millivolt. Its range of adjustment is limited to only ±5 mV at the power amp output. The component values are those used in a production amplifier – the negative-feedback network had the values as shown in Figure 16.2b – note, however, that the integrator resistors R1, R2 have been changed to 180k and the servo injection resistor R_{inj} has been changed from 22k to 2k2 to obtain the desired LF roll-off frequency.

Figure 16.6 does not include any filter components to prevent noise or hum on the ±15 V rails from entering the servo; details on how to do this are given in Figure 16.1. It should be pointed out that while this sort of external nulling is usually quite satisfactory, it will not perform as well over a wide temperature range as using the official offset-null pins if they are available.

It can be concluded that in most cases the 2C integrator is superior to the 1C version. It is true that two capacitors are needed instead of one, but in many cases the price is well worth paying for better and more predictable servo performance.

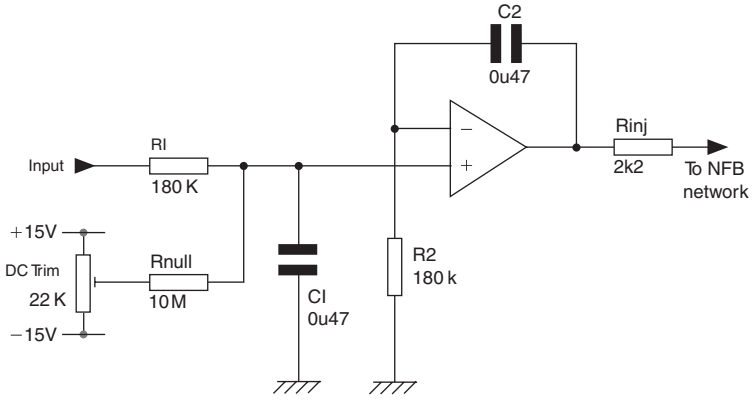


Figure 16.6: Adding a DC trimming network to the 2C servo integrator

Table 16.1: Op-amp specs compared

Type	Offset at 25°C (mV)	Offset over −40 to +85°C (mV)	Relative cost
TL051	1.5	2.5	1.00
OPA134	2	3	1.34
AD711JN	2	3	1.48
OPA627AP	0.28	0.5	16.0

Note that the TL051 looks like quite a bargain, and going for a serious improvement on this with the OPA627AP will cost you deep in the purse.

Choice of Op-Amps

All of these integrator circuits use high resistor values to keep the size of the capacitors down. It is essential to use FET-input op-amps, with their near-zero bias and offset currents. Bipolar op-amps have many fine properties, but they are not useful here. You will need a reasonably high-quality FET op-amp to beat non-servo power amplifiers, which can be designed so their output offset does not exceed ±15 mV offset at the output.

Some prime candidates are given in Table 16.1, giving the maximum±offset voltages, and the relative cost at the time of writing.

In many designs a dual op-amp is the best choice, the remaining section being used as a comparator driven by an over-temperature detection device such as a thermistor. If the op-amp is accurate enough to do its job as a servo, it will almost certainly be good enough for temperature detection.

Servo Authority

The phrase ‘servo authority’ refers to the amount of control that the DC servo system has over the output DC level of the amplifier. It is, I hope, clear that the correct approach is to design a good input stage that gives a reasonably small DC offset unaided, and then add the servo system to correct the last few dozen millivolts, rather than to throw together something that needs to be hauled into correct operation by brute-force servo action.

In the latter case, the servo must have high authority in order to do its job, and if the servo op-amp dies and its output hits one of its rails, the amplifier will follow suit. The DC offset protection should come into action to prevent disaster, but it is still an unhappy situation.

However, if the input stage is well designed, so the servo is only called upon to make fine adjustments, it is possible to limit the servo authority, by proportioning the circuit values so that R3 in Figure 16.2 is relatively high. Then, even if the op-amp fails, the amplifier offset will be modest. In many cases it is possible for the amplifier to continue to function without any ill effects on the loudspeakers. This might be valuable in sound-reinforcement applications and the like.

Calculating the effects of op-amp failure in the circuit of Figure 16.2 is straightforward. The system appears as a shunt-feedback amplifier where R3 is the input resistance and R1 is the feedback resistance. Thus if the op-amp is working from ±15 rails, then, ignoring saturation effects, the main amplifier output will be displaced by ±1.5 V.

When limiting servo authority, it is of course essential to allow enough adjustment to deal with any combination of component tolerances that may happen along. Do not limit it too much.

Design of LF Roll-Off Point

Calculating the frequency response of the servo-controlled system is surprisingly easy. The −3 dB point will occur where the feedback through the normal network and the integrating servo path are equal in amplitude; it is −3 dB rather than −6 dB because the two signals are displaced in phase by 90°. This is exactly the same as the −3 dB point obtained with an RC circuit, which happens at the frequency where the impedance of the R and C are equal in magnitude, though displaced in phase by 90°.

As a first step, decide what overall gain is required; this sets the ratio of R1 and R2. Next determine how low R1 can conveniently be made to minimize the noise contribution of R2. This establishes the actual values of R1 and R2. It is important to remember that the servo injection resistor R3, being connected to an effective AC ground at the servo op-amp output, is effectively in parallel with R2 and has a small influence on the main amplifier gain. Third, decide how low a −3 dB point you require for the overall system, and what servo authority you are prepared to allow. I shall take 0.2 Hz as an example, to demonstrate how a servo system makes such a low value easy to attain. Using the values shown in Figure 16.2, the section above demonstrates that the servo authority is more than enough to deal with any possible offset errors, while not being capable of igniting the loudspeakers if the worst happens. R3 is therefore 22k, which is 10 times R2, so at the −3 dB point the integrator output must be 10 times the main amplifier output; in other words it must have a gain of 10 at 0.2 Hz.

The next step is to choose the integrator type; the one op-amp, one capacitor version of Figure 16.5 is clearly the most economical so we will use that. The frequency-response equation given above is then used to set suitable values for R1, R2, and C1 in Figure 16.5. Non-electrolytic capacitors of 470 nF are reasonably priced and this gives a value for R1, R2 of 338 kΩ; the preferred value of 330k is quite near enough.

Servo Overload

The final step is to check that the integrator will not be overdriven by the audio-frequency signals at the amplifier output, bearing in mind that the op-amp will be running off lower supply rails that are half or less of the main amplifier rail voltages. Here I will assume the amplifier rails are ±45 V, i.e. three times the ±15 V op-amp rails. Hence the integrator will clip with full amplifier output at the frequency where integrator gain is 1/3. The integrator we have just designed has a gain of 10 times at 0.2 Hz and a slope of −6 dB/octave, so its gain will have fallen to unity at 2 Hz, and to 1/3 at 6 Hz. Hence the integrator can handle any amplifier output down to maximum power at 6 Hz, which is somewhat below the realm of audio, and all should be well.

Servo Testing

One problem with servo designs of this type is that they are difficult to test; frequencies of 0.2 Hz and below are well outside the capabilities of normal audio test equipment. It is not too hard to find a function generator that will produce the range 0.1–1.0 Hz, but measuring levels to find the −3 dB

Table 16.2: Tilt on 20 Hz square wave with different LF roll-off frequencies

−3 dB point (Hz)	Tilt (%)
0.15	2.5
0.23	3.5
0.32	5.0
0.50	7.4
0.70	10.5
1.0	15.2
1.4	20
2.1	28

Note that the tilt is expressed as a percentage of the zero to peak voltage, not peak to peak.

point is difficult. A storage oscilloscope will give approximate results if you have one; the accuracy is not usually high.

One possibility is the time-honored method of measuring the tilt on a low-frequency square wave. Accuracy is still limited, but you can use an ordinary oscilloscope. Even very-low-frequency roll-offs put an easily visible tilt on a 20 Hz square wave, and this should be fast enough to give reasonable synchronization on a non-storage oscilloscope. A rough guide is shown in Table 16.2.

Performance Issues

The advantages of using a DC servo have been listed above, without mentioning any disadvantages, apart from the obvious one that more parts are required and a little power is needed to run the op-amp. It could easily be imagined that another and serious drawback is that the presence of an op-amp in the negative-feedback network of an amplifier could degrade both the noise and distortion performance. However, this is not the case. When the system in Figure 16.2b is tested with a Load-Invariant amplifier and an OPA134 op-amp as a servo, there is no measurable effect on either quantity.

The distortion performance is unaffected because the servo integrator passes very little signal at audio frequencies. The noise performance is preserved because the integrators are very quiet due to their falling frequency response, and with the long integration constants used here they are working at a noise gain of unity at audio frequencies. Both parameters benefit from the fact that the servo feedback path via R3 has one-tenth of the gain of the main feedback path through R1.

Multi-Pole Servos

All the servos shown above use an integrator and therefore have a single pole. It is possible to make servos that have more than one pole, and they have been used in some designs, though the motivation for doing it is somewhat unclear. The usual arrangement has a single-op-amp non-inverting integrator followed by a simple RC lag network that feeds into the feedback point. Naturally, once you have more than one pole in a system there is the possibility of an underdamped response and gain peaking, so this approach demands careful design, not least because measuring gain peaking at 0.1 Hz is not that easy.

CHAPTER 17

Amplifier and Loudspeaker Protection

Categories of Amplifier Protection

The protection of solid-state amplifiers against overload is largely a matter of safeguarding them from load impedances that are too low and endanger the output devices, the most common and most severe condition being a short across the output. This must be distinguished from the casual use of the word 'overload' to mean excessive signal that causes clipping and audible distortion.

Overload protection is not the only safety precaution required. An equally vital requirement is DC offset protection – though here it is the loudspeaker load that is being protected from the amplifier, rather than the other way round.

Similarly, thermal protection is also required for a fully equipped amplifier. Since a well-designed product will not overheat in normal operation, this is required to deal with two abnormal conditions:

1. The amplifier heat-sinking is designed to be adequate for the reproduction of speech and music (which has a high peak-to-volume ratio, and therefore brings about relatively small dissipation) but cannot sustain long-term sine-wave testing into the minimum specified load impedance without excessive junction temperatures. Heat-sinking forms a large part of the cost of any amplifier, and so economics makes this a common state of affairs. Similar considerations apply to the rating of amplifier mains transformers, which are often designed to indefinitely supply only 70% of the current required for extended sine-wave operation. Some form of thermal cut-out in the transformer itself then becomes essential (see Chapter 9).
2. The amplifier is designed to withstand indefinite sine-wave testing, but is vulnerable to having ventilation slots, etc. blocked, interfering either with natural convection or fan operation.

Finally, all amplifiers require internal fusing to minimize the consequences of a component failure – i.e. protecting the amplifier from itself – and to maintain safety in the event of a mains wiring fault.

Semiconductor Failure Modes

Solid-state output devices have several main failure modes, including excess current, excess power dissipation, and excess voltage. These are specified in manufacturer's data sheets as Absolute

Maximum Ratings, usually defined by some form of words such as 'exceeding these ratings even momentarily may cause degradation of performance and/or reduction in operating lifetime'. For semiconductor power devices ratings are usually plotted as a safe operating area (SOA) that encloses all the permissible combinations of voltage and current. Sometimes there are extra little areas, notably those associated with second breakdown in BJTs, with time limits (usually in microseconds) on how long you can linger there before something awful happens.

It is of course also possible to damage the base–emitter junction of a BJT by exceeding its current or reverse voltage ratings, but this is unlikely in power amplifier applications. In contrast the insulated gate of an FET is more vulnerable and Zener clamping of gate to source is usually considered mandatory, especially since FET amplifiers often have separate higher supply rails for their small-signal sections.

BJTs have an additional important failure mode known as second breakdown, which basically appears as a reduction in permissible power dissipation at high voltages, due to local instability in current flow. The details of this mechanism may be found in any textbook on transistor physics.

Excessive current usually causes failure when the I^2R heating in the bond wires becomes too great and they fuse. This places a maximum on the current handling of the device no matter how low the voltage across it, and hence the power dissipation. In a TO-3 package only the emitter bond wire is vulnerable, as the collector connection is made through the transistor substrate and flange. If this wire fails with high excess current then on some occasions the jet of vaporized metal will drill a neat hole through the top of the TO-3 can – an event that can prove utterly mystifying to those not in the know.

Any solid-state device will fail from excess dissipation, as the internal heating will raise the junction temperatures to levels where permanent degradation occurs.

Excess emitter-collector or source-drain voltage will also cause failure. This failure mode does not usually require protection as such, because designing against it should be fairly easy. With a resistive load the maximum voltage is defined by the power-supply rails, and when the amplifier output is hard against one rail the voltage across the device that is turned off will be the sum of the two rails, assuming a DC-coupled design. If devices with a $V_{ce(max)}$ greater than this is selected there should be no possibility of failure. However, practical amplifiers will be faced with reactive load impedances, and this can double the V_{ce} seen by the output devices. It is therefore necessary to select a device that can withstand at least twice the sum of the HT rail voltages, and allow for a further safety margin on top of this. Even greater voltages may be generated by abrupt current changes in inductive loads, and these may go outside the supply-rail range, causing device failure by reverse biasing. This possibility is usually dealt with by the addition of *catching* diodes to the circuit (see below) and does not in itself affect the output device specification required.

Power semiconductors have another failure mode initiated by repeated severe temperature changes. This is usually known as *thermal cycling* and results from stresses set up in the silicon by the differing expansion coefficients of the device chip and the header it is bonded to. This constitutes

the only real wear-out mechanism that semiconductors are subject to. The average lifetimes of a device subjected to temperature variations ΔT can be approximately predicted by:

$$N = 10^7 \cdot e^{-0.05 \cdot \Delta T} \qquad \text{Equation 17.1}$$

where N = cycles to failure and ΔT is the temperature change.

This shows clearly that the only way open to the designer to minimize the risk of failure is to reduce the temperature range or the number of temperature cycles. Reducing the junction temperature range requires increasing heat-sink size or improving the thermal coupling to it. Thermal coupling can be quickly improved by using high-efficiency thermal washers, assuming their increased fragility is acceptable in production, and this is much more cost-effective than increasing the weight of heat-sink. The number of cycles can only be minimized by leaving equipment (such as Class-A amplifiers) powered long term, which has distinct disadvantages in terms of energy consumption and possibly safety.

Overload Protection

Solid-state output devices are much less tolerant to overload conditions than valves, and often fail virtually instantaneously. Some failure modes (such as overheating) take place slowly enough for human intervention, but this can never be relied upon. Overload protection is therefore always an important issue, except for specialized applications such as amplifiers built into powered loudspeakers, where there are no external connections and no possibility of inadvertent short-circuits.

Driven by necessity, workable protection systems were devised relatively early in the history of solid-state amplifiers (see Bailey[1], Becker[2], and Motorola[3]). Part of the problem is defining what constitutes adequate current delivery into a load. Otala[4] has shown that a complex impedance, i.e. containing energy-storage elements, can be made to draw surprisingly large currents if specially optimized pulse waveforms are used that catch the load at the worst part of the cycle; however, it seems to be the general view that such waveforms rarely, if ever, occur in real life.

Verifying that overload protection works as intended over the wide range of voltages, currents, and load impedances possible is not a light task. Peter Baxandall introduced a most ingenious method of causing an amplifier to plot its own limiting lines[5].

Overload Protection by Fuses

The use of fuses in series with the output line for overload protection is no longer considered acceptable, as it is virtually impossible to design a fuse that will blow fast enough to protect a semiconductor device, and yet be sufficiently resistant to transients and turn-on surges. There are also the obvious objections that the fuse must be replaced every time the protection is brought into action, and there is every chance it will be replaced by a higher-value fuse that will leave the amplifier completely vulnerable. Fuses can react only to the current flowing through them, and are unable to take account of other important factors such as the voltage drop across the device protected.

Series output fuses are sometimes advocated as a cheap means of DC offset protection, but they are not dependable in this role.

Placing a fuse in series with the output will cause low-frequency distortion due to cyclic thermal changes in the fuse resistance. The distortion problem can, in theory at least, be sidestepped by placing the fuse inside the global feedback loop; however, what will the amplifier do when its feedback is abruptly removed when the fuse blows? (See also page 456 on DC offset protection below.)

One way of so enclosing fuses that I have seen advocated is to use them instead of output emitter resistors Re; I have no personal experience of this technique, but since it appears to add extra time-dependent thermal uncertainties (due to the exact fuse resistance being dependent upon its immediate thermal history) to a part of the amplifier where they already cause major difficulties, I do not see this as a promising path to take. There is the major difficulty that the failure of only one fuse will generate a maximal DC offset, so we may have dealt with the overload, but there is now a major DC offset to protect the loudspeaker from. The other fuse may blow as a consequence of the large DC current flow, but sizing a fuse to protect properly against both overload and DC offset may prove impossible.

Amplifier circuitry should always include fuses in each HT line. These are not intended to protect the output devices, but to minimize the damage when the output devices have already failed. They can and should therefore be of the slow-blow type, and rated with a good safety margin, so that they are entirely reliable; a fuse operated anywhere near its nominal fusing current has a short lifetime, due to heating and oxidation of the fuse wire. HT fuses cannot save the output devices, but they do protect the HT wiring and the bridge rectifier, and prevent fire. There should be separate DC fuses for each channel, as this gives better protection than one fuse of twice the size, and allows one channel to keep working in an emergency.

Similarly, the mains transformer secondaries should also be fused. If this is omitted, a failure of the rectifier will inevitably cause the mains transformer to burn out, and this could produce a safety hazard. The secondary fuses can be very conservatively rated to make them reliable, as the mains transformer should be able to withstand a very large fault current for a short time. The fuses must be of the slow-blow type to withstand the current surge into the reservoir capacitors at switch-on.

The final fuse to consider is the mains fuse. The two functions of this are to disconnect the live line if it becomes shorted to chassis, and to protect against gross faults such as a short between live and neutral. This fuse must also be of the slow-blow type, to cope with the transformer turn-on current surge as well as charging the reservoirs. In the UK, there will be an additional fuse in the molded mains plug. This does not apply to mains connectors in other countries and so a mains fuse built into the amplifier itself is absolutely essential.

Electronic Overload Protection

There are various approaches possible to overload protection. The commonest form (called electronic protection here to distinguish it from fuse methods) uses transistors to detect the current

and voltage conditions in the output devices, and shunts away the base drive from the latter when the conditions become excessive. This is cheap and easy to implement (at least in principle), and since it is essentially a clamping method requires no resetting. Normal output is resumed as soon as the fault conditions are removed. The disadvantage is that a protection scheme that makes good use of the device SOA may allow substantial dissipation for as long as the fault persists undetected, and while this should not cause short-term failure if the protection has been correctly designed, the high temperatures generated may impair long-term reliability. In my recent designs a microcontroller detects when SOA limiting is happening and opens the output relay if it persists.

An alternative approach does not limit at all on a cycle-by-cycle basis, but simply drops out the DC protection relay when overload is detected. This will clearly only work if the output stage can survive uncontrolled overload dissipation for long enough for the circuitry to act and the mechanical parts of the relay to move.

In either case the output relay may either be opened for a few seconds delay, after which it resets, or stay latched open until the protection circuit is reset. This is normally done by cycling the mains power on and off, to avoid the expense of a reset button that would rarely be used.

If the equipment is essentially operated unattended, so that an overload condition may persist for some time, the self-resetting system will subject the output semiconductors to severe temperature changes, which may shorten their operational lifetime.

Plotting the Protection Locus

The standard method of representing the conditions experienced by output devices, of whatever technology, is to draw loadlines onto a diagram of the component's SOA, to determine where they cross the limits of the area. This is shown in Figure 17.1, for an amplifier with ±40V HT rails, which would give 100W into 8Ω and 200W into 4Ω, ignoring losses; the power transistor is a Motorola MJ15024. You do not need to fix the HT voltage before drawing most of the diagram; the position of the SOA limits is fixed by the device characteristics. The line AB represents the maximum current rating of 16A, and the reciprocal curve BC the maximum power dissipation of 250W. The maximum V_{ce} is 250V, and is far off the diagram to the right. Line CD defines the second-breakdown region, effectively an extra area removed from the high-voltage end of the power-limited region. Second breakdown is an instability phenomenon that takes a little time to develop, so manufacturer's data often allows brief excursions into the region between the second-breakdown line and the power limit. The nearer these excursions go towards the power limit, the briefer they must be if the device is to survive, and trying to exploit this latitude in amplifiers is living dangerously, because the permitted times are very short (usually tens of microseconds) compared with the duration of audio waveforms.

The resistive loadline XY represents an 8Ω load, and as a point moves along it, the coordinates show the instantaneous voltage across the output device and the current through it. At point X, the current is maximal at 5.0A with zero voltage across the device, as $V_{ce(sat)}$ values and the like can be ignored without significant error. The power dissipated in the device is zero, and what matters is that point X is well below the current-limit line AB. This represents conditions at clipping.

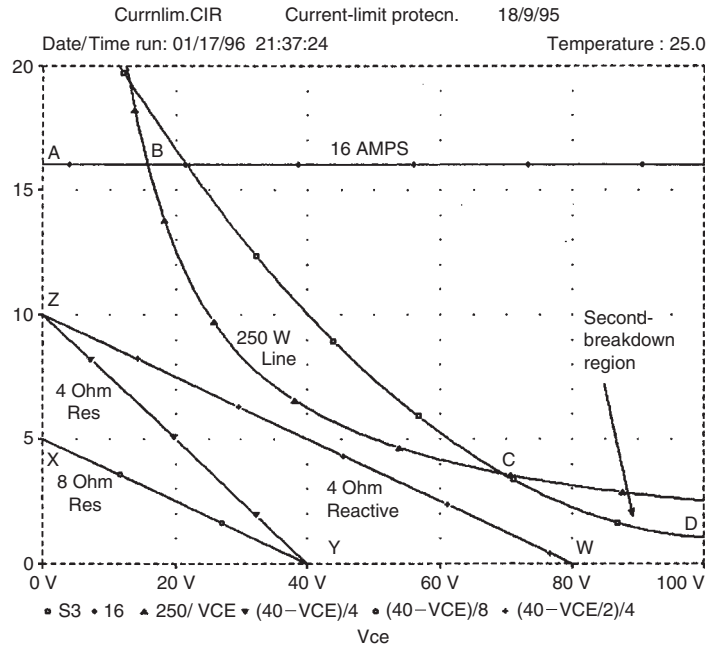


Figure 17.1: The safe operating area (SOA) of a typical TO-3 high-power transistor, in this case the Motorola MJ15024

At the other end, at Y, the loadline has hit the *X*-axis and so the device current is zero, with one rail voltage (40 V) across it. This represents the normal quiescent state of an amplifier, with zero volts at its output, and zero device dissipation once more. So long as Y is well to the left of the maximum-voltage line all is well. Note that while you do not need to decide the HT voltage when drawing the SOA for the device, you must do so before the loadlines are drawn, as all lines for purely resistive loads intersect the *X*-axis at a voltage representing one of the HT rails.

Intermediate points along XY represent instantaneous output voltages between 0 V and clipping; voltage and current coexist and so there is significant device dissipation. If the line cuts the maximum-power rating curve BC, the dissipation is too great and the device will fail.

Different load resistances are represented by lines of differing slope; ZY is for a 4 Ω load. The point Y must be common to both lines, for the current is zero and the rail voltage unchanged no matter what load is connected to a quiescent amplifier. Point Z is, however, at twice the current, and there is clearly a greater chance of this low-resistance line intersecting the power limit BC. Resistive loads cannot reach the second-breakdown region with these rails.

Unwelcome complications are presented by reactive loading. Maximum current no longer coincides with the maximum voltage, and vice versa. A typical reactive load turns the line XY into an ellipse, which gets much nearer to the SOA limit. The width (actually the minor axis, to be mathematical) of the ellipse is determined by the amount of reactance involved, and since this is another independent variable, the diagram could soon become over-complex. The solution is to take the worst case for all possible reactive loads of the form $R + jX$, and instead of trying to

draw hundreds of ellipses, to simply show the envelope made up of all their closest approaches to the SOA limit. This is another straight line, drawn from the same maximum current point Z to a point W at twice the rail voltage. There is clearly a much greater chance that the ZW line will hit the power-limit or second-breakdown lines than the 4Ω resistive line ZY, and the power devices must have an SOA large enough to give a clear safety margin between its boundary and the reactive envelope line for the lowest rated load impedance. The protection locus must fit into this gap, so it must be large enough to allow for circuit tolerances.

The final step is plot the protection locus on the diagram. This locus, which may be a straight line, a series of lines or an arbitrary curve, represents the maximum possible combinations of current and voltage that the protection circuitry permits to exist in the output device. Most amplifiers use some form of VI limiting, in which the permitted current reduces as the voltage across the device increases, putting a rough limit on device power dissipation. When this relationship between current and voltage is plotted, it forms the protection locus.

This locus must always be above and to the right of the reactive envelope line for the lowest rated load, or the power output will be restricted by the protection circuitry operating prematurely. It must also always be to the left and below the SOA limit, or it will allow forbidden combinations of voltage and current that will cause device failure.

Simple Current Limiting

The simplest form of overload protection is shown in Figure 17.2, with both upper and lower sections shown. For positive output excursions, R1 samples the voltage drop across emitter resistor Re1, and when it exceeds the V_{be} of approximately 0.6V, TR1 conducts and shunts current away from TR2 base. The component values in Figure 17.2 give a 5.5A constant-current regime, as shown in Figure 17.3, which was simulated using a model like that in Figure 17.9 below. The loadlines shown represent 8 and 4Ω resistive, and 4Ω worst-case reactive (ZW). The current-limit line is exactly horizontal, though it would probably show a slight slope if the simulation was extended to include more of the real amplifier, such as real current sources, etc.

The value of Re1 is usually determined by the requirements of efficiency or quiescent stability, and so the threshold of current limiting is set by R1 and R2. This circuit can only operate at a finite speed, and so R1 must be large enough to limit TR1 base current to a safe value; 100Ω seems sufficient in practice. Re1 is usually the output emitter resistor, as well as current sensor, and so does double duty.

The current drawn by TR1 in shunting away TR2 base drive is inherently limited by *I*, the constant-current load of the VAS. There is no such limit on TR4, which can draw large and indeterminate currents through VAS transistor TR7. If this is a TO92 device it will probably fail. It is therefore essential to limit the VAS current in some way, and a common approach is shown in Figure 17.2. There is now a secondary layer of current limiting, with TR8 protecting TR7 in the same way that TR1 protects TR2, TR3. The addition of R_s to sense the VAS current does not significantly affect VAS operation, and does not constitute local negative feedback. This is because the input to TR7 is

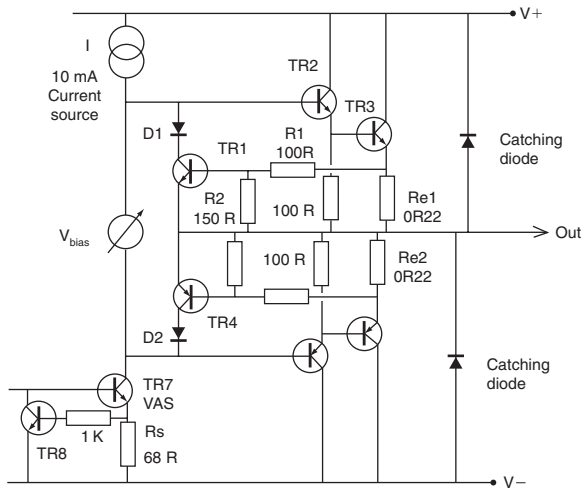


Figure 17.2: Simple current-limit circuit

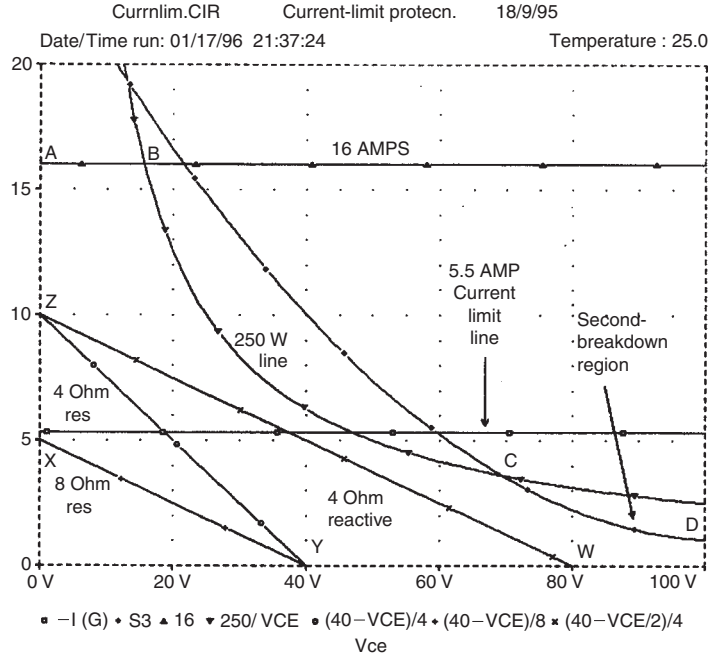


Figure 17.3: Current limiting with ±40 V HT rails

a current from the input stage, and not a voltage; the development of a voltage across R_s does not affect the value of this current, as it is effectively being supplied from a constant-current source.

It has to be faced that this arrangement often shows signs of HF instability when current limiting, and this can prove difficult or impossible to eradicate completely. (This applies to single- and double-slope VI limiting also.) The basic cause appears to be that under limiting conditions

there are two feedback systems active, each opposing the other. The global voltage feedback is attempting to bring the output to the demanded voltage level, while the overload protection must be able to override this to safeguard the output devices. HF oscillation is often a danger to BJT output devices, but in this case it does not seem to adversely affect survivability. Extensive tests have shown that in a conventional BJT output stage, the oscillation seems to reduce rather than increase the average current through the output devices, and it is arguable that it does more good than harm. It has to be said, however, that the exact oscillation mechanism remains obscure despite several investigations, and the state of our knowledge in this area is far from complete.

The diodes D1, D2 in the collectors of TR1, TR4 prevent them conducting in the wrong half-cycle if the Re voltage drops are large enough to make the collector voltage go negative. Under some circumstances you may be able to omit them, but the cost saving is negligible.

The *loadline* for an output short-circuit on the SOA plot is a vertical line, starting upwards from Y, the HT rail voltage on the *X*-axis, and showing that current increases indefinitely without any reduction of the voltage drop across the output devices. An example is shown in Figure 17.3 for ±40 V rails. When the short-circuit line is prolonged upwards it hits the 5.5 A limiting locus at 40 V and 5.5 A; at 220 W this is just inside the power-limit section of the SOA. The devices are therefore safe against short-circuits; however, the 4 Ω resistive loadline also intersects the 5.5 A line, at $V_{ce} = 18$ V and $I_c = 5.5$ A, limiting the 4 Ω output capability to 12 V peak. This gives 18 W rather than 200 W in the load, despite the fact that full 4 Ω output would in fact be perfectly safe. The full 8 Ω output of 100 W is possible as the whole of XY lies below 5.5 A.

With 4 Ω reactive loads the situation is worse. The line ZW cuts the 5.5 A line at 38 V, leaving only 2 V for output, and limiting the power to a feeble 0.5 W.

The other drawback of constant-current protection is that if the HT rails were increased only slightly, to ±46 V, the intersection of a vertical line from Y to the *X*-axis centre would hit the power-limit line, and the amplifier would no longer be short-circuit-proof unless the current limit was reduced.

Single-Slope VI Limiting

Simple current limiting makes very poor use of the device SOA; single-slope VI limiting is greatly superior because it uses more of the available information to determine if the output devices are endangered. The V_{ce} as well as the current is taken into account. The most popular circuit arrangement is seen in Figure 17.4, where R3 has been added to reduce the current-limit threshold as V_{ce} increases. This simple summation of voltage and current seems crude at first sight, but Figure 17.5 shows it to be an enormous improvement over simply limiting the current.

The protection locus has now a variable slope, making it much easier to fit between reactive load lines and the SOA boundary; the slope is set by R3. In Figure 17.5, Locus 1 is for R3 = 15 k and Locus 2 for R3 = 10 k. If Locus 2 is chosen the short-circuit current is reduced to 2 A, while still allowing the full 4 Ω resistive output.

Current capability at $V_{ce} = 20$ V is increased from 5.5 to 7.5 A.

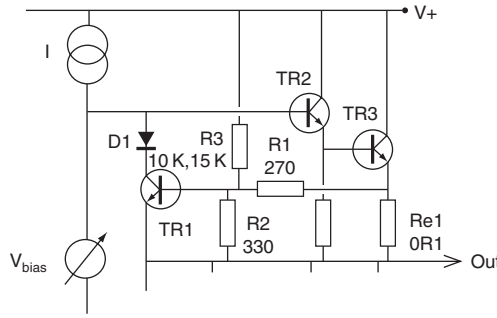


Figure 17.4: Single-slope VI limiter circuit

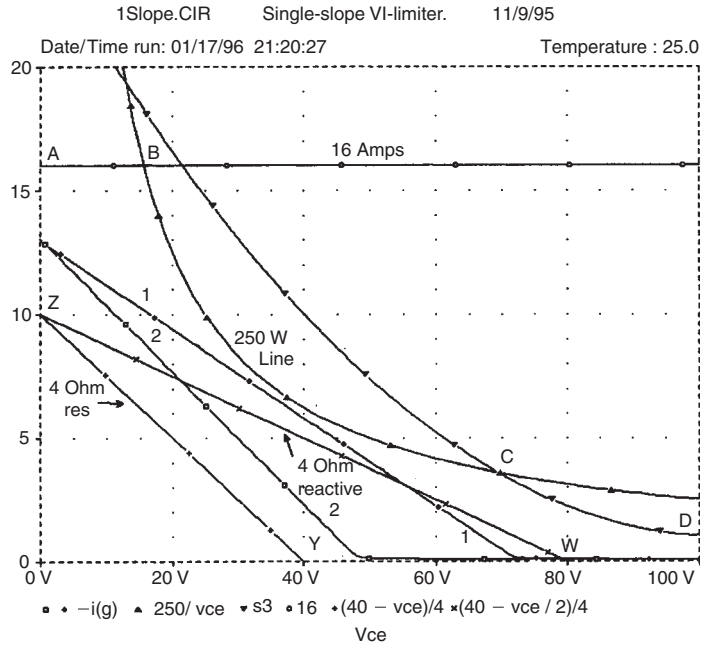


Figure 17.5: Single-slope locus plotted on MJ15024 SOA

Dual-Slope VI Limiting

The motivation for more complex forms of protection than single-slope VI limiting is usually the saving of money, by exploiting more of the output device SOA. In a typical amplifier required to give 165 W into 8 Ω and 250 W into 4 Ω (assuming realistic losses) the number of device pairs in the output stage can be reduced from three to two by the use of dual-slope protection, and the cost saving is significant. The single-slope limiting line is made dual-slope by introducing a breakpoint in the locus so it is made of two straight-line sections as in Figure 17.7, allowing it to be moved closer to the curved SOA limit; the current delivery possible at low device voltages is further increased.

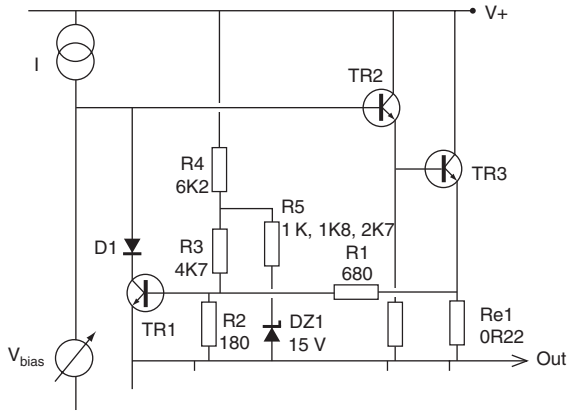


Figure 17.6: Dual-slope locus plotted on MJ15024 SOA

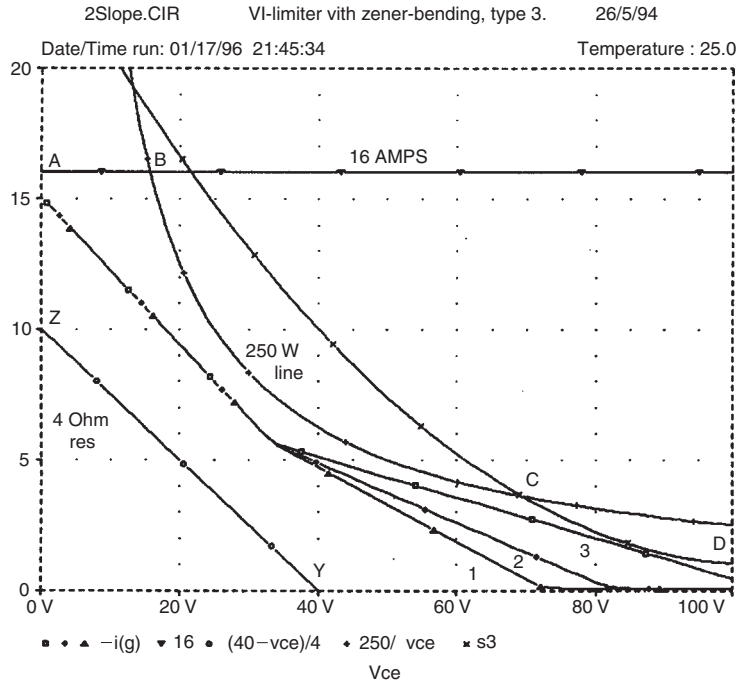


Figure 17.7: Dual-slope VI limiter circuit

A dual-slope system is shown in Figure 17.6. The action of the V_{ce} component on sensing transistor TR1 is reduced when V_{ce} is high enough for Zener diode DZ1 to conduct. The series combination of R4 and R1 is chosen to give the required initial slope with low V_{ce} (i.e. the left-hand slope) but as the voltage increases the Zener conducts and diverts current through R5, whose value controls the right-hand slope of the protection locus. Loci 1, 2, and 3 are for R5 = 2k7, 1k8, and 1k respectively.

Current capability at $V_{ce} = 20$V is further increased from 7.5 to 9.5A.

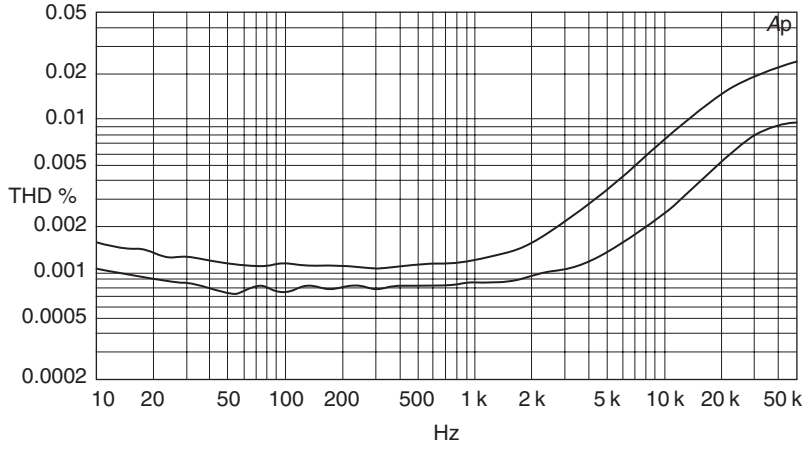


Figure 17.8: The effect on distortion of a prematurely active VI limiter; the lower trace has the limiter disabled. Output power 150 W/8 Ω

VI Limiting and Temperature Effects

The component values for the VI limiters, of whatever type, are most conveniently determined by use of a SPICE simulator and a certain amount of cut-and-try. However, when the values settled on are put into practice, the results are often disappointing, with the amplifier distortion performance being degraded by the VI limiters starting to act when they should in theory be firmly off. The effect is demonstrated in Figure 17.8, where it can be seen that distortion is roughly tripled when it rises above the noise.

VI limiters of the straightforward kind with the simple circuitry shown above depend on a voltage exceeding the sense transistor V_{be} to make them operate. They are therefore somewhat temperature sensitive, and this can be a real problem. The overload protection circuitry will almost always be close to the output devices being protected, and therefore near the heat-sink. It is therefore very likely to get hotter than other parts of the amplifier, and the VI limiters will begin to act much earlier than expected. A SPICE simulator, unless told otherwise, will default all its component temperatures to a nice comfy 25°C, and will not give warning of the problem.

Taking this temperature into account, and also ensuring that the VI limiters are firmly off in normal operation, so distortion performance is not degraded, means that the VI limiter component values differ significantly from those derived from room-temperature simulations. This implies that if the circuitry is designed not to come on early when the unit is hot, it will operate late when cold, if a fault occurs just after it has been switched on. The output stage must have enough capability to handle this without damage.

It is of course always possible to design more complex VI limiting circuitry that is less temperature sensitive. For example, the single sensing transistor could be replaced by a differential pair, which would be much less temperature sensitive.

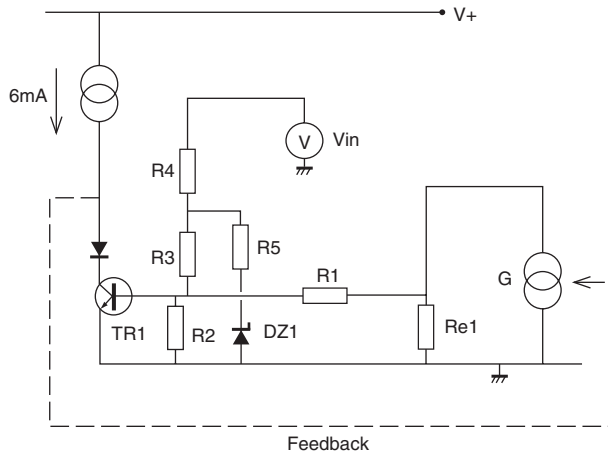


Figure 17.9: A conceptual model of an overload protection circuit that implements dual-slope limiting

More complex systems have been conceived, for example by Crown, who have patented several sophisticated systems that are effectively analog computers, but this approach has not been widely adopted. In systems designed to deal with fault conditions, simplicity, and therefore hopefully reliability, is a great virtue.

Simulating Overload Protection Systems

The calculations for protection circuitry can be time-consuming. Simulation is quicker; Figure 17.9 shows a conceptual model of a dual-slope VI limiter, which allows the simulated protection locus to be directly compared with the loadline and the SOA. The amplifier output stage is reduced to one-half (the positive or upper half) by assuming symmetry, and the combination of the actual output device and the load represented by voltage-controlled current source G. The output current from controlled source G is the same as the output device current in reality, and passes through current-sense resistor Re1.

The 6 mA current source *I* models the current from the previous stage that TR1 must shunt away from the output device. Usually this is an accurate model because the VAS collector load will indeed be a current source.

The feedback loop is closed by making the voltage at the collector of TR1 control the current flowing through G and hence Re1.

In this version of VI protection the device voltage is sensed by R4 and the current thus engendered is added to that from R1 at the base of TR1. This may seem a crude way of approximating a constant power curve, and indeed it is, but it provides very effective protection for low- and medium-powered amplifiers.

V_{in} models the positive supply rail, and exercises the simulation through the possible output voltage range. In reality the emitter of TR1 and Re1 would be connected to the amplifier output, which

would be moved up and down to vary the voltage across the output devices, and hence the voltage applied across R1, R2. Here it is easier to alter the voltage source V, as the only part of the circuit connected to HT + . V+ is fixed at a suitable HT voltage, e.g. +50V.

The simulation only produces the protection locus, and the other lines making up the SOA plot are added at the display stage. $I_{c(max)}$ is drawn by plotting a constant to give a horizontal line at 16A. $P_{(max)}$ is drawn as a line of a constant power, by using the equation $250/V_{ce}$ to give a 250W line. In PSPICE there seems to be no way to draw a strictly vertical line to represent $V_{ce(max)}$, but in the case of the MJ15024 this is 250V, and is for most practical purposes off the right-hand end of the graph anyway. The second-breakdown region is more difficult to show, for in the manufacturer's data the region is shown as bounded by a nonlinear curve. The voltage/current coordinates of the boundary were read from manufacturer's data, and approximately modeled by fitting a second-order polynomial. In this case it is:

$$I = 24.96 - 0.463 \cdot V_{ce} + 0.00224 \cdot V_{ce}^2 \qquad \text{Equation 17.2}$$

This is only valid for the portion that extends below the 250W constant-power line, at the bottom right of the diagram.

As previously mentioned, simulation results for protection circuitry must be carefully checked against reality because of temperature effects.

Testing the Overload Protection

One of the more nerve-wracking aspects of amplifier testing is the verification of the overload protection system. This is best done by slowly reducing the test load resistance from the rated value to one that is expected to trigger the overload, rather than wading straight in with a crowbar across the output terminals. If an amplifier has a rated load of 8Ω, then the protection might be expected to act at 2Ω, or perhaps 1Ω if the design is intended to deal authoritatively with deep dips in loudspeaker impedance.

Obviously there needs to be some way of monitoring that the VI limiters are beginning to act, and this may be as simple as observing the output waveform on an oscilloscope; when the peaks of the sine wave are starting to get clipped then limiting is occurring. You need to make sure you are not seeing voltage clipping because the supply rails have been dragged down. When you are sure that the VI limiters are coming in as expected, then, and not before, is it time to start applying short-circuits to the output. This approach minimizes the likelihood of output device damage. Blown output transistors are time-consuming to replace, with often quite a bit of dismantling involved, and there is also the likelihood of collateral damage to drivers and so on that has to be checked for and diagnosed; it is *much* more time efficient to take a gradualist approach to overloading the output stage. I have recently pursued this method with four different commercial amplifiers I designed, and in each case the verification procedure was completed without destroying a single transistor, a record of which I am mildly proud.

Complete verification includes overload testing with a 10% high mains supply voltage, and possibly at elevated ambient temperatures.

Speaker Short-Circuit Detection

Some amplifiers test the speaker outputs for short-circuits before unmuting and connecting the power amplifiers to them. This usually entails using a changeover contact configuration for the output muting relay, with the external load connected to the moving contact and the power amplifier output connected to the normally open contact. When the amplifier is muted, a very small current, too small to cause audible clicks, is passed through the normally closed contact and into the loudspeaker load. The resulting voltage is applied to a comparator and if it is too low the amplifier is inhibited from unmuting. This prevents the output devices from being unnecessarily stressed by a short-circuit. Since the test is only made when the amplifier is muted, the normal overload protection system must still be provided to cope with short-circuits that occur while the amplifier is unmuted.

An interesting failure mode with this scheme can occur if the test current is made too small. Loudspeakers can also turn sound into electricity instead of vice versa, and I know of one design that would refuse to start up in a noisy environment.

Catching Diodes

These are reverse-biased power diodes connected between the supply rails and the output of the amplifier, to allow it to absorb transients generated by fast current changes into an inductive load. They are also known as clamp diodes or clamping diodes. All moving-coil loudspeakers present an inductive impedance over some frequencies.

When an amplifier attempts to rapidly change the current flowing in an inductive load, the inductance can generate voltage spikes that drive the amplifier output outside its HT rail voltages; in other words, if the HT voltage is ± 50 V, then the output might be forced by the inductive back-EMF to 80 V or more, with the likelihood of failure of the reverse-biased output devices. Catching diodes prevent this by conducting and clamping the output so it cannot move more than about 1 V outside the HT rails. These diodes are presumably so called because they catch the output line if it attempts to move outside the rails.

So how can the output rail move outside the supply rails, no matter how fast the voltage change applied to a reactive load, if it is firmly held by the amplifier negative-feedback loop? The answer is that a flyback pulse typically occurs when the amplifier is suddenly *disconnected* from a reactive load, rather than when there is a sudden change in the signal. This happens when VI limiters cut in, turning off the half of the output stage that was until then driving the load. Now neither of the output devices are conducting, and this is when the voltage spike occurs and the clamp diodes justify their cost.

This sounds like a sharp crack of high amplitude; it is not a nice noise, but sometimes can be difficult to identify as it tends to happen during signal peaks. It can usually be more easily

diagnosed by looking at an oscilloscope, as the sudden voltage excursion is much steeper than the signal waveforms. The only way to avoid these noises – for the catching diodes only limit the spike amplitude rather than suppressing it altogether – is to make sure that the output stage is big enough for its task, so the VI limiters can be designed with a big margin between normal use with likely loads, and the fault conditions that make it essential for them to act.

The diode current rating should be not less than 2A, and the PIV 200V or greater, and at least twice the sum of the HT rails. I usually specify 400 PIV 3A diodes, and they never seem to fail.

DC Offset Protection

In some respects, any DC-coupled power amplifier is an accident waiting to happen. If the amplifier suffers a fault that causes its output to sit at a significant distance away from ground, then a large current is likely to flow through the loudspeaker system. This may cause damage either by driving the loudspeaker cones beyond their mechanical limits or by causing excessive thermal dissipation in the voice-coils, the latter probably being the most likely. In either case the financial loss is likely to be serious. There is also a safety issue, in that overheating of voice-coils or crossover components could cause a fire.

Since most power amplifiers consist of one global feedback loop, there are many possible component failures that could produce a DC offset at the output, and in most cases this will result in the output sitting at one of the HT rail voltages. The only way to save the loudspeaker system from damage is to remove this DC output as quickly as possible. The DC protection system must be functionally quite separate from the power amplifier itself or the same fault may disable both.

There are several possible ways to provide DC protection:

1. By fusing in the output line, the assumption being that a DC fault will give a sustained current flow that will blow the fuse when music-type current demands will not.
2. By means of a relay in the output line, which opens when a DC offset is detected.
3. By triggering a crowbar that shunts the output line to ground, and blows the HT fuses. The crowbar device is usually a triac, as the direction of offset to be dealt with is unpredictable.
4. By shutting down the power supply when a DC fault is detected. This can be done simply by an inhibit input if a switched-mode PSU is used. Conventional supplies are less easy.

DC Protection by Fuses

Fuses in series with the output line are sometimes recommended for DC offset protection, but their only merit is cheapness. It may be true that they have a slightly better chance of saving expensive loudspeakers than the HT fuses, but there are at least three snags:

1. Selection of the correct fuse size is not at all easy. If the fuse rating is small and fast enough to provide some real loudspeaker protection, then it is likely to be liable to nuisance blowing on

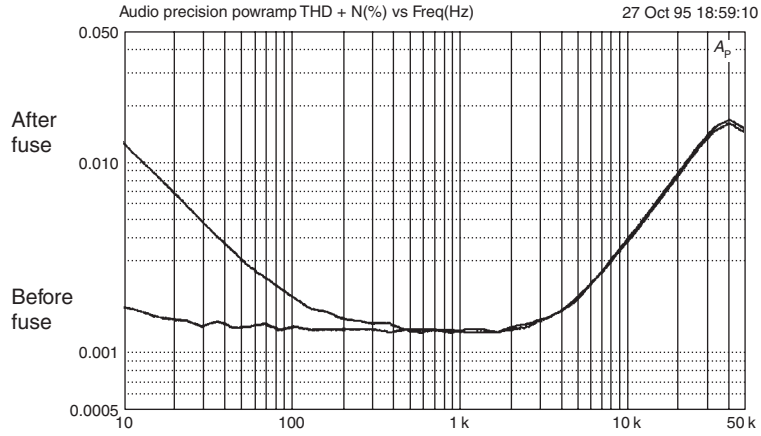


Figure 17.10: Fuse distortion. THD measured before and after the fuse at 25 W into 8 Ω

large bass transients. A good visual warning is given by behavior of the fuse wire; if this can be seen sagging on transients, then it is going to fail sooner rather than later. At least one writer on DIY Class-A amplifiers gave up on the problem, and coolly left the tricky business of fuse selection to the constructor!

2. Fuses running within sight of their nominal rated current generate distortion at LF due to cyclic changes in their resistance caused by I^2R heating; the THD would be expected to rise rapidly as frequency falls, and Greiner[6] states that harmonic and intermodulation distortion near the burn-out point can reach 4%. It should be possible to eradicate this by including the fuse inside the global feedback network, for the distortion will be generated at low frequencies where the feedback factor is at its greatest, but there are problems with amplifier behavior after the fuse has blown.

In my tests, the distortion generated was fairly pure third harmonic. Figure 17.10 shows the THD measured before and after a T1A (slow-blow) fuse in series with an 8 Ω load at 25 W. Below 100 Hz the distortion completely swamps that produced by the amplifier, reaching 0.007% at 20 Hz. The distortion rises at rather less than 6 dB/octave as frequency falls. The fuse in this test is running close to its rating, as increasing the power to 30 W caused it to blow.

3. Fuses obviously have significant resistance (otherwise they would not blow) so putting one in series with the output will degrade the theoretical damping factor. However, whether this is of any audible significance is very doubtful.

Note that the HT rail fuses, as opposed to fuses in the output line, are intended only to minimize amplifier damage in the event of output device failure. They must not be relied upon for speaker protection against DC offset faults. Often when one HT fuse is caused to blow the other also does so, but this cannot be relied upon, and obviously asymmetrical HT fuse blowing will in itself give rise to a large DC offset.

Relay Protection and Muting Control

Relay protection against DC offsets has the merit that, given careful relay selection and control-circuitry design, it is virtually foolproof. The relay should be of the normally open type so that if the protection fails it will be to a safe condition.

The first problem is to detect the fault condition as soon as possible. This is usually done by low-pass filtering the audio output, to remove all signal frequencies, before the resulting DC level is passed to a comparator that trips when a set threshold is exceeded. This is commonly in the range of 1–2 V, well outside any possible DC offsets associated with normal operation; these will almost certainly be below 100 mV. Any low-pass filter must introduce some delay between the appearance of the DC fault and the comparator tripping, but with sensible design this will be too brief to endanger normal loudspeakers. There are other ways of tackling the fault-detection problem, for example by detecting when the global negative feedback has failed, but the filtering approach appears to be the simplest method and is generally satisfactory. First-order filtering seems to be quite adequate, though at first sight a second-order active filter would give a faster response time for the same discrimination against false triggering on bass transients. In general there is much to be said for keeping protection circuitry as simple and reliable as possible.

Let us now examine DC offset detection circuitry in more detail. The problem falls neatly into two halves – distinguishing between acceptable large AC signals of up to 30 V rms or more, and DC offsets that may only be a volt or so before stern action is desired, and applying the result to a circuit that can detect both positive and negative transgressions. To perform the first task, relatively straightforward low-pass filtering is often adequate, but the bidirectional detection can be tackled in many ways, and sometimes presents a few unexpected problems.

At this point we might consider how quickly the DC offset protection must operate to be effective. Clearly there will always be some delay, as we are discriminating against normal high-amplitude bass information, but otherwise the quicker the better if the loudspeaker is to be saved. My experience of deliberately setting fire to loudspeaker elements is limited (and I hasten to point out that I have so far never set fire to one accidentally) but here is one test I can report.

I once had the entertaining task of determining just how long a speaker element – the LF unit, obviously, as the tweeter was protected by the crossover from any DC – could sustain an amplifier DC fault. The tests, which were conducted outdoors to avoid triggering the fire alarms, showed that a well-designed and conservatively rated loudspeaker could be turned into smouldering potential landfill in less than a second. The loudspeaker unit in question was a high-quality LF unit with the relatively small diameter of 5 inches, made by a respected manufacturer. The test involved applying +40 V to it, as if its accompanying amplifier had failed. The cone and voice-coil assembly shot out of the magnetic gap as if propelled by explosives, and then burst into flames in less than a second. All we could really conclude as the smoke cleared was that a second was way too long a reaction time for a protection system.

Filtering for DC Protection

A good DC protection filter is that which discriminates best between powerful low-frequency signals and a genuine amplifier problem. It is easy to make the filter time-constant so long that it will never be false-triggered by a thumping great bass note, but then its time-domain response will be so slow that your precious loudspeakers will be history before the amplifier reacts to protect them.

The simplest possible filter is a single-pole circuit that requires only one RC time-constant; in many cases this is quite good enough, but some more sophisticated approaches are also described here.

The single RC filter

The time-constant needs to be long enough to filter out the lowest frequency anticipated, at the full voltage output of the amplifier. The ability to sustain 10 Hz at the onset of clipping is usually adequate for audio, but if you are designing subsonic amplifiers to drive vibration tables, you will need to go a bit lower. Figure 17.11a shows the single-pole filter with typical values of 47 k and 47 μF that give a −3 dB point at 0.07 Hz. This is appropriate for low to medium amplifier powers, when feeding a later bidirectional detector that will trigger an offset of the order of 1 V. The value of R1 is set by the current demands of this later stage – these can be significant, as we will see in the next section. The value of C1 is then determined by the required −3 dB frequency, and this means that it will be an electrolytic. It is important to remember at this point that DC offsets may arrive with either polarity, and may persist for long periods before someone notices there is a problem, so C1 needs to be either a nonpolar electrolytic or constructed from two ordinary electrolytics connected back to back in the time-honored fashion. Both methods are effective so it comes down to the fine details of the economics of component sourcing. Some amplifiers remove the supply from the power amplifier sections, so the offset does not persist and this precaution may seem unnecessary; however, there is no point in trying to save fractions of a penny by possibly compromising the reliability of something as important as the DC offset protection. C1 should have a voltage rating at least equal to the supply rails of the amplifier concerned.

The single-pole filter in Figure 17.11a is −3 dB at 0.07 Hz. To evaluate it, it was fed from a power amplifier giving 55 V peak, and the filter output connected to a bidirectional detector that had trip points at ±2.0 V. This setup triggered at 2.0 Hz when a 55 V peak signal starting at 50 Hz was slowly reduced in frequency. This corresponds to a filter attenuation of −28.8 dB at 2.0 Hz, and this frequency was used as the criterion for bass rejection thereafter. When a fault was simulated so the input to the filter shot up to +55 V, and stayed there, the detector gave a DC offset indication after 78 ms.

This circuit is easily adapted to stereo usage by having two resistors feeding into it, as in Figure 17.11b. If the resistors remain the same value, then the resistance seen by C is halved, and its capacitance must be doubled to maintain the same roll-off frequency. The incoming DC offset is also halved, so the detector sensitivity must be doubled if it is to trigger from the same level of offset on one of the stereo amplifier outputs. You could also object that a positive offset on one channel might be canceled out by a negative offset on the other; this seems laughably unlikely

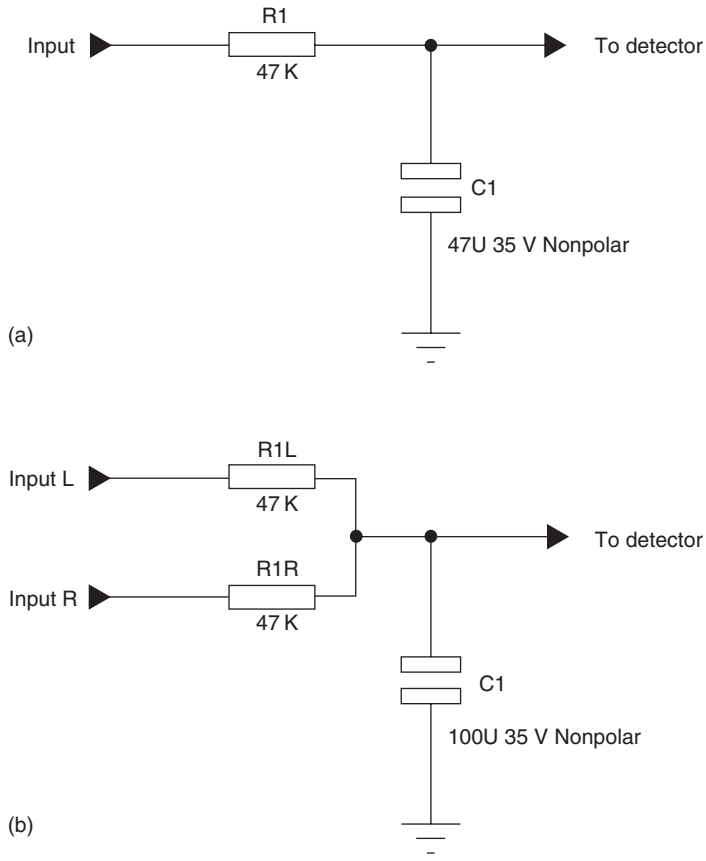


Figure 17.11: Mono and stereo single-pole filters for offset protection. The −3 dB point is 0.07 Hz

until you recall that bridged amplifiers are driven with input signals that are in anti-phase, so a DC error in the drive circuitry could present just this situation. More sophisticated circuits provide two independent inputs that do not interact, avoiding this problem. More on this later, in the section on detectors.

The dual RC filter

The thinking behind the use of more complicated filtering is that a faster response roll-off will give better discrimination against high-amplitude bass events, so a higher −3 dB frequency can be used with (hopefully) a quicker response in the time domain.

The simplest method is to cascade two single-pole RC filters, as shown in Figure 17.12. This obviously gives a rather soggy roll-off, but has the merit of not introducing any more semiconductors that might fail. The non-standard capacitor values shown give the same attenuation of −28.8 dB at 2.0 Hz as the previous circuit. The only real snag to this scheme is that it does not work. The time to react was 114 ms, half as long again as the simple filter above. However, I have seen it used in several designs, so you might come across it.

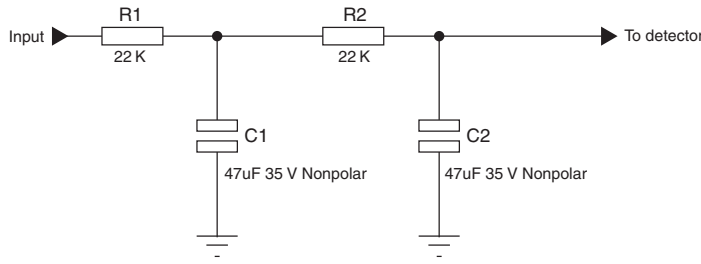


Figure 17.12: A dual RC filter offset protection

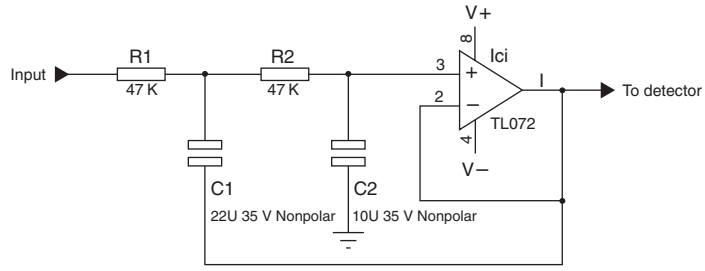


Figure 17.13: A second-order Sallen-and-Key filter input for offset protection

The second-order active filter

Some amplifier designs use an active filter to separate the bass from the breakdowns. This obviously allows a nice sharp roll-off, and gives the freedom to set the filter damping factors and so on. But does it deliver? I tested the circuit of Figure 17.13, a Sallen-and-Key configuration that with the values shown gives a second-order Butterworth (maximal flatness) characteristic, with a −3 dB point at 0.23 Hz; due to the increased filter slope the attenuation is once more −28.8 dB at 2.0 Hz. The reaction time is 109 ms, which is better than the dual RC filter yet somewhat inferior to the single-pole filter of Figure 17.11a – most disappointing. The Bessel filter characteristic is noted for a better response in the time domain, at the expense of a sharp roll-off, so I tried that. The component values in Figure 17.13 are now R1 = R2 = 35 k, C1 = 13.3 μF, and C2 = 10 μF. The reaction time is actually worse, at 131 ms, which was rather a surprise.

Building active filters usually means using op-amps. Putting an op-amp into the system creates a need for low-voltage supplies within the power amplifier, which is highly inconvenient if they do not exist already. Most protection designs use discrete transistors throughout, and one of the advantages of the Sallen-and-Key configuration is that it can be realized using a simple emitter-follower.

An important consideration is that op-amps have a limited common-mode voltage capability, and they will not appreciate having the full-power amplifier supply rail applied to them directly. It will be necessary to scale down the incoming voltages and allow for this when setting the detector thresholds.

The conclusion seems inescapable that, for once, the simplest circuit is the best: the single-pole filter is the way to go.

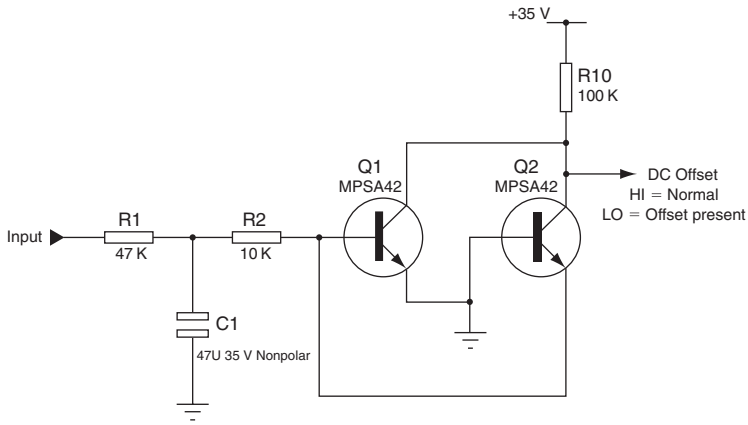


Figure 17.14: A common bidirectional detect circuit, giving very different thresholds for positive and negative inputs

Bidirectional DC Detection

There are many, many ways to construct circuits that will respond to both positive and negative signals of a defined level, and here some of the more common and more useful ones are examined.

The conventional two-transistor circuit

The circuit in Figure 17.14 is probably the most common approach to bidirectional detection. When the input exceeds +0.6 V, Q1 turns on and the output voltage falls while Q2 stays off. When the input goes negative, Q2 operates in common-base mode, and conducts, Q1 remaining off as its base–emitter junction is reverse-biased. In either case current is drawn through R10 and the output voltage drops to signal an offset. There is a certain elegance in the way that the conducting base–emitter junction protects its neighbor from excess reverse bias, but this circuit has one great disadvantage. Since Q2 operates in common-base mode, it has near-unity current gain, as opposed to Q1, which is in common-emitter mode and therefore has current gain equal to the device beta.

This makes the two thresholds very asymmetrical. When the detector is driven from a single-pole RC filter with R1 = 47 k, the positive threshold is +1.05 V but the negative threshold is −5.5 V. To reduce this asymmetry R1 needs to be kept low, which leads to inconveniently large values of C1.

The one-transistor version

Figure 17.15 shows a variation on this theme, saving a transistor by adding diodes and resistors. With current component pricing the economic benefit is trivial, but it is still a circuit that has seen a great deal of use in Japanese amplifier designs. For positive inputs, D1, Q1, and D2 conduct. For negative inputs, D3 and Q1 conduct, the latter getting its base current through R2. As for the previous circuit, the current-gain differences between common-base and common-emitter modes of transistor operation give asymmetrical thresholds, slightly less so because of the effect of D2 during positive inputs.

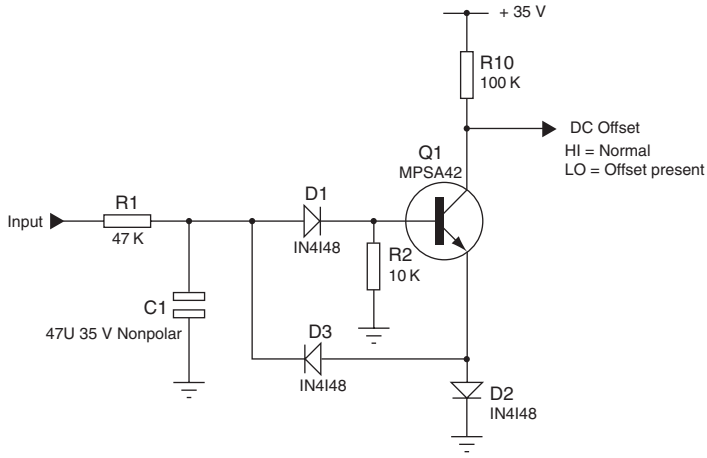


Figure 17.15: Another implementation of the same principle, saving a transistor but retaining the problem of asymmetrical thresholds

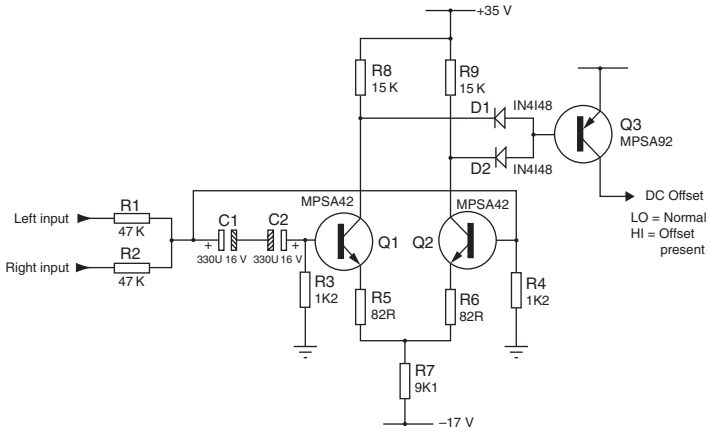


Figure 17.16: The differential detector, which can have very low thresholds. It uses a high-pass rather than a low-pass filter

The differential detector

The interesting circuit of Figure 17.16, which has also seen use in Japanese hi-fi equipment, is based on a differential pair. This removes the objection to all the other circuits here, which is that it takes 0.6V on the base to turn on a transistor directly, and so the detection thresholds will be that or more, due to extra diodes and so on. In this circuit the differential pair Q1, Q2 cancels out the 0.6V V_{be} drop, and sensitivity can be much higher; under what conditions this is actually necessary is a moot point. There is no low-pass filter as such; instead the same effect is achieved by high-pass filtering the signal, to remove DC and information. The result is then subtracted from the unfiltered signal by the differential pair, so only the DC and low-frequency signals remain.

It works like this: for positive inputs Q2 turns on more and Q1 less, so the voltage on Q2 collector falls and Q3 is turned on via D2, and passes an offset signal to the rest of the system. For negative

inputs Q1 turns on more and Q2 less, so the voltage on Q1 collector falls and Q3 is turned on via D1. The thresholds depend on the gain of the pair, set by the ratio of R5, R6 to R8, R9, and whatever voltage is set up on the Q3 emitter. The circuit gives excellent threshold symmetry.

The Self detector

Figure 17.17 shows my own version of a bidirectional detector. This has two advantages: it is symmetrical in its thresholds, and can be handily converted to an economical stereo or multichannel form without any loss of sensitivity. The only downside is that the thresholds are relatively high at about ±2.1 V with the component values shown. This is actually quite low enough to protect loudspeakers and, in any case, your typical serious amplifier fault smacks the output hard against the supply rails, and detecting this is not very hard. The exactness of the threshold symmetry depends on the properties of the transistors used, but is more than good enough to eliminate any problems. It works well with transistors such as MPSA42/MPS92, which are designed for high-voltage applications and therefore have low beta.

For positive inputs, D1, Q1, and Q2 conduct, with D4 supplying the base current for Q2. With a negative input, D2, Q2, and Q1 conduct, D3 now supplying base current for Q1. In each case there are two diode drops and two V_{be} drops in series, which if each one was a nominal 0.6 V would give thresholds of ±2.4 V; in practice the diode supplying the much smaller base current has a lesser voltage across it, and the real thresholds come to ±2.1 V. Note that R11 is very definitely required to limit the current flowing through Q1 and Q2 when the input goes negative; R10 inherently limits it for positive inputs.

Figure 17.18 shows the stereo version, which uses separate filters for each channel, and two more diodes. The operation is exactly as before for each channel, and so the thresholds are unchanged.

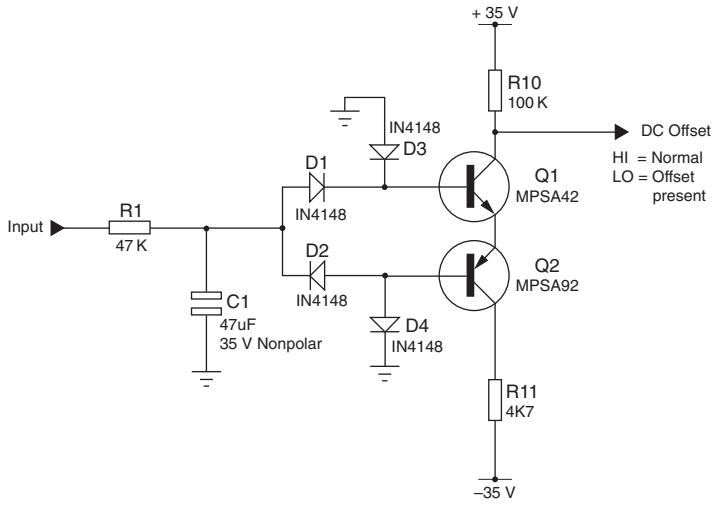


Figure 17.17: The Self detector – good symmetry and easily expandable for more channels

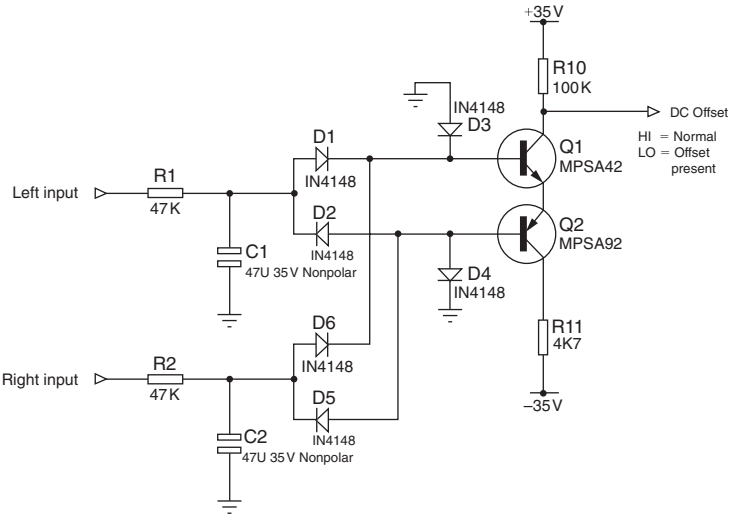


Figure 17.18: The stereo version of Figure 17.17

Equal-value positive and negative offsets on the two inputs do not cancel, and an offset is always clearly signaled.

Having paid for a DC protection relay, it seems only sensible to use it for system muting as well, to prevent thuds and bangs from the upstream parts of the audio system from reaching the speakers at power-up and power-down. Most power amplifiers, being dual-rail (i.e. DC-coupled), do not generate enormous thumps themselves, but they cannot be guaranteed to be completely silent and will probably produce an audible turn-on thud.

An amplifier relay-control system should:

- Leave the relay de-energized when muted. At power-up, there should be a delay of at least 1 second before the relay closes. This can be increased if required.
- Drop out the relay as fast as is possible at power-down, to stop the dying moans of the pre-amp, etc. from reaching the outside world (see the section on mains-fail detection below for more details).
- Drop out the relay as fast as is possible when a DC offset of more than 1–2 V, in either direction, is detected at the output of either power amp channel; the exact threshold is not critical. This is normally done by low-pass filtering the output (47 k and 47 μF works OK) and applying it to some sort of absolute-value circuit to detect offsets in either direction. The resulting signal is then OR-ed in some way with the muting signal mentioned above.
- Do not forget that the contacts of a relay have a much lower current rating for breaking DC rather than AC. This is an issue that does not seem to have attracted the attention it deserves.

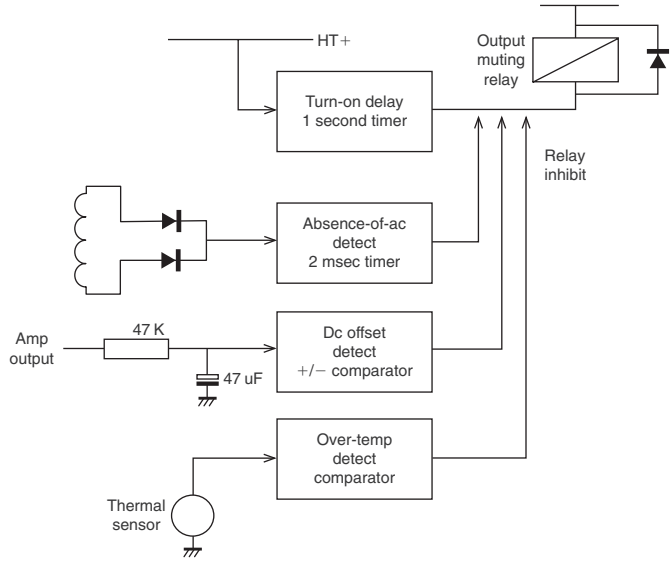


Figure 17.19: Output relay control combining DC offset protection and power-on/off muting

A block diagram of a relay control system meeting the above requirements is shown in Figure 17.19, which includes over-temperature protection. Any of the three *inhibit* signals can override the turn-on delay and pull out the relay.

Distortion in Output Relays

Relays remain the only simple and effective method of disconnecting an amplifier from its load. The contacts can carry substantial currents, and it has been questioned whether they can introduce nonlinearities.

My experience is that silver-based contacts in good condition show effectively perfect linearity. Take atypical relay intended by its manufacturer for output-switching applications, with 'silver alloy' contacts – whatever that means – rated at 10 A. Figure 17.20 shows THD before and after the relay contacts while driving an 8 Ω load to 91 W, giving a current of 3.4 A rms. There is no significant difference; the only reason that the lines do not fall exactly on top of each other is because of the minor bias changes that Class-B is prone to. This apparently perfect linearity can be badly degraded if the contacts have been maltreated by allowing severe arcing – typically while trying and failing to break a severe DC fault.

Not everyone is convinced of this. If the contacts were nonlinear for whatever reason, an effective way of dealing with it would be to include them in the amplifier feedback loop, as shown in Figure 17.21. R1 is the main feedback resistor and R2 is a subsidiary feedback path that remains closed when the relay contacts open, and hopefully prevents the amplifier from going completely berserk. With the values shown the normal gain is 15.4 times, and with the contacts open it is 151 times. There is a feedback factor of about 10 to linearize any relay problems.

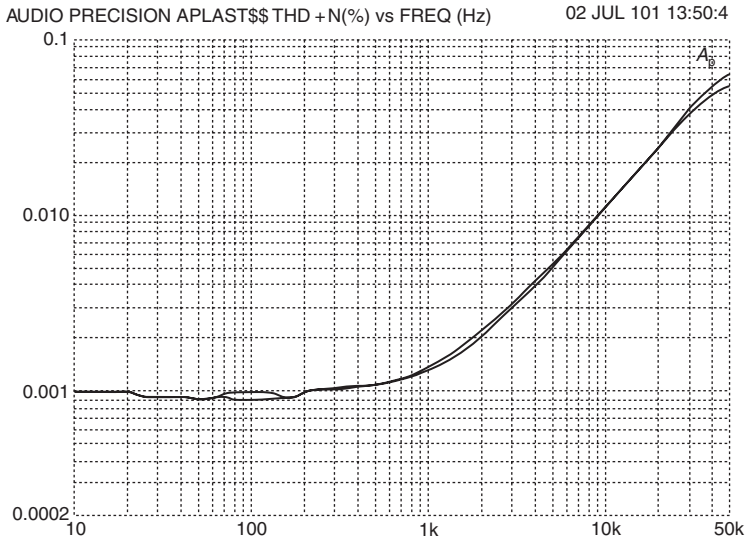


Figure 17.20: Relay contacts in themselves are completely distortion-free. Current through contacts was 3.4 A rms

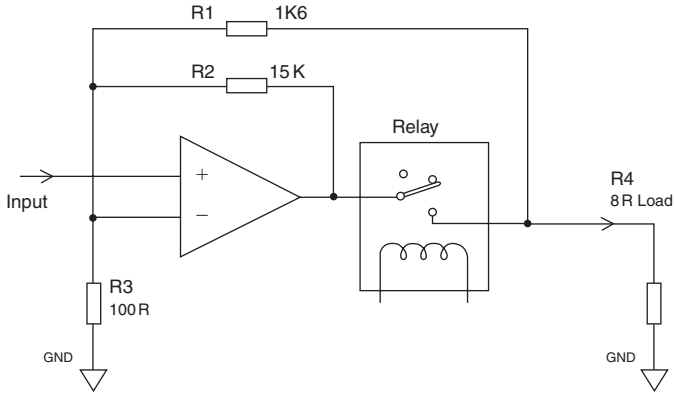


Figure 17.21: How to enclose relay contacts in the feedback loop. The gain shoots up when the relay contacts open, so muting the input signal is desirable

The problem of course is that if there is to be a healthy amount of NFB wrapped around the relay contacts, R2 must be fairly high and so the closed-loop gain shoots up. If there is still an input signal, then the amplifier will be driven heavily into clipping. Some designs object to this, but even if the amplifier does not fail it is likely to accumulate various DC offsets on its internal time-constants as a result of heavy clipping, and these could cause unwanted noises when the relay contacts close again. One solution to this is a muting circuit at the amplifier input that removes the signal entirely and prevents clipping. This need not be a sophisticated circuit, as huge amounts of muting are not required; −40 dB should be enough. It must, however, pass the signal cleanly when not muting.

An all too real form of nonlinearity can occur if the relay is constructed so that its frame makes up part of the switched electrical circuit as well as the magnetic circuit. (This is not the case with the

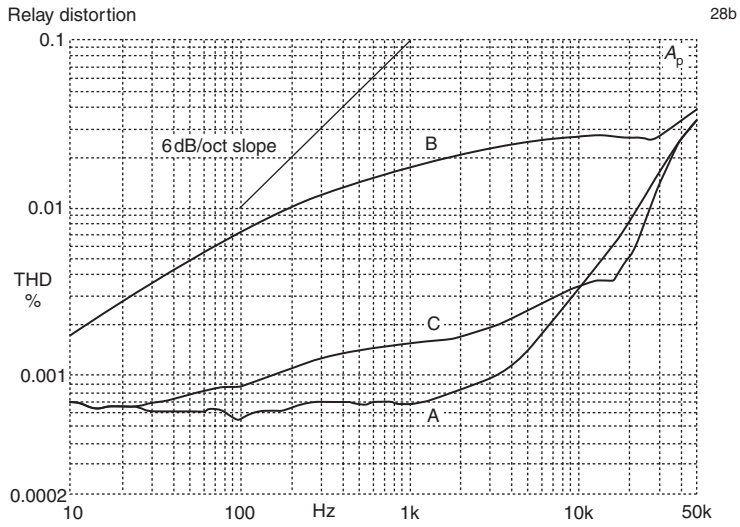


Figure 17.22: Trace A is amplifier distortion alone, B total distortion with power relay in circuit. Trace C shows that enclosing the relay in the feedback loop is not a complete cure

audio application relay discussed above.) A relay frame is made of soft iron, to prevent it becoming permanently magnetized, and this appears to present a nonlinear resistance to a loudspeaker level signal, presumably due to magnetization and saturation of the material. (It should be said at once that this is described by the manufacturer as a 'power relay' and is apparently not intended for audio use.) A typical example of this construction has massive contacts of silver/cadmium oxide, rated at 30 amps AC, which in themselves appear to be linear. However, used as an amplifier output relay, this component generates more – much more – distortion than the power amplifier it is associated with.

The effect increases with increasing current; 4.0 A rms passing through the relay gives 0.0033% THD and 10 A rms gives 0.018%. The distortion level appears to increase with the square of the current. Experiment showed that the distortion was worst where the frame width was narrowest, and hence the current density greatest.

Figure 17.22 shows the effect at 200 W rms/2 Ω (i.e. with 10 A rms through the load) before and after the relay. Trace A is the amplifier alone. This is a Blameless amplifier and so THD is undetectable below 3 kHz, being submerged in the noise floor, which sets a measurement limit of 0.0007%.

Trace B adds in the extra distortion from the relay. It seems to be frequency-dependent, but rises more slowly than the usual slope of 6 dB/octave. Trace C shows the effect of closing the relay in the NFB loop using the circuit and component values of Figure 17.21; the THD drops to about a tenth, which is what simple NFB theory would predict. Note that from 10 to 35 kHz the distortion is now lower than before the relay was added; this is due to fortuitous cancelation of amplifier and relay distortion.

Figure 17.23 was obtained by sawing a 3 mm by 15 mm piece from a relay frame and wiring it in series with the amplifier output, by means of copper wires soldered at each end. As before the level

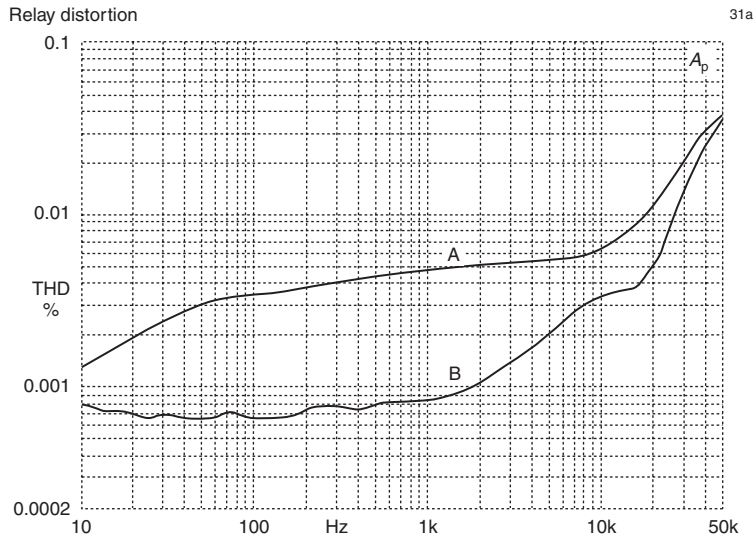


Figure 17.23: Trace A here is total distortion with a sample of the power relay frame material wired in circuit. Trace B is the same, enclosed in the feedback loop as before

was 200W rms/2Ω, i.e. 10A rms. Trace A is the raw extra distortion; this is lower than shown in Figure 17.22 because the same current is passing through less of the frame material. Trace B is the result of enclosing the frame fragment in the NFB loop exactly as before. This removes all suspicion of interaction with coil or contacts and proves it is the actual frame material itself that is nonlinear.

Wrapping feedback around the relay helps but, as usual, is not a complete cure. Soldering on extra wires to the frame to bypass as much frame material as possible is also useful, but it is awkward and there is the danger of interfering with proper relay operation. No doubt any warranties would be invalidated. Clearly it is best to avoid this sort of relay construction if you possibly can, but if high-current switching is required, more than an audio-intended relay can handle, the problem may have to be faced.

Output Crowbar DC Protection

Since relays are expensive and require control circuitry, and fuse protection is very doubtful, there has for at least two decades been interest in simpler and wholly solid-state solutions to the DC-protection problem. The circuit of Figure 17.24 places a triac across the output, the output signal being low-pass filtered by R and C. If sufficient DC voltage develops on C to fire the diac, it triggers the triac, shorting the amplifier output to ground.

While this approach has the merit of simplicity, in my experience it has proved wholly unsatisfactory. The triac needs to be very big indeed if it is to work more than once, because it must pass enough current to blow the HT rail fuses. If these fuses were omitted the triac would have to dump the entire contents of a power-supply reservoir capacitor to ground through a low total resistance, and the demands on it become quite unreasonable.

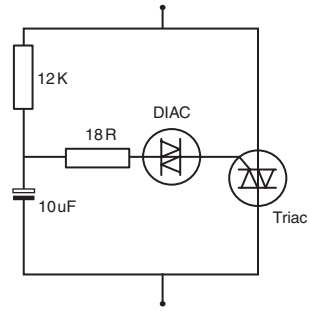


Figure 17.24: Output crowbar DC protection

An output crowbar is also likely to destroy the output devices; the assumption behind this kamikaze crowbar system is presumably that the DC offset is due to blown output devices, and a short across the output can do no more harm. This is quite wrong, because any fault in the small-signal part of the amplifier will also cause the output to saturate positive or negative, with the output devices in perfect working order. The operation of the crowbar under these circumstances may destroy the output devices, for the overload protection may not be adequate to cope with such a very direct short-circuit.

Protection by Power-Supply Shutdown

Conventional transformer power supplies can be shut down quickly by firing crowbar SCRs across the supply rails; this overcomes one of the objections to output crowbars, as collateral damage to other parts of the circuit is unlikely, assuming of course you are correctly trying to blow the DC rail fuses and not the transformer secondary fuses. The latter option would severely endanger the bridge rectifier, and the crowbar circuitry would have to handle enormous amounts of energy as it emptied the reservoir capacitors. Even blowing the DC fuses will require SCRs with a massive peak-current capability.

A conventional power supply can also be shut down by using relays to open the DC power supply rails. In a two-channel amplifier sharing one power supply and one offset-detect circuit this is relatively straightforward, and one two-make relay can do the job. If the two channels have separate power supplies then they will also need separate offset-detect circuits and separate rail-switching relays to avoid the need for a four-make relay, which is not a common component in high-current sizes.

Amplifiers have been designed with power switching relays between the transformer secondary and the rectifier. This is not a good plan as this circuit contains large charging-current pulses, which are likely to cause excessive I^2R heating of the relay contacts.

If your amplifier is powered by a switch-mode supply, it may well have a logic input that gives the option of near-instant shutdown. This can be connected to a DC-detect low-pass filter, and the occurrence of a DC error then gives an apparently foolproof shutdown of everything.

There are (as usual) snags to this. Firstly, the high relative cost of switch-mode supplies means that almost certainly a single supply will be shared between two or more amplifier channels, and so

both channels are lost if one fails. This is not a serious problem for domestic use, as few people are going to want to carry on listening to one channel of a stereo source. It may, however, be a disadvantage in sound-reinforcement applications. Secondly, and more worryingly, this provides very dubious protection against a fault in the supply itself. If such a fault causes one of the HT rails to collapse, then it may well also disable the shutdown facility, and all protection is lost.

Thermal Protection

This section deals only with protecting the output semiconductors against excessive junction temperature; the thermal safeguarding of the mains transformer is dealt with in Chapter 9.

Output devices that are fully protected against excess current, voltage, and power are by no means fully safeguarded. Most electronic overload protection systems allow the devices to dissipate much more power than in normal operation; this can and should be well inside the rated capabilities of the component itself, but this gives no assurance that the increased dissipation will not cause the heat-sink to eventually reach such temperatures that the crucial junction temperatures are exceeded and the device fails. If no temperature protection is provided this can occur after only a few minutes' drive into a short. Heat-sink over-temperature may also occur if ventilation slots, etc. are blocked, or heat-sink fins covered up.

The solution is a system that senses the heat-sink temperature and intervenes when it reaches a preset maximum. This intervention may be in the form of:

1. Causing an existing muting/DC-protection relay to drop out, breaking the output path to the load. If such a relay is fitted, then it makes sense to use it.
2. Muting or attenuating the input signal so the amplifier is no longer dissipating significant power.
3. Removing the power supply to the amplifier sections. This normally implies using a bimetallic thermal switch to break the mains supply to the transformer primary, as anywhere downstream of here requires two lines to be broken simultaneously, e.g. the positive and negative HT rails.

Each of these actions may be either self-resetting or latching, requiring the user to initiate a reset. The possibility that a self-resetting system will cycle on and off for long periods, subjecting the output semiconductors to severe temperature changes, must be borne in mind. Such thermal cycling can greatly shorten the life of semiconductors. In an attempt to address this issue, some IC power amplifiers mute and unmute very rapidly, almost on a per-cycle basis. The rationale behind this is that while the output devices never have time to cool down much, this is actually a good thing as sustained high temperatures are less damaging to device reliability than thermal cycling.

The two essential parts of a thermal protection system are the temperature-sensing element and whatever arrangement performs the intervention. While temperature can be approximately sensed by silicon diodes, transistor junctions, etc., these typically require some sort of setup or calibration procedure, due to manufacturing tolerances. This is wholly impractical in production, for it

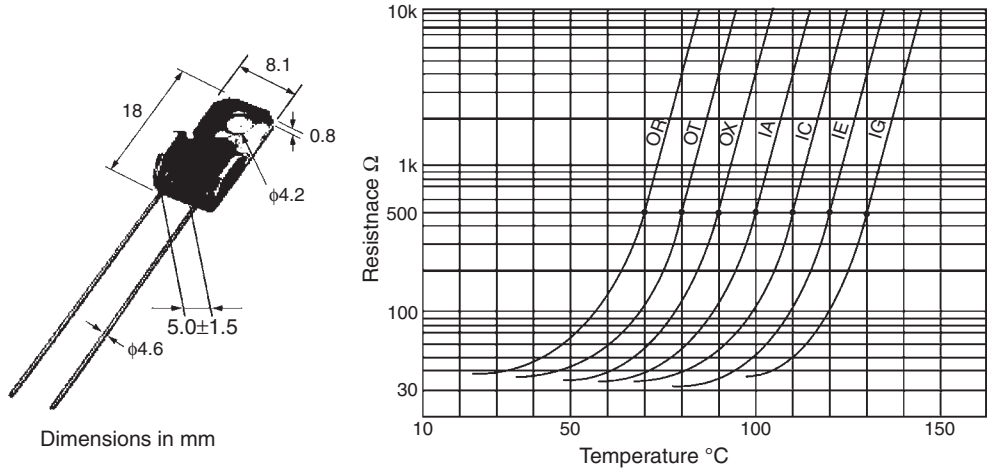


Figure 17.25: The physical package and the resistance–temperature curves of a suitable thermistor for over-temperature protection

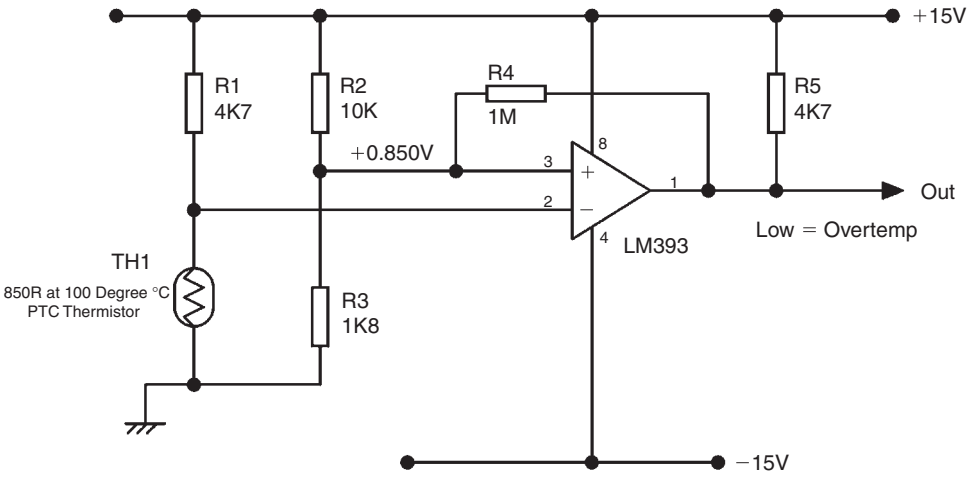


Figure 17.26: Over-temperature protection circuit using a positive thermistor and comparator

requires the heat-sink (which normally has substantial thermal inertia) to be brought up to the critical temperature before the circuit is adjusted. This not only takes considerable time, but also requires the output devices to reach a temperature at which they are somewhat endangered.

Until relatively recently I would have put thermistors into the same category, but they have now improved to the point where thermistor-based temperature protection systems can be made sufficiently accurate without worrying about calibration. They can also be obtained with a screw-on tab that makes mounting much easier.

Figure 17.25 shows the physical package and the resistance–temperature curves for a range of positive thermistors (i.e. parts where the resistance increases with temperature) made by Nichicon. A standard negative thermistor reduces its resistance as temperature increases. Figure 17.26 shows

a comparator-based over-temperature detector using a positive thermistor. This circuit uses an LM393 comparator; it is configured for the output to go low when the set temperature is exceeded. The thermistor is in the bottom arm of the potential divider R1, TH1, and the reference voltage is established by potential divider R2, R3 at +0.85 V. With the thermistor chosen this represents a cut-out temperature of 100°C. R4 provides a little positive feedback to the reference divider, to create hysteresis around the switching point. The application of hysteresis is particularly important when comparators rather than op-amps are used as they have no internal compensation and are more thus prone to oscillation. Good rail decoupling is also important.

The LM393 comparator has an open-collector output and so the pull-up resistor R5 is required. If it would be more convenient for the output to go high when the set temperature is exceeded, the positions of the thermistor and reference divider could be swapped, but there is then a problem with the application of hysteresis as the resistance of the thermistor varies widely over the operating temperature range, and so the source resistance of the potential divider R1, TH1 also varies widely, and the amount of hysteresis is therefore temperature-dependent, which is not helpful to the design process.

The strong variation of thermistor resistance with temperature means that it is hard to significantly change the temperature cut-out setting by changing the reference voltage. For example, if the rightmost thermistor type in Figure 17.25 was used to implement cut-out at 150°C, it would be difficult to change that to 90°C because at that temperature the thermistor resistance has flattened out at around 30 Ω, and shows little variation with temperature. It would be necessary to change the thermistor type, which is fine unless you already have 100,000 in stock.

If greater accuracy or more flexibility is required, the best method is the use of integrated temperature sensors that do not require any calibration. A good example is the National Semiconductor LM35DZ, a three-terminal device that outputs 10 mV for each °C above zero. Without any calibration procedure, the output voltage may be compared against a fixed reference. In Figure 17.27 an op-amp is used as a comparator, and this time the circuit is configured for the output to go high when the set temperature is exceeded. The resulting over-temperature signal is used to pull out the muting relay.

The output of the LM35 is applied to the non-inverting input of the op-amp via R3, which in combination with R4 gives a little positive feedback that introduces hysteresis into the switching point and prevents dithering or oscillation of the output; the amount of hysteresis is constant because the output impedance of the LM35 is low and not temperature-dependent, unlike the thermistor sensor used in the previous circuit. When the output from the LM35 exceeds the reference voltage set up by the divider R1, R2, the op-amp output switches from low to high, and this signal goes off to the rest of the control circuitry. The reference in this example is set to a voltage of +0.95 V, corresponding to a cut-out temperature of 95°C; because of the linear sensor output this can be easily altered over a very wide range by changing the value of R2. The assumption is made here that the op-amp supply rails will be regulated by the usual methods, and if so this will be more than accurate enough for setting the reference voltage.

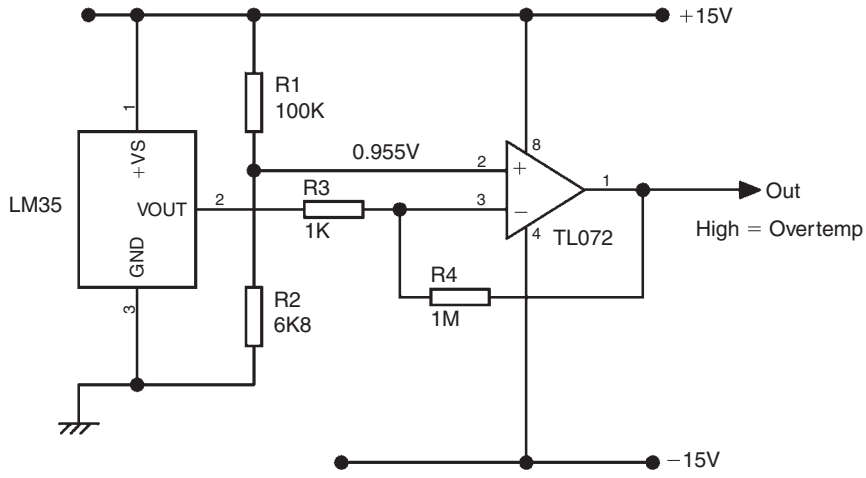


Figure 17.27: Over-temperature protection circuit using an LM35 temperature sensor and an op-amp

If the circuit is required to operate in the other sense (i.e. output low = over-temp.) to interface with the rest of the protection system, then the op-amp inputs can simply be interchanged, and the hysteresis must be applied to the reference divider instead. The LM35 approach gives the most accurate and trouble-free temperature protection in my experience. It allows a wide variation in setting without changing components. The downside is that IC temperature sensors are significantly more expensive than thermistors, etc., and harder to mount.

Another pre-calibrated type of temperature sensor is the thermal switch, which usually operates on the principle of a bistable bimetallic element. These should not be confused with thermal fuses, which are once-only components that open the circuit by melting an internal fusible alloy; they are in common use for transformer thermal protection, as they are cheaper than the self-resetting thermal switches. The trouble with thermal fuses is that they are relatively uncommon, and the chance of a blown thermal fuse being replaced with the correct component in the field is not high.

The physical positioning of the temperature sensor requires some thought. In an ideal world we would judge the danger to the output devices by assessing the actual junction temperature; since this is difficult to do the sensor must get as close as it can. It is shown elsewhere that the top of a TO-3 transistor can get hotter than the flange, and as for quiescent biasing sensors, the top is the best place for the protection sensor. This does, however, present some mechanical problems in mounting. This approach may not be equally effective with plastic flat-pack devices such as TO-3P, for the outer surface is an insulator; however, it still gets hotter than the immediately adjacent heat-sink, and this sensor position undoubtedly gives the fastest response.

Alternatively, the protection sensor can be mounted on the main heat-sink, which is mechanically much simpler but imposes a considerable delay between the onset of device heating and the sensor reacting. For this reason a heat-sink-mounted sensor will normally need to be set to a lower trip temperature, usually in the region of 80°C, than if it is device-mounted. The more closely the

sensor is mounted to the devices, the better they are protected. If two amplifiers share the same heat-sink, the sensor should be placed between them; if it was placed at one end the remote amplifier would suffer a long delay between the onset of excess heating and the sensor acting.

One well-known make of PA amplifiers implements (or used to) temperature protection by mounting a thermal switch in the live mains line on top of one of the TO-3 cans in the output stage. This gains the advantage of fast response to dangerous temperatures, but there is the obvious objection that lethal mains voltages are brought right into the center of the amplifier circuitry, where they are not normally expected, and this represents a real hazard to service personnel.

Mains-Fail Detection

In the section on relay control above it was noted that one of the functions of the control circuitry is to mute the amplifier at power-down by opening the output relays. Almost all audio circuitry will produce thumps and bangs at some point as the supply rails collapse; if large reservoir capacitors are involved this can be quite a lengthy process. It is therefore important to include a circuit that can detect loss of mains power quickly – before the voltage on the reservoirs in either the main amplifier or auxiliary op-amp supplies can fall significantly.

The most common method of mains-fail detection is to feed a small capacitor via a rectifier from the transformer secondary. The capacitor is shunted with a resistance and the resulting RC time-constant sized so that the capacitor voltage falls much more quickly than that on any of the main power-supply reservoirs. The principle is shown in Figure 17.28. In this case the main-fail circuit was driven from a transformer secondary that powered a +5V regulator powering a housekeeping PIC microcontroller. The mains-fail signal was fed directly to a PIC input port pin.

Note that half-wave rectification is used. R1, R2, and R3 are chosen so that Q1 stays on at least down to the minimum mains operating voltage (which is usually determined by the power supply regulators dropping out) but turns off promptly when the mains supply is removed. It is important

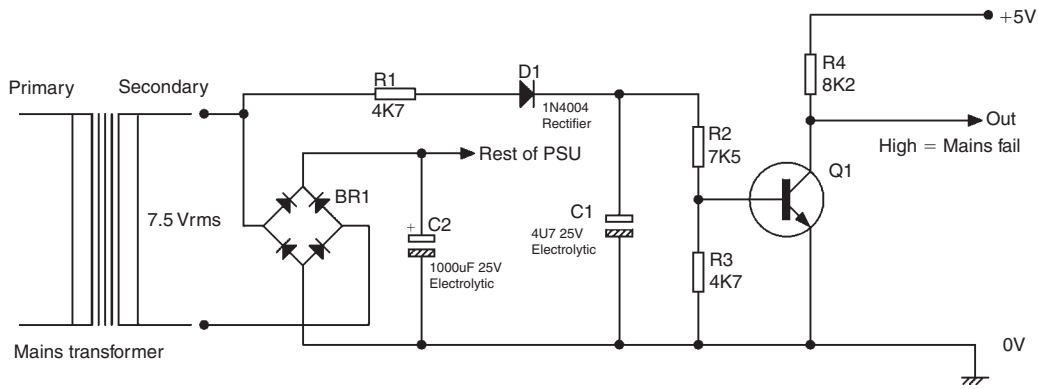


Figure 17.28: A simple mains-fail detection circuit, as used in a commercial power amplifier I designed a while ago

to remember that there will be a substantial ripple voltage on the mains-fail capacitor C1, and if it is an electrolytic you must check that the ripple current will not lead to excessive heating.

A more sophisticated technique for mains-fail detection is a 2 ms (or thereabouts) timer that is held reset by the AC on the mains transformer secondary, except for a brief period around the AC zero-crossing, which is not long enough for the timer to trigger. When the incoming AC disappears, the near-continuous reset is removed, the timer fires, and the relay is dropped out within 10 ms. This will be long before the various reservoir capacitors in the system can begin to discharge. However, if the mains switch contacts are generating RF that is in turn reproduced as a click by the pre-amp, then even this method may not be fast enough to mute it.

A simple way to implement this is shown in Figure 17.29. C4, D3, and Q3 implement the output relay turn-on delay. Q1 is always on except during zero-crossings, and keeps C3 discharged. If the mains fails C3 charges rapidly via R5 and turns on Q2, which quickly discharges C4, turning off the output relays. Note that this circuit is designed to run off one of the power amplifier supply rails and does not require an auxiliary power supply.

The Zeners across the output relays require a word of explanation. When a relay is driven by a transistor, a reversed diode across the coil is required to prevent the abrupt turn-off of current making the coil voltage reverse, driving the collector more negative. I have measured −120 V, enough to destructively exceed the V_{ceo} of most transistors.

This protection conceals a lurking snag; relay drop-out time is hugely increased by the reversed diode, as it provides a path for coil current to circulate while the magnetic field decays. It takes roughly five times longer, which is really not what we need here. This is a good point to stop and consider exactly what we need to do: the aim is not 'suppress all back-EMF' but 'protect the transistor'. If the back-EMF is clamped to say −27 V for a 24 V relay by a suitable Zener diode in series with the reverse diode, the circulating current stops much sooner, and drop-out is almost as fast as for the non-suppressed relay. It is speeded up by a factor of about 4 on moving from

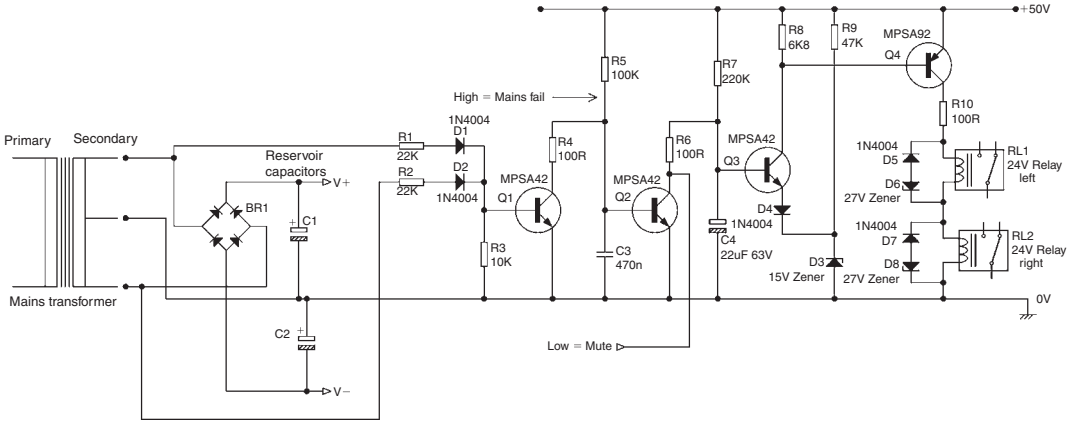


Figure 17.29: A more sophisticated mains-fail detection circuit that gives a faster response to the removal of the power

conventional protection to Zener clamping. For relays of the usual size, a 500 mW Zener appears to be adequate. If this circuit is interfaced to a DC offset detector and an over-temperature circuit, via the input to Q2 collector, it implements the complete protection system in Figure 17.19.

The surprisingly complex and subtle subject of relay control is dealt with in much more detail in my book *Self On Audio*[7].

Very fast mains-fail detection does carry dangers. It can react to transient voltage drops on the mains that would otherwise be ignored by the equipment, enforcing the full power-up delay every time a glitch comes along. This will not be well received by the user. In one case, the prototype of a power amplifier of my design was being evaluated at home by one of my colleagues, and it showed just this behavior, shutting down and restarting at fairly regular intervals. It was installed in an old house with an old refrigerator in the kitchen, and a little investigation showed that every time the rather large refrigerator motor started, it very briefly dragged the mains voltage down to less than half its normal value, due to the high resistance of the elderly house wiring. Fortunately the product in question did not emit noises the instant the power was removed, and reverting to a simpler and slower mains-fail detector as shown in Figure 17.28 solved the problem (and saved a few pennies in parts).

On the whole, it is best, if possible, to design equipment so that it does not produce prompt noises on power-down, allowing the more tolerant (and cheaper and simpler) means of mains-fail detection to be used.

Powering Auxiliary Circuitry

Whenever it is necessary to power auxiliary circuitry, such as the relay control system described above, there is an obvious incentive to use the main HT rails. A separate PSU requires a bridge rectifier, reservoir capacitor, fusing, and an extra transformer winding, all of which will cost a significant amount of money.

The main disadvantage is that the HT rails are at an inconveniently high voltage for powering control circuitry. For low-current sections of this circuitry, such as relay timing, the problem is not serious as the same high-voltage small-signal transistors can be used as in the amplifier small-signal sections, and the power dissipation in collector loads, etc. can be controlled simply by making them higher in value. The biggest problem is the relay energizing current; many relay types are not available with coil voltages higher than 24 V, and this is not easy to power from a 50 V HT rail without wasting power in a big dropper resistor. This causes unwanted heating of the amplifier internals, and provides a place for service engineers to burn themselves.

One solution in a stereo amplifier is to run the two relays in series; the snag (and for sound reinforcement work it may be a serious one) is that both relays must switch together, so if one channel fails with a DC offset, both are muted. In live work independent relay control is much to be preferred, even though most of the relay control circuitry must be duplicated for each channel.

If the control circuitry is powered from the main HT rails, then its power must be taken off *before* the amplifier HT fuses. The control circuitry will then be able to mute the relays when appropriate, no matter what faults have occurred in the amplifiers themselves.

If there is additional signal circuitry in the complete amplifier it is not advisable to power it in this way, especially if it has high gain, e.g. a microphone preamplifier. When such signal circuits are powered in this way, it is usually by ±15V regulators from the HT rails, with series dropper resistors to spread out some of the dissipation. However, bass transients in the power amplifiers can pull down the HT rails alarmingly, and if the regulators drop out large disturbances will appear on the nominally regulated low-voltage rails, leading to very-low-frequency oscillations that will be extremely destructive to loudspeakers. In this case the use of wholly separate clean rails run from an extra transformer secondary is strongly recommended. There will be no significant coupling between the supplies due to the use of a shared transformer primary.

References

[1] A. Bailey, Output transistor protection in AF amplifiers, Wireless World (June 1968) p. 154.

[2] R. Becker, High-power audio amplifier design, Wireless World (February 1972) p. 79.

[3] Motorola, High power audio amplifiers with sort circuit protection, Motorola Application Note AN-485, 1972.

[4] M. Otala, Peak current requirement of commercial loudspeaker systems, JAES 35 (June 1987) p. 455.

[5] P. Baxandall, Technique for displaying current and voltage capability of amplifiers, JAES 36 (January/February 1988) p. 3.

[6] R. Greiner, Amplifier–loudspeaker interfacing, JAES 28 (5) (May 1980) pp. 310–315.

[7] D. Self, Self On Audio, second ed., Newnes, 2006, p. 421.

CHAPTER 18

Grounding, Cooling, and Layout

Audio Amplifier PCB Design

This section addresses the special PCB design problems presented by power amplifiers, particularly those operating in Class-B. All power amplifier systems contain the power-amp stages themselves, and usually associated control and protection circuitry; most also contain small-signal audio sections such as balanced input amplifiers, subsonic filters, output meters, and so on.

Other topics that are related to PCB design, such as grounding, safety, reliability, etc., are also dealt with.

The performance of an audio power amplifier depends on many factors, but in all cases the detailed design of the PCB is critical, because of the risk of inductive distortion due to crosstalk between the supply rails and the signal circuitry; this can very easily be the ultimate limitation on amplifier linearity, and it is hard to overemphasize its importance. The PCB design will to a great extent define both the distortion and crosstalk performance of the amplifier.

Apart from these performance considerations, the PCB design can have considerable influence on ease of manufacture, ease of testing and repair, and reliability. All of these issues are addressed below.

Successful audio PCB layout requires enough electronic knowledge to fully appreciate the points set out below, so that layout can proceed smoothly and effectively. It is common in many electronic fields for PCB design to be handed over to draughtspersons who, while very skilled in the use of CAD, have little or no understanding of the details of circuit operation. In some fields this works fine; in power amplifier design it will not, because basic parameters such as crosstalk and distortion are so strongly layout-dependent. At the very least the PCB designer should understand the points set out below.

Crosstalk

All crosstalk has a transmitting end (which can be at any impedance) and a receiving end, usually either at high impedance or virtual earth. Either way, it is sensitive to the injection of small currents. When interchannel crosstalk is being discussed, the transmitting and receiving channels are usually called the speaking and nonspeaking channels, respectively.

Crosstalk comes in various forms:

- Capacitive crosstalk is due to the physical proximity of different circuits, and may be represented by a small notional capacitor joining the two circuits. It usually increases at the rate of 6 dB/octave, though higher dB/octave rates are possible. Screening with any conductive material is a complete cure, but physical distance is usually cheaper.
- Resistive crosstalk usually occurs simply because ground tracks have a non-zero resistance. Copper is not a room-temperature superconductor. Resistive crosstalk is constant with frequency.
- Inductive crosstalk is rarely a problem in general audio design; it might occur if you have to mount two uncanned audio transformers close together, but otherwise you can usually forget it. The notable exception to this rule is the Class-B audio power amplifier, where the rail currents are half-wave sines that seriously degrade the distortion performance if they are allowed to couple into the input, feedback or output circuitry.

In most line-level audio circuitry the primary cause of crosstalk is unwanted capacitive coupling between different parts of a circuit, and in most cases this is defined solely by the PCB layout. Class-B power amplifiers, in contrast, should suffer very low or negligible levels of crosstalk from capacitive effects, as the circuit impedances tend to be low and the physical separation large; a much greater problem is inductive coupling between the supply-rail currents and the signal circuitry. If coupling occurs to the same channel it manifests itself as distortion, and can dominate amplifier nonlinearity. If it occurs to the other (nonspeaking) channel it will appear as crosstalk of a distorted signal. In either case it is thoroughly undesirable, and precautions must be taken to prevent it.

The PCB layout is only one component of this, as crosstalk must be both emitted and received. In general the emission is greatest from internal wiring, due to its length and extent; wiring layout will probably be critical for best performance and needs to be fixed by cable ties, etc. The receiving end is probably the input and feedback circuitry of the amplifier, which will be fixed on the PCB. Designing these sections for maximum immunity is critical to good performance.

Rail Induction Distortion

The supply rails of a Class-B power-amp carry large and very distorted currents. As previously outlined, if these are allowed to crosstalk into the audio path by induction the distortion performance will be severely degraded. This applies to PCB conductors just as much as cabling, and it is sadly true that it is easy to produce an amplifier PCB that is absolutely satisfactory in every respect but this one, and the only solution is another board iteration. The effect can be completely prevented but in the present state of knowledge I cannot give detailed guidelines to suit every constructional topology. The best approach is the following.

Minimize radiation from the supply rails by running the V+ and V− rails as close together as possible. Keep them away from the input stages of the amplifier and the output connections; the

best method is to bring the rails up to the output stage from one side, with the rest of the amplifier on the other side. Then run tracks from the output to power the rest of the amp; these carry no half-wave currents and should cause no problems.

Minimize pickup of rail radiation by keeping the area of the input and feedback circuits to a minimum. These form loops with the audio ground and these loops must be as small in area as possible. This can often best be done by straddling the feedback and input networks across the audio ground track, which is taken across the center of the PCB from input ground to output ground.

Induction of distortion can also occur into the output and output-ground cabling, and even the output inductor. The latter presents a problem as it is usually difficult to change its orientation without a PCB update.

Mounting Output Devices on the Main PCB

The most important decision is whether or not to mount the power output devices directly on the main amplifier PCB. There are strong arguments for doing so, but it is not always the best choice.

Advantages

- The amplifier PCB can be constructed so as to form a complete operational unit that can be thoroughly tested before being fixed into the chassis. This makes testing much easier, as there is access from all sides; it also minimizes the possibility of cosmetic damage (scratches, etc.) to the metalwork during testing.
- It is impossible to connect the power devices wrongly, providing you get the right devices in the right positions. This is important for such errors usually destroy both output devices and cause other domino-effect faults that are very time-consuming to correct.
- The output device connections can be very short. This seems to help stability of the output stage against HF parasitic oscillations.

Disadvantages

- If the output devices require frequent changing (which obviously indicates something very wrong somewhere) then repeated resoldering will damage the PCB tracks. However, if the worst happens the damaged track can usually be bridged out with short sections of wire, so the PCB need not be scrapped; make sure this is possible.
- The output devices will probably get fairly hot, even if run well within their ratings; a case temperature of 90°C is not unusual for a TO-3 device. If the mounting method does not have a degree of resilience, then thermal expansion may set up stresses that push the pads off the PCB.
- The heat-sink will be heavy, and so there must be a solid structural fixing between this and the PCB. Otherwise the assembly will flex when handled, putting stress on soldered connections.

Single- and Double-Sided PCBs

Single-sided PCBs are the usual choice for power amplifiers, because of their lower cost; however, the price differential between single- and double-sided plated-through-hole (PTH) is much less than it used to be. It is not usually necessary to go double-sided for reasons of space or convoluted connectivity, because power amplifier components tend to be physically large, determining the PCB size, and in typical circuitry there are a large number of discrete resistors, etc., that can be used for jumping tracks.

Bear in mind that single-sided boards need thicker tracks to ensure adhesion in case desoldering is necessary. Adding one or more ears to pads with only one track leading to them gives much better adhesion, and is highly recommended for pads that may need resoldering during maintenance; unfortunately it is a very tedious task with most CAD systems.

The advantages of double-sided PTH for power amplifiers are as follows:

- No links are required.
- Double-sided PCBs may allow one side to be used primarily as a ground plane, minimizing crosstalk and EMC problems.
- Much better pad adhesion on resoldering as the pads are retained by the through-hole plating.
- There is more total room for tracks, and so they can be wider, giving less voltage drop and PCB heating.
- The extra cost is small.

Power-Supply PCB Layout

Power-supply subsystems have special requirements due to the very high capacitor-charging currents involved:

- Tracks carrying the full supply-rail current must have generous widths. The board material used should have not less than 2-oz copper. Four-ounce copper can be obtained but it is expensive and has long lead times – not really recommended.
- Reservoir capacitors must have the incoming tracks going directly to the capacitor terminals; likewise the outgoing tracks to the regulator must leave from these terminals. In other words, do not run a tee off to the cap. Failure to observe this puts sharp pulses on the DC and tends to worsen the hum level.
- The tracks to and from the rectifiers carry charging pulses that have a considerably higher peak value than the DC output current. Conductor heating is therefore much greater due to the higher value of I^2R. Heating is likely to be especially severe at PC-mount fuseholders. Wire links may also heat up and consideration should be given to two links in parallel; this sounds crude but actually works very effectively.

Track heating can usually be detected simply by examining the state of the solder mask after several hours of full-load operation; the green mask materials currently in use discolor to brown on heating. If this occurs then as a very rough rule the track is too hot. If the discoloration tends to dark brown or black then the heating is serious and must definitely be reduced.

- If there are PCB tracks on the primary side of the mains transformer, and this has multiple taps for multi-country operation, then remember that some of these tracks will carry much greater currents at low-voltage tappings; mains current drawn on 90 V input will be nearly three times that at 240 V.

Be sure to observe the standard safety spacings for creepage and clearance between mains tracks and other conductors (see Chapter 19 for the spacings required).

(This applies to all track–track, track–PCB edge, and track–metal fixing spacings.)

In general PCB tracks carrying mains voltages should be avoided, as presenting an unacceptable safety risk to service personnel. If it must be done, then warnings must be displayed very clearly on both sides of the PCB. Mains-carrying tracks are unacceptable in equipment intended to meet UL regulations in the USA, unless they are fully covered with insulating material that is non-flammable and can withstand at least 120 °C (e.g. polycarbonate).

Power Amplifier PCB Layout Details

A simple unregulated supply is assumed.

- Power amplifiers have heavy currents flowing through the circuitry, and all of the requirements for power supply design also apply here. Thick tracks are essential and 2-oz copper is highly desirable, especially if the layout is cramped.

If attempting to thicken tracks by laying solder on top, remember that ordinary 60:40 solder has a resistivity of about six times that of copper, so even a thick layer may not be very effective.

- The positive and negative rail reservoir caps will be joined together by a thick earth connection; this is called reservoir ground (RG). *Do not* attempt to use any point on this track as the audio-ground star-point, as it carries heavy charging pulses and will induce ripple into the signal. Instead take a thick tee from the center of this track (through which the charging pulses will not flow) and use the end of this as the star-point.
- Low-value resistors in the output stage are likely to get very hot in operation – possibly up to 200°C. They must be spaced out as much as possible and kept from contact with components such as electrolytic capacitors. Keep them away from sensitive devices such as the driver transistors and the bias-generator transistor.
- Vertical power resistors. The use of these in power amplifiers appears at first attractive, due to the small amount of PCB area they take up. However, the vertical construction means that any impact on the component, such as might be received in normal handling, puts a

very great strain on the PCB pads, which are likely to be forced off the board. This may result in it being scrapped. Single-sided boards are particularly vulnerable, having much lower pad adhesion due to the absence of bias.

- Solderable metal clips to strengthen the vertical resistors are available in some ranges (e.g. Vitrohm) but this is not a complete solution, and the conclusion must be that horizontal-format power resistors are preferable.
- Rail-decoupler capacitors must have a separate ground return to the reservoir ground. This ground must *not* share any part of the audio ground system, and must *not* be returned to the star-point (see Figure 18.1).
- The exact layout of the feedback take-off point is critical for proper operation. Usually the output stage has an 'output rail' that connects the emitter power resistors together. This carries the full output current and must be substantial. Take a tee from this track for the output connection, and attach the feedback take-off point to somewhere along this tee. *Do not* attach it to the track joining the emitter resistors.
- The input stages (usually a differential pair) should be at the other end of the circuitry from the output stage. Never run input tracks close to the output stage. Input stage ground and the ground at the bottom of the feedback network must be the same track running back to the star-point. No decoupling capacitors, etc. may be connected to this track, but it seems to be permissible to connect input bias resistors, etc. that pass only very small DC currents.
- Put the input transistors close together. The closer the temperature match, the less the amplifier output DC offset due to V_{be} mismatching. If they can both be hidden from 'seeing' the infrared radiation from the heat-sink (for example by hiding them behind a large electrolytic) then DC drift is reduced.
- Most power amplifiers will have additional control circuitry for muting relays, thermal protection, etc. Grounds from this must take a separate path back to reservoir ground, and *not* the audio star-point.
- Unlike most audio boards, power amps will contain a mixture of sensitive circuitry and a high-current power supply. Be careful to keep bridge-rectifier connections, etc. away from input circuitry.
- Mains/chassis ground will need to be connected to the power amplifier at some point. Do not do this at the transformer center-tap as this is spaced away from the input ground voltage by the return charging pulses, and will create severe ground loop hum when the input ground is connected to mains ground through another piece of equipment.

Connecting mains ground to the star-point is better, as the charging pulses are excluded, but the track resistance between input ground and star will carry any ground-loop currents and induce a buzz.

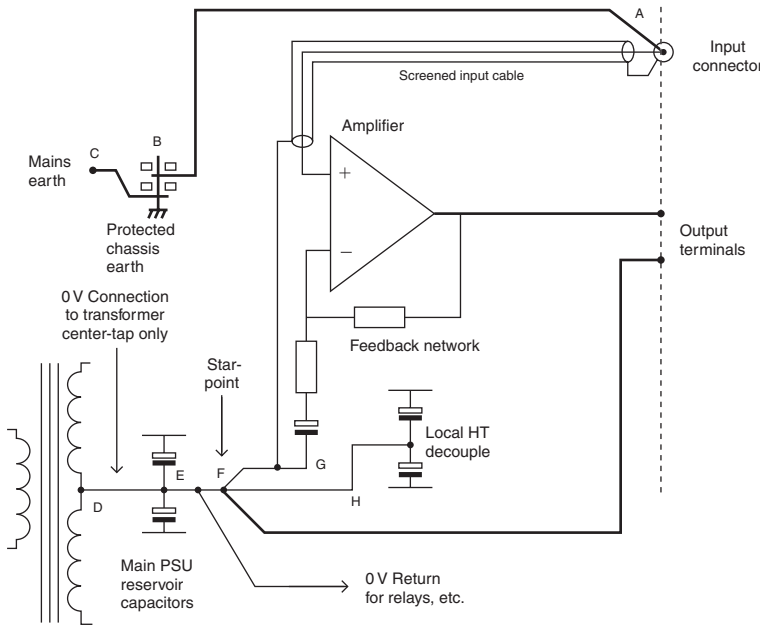


Figure 18.1: Grounding system for a typical power amplifier

Connecting mains ground to the input ground gives maximal immunity against ground loops.

- If capacitors are installed the wrong way round the results are likely to be explosive. Make every possible effort to put all capacitors in the same orientation to allow efficient visual checking. Mark polarity clearly on the PCB, positioned so it is still visible when the component is fitted.
- Drivers and the bias generator are likely to be fitted to small vertical heat-sinks. Try to position them so that the transistor numbers are visible.
- All transistor positions should have emitter, base, and collector or whatever marked on the top print to aid fault-finding. TO-3 devices need also to be identified on the copper side, as any screen-printing is covered up when the devices are installed.
- Any wire links should be numbered to make it easier to check they have all been fitted.

The Audio PCB Layout Sequence

PCB layout must be considered from an early stage of amplifier design. For example, if a front-facial layout shows the volume control immediately adjacent to a loudspeaker routing switch, then a satisfactory crosstalk performance will be difficult to obtain because of the relatively high impedance of the volume-control wipers. Shielding metalwork may be required for satisfactory performance and this adds cost. In many cases the detailed electronic design has an effect on crosstalk, quite independently from physical layout.

(a) Consider implications of facia layout for PCB layout.

(b) Circuitry designed to minimize crosstalk. At this stage try to look ahead to see how op-amp halves, switch sections, etc. should be allocated to keep signals away from sensitive areas. Consider crosstalk at above-PCB level; for example, when designing a module made up of two parallel double-sided PCBs, it is desirable to place signal circuitry on the inside faces of the boards, and power and grounds on the outside, to minimize crosstalk and maximize RF immunity.

(c) Facia components (pots, switches, etc.) placed to partly define available board area.

(d) Other fixed components such as power devices, driver heat-sinks, input and output connectors, and mounting holes placed. The area left remains for the purely electronic parts of the circuitry that do not have to align with metalwork, etc. and so may be moved about fairly freely.

(e) Detailed layout of components in each circuit block, with consideration towards manufacturability.

(f) Make efficient use of any spare PCB area to fatten grounds and high-current tracks as much as possible. It is not wise to fill in every spare corner of a prototype board with copper as this can be time-consuming (depending on the facilities of your PCB CAD system) and some of it will probably have to be undone to allow modifications.

Ground tracks should always be as thick as practicable. Copper is free (once you've bought the laminate, that is).

Miscellaneous Points

- On double-sided PCBs, copper areas should be solid on the component side, for minimum resistance and maximum screening, but will need to be cross-hatched on the solder side to prevent distortion if the PCB is flow-soldered. A common standard is 10 thou wide non-copper areas, i.e. mostly copper with small square holes; this is determined in the CAD package. If in doubt consult those doing the flow-soldering.
- Do not bury component pads in large areas of copper, as this causes soldering difficulties.
- There is often a choice between running two tracks into a pad, or taking off a tee so that only one track reaches it. The former is better because it holds the pad more firmly to the board if desoldering is necessary. This is *particularly important* for components like transistors that are relatively likely to be replaced; for single-sided PCBs it is absolutely vital.
- If two parallel tracks are likely to crosstalk, then it is beneficial to run a grounded screening track between them. However, the improvement is likely to be disappointing, as electrostatic lines of force will curve over the top of the screen track.
- Jumper options must always be clearly labeled. Assume everyone loses the manual the moment they get it.

- Label pots and switches with their function on the screen-print layer, as this is a great help when testing. If possible, also label circuit blocks, e.g. 'DC offset detect'. The labels must be bigger than component ident text to be clearly readable.

Amplifier Grounding

The grounding system of an amplifier must fulfill several requirements, amongst which are:

- The definition of a *star-point* as the reference for all signal voltages.
- In a stereo amplifier, grounds must be suitably segregated for good crosstalk performance. A few inches of wire as a shared ground to the output terminals will probably dominate the crosstalk behavior.
- Unwanted AC currents entering the amplifier on the signal ground, due to external ground loops, must be diverted away from the critical signal grounds, i.e. the input ground and the ground for the feedback arm. Any voltage difference between these last two grounds appears directly in the output.
- Charging currents for the PSU reservoir capacitors must be kept out of all other grounds.

Ground is the point of reference for all signals, and it is vital that it is made solid and kept clean; every ground track and wire must be treated as a resistance across which signal currents will cause unwanted voltage drops. The best method is to keep ground currents apart by means of a suitable connection topology, such as a separate ground return to the star-point for the local HT decoupling, but when this is not practical it is necessary to make every ground track as thick as possible, and fattened up with copper at every possible point. It is vital that the ground path has no necks or narrow sections, as it is no stronger than the weakest part. If the ground path changes board side then a single via-hole may be insufficient, and several should be connected in parallel. Some CAD systems make this difficult, but there is usually a way to fool them.

Power amplifiers rarely use double-insulated construction and so the chassis and all metalwork must be permanently and solidly grounded for safety; this aspect of grounding is covered in Chapter 19. One result of permanent chassis grounding is that an amplifier with unbalanced inputs may appear susceptible to ground loops. One solution is to connect audio ground to chassis only through a 10 Ω resistor, which is large enough to prevent loop currents becoming significant. This is not very satisfactory as:

- The audio system as a whole may thus not be solidly grounded.
- If the resistor is burnt out due to misconnected speaker outputs, the audio circuitry is floating and could become a safety hazard.
- The RF rejection of the power amplifier is likely to be degraded. A 100 nF capacitor across the resistor may help.

A better approach is to put the audio-chassis ground connection at the input connector, so in Figure 18.1, ground-loop currents must flow through A–B to the protected earth at B, and then to mains ground via B–C. They cannot flow through the audio path E–F. This topology is very resistant to ground loops, even with an unbalanced input; the limitation on system performance in the presence of a ground loop is now determined by the voltage drop in the input cable ground, which is outside the control of the amplifier designer. A balanced input could in theory cancel out this voltage drop completely.

Figure 18.1 also shows how the other grounding requirements are met. The reservoir charging pulses are confined to the connection D–E, and do not flow through E–F, as there is no other circuit path. E–F–H carries ripple, etc., from the local HT decouplers, but likewise cannot contaminate the crucial audio ground A–G.

Ground Loops: How They Work and How to Deal with Them

A ground loop is created whenever two or more pieces of mains-powered equipment are connected together, so that mains-derived AC flows through shields and ground conductors, degrading the noise floor of the system. The effect is worst when two or more units are connected through mains ground as well as audio cabling, and this situation is what is normally meant by the term 'ground loop'. However, ground currents can also flow in systems that are not galvanically grounded; they are of lower magnitude but can still degrade the noise floor, so this scenario is also considered here.

The ground currents may either be inherent in the mains supply wiring (see 'Hum injection by mains grounding currents' below) or generated by one or more of the pieces of equipment that make up the audio system (see sections 'Hum injection by transformer stray magnetic fields' and 'Hum injection by transformer stray capacitance' below).

Once flowing in the ground wiring, these currents will give rise to voltage drops that introduce hum and buzzing noises. This may occur either in the audio interconnects, or inside the equipment itself if it is not well designed (see section 'Ground currents inside equipment' on page 492).

Here I have used the word 'ground' for conductors and so on, while 'earth' is reserved for the damp crumbly stuff into which copper rods are thrust.

Hum Injection by Mains Grounding Currents

Figure 18.2 shows what happens when a so-called 'technical ground' such as a buried copper rod is attached to a grounding system that is already connected to 'mains ground' at the power distribution board. The latter is mandatory both legally and technically, so one might as well accept this and denote it as the reference ground. In many cases this 'mains ground' is actually the neutral conductor, which is only grounded at the remote transformer substation. A–B is the cable from substation to consumer, which serves many houses from connections tapped off along its length. There is substantial current flowing down the N + E conductor, so point B is often 1 V rms or more above earth. From B onwards, in the internal house wiring, neutral and ground are always separate (in the UK, anyway).

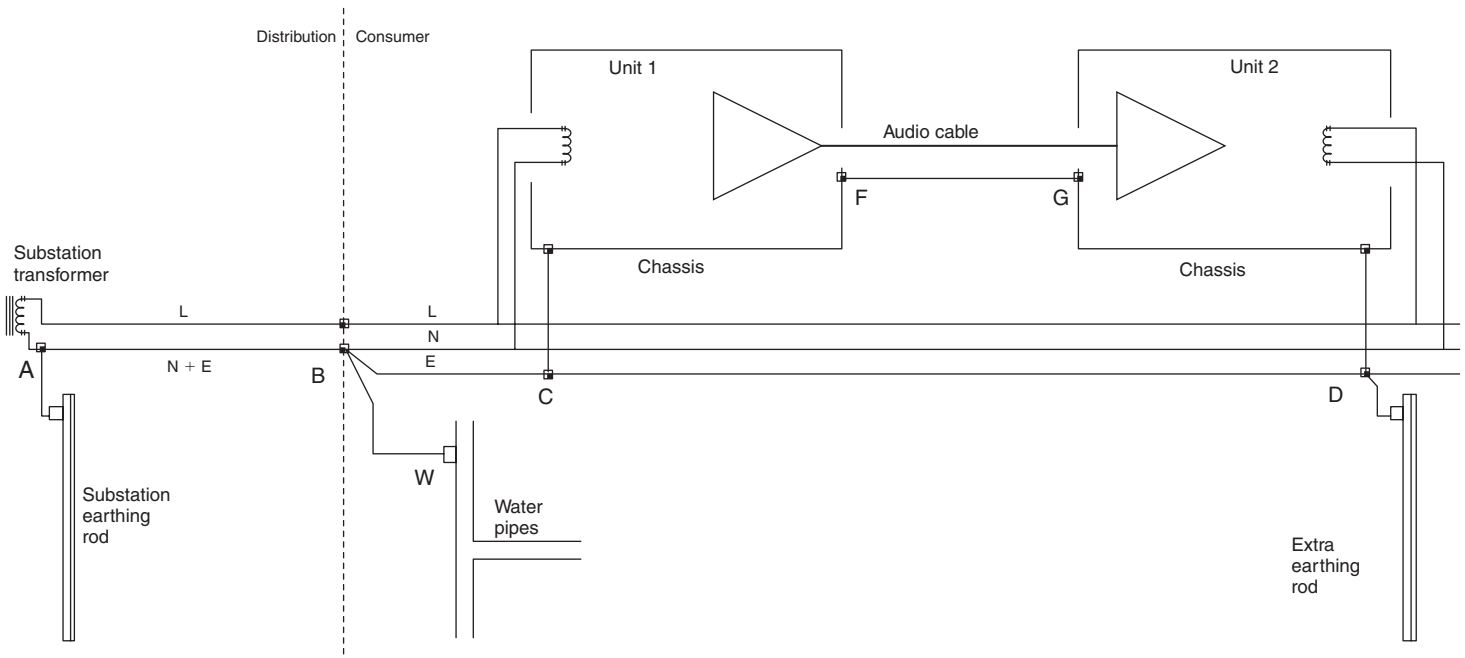


Figure 18.2: The pitfalls of adding a 'technical ground' to a system that is already grounded via the mains

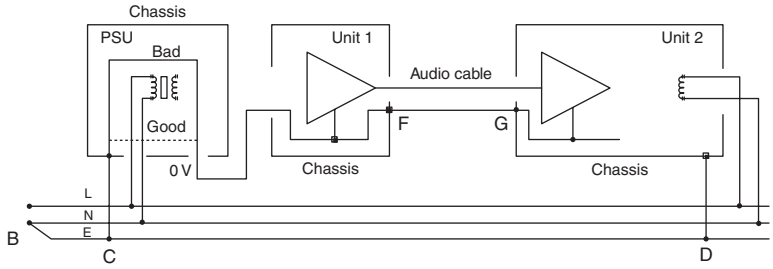


Figure 18.3: Poor cable layout in the PSU at left wraps a loop around the transformer and induces ground currents

Two pieces of audio equipment are connected to this mains wiring at C and D, and joined to each other through an unbalanced cable F–G. Then an ill-advised connection is made to earth at D; the 1 V rms is now impressed on the path B–C–D, and substantial current is likely to flow through it, depending on the total resistance of this path. There will be a voltage drop from C to D, its magnitude depending on what fraction of the total BCDE resistance is made up by the section C–D. The earth wire C–D will be of at least 1.5 mm^2 cross-section, and so the extra connection F–G down the audio cable is unlikely to reduce the interfering voltage much.

To get a feel for the magnitudes involved, take a plausible ground current of 1 A. The 1.5 mm^2 ground conductor will have a resistance of 0.012 Ω/m, so if the mains sockets at C and D are 1 m apart, the voltage C–D will be 12 mV rms. Almost all of this will appear between F and G, and will be indistinguishable from wanted signal to the input stage of Unit 2, so the hum will be severe, probably only 30 dB below the nominal signal level.

The best way to solve this problem is not to create it in the first place. If some ground current is unavoidable then the use of balanced inputs (or ground-cancel outputs – it is not necessary to use both) should give at least 40 dB of rejection at audio frequencies.

Figure 18.2 also shows a third earthing point, which fortunately does not complicate the situation. Metal water pipes are bonded to the incoming mains ground for safety reasons, and since they are usually electrically connected to an incoming water supply, current flows through B–W in the same way as it does through the copper rod link D–E. This water-pipe current does not, however, flow through C–D and cannot cause a ground-loop problem. It may, however, cause the pipes to generate an AC magnetic field that is picked up by other wiring.

Hum Injection by Transformer Stray Magnetic Fields

Figure 18.3 shows a thoroughly bad piece of physical layout that will cause ground currents to flow even if the system is correctly grounded to just one point.

Here Unit 1 has an external DC power supply; this makes it possible to use an inexpensive frame-type transformer that will have a large stray field. But note that the wire in the PSU that connects mains ground to the outgoing 0 V takes a half-turn around the transformer, and significant current will be

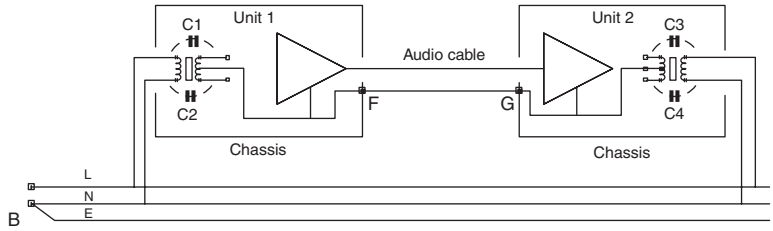


Figure 18.4: The injection of mains current into the ground wiring via transformer interwinding capacitance

induced into it, which will flow round the loop C–F–G–D, and give an unwanted voltage drop between F and G. In this case reinforcing the ground of the audio interconnection is likely to be of some help, as it directly reduces the fraction of the total loop voltage that is dropped between F and G.

It is difficult to put any magnitudes to this effect because it depends on many imponderables such as the build quality of the transformer and the exact physical arrangement of the ground cable in the PSU. If this cable is rerouted to the dotted position in the diagram, the transformer is no longer enclosed in a half-turn, and the effect will be much smaller.

Hum Injection by Transformer Stray Capacitance

It seems at first sight that the adoption of Class-II (double-insulated) equipment throughout an audio system will give inherent immunity to ground-loop problems. Life is not so simple, though it has to be said that when such problems do occur they are likely to be much less severe. This mains transformer problem afflicts all Class-II equipment to a certain extent.

Figure 18.4 shows two Class-II units connected together by an unbalanced audio cable. The two mains transformers in the units have stray capacitance from both live and neutral to the secondary. If these capacitances were all identical no current would flow, but in practice they are not, so 50 Hz currents are injected into the internal 0 V rail and flow through the resistance of F–G, adding hum to the signal. A balanced input or ground-canceling output will remove or render negligible the ill effects.

Reducing the resistance of the interconnect ground path is also useful – more so than with other types of ground loop, because the ground current is essentially fixed by the small stray capacitances, and so halving the resistance F–G will dependably halve the interfering voltage. There are limits to how far you can take this – while a simple balanced input will give 40 dB of rejection at low cost, increasing the cross-sectional area of copper in the ground of an audio cable by a factor of 100 times is not going to be either easy or cheap. Figure 18.4 shows equipment with metal chassis connected to the 0 V (this is quite acceptable for safety approvals – what counts is the isolation between mains and everything else, not between low-voltage circuitry and touchable metalwork); note the chassis connection, however, has no relevance to the basic effect, which would still occur even if the equipment enclosure was completely nonconducting.

Table 18.1: Typical ground currents for different sorts of equipment

Equipment type	Power consumption	Ground current
Turntable, CD, cassette deck	20 W or less	5 μA
Tuners, amplifiers, small TVs	20–100 W	100 μA
Big amplifiers, subwoofers, large TVs	More than 100 W	1 mA

The magnitude of ground current varies with the details of transformer construction, and increases as the size of the transformer grows, as shown in Table 18.1. Therefore the more power a unit draws, the larger the ground current it can sustain. This is why many systems are subjectively hum-free until the connection of a powered subwoofer, which is likely to have a larger transformer than other components of the system.

Ground Currents Inside Equipment

Once ground currents have been set flowing, they can degrade system performance in two locations: outside the system units, by flowing in the interconnect grounds, or inside the units, by flowing through internal PCB tracks, etc. The first problem can be dealt with effectively by the use of balanced inputs, but the internal effects of ground currents can be much more severe if the equipment is poorly designed.

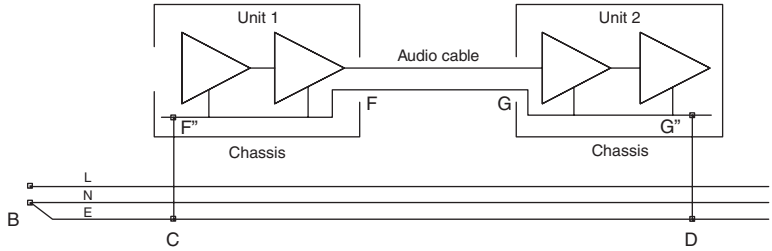


Figure 18.5: If ground current flows through the path F′–F–G–G′ then the relatively high resistance of the PCB tracks produces voltage drops between the internal circuit blocks

Figure 18.5 shows the situation. There is, for whatever reason, ground current flowing through the ground conductor CD, causing an interfering current to flow round the loop C–F–G–D as before. Now, however, the internal design of Unit 2 is such that the ground current flowing through F–G also flows through G–G′ before it encounters the ground wire going to point D. G–G′ is almost certain to be a PCB track with higher resistance than any of the cabling, and so the voltage drop across it can be relatively large, and the hum performance correspondingly poor. Exactly similar effects can occur at signal outputs; in this case the ground current is flowing through F–F′.

Balanced inputs will have no effect on this; they can cancel out the voltage drop along F–G, but if internal hum is introduced further down the internal signal path, there is nothing they can do about it.

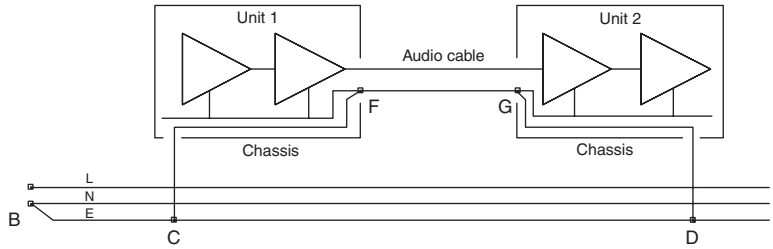


Figure 18.6: The correct method of dealing with ground currents; they are diverted away from internal circuitry

The correct method of handling this is shown in Figure 18.6. The connection to mains ground is made right where the signal grounds leave and enter the units, and are made as solidly as possible. The ground current no longer flows through the internal circuitry. It does, however, still flow through the interconnection at F–G, so either a balanced input or a ground-canceling output will be required to deal with this.

Balanced Mains Power

There has been speculation in recent times as to whether a balanced mains supply is a good idea. This means that instead of live and neutral (230 and 0V) you have live and the other live (115–0–115V) created by a center-tapped transformer with the tap connected to neutral (see Figure 18.7).

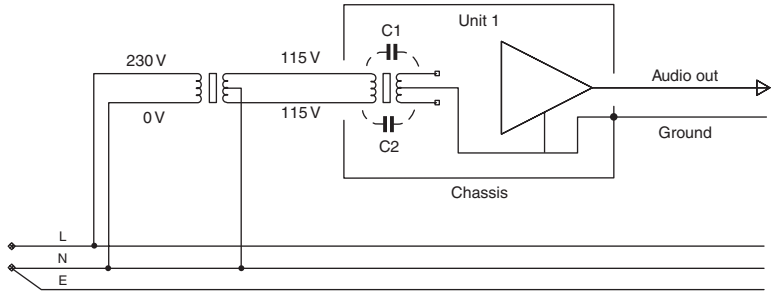


Figure 18.7: Using a balanced mains supply to cancel ground currents stemming from interwinding capacitance in the mains transformer – an expensive solution

It has been suggested that balanced mains has miraculous effects on sound quality, makes the sound stage 10-dimensional, etc. This is obviously nonsense. If a piece of gear is that fussy about its mains (and I do not believe any such gear exists) then dispose of it.

If there is severe RFI on the mains, an extra transformer in the path may tend to filter it out. However, a proper mains RFI filter will almost certainly be more effective – it is designed for the job, after all – and will definitely be much cheaper.

Where you might gain a real benefit is in a Class-II (i.e. double-insulated) system with very feeble ground connections. Balanced mains would tend to cancel out the ground currents caused

by transformer capacitance (see Figure 18.4 and above for more details on this) and so reduce hum. The effectiveness of this will depend on C1 being equal to C2 in Figure 18.7, which is determined by the details of transformer construction in the unit being powered. I think that the effect would be small with well-designed equipment and reasonably heavy ground conductors in interconnects. Balanced audio connections are a much cheaper and better way of handling this problem, but if none of the equipment has them then beefing up the ground conductors should give an improvement. If the results are not good enough then, as a last resort, balanced mains may be worth considering.

Finally, bear in mind that any transformer you add must be able to handle the maximum power drawn by the audio system at full throttle. This can mean a large and expensive component.

I would not be certain about the whole of Europe, but to the best of my knowledge it is the same as the UK, i.e. not balanced. The neutral line is at earth potential, give or take a volt, and the live is 230 V above this. The three-phase 11 kV distribution to substations is often described as 'balanced', but this just means that power delivered by each phase is kept as near equal as possible for the most efficient use of the cables.

It has often occurred to me that balanced mains 115–0–115 V would be a lot safer. Since I am one of those people that put their hands inside live equipment a lot, I do have a kind of personal interest here.

Class-I and Class-II

Mains-powered equipment comes in two types: grounded and double-insulated. These are officially called Class-I and Class-II respectively.

Class-I equipment has its external metalwork grounded. Safety against electric shock is provided by limiting the current the live connection can supply with a fuse. Therefore, if a fault causes a short-circuit between live and metalwork, the fuse blows and the metalwork remains at ground potential. A reasonably low resistance in the ground connection is essential to guarantee the fuse blows. A three-core mains lead is mandatory. Two-core IEC mains leads are designed so they cannot be plugged into three-pin Class-I equipment. Class-I mains transformers are tested to 1.5 kV rms.

Class-II equipment is not grounded. Safety is maintained not by interrupting the supply in case of a fault, but by preventing the fault happening in the first place. Regulations require double-insulation and a generally high standard of construction to prevent any possible connection between live and the chassis. A two-core IEC mains lead is mandatory; it is not permitted to sell a three-core lead with a Class-II product. This would present no hazard in itself, but is presumably intended to prevent confusion as to what kind of product is in use. Class-II mains transformers are tested to 3 kV rms, to give greater confidence against insulation breakdown.

Class-II is often adopted in an attempt to avoid ground loops. Doing so eliminates the possibility of major problems, at the expense of throwing away all hope of fixing minor ones. There is no way to prevent capacitance currents from the mains transformer flowing through the ground connections

(see section 'Ground loops: how they work and how to deal with them'). It is also no longer possible to put a grounded electrostatic screen between the primary and secondary windings. This is serious as it deprives you of your best weapon against mains noise coming in and circuit RF emissions getting out. In Class-II the external chassis may be metallic, and connected to signal 0 V as often as you like.

If a Class-II system is not connected to ground at any point, then the capacitance between primaries and secondaries in the various mains transformers can cause its potential to rise well above ground. If it is touched by a grounded human, then current will flow, and this can sometimes be perceptible, though not directly, as a painful shock like static electricity. The usual complaint is that the front panel of equipment is 'vibrating', or that it feels 'furry'. The maximum permitted touch current (flowing to ground through the human body) permitted by current regulations is 700 μA, but currents well below this are perceptible. It is recommended, though not required, that this limit be halved in the tropics where fingers are more likely to be damp. The current is measured through a 50 k resistance to ground.

When planning new equipment, remember that the larger the mains transformer, the greater the capacitance between primary and secondary, and the more likely this is to be a problem. To put the magnitudes into perspective, I measured a 500 VA toroid (intended for Class-II usage and with no interwinding screen) and found 847 pF between the windings. At 50 Hz and 230 V this implies a maximum current of 63 μA flowing into the signal circuitry, the actual figure depending on precisely how the windings are arranged. A much larger 1500 VA toroidal transformer had 1.3 nF between the windings, but this was meant for Class-I use and had a screen, which was left floating to get the figure above.

Warning

Please note that the legal requirements for electrical safety are always liable to change. This book does not attempt to give a complete guide to what is required for compliance. The information given here is correct at the time of writing, but it is the designer's responsibility to check for changes to compliance requirements. The information is given here in good faith but the author accepts no responsibility for loss or damage under any circumstances.

Cooling

All power amplifiers will have a heat-sink that needs cooling, usually by free convection, and the mechanical design is often arranged around this requirement. There are three main approaches to the problem:

(a) The heat-sink is entirely internal, and relies on convected air entering the bottom of the enclosure and leaving near the top (passive cooling).

Advantages. The heat-sink may be connected to any voltage, and this may eliminate the need for thermal washers between power device and sink. On the other hand, some sort of conformal

material is still needed between transistor and heat-sink. A thermal washer is much easier to handle than the traditional white oxide-filled silicone compound, so you will probably be using them anyway. With this form of construction there are no safety issues as to the heat-sink temperatures.

Disadvantages. This system is not suitable for large dissipations, due to the limited fin area possible inside a normal-sized box, and the relatively restricted convection path.

(b) The heat-sink is partly internal and partly external, as it forms one or more sides of the enclosure. Advantages and disadvantages are much as above; if any part of the heat-sink can be touched then the restrictions on temperature and voltage apply. Greater heat dissipation is possible.

(c) The heat-sink is primarily internal, but is fan-cooled (active cooling). Fans always create some noise, and this increases with the amount of air they are asked to move. Fan noise is very unwelcome in a domestic hi-fi environment; it seems pointless to strive for beautifully quiet electronics if there is a fan whirring away. Fan noise is of little importance in most PA applications, but could be an issue in small venues.

This allows maximal heat dissipation, but requires an inlet filter to prevent the build-up of dust and fluff internally. Persuading people to clean such filters regularly is near impossible.

The internal space in the enclosure will require some ventilation to prevent heat build-up; slots or small holes are desirable to keep foreign bodies out. Avoid openings on the top surface if you can as these will allow the entry of spilled liquids and increase dust entry. BS415 is a good starting point for this sort of safety consideration, and this specifies that slots should be no more than 3 mm wide.

Reservoir electrolytics, unlike most capacitors, suffer significant internal heating due to ripple current. Electrolytic capacitor life is very sensitive to temperature, so mount them in the coolest position available, and if possible leave room for air to circulate between them to minimize the temperature rise.

Convection Cooling

Efficient passive heat removal requires extensive heat-sinking with a free convective air flow, and this often indicates putting the sinks on the side of the amplifier; the front of the unit will carry at least the mains switch and power indicator light, while the back carries the in/out and mains connectors, leaving only the sides completely free.

It is important to realize that the buoyancy forces that drive natural convection are very small, and even small obstructions to flow can seriously reduce the rate of flow and hence the cooling. If ventilation is by slots in the top and bottom of an amplifier case, then the air must be drawn under the unit, and then execute a sharp right-angled turn to go up through the bottom slots. This change of direction is a major impediment to air flow, and if you are planning to lose a lot of heat then it feeds into the design of something so humble as the feet the unit stands on – the higher the better,

for air flow. In one instance the amplifier feet were made 13 mm taller and all the internal amplifier temperatures dropped by 5°C. Standing such a unit on a thick-pile carpet can be a really bad idea, but someone is bound to do it (and then drop their coat on top of it); hence the need for over-temperature cut-outs if amplifiers are to be fully protected.

Heat-sink materials

Heat-sinks in the audio business are almost always made of aluminum. It is cheaper than copper but of comparable conductivity, and has the very useful property that it can be extruded to form shapes with fins, and can be anodized in all sorts of colors. Aluminum is never used as a pure element because it is too soft to be practical; it is produced as an alloy with small amounts of copper and silicon. Even then it is still a relatively soft material, and there can be problems with tapped screw holes.

Copper heat-sinks have become quite common in the specialized field of cooling overclocked computer processors, where performance is critical and people are prepared to pay for it. It has occasionally appeared in audio amplifiers. A copper heat-sink may cost three times that of the same-sized aluminum part because of the higher material and fabrication costs. Copper cannot be extruded, so copper heat-sinks must be machined, and the machining process is more demanding than for aluminum. Due to density and its abrasive nature, machining holes and other details in copper takes significantly longer and wears your tooling faster.

The cost of the copper raw material is about the same as aluminum by weight but it has three times the density of aluminum. Thus, to make a heat-sink of a given size, the raw material cost is three times that of aluminum.

Silver is a better heat-sink material than copper – it is in fact the only one, as it is the element with the highest thermal conductivity – but the difference is only 7%, and the extra cost rules it out for all but the highest of high-end products. Silver heat-sinks have been used for CPU cooling – so has silver-plated copper, the plating being purely for the visual effect. It would of course stop the copper oxidizing, but silver is subject to blackening due to sulfide formation, so you may not be gaining much aesthetically. The surface of the silver can be lacquered to prevent chemical action, but this introduces an extra layer with much worse conductivity than metal.

Spending money on more precious metals is worse than pointless, for gold is a worse conductor than copper and not much superior to aluminum. Platinum is very much worse, though I would not rule out the possibility that someone, somewhere, has made a heat-sink out of it.

It is worth noting that aluminum is not the cheapest of metals, despite being relatively common, because it is produced from its ore by electrolytic refining, which uses huge amounts of electricity. This is why aluminum has been called ‘congealed electricity’. Steel – which is cheap – is unfortunately the bottom metal in our conductivity table, being nine times worse than copper and five times worse than aluminum. This is why you don’t see steel heat-sinks. I did know of one company that tried to use them, but they didn’t get far with it. Apart from the problems of machining a harder material, steel rusts readily and so requires a protective coating that further impairs its thermal performance.

Table 18.2 gives the thermal conductivity of various substances and also some bulk prices. These prices are of course subject to variation due to market conditions, and at the time of writing the trend is very much upwards due to the increasing demand for materials in the Far East. At the bottom of the table are included some bulk thermal conductivities for the Warth/Laird materials used in insulating thermal washers, which are described below.

Table 18.2: The thermal conductivity of various substances, some being more useful in amplifier design than others

Material	W/m-K	£/tonne 2005
Helium II	100,000	
Diamond	2500	
Graphite	470	
Silver	429	816,670
Copper	401	2576
Gold	310	41,029,324
Aluminum	250	1044
Beryllium	177	
Tungsten	174	
Brass	109	
Platinum	70	21,233,329
Steel	46	400
Warth K381	3.8	
Warth K200	1.70	
Warth K177	0.79	

Diamond makes a truly excellent heat-sink, as it has extremely high thermal conductivity but is also an excellent electrical insulator. It has often been used for cooling exotic semiconductors such as microwave transistors and laser diodes. Diamond heat-sinks are not of course big things with fins, but small flat pieces used to spread the heat out into cheaper materials. Flat bits of synthetic diamond for this purpose can be bought commercially with W/m-K values from 1000 to 1800. The drawbacks are the very high price and the very great difficulty of machining the material.

Discussing the use of diamond for amplifier heat-sinks may appear frivolous, but given the remarkable advances in carbon chemistry in the last few years, it is not impossible that one day we may see big finned heat-sinks grown from solid diamond – which would, in all senses of the word, be truly cool.

Just to put things in perspective, the top entry in the table is for helium II, which shows the highest thermal conductivity of any known substance as heat conduction in it occurs by an exceptional quantum mechanical mechanism. Helium II is a liquid and only exists below 2.2 K, within sight of absolute zero, and it must be regretfully concluded that it does not make a very practical heat-sink for general use.

Heat-sink compounds

Heat-sink compounds, in conjunction with the traditional mica washers, are still in common use by Far East manufacturers at the time of writing, though their use is definitely declining. The most common heat-sink compound is the white-colored paste or thermal grease, typically silicone oil filled with aluminum oxide, zinc oxide, or boron nitride. Some exotic brands of thermal interfaces (notably Arctic Silver) use micronized or pulverized silver.

While heat-sink compounds can provide good thermal performance, they are horribly messy to apply and create problems when servicing equipment, as the force required to break the suction between the compound and heat-sink is often enough to delaminate mica washers – avoid them if you can.

Thermal washers

These are typically made of thin silicone rubber loaded with highly thermally conductive but electrically insulating compounds such as aluminum oxide or boron nitride, and are usually reinforced by an inner weave of fiberglass to resist tearing. They are very much easier to apply than heat-sink compounds.

The disadvantages are that because they are made very thin (0.25 mm at most, and sometimes much less) to increase thermal performance, they are vulnerable to being punctured by even very small pieces of metal contamination. Areas used for their assembly into equipment must be kept scrupulously clean of swarf and metal dust. Holes must also be carefully deburred to prevent cutting through the material. In general, the higher the performance of the washer, the more fragile it is and the more carefully it must be handled. This is because the increased proportion of particles of the thermally conductive compound in the silicone elastomer makes it harder and more brittle and crumbly. It is good practice to always design in the cheapest and toughest version of the thermal washers; these will also be the least efficient. If things do go a bit wrong thermally, you then have the option of switching to a more effective but costlier material to bring down device temperatures. In one power amplifier I was associated with, changing to more efficient thermal washers brought down some worryingly high junction temperatures by about 10°C, and we all slept better at night after that.

Many standard shapes are available for common semiconductor packages. For a TO220 package and a mounting pressure of 50 psi, thermal resistance ranges from 1.5 to 3.4°C/W. Some of the bulk thermal conductivities for the Warth/Laird materials are included at the bottom of Table 18.2, and it can be seen that there is quite a difference between the standard K177 and the high-performance K381. You can also see that even the best thermal washer materials have much less thermal conductivity than the worst of metals, and this is why thermal washers have to be so very thin.

A special type is the phase-change thermal washer. These are made of a thermally conductive material that melts when it reaches its operating temperature, and therefore makes superior contact with the metal surface of the heat-sink or semiconductor. A typical melting temperature is 65°C. They are made from a phase-change compound coated on a fiberglass web, and can be handled just like ordinary thermal washers at room temperature. The downside is in servicing; the melting causes the thermal material to stick to the metal, and the washer cannot be easily removed and certainly not reused.

Sometimes you don't need electrical isolation, but you do need a conformal material between two surfaces to fill in the minor irregularities and so improve heat flow. In this case graphite foil, which is simply thin layers of graphite reinforced with a small amount of fiberglass to give it some (but not much) mechanical integrity, is highly effective. It is, however, fragile and needs careful handling. Like other thermal materials, tooling for custom shapes is relatively inexpensive. There is more on graphite foil in Chapter 15.

Fan Cooling

Fan cooling allows much more heat to be removed than relying on the very slow movements of natural convection, but it has disadvantages that mean you should think very hard before adopting it. A fan costs a significant amount of money and introduces an electromechanical component that may well be the least reliable part of the unit. Fans create acoustic noise, and tend to fill up the enclosure with dust and dirt.

However, sometimes they are unavoidable. In hi-fi applications, big Class-A amplifiers will need fans to remove the prodigious heat output. In the sound-reinforcement world, fans are useful because they mean amplifiers can be made lighter and more compact.

The fans used in audio applications are usually of the axial DC type, which are available in a wide range of diameters and airflow capabilities. The DC powering makes control of fan speed much simpler when it is required. The operating voltages generally available are nominally 5, 6, 12, 24, and 48 V, though most fans will operate over quite a wide range of reduced voltage. The 12 V versions are by far the most common and give the greatest choice of airflow/noise trade-offs and so on, with the 24 V versions coming second.

If fan cooling is essential, there are certain things to be taken into consideration to make it as trouble-free as possible, as outlined below.

Firstly, the fan should be fitted so as to push air into the case rather than suck it out. Push operation means that all the air entering the case has passed through the filter, even if the case has a few leaks, which it almost certainly will have. The suction method may pull most of the air in through the filter, but some will be coming in through the leaks and will bring dirt and dust with it.

There are two schools of thought on the provision of air filters. The first regards an air filter as essential to keep dust, etc. out, and issues stern warnings that the filter must be cleaned regularly or dire things will come to pass, as the reduced air flow leads to overheating. The second recognizes that in general people just don't clean air filters, no matter how much you threaten them, and so no filter is fitted, the obvious downside to this being that the dust brought in with the air soon causes the internals of the equipment to resemble the inside of a neglected vacuum cleaner, and eventually there will be problems.

Secondly, a DC fan motor uses current in quite hefty pulses and generates a corresponding magnetic field, so keep it away from sensitive circuitry. The two wires to the fan must be twisted together or otherwise kept adjacent or otherwise the resulting loop will radiate nasty sharp spikes.

Obviously the fan ground return must be kept quite separate from audio ground. A quick look at a particular 80-mm-diameter 12 V fan shows that it draws 170 mA and produces pulses at 140 Hz; the frequency falls as the input voltage is reduced and the fan runs more slowly.

Thirdly, most DC fans will operate over quite a wide voltage range, though the rated voltage should not be exceeded. As the voltage is reduced the fan turns more slowly and moves less air, but it also generates less acoustic noise. A large fan running below full speed may shift the same volume of air as a smaller fan at full throttle, but in most cases will make less noise about it. However, the large fan may of course be more expensive.

If you are planning to run a fan at the lower end of its voltage range to keep it quiet then be very wary. As the bearings age and friction increases the fan may stop turning altogether; it is difficult or impossible to predict how soon this is likely to be a problem.

Fan control systems

The simplest way to apply a fan is to select a model with enough flow rate for the most demanding conditions and have it running full speed continuously. Naturally this means that most of the time it is providing more cooling than is necessary and more noise than is desirable.

A fan can of course be controlled by a simple on/off thermostat, giving what in the world of control theory is called 'bang-bang' control, but this is crude and more audibly intrusive than having the fan running at full speed all the time.

Most fan applications have a proportional control system that aims to keep the fan running at a relatively constant, or at any rate very slowly varying, speed that matches the power being dissipated by the heat-sinks. This is usually a simple proportional servo circuit, with no PID complications, and the temperature control will not be very accurate as the loop gain has to be kept low. The heat-sinks have a large thermal mass and the fan speed only affects their temperature slowly, so a high value of loop gain will lead to slow oscillations in fan speed, otherwise known as hunting. This is aurally very distracting and must be carefully avoided. Fortunately there is no need for great accuracy of temperature control so a low loop gain is not normally a problem. Figure 18.8 shows a simple proportional servo circuit to control a DC fan, showing how it interfaces to the power supply; the +15 V positive regulator is not shown. It assumes that supply rails of ±15 V or thereabouts are being derived to power op-amps.

The LM35 temperature sensor IC is a very attractive component for this application. It puts out a voltage that is linearly dependent on temperature, so that its output at 25°C is +250 mV and at 50°C is +500 mV. It uses very little power, is trimmed by the manufacturer at wafer level so it requires no trimming or calibration, and has a low output impedance. The downside is that it is relatively expensive, and in the TO92 package, which is the most commonly available, is not easy to mount onto a heat-sink surface without the use of either glue or a special clip. Thermistors are cheaper but have the disadvantage that they are strongly nonlinear in their temperature–resistance relation, which means the loop gain of the servo depends on temperature, rather complicating the design of a stable control loop.

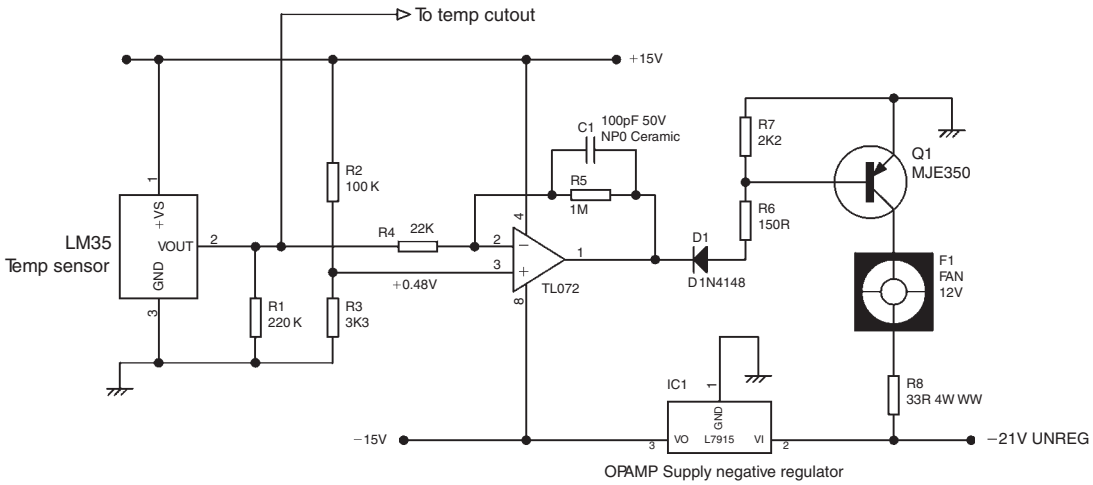


Figure 18.8: Fan control servo circuit using an LM35 temperature sensor

The servo circuit is essentially just an inverting amplifier with its gain set by the ratio of R4, R5. The set point is defined as 48°C by the reference divider R2, R3. Everything happens slowly in this circuit and decoupling this divider was not necessary. The fan drive circuitry may look a bit more complicated than necessary, but it was configured that way for a very good reason; I'm sure you never doubted that. The product in which this circuit was used was a powered mixer with a built-in DSP effects module for creating reverberation and so on. This was by today's standards a rather power-hungry device, and it took its +5V supply via a regulator IC from the +15V rail. In the interests of sharing the loading on the mains transformer secondaries, it was therefore desirable to run the fan between the unregulated −22V rail and 0V.

When the heat-sink temperature exceeds the set point, the op-amp output moves in the negative direction and turns on Q1 more via D1, R6, and R7. D1 prevents the base–emitter junction of Q1 from being reverse-biased when the op-amp output is high. Q1 turns on more and the fan turns faster. The dropper resistor R8 looks like it simply wastes power, which in a sense it does. It is present solely because 12V fans are much the most common, and fans running off 21V do not exist. A 24V model could have been used but it would have been more expensive and would never have run at its full capacity, so the resistive dropper solution was chosen. Running the fan off the regulated −15V supply was not considered a good plan as it would greatly increase the dissipation of the −15V regulator, and the pulsed current taken by the fan motor would have made the −15V rail noisy.

Even with effective proportional control, the noise from a fan may be disturbing in the silences between the music. An excellent way to cure this is to arrange for the fan to stop when the output of the power amplifiers falls below a certain threshold, on the basis that if the amplifier is no longer dissipating significant heat, the output device temperature cannot rise and there is no immediate need for further cooling. I used this philosophy with great success in a series of powered mixers intended for small venues, giving up to 250W/8Ω from two channels. In this application the need

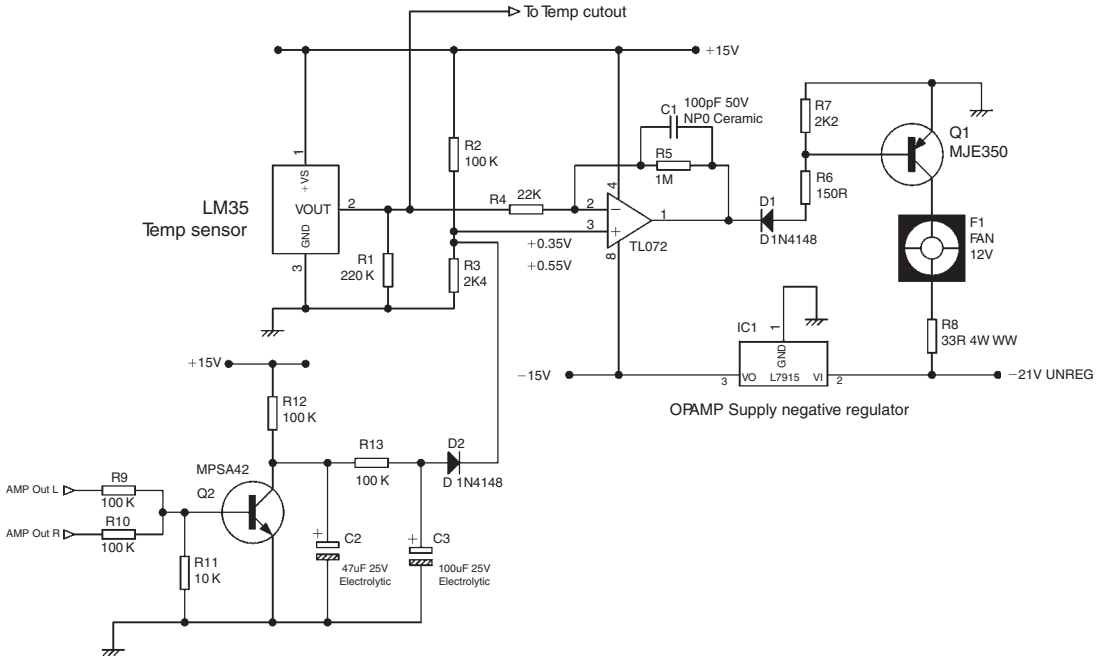


Figure 18.9: Improved servo circuit stops the fan when there is no audio output

for relatively small and light heat-sinks made fan cooling essential, but it was probable that fan noise would be intrusive – think about acoustic guitar music.

A somewhat more sophisticated version of this approach, which was adopted in the production versions, does not force the fan to immobility in the absence of signal, but instead shifts the set point upwards to a rather higher temperature. This recognizes that in this sort of sound-reinforcement application, ultimately cooling is more important than silence. Now, if the heat-sink is on the hot side, the fan will continue to run when the music stops, though because of the proportional nature of the fan control it may well do so at less than full throttle. A circuit I have used for this approach can be seen in Figure 18.9.

The fan control servo itself is unchanged from Figure 18.8, apart from the lowering of the set point (without the music-sensing circuitry activated) to 35°C by reducing the value of R9. This ensures that the heat-sink gets as much cooling as possible while the amplifier output masks the sound of a fan running quickly.

The output from the two channels of power amplification is summed at the base of Q2, and when there is any significant volume, Q2 is held hard on for at least half the time. The voltage on the cathode of D2 is therefore only about 200 mV (the $V_{ce(sat)}$ of Q2) above 0 V, D2 is reverse-biased, and the set point voltage at the junction of R2, R3 is unaffected. When, however, the amplifier output falls below the threshold, Q2 now turns off and its collector voltage rises. After a delay set by R12, C2 and R13, C3, D2 conducts and raises the set point voltage of the servo; with the resistor values shown it is increased from 35 to 55°C. These temperatures can easily be modified. The time

delays ensure that the fan set point is not being modulated at signal frequency or syllabic speed and prevent the fan turning on and off during very short pauses in the music.

It could be observed that what is really needed here is a maximum-selector circuit, such as a peak-rectifier driven by separate diodes, so that a positive input from one channel cannot be canceled out by a negative input from the other. In practice this extra complication is quite unnecessary. The greatest amplitudes in music are always in the bass register, and bass instruments are usually panned to the center to get the benefit of both channels of amplification, and so will almost always be in phase; this is also important if you plan to cut a master for retro-vinyl, as large anti-phase amplitudes will upset stylus tracking.

Fans do not stop turning immediately when the power is removed, but this is usually masked by the fade-out of the music. If the audio does stop abruptly, you may well hear the sound of the fan spooling down. It is not easy to think of a fix for this; fans have electronic commutation built in and so shorting their terminals does not induce dynamic braking.

Fan failure safety measures

Most amplifiers have over-temperature sensors that initiate shutdown if the heat-sink or power devices get too hot. This will eventually be triggered if the fan cooling fails, but heating things up to this extent is likely to stress the output devices and should be avoided if possible. To this end fan-fail detectors are sometimes fitted. It is fairly simple to design a circuit that monitors the pulsating current drawn by a DC fan motor, and raises a warning signal if the pulses cease because the fan has stopped turning, because of bearing failure or the blades being choked with debris. This gives some protection but does not of course detect the likeliest problem, which is the air filter becoming choked or the inlet port being blocked, so that insufficient air is being delivered although the fan is still flailing with all its might. A light vane placed in the air flow, its position being monitored by optoelectronic means, will give a warning against these conditions but is itself liable to jamming by dust and dirt, and it is rarely worth the cost and complication. A better method of air-flow detection is the use of two thermistors, arranged so they significantly self-heat by virtue of the current passing through them. If one thermistor is placed in the air flow and the other shielded from it, the first only will be cooled and the existence of a temperature difference, and hence a difference in electrical resistance, gives assurance that air is actually moving.

Heat Pipes

A heat pipe is a very effective method of transporting large quantities of heat from one place to another with a very small difference in temperature between the hot and cold zones. Typically it shifts heat from somewhere in the middle of a piece of equipment to an outside surface, where it is much more convenient to place a large finned heat-sink. A heat pipe is simply a sealed length of pipe containing a small amount of fluid that boils at the hot end. The vapor moves to the cold end and condenses, and capillary action in a wick structure lining the pipe, or gravity, then returns the fluid to the hot end. Heat pipes are relatively expensive but their unique advantage is their great efficiency in transferring heat over relatively long distances. They are a much better heat conductor than the equivalent cross-section of solid copper.

Most heat pipes use either ammonia or water as their working fluid, the boiling temperature being controlled by the pressure set up in the sealed pipe at manufacture. Water has a useful range of 30–200°C, which covers pretty much all electronic applications.

To take a specific example, a 6-mm-diameter heat pipe with a sintered metal wick transferring 10 W over a distance of 100 mm would have a thermal resistance of only 2.1°C/W. Doubling the distance only increases this to 2.5°C/W. In contrast, if the heat was being transferred through a solid block of metal, doubling the distance would naturally double the thermal resistance.

Heat pipes are now quite often used for cooling computer CPUs and high-end graphics processors; this is bringing prices down and the introduction of heat pipes into audio amplifiers in significant numbers looks very likely.

Mechanical Layout and Design Considerations

The mechanical design adopted depends very much on the intended market, and production and tooling resources, but I offer below a few purely technical points that need to be taken into account:

Wiring Layout

There are several important points about the wiring for any power amplifier:

- Keep the + and − HT supply wires to the amplifiers close together. This minimizes the generation of distorted magnetic fields that may otherwise couple into the signal wiring and degrade linearity. Sometimes it seems more effective to include the 0 V line in this cable run; if so it should be tightly braided to keep the wires in close proximity. For the same reason, if the power transistors are mounted off the PCB, the cabling to each device should be configured to minimize loop formation.
- The rectifier connections should go direct to the reservoir capacitor terminals, and then away again to the amplifiers. Common impedance in these connections superimposes charging pulses on the rail ripple waveform, which may degrade amplifier PSRR.
- Do not use the actual connection between the two reservoir capacitors as any form of star-point. It carries heavy capacitor-charging pulses that generate a significant voltage drop even if thick wire is used. As Figure 18.1 shows, the ‘star-point’ is teed off from this connection. This is a star-point only insofar as the amplifier ground connections split off from here, so do not connect the input grounds to it, as distortion performance will suffer.

Semiconductor Installation

Driver transistors are usually in the TO-225AA (e.g. the MJE340, MJE350) or the TO-220 package. Power transistors are commonly in the TO-3P or MT200 packages, though the venerable TO-3 all-metal package is still occasionally encountered. For each package the effective cooling area (i.e. the metal part of the package) is given in Table 18.3, as this is needed to derive the temperature difference between the device and the heat-sink, given the thermal conductivity of the thermal washer used and the power dissipated. The area lost to the mounting hole or holes has been allowed

Table 18.3: The thermal parameters of common transistor packages

Package	Metal area (mm^2)	Body area (mm^2)	Metal area (%)	Therm res. junction case (°C/W)
TO-225AA	27	84	32	6.25
TO-220	91	154	59	1.66
TO-3P	313	533	58	0.83
TO264	327	520	63	0.69
MT200	626	779	80	0.63
TO-3	577	625	92	0.70

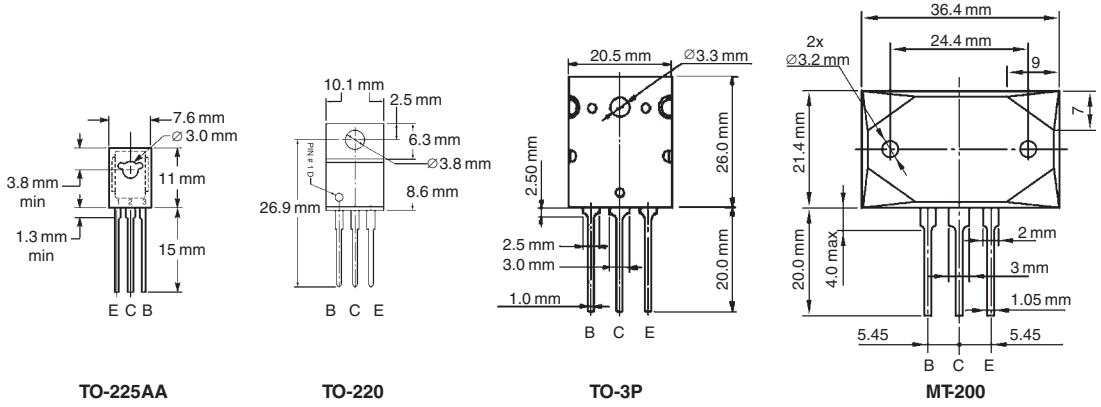


Figure 18.10: Dimensions of the most popular plastic transistor packages: TO-225AA, TO-220, TO-3P, and MT200. Drawing is only approximately to scale

for in calculating effective area. A comparison is also made between the metal area and the total area, which gives some sort of notion of how effective the package is. Thermal resistance from junction to case is also given; this is an average for different devices in the same package. The most popular packages are illustrated in Figure 18.10.

It is clear that in these terms the TO-225AA is not very area efficient, but then it is not intended for high-power usage. The MT200 is obviously the best plastic package, being beaten only by the TO-3 with its all-metal mounting flange – and all its mounting difficulties. The TO264 package is important because it is used for the intriguing new five-leg Onsemi ThermalTrak transistors with integral temperature-sensing diodes (see Chapter 15); the package is less efficient than the MT200 but the thermal resistance junction-to-case is still quite respectable.

- *TO-225AA driver transistor installation*. These devices are usually mounted onto separate heat-sinks that are light enough to be soldered into the PCB without further fixing. Silicone thermal washers ensure good thermal contact, and spring clips are used to hold the package firmly against the sink. Electrical isolation between device and heat-sink is not normally essential, as the PCB need not make any connection to the heat-sink fixing pads. If spring clips are not used, there is a single mounting hole for a bolt.

- *TO220 driver transistor installation*. As for the TO-225AA package.
- *TO-3P and TO264 power transistor installation*. These large flat plastic devices are best mounted on to the main heat-sink with special spring clips, which are not only very rapid to install, but also generate less mechanical stress in the package than bolting the device down by its mounting hole. They also give a more uniform pressure onto the thermal washer material. Nonetheless, it is not always convenient to use the specially shaped extrusions required to fit the spring clips, and screws are satisfactory so long as the appropriate torque setting is used. A good-sized washer should always be fitted under the screw head to spread the stress as much as possible. Being flat plastic devices, these transistors can be mounted directly onto the thick web of a heat-sink, using blind tapped holes. Holes tapped into aluminum are not very durable and you have to be careful when installing or replacing devices.
- *MT200 power transistor installation*. These even larger flat plastic devices have two mounting holes and once again can be mounted directly onto a large mass of metal. Once again, use good-sized washers under the screw heads. There is also an MT100 package, which is somewhat smaller than the TO-3P, but it does not seem to have been so widely adopted.
- *TO-3 power transistor installation*. The rather dated TO-3 package is extremely efficient at heat transfer, but notably more awkward to mount, as they have to be bolted to a relatively thin piece of metal so that the base and emitter pins can protrude through the other side. Examples of TO-3 mounting are shown in Chapter 15.

When the first edition of this book was produced, TO-3 packages were still fairly common in demanding power amplifier applications, but now they have pretty much fallen out of use, being replaced by plastic packages such as TO-3P and the larger MT200, which offer much easier mounting. The section below has been allowed to stand as it will be useful to those rebuilding old equipment. There is also the point that audio fashion is an unpredictable thing, and for all I know there will be a sudden wave of TO-3 nostalgia. I understand some people still use valves …

My preference is for TO-3s to be mounted on an aluminum thermal coupler, which is bolted against the component side of the PCB. The TO-3 pins may then be soldered directly on the PCB solder side. The thermal coupler is drilled with suitable holes to allow M3.5 fixing bolts to pass through the TO-3 flange holes, through the flange, and then be secured on the other side of the PCB by nuts and crinkle washers, which will ensure good contact with the PCB mounting pads. For reliability the crinkle washers must cut through the solder-tinning into the underlying copper; a solder contact alone will creep under pressure and the contact force decay over time.

Insulating sleeves are essential around the fixing bolts where they pass through the thermal coupler; nylon is a good material for these as it has a good high-temperature capability. Depending on the size of the holes drilled in the thermal coupler for the two TO-3 package pins (and this should be as small as practicable to maximize the area for heat transfer), these are also likely to require insulation; silicone rubber sleeving carefully cut to length is very suitable.

An insulating thermal washer must be used between TO-3 and flange; these tend to be delicate and the bolts must not be over-tightened. If you have a torque-wrench, then 10 Nw/m is an approximate upper limit for M3.5 fixing bolts. *Do not* solder the two transistor pins to the PCB until the TO-3 is firmly and correctly mounted, fully bolted down, and checked for electrical isolation from the heat-sink. Soldering these pins and *then* tightening the fixing bolts is likely to force the pads from the PCB. If this should happen then it is quite in order to repair the relevant track or pad with a small length of stranded wire to the pin; 7/02 size is suitable for a very short run.

Alternatively, TO-3s can be mounted off-PCB (e.g. if you already have a large heat-sink with TO-3 drillings) with wires taken from the TO-3 pads on the PCB to the remote devices. These wires should be fastened together (two bunches of three is fine) to prevent loop formation (see above). I cannot give a maximum safe length for such cabling, but certainly 8 inches causes no HF stability problems in my experience. The emitter and collector wires should be substantial, e.g. 32/02, but the base connections can be as thin as 7/02. The cable routing will need to be carefully chosen to avoid Distortion 6 – the inductive coupling of half-wave-rectified sine pulses into sensitive parts of the circuitry (see Chapter 7 for more on this).

There are (or used to be) sockets for TO-3 transistors. These were retained by the two mounting bolts and gripped the two pins with spring-loaded contacts, the idea being that device replacement was simpler – don't touch them. The contacts are in general not adequate to handle large currents through the emitter pin. However, in my experience there were worse problems than that; it was found that even in a clean domestic environment the contacts would corrode and become intermittent after a couple of years.

CHAPTER 19

Testing and Safety

'Out of this nettle, danger, we pluck this flower, safety.'

William Shakespeare, Henry IV Part I

Testing and Fault-Finding

Testing power amplifiers for correct operation is relatively easy; fault-finding when something is wrong is not. I have been professionally engaged with power amplifiers for a long time, and I must admit I still sometimes find it to be a difficult and frustrating business.

There are several reasons for this. Firstly, almost all small-signal audio stages are IC based, so the only part of the circuit likely to fail can be swiftly replaced, so long as the IC is socketed. A power amplifier is the only place where you are likely to encounter a large number of components all in one big negative-feedback loop. The failure of any components may (if you are lucky) simply jam the amplifier output hard against one of the rails, or (if you are not) cause simultaneous failure of all the output devices, possibly with a domino-theory trail of destruction winding through the small-signal section. A certain make of high-power amplifier in the mid-1970s was a notorious example of the domino effect, and when it failed (which was often) the standard procedure was to replace *all* of the semiconductors, back to and including the bridge rectifier in the power supply.

The advice given here is aimed primarily at the power amplifier designs included in this book, but are general enough to apply to most semiconductor power amplifiers.

By far the most important step to successful operation is a careful visual inspection before switch-on. As in all power amplifier designs, a wrongly installed component may easily cause the immediate failure of several others, making fault-finding difficult and the whole experience generally less than satisfactory. It is therefore most advisable to meticulously check:

- That the supply and ground wiring is correct.
- That all transistors are installed in the correct positions.
- That the drivers and TO-3 output devices are not shorted to their respective heat-sinks through faulty insulating washers.
- That the circuitry around the bias generator transistor in particular is correctly built. An error here that leaves this transistor turned off will cause large currents to flow through the output devices and may damage them before the rail fuses can act.
- That the bias adjustment is set to minimum.

For the Trimodal amplifier in Chapter 10, I recommend that the initial testing is done in Class-B mode. There is the minimum amount of circuitry to debug (the Class-A current controller can be left disconnected or not built at all until later) and at the same time the Class-B bias generator can be checked for its operation as a safety circuit on Class-A/AB mode.

The second step is to obtain a good sine-wave output with no load connected. A fault may cause the output to sit hard up against either rail; this should not in itself cause any damage to components. Since a power amplifier consists of one big feedback loop, localizing a problem can be difficult. The best approach is to take a copy of the circuit diagram and mark on it the DC voltage present at every major point. It should then be straightforward to find the place where two voltages fail to agree; e.g. a transistor installed backwards usually turns fully on, so the feedback loop will try to correct the output voltage by removing all drive from the base. The clash between 'full on' and 'no base drive' signals the error.

When checking voltages in circuit, bear in mind that in my designs the feedback network capacitor is protected against reverse voltage in both directions by diodes that will conduct if the amplifier saturates in either direction.

This DC-based approach can fail if the amplifier is subject to high-frequency oscillation, as this tends to cause apparently anomalous DC voltages. In this situation the use of an oscilloscope is really essential. An expensive oscilloscope is not necessary, but a bandwidth of at least 50 MHz is essential to avoid missing some kinds of parasitic oscillation (though they will certainly make their presence felt by their effect on the THD residual). A digital scope is at a serious disadvantage here, because HF oscillation is likely to be aliased into nonsense and be hard to interpret.

The third step is to obtain a good sine wave into a suitable high-wattage load resistor. It is possible for faults to become evident under load that are not shown up in the second step above.

Setting the quiescent conditions for any Class-B amplifier can only be done accurately by using a distortion analyzer. If you do not have access to one, the best compromise is to set the quiescent voltage drop across both emitter resistors to 10 mV when the amplifier is at working temperature; disconnect the output load to prevent DC offsets causing misleading current flow. This should be close to the correct value, and the inherent distortion of the designs is so low that minor deviations are not likely to be very significant. This implies a quiescent current of approximately 50 mA.

It may simplify fault-finding if the diodes in the collectors of the protection transistors are not installed until the basic amplifier is working correctly, as errors in the SOAR protection cannot then confuse the issue. This demands some care in testing, as there is then no short-circuit protection.

I insert here a few precautions learnt the hard way; as Benjamin Franklin put it, experience keeps a dear school, but fools will learn in no other.

- Make sure the reservoir capacitors are *fully* discharged before applying your soldering iron to the circuitry. Use a 10 Ω power resistor, not a screwdriver.
- If you have one of those handy trimming tools for bias adjustment, take off the metal clip and throw it away *before* you drop the tool into a 200 W amplifier.

- If the THD is too high check that the output connections are tight, if they are in the form of binding posts.

Powering up for the First Time

Testing an amplifier by hitting it with its full operating voltage is a risky business. Slowly winding up a variable-voltage transformer from zero is usually much safer, but not always – some amplifiers with complicated compensation schemes are not unconditionally stable (in the true sense of the word, as explained in Chapter 8) and are in fact very unstable when operated from supply rails that are much below normal. I am thinking here particularly of the four-stage Otala architecture described in Chapter 2. This never seems to be a problem with three-stage amplifiers using straightforward dominant-pole compensation.

It is clearly desirable for an amplifier to start working, even if imperfectly, as soon as possible when the supply rails are being wound up from zero. The sooner you can find out if something is amiss, the better chance you have of avoiding damage. Taking power transistors off heat-sinks and replacing them is a time-consuming business, and it doesn't take a lot of that sort of thing to knock a hole in your profit margins.

How low a voltage can a power amplifier be expected to work from, and give a clean-looking sine wave that is very good evidence that all is well? The measurements in Table 19.1 were made on a three-stage power amplifier with a very similar circuit to the Blameless amplifiers, designed to work from ±65 V rails. It first began to show signs of life at a rail voltage of ±3.8 V, though the maximum unclipped output level was very restricted at 128 mV.

Even so, a visually good sine wave was obtained with no load; a sine wave with 0.51% THD is visually perfect, and even the 5% distortion in the loaded condition, which is of course a more searching test that all is well, is not that easy to see as it is all crossover distortion due to the bias being set to minimum. As the supply rail is raised further, the maximum output increases rapidly and the distortion, both unloaded and loaded, falls rapidly. By ±5 V you can be pretty sure that all is as it should be, especially if you have a note of what output and distortion to expect under these conditions. Continuous checking of device temperatures with a questing finger is a wise precaution; if something is getting disquietingly hot at less than one-tenth of the intended supply voltage some prompt investigation is called for. After that, a quick check at perhaps half rail voltage and you can then apply full volts with some confidence.

The very great desirability of this gradual start-up procedure has implications for the design of the power amplifier. Don't go for current-source biasing schemes that need a lot of volts to work

Table 19.1: Amplifier distortion levels at very low rail voltages

Supply rail (V)	Output level rms (mV)	THD no load (%)	THD into 8 Ω (%)
±5.4	640	0.038	0.75
±4.1	256	0.17	1.6
±3.8	128	0.51	5

properly. And *don't* put a resistor in the tail of the input stage pair, like I have done in the past. In past editions both the Load-Invariant amplifier (2k2) and the Trimodal amplifier (1 k) had these resistors; I put my hands up – it was a mistake. The notional function of the resistor in the tail was to minimize the damage if the tail current-source transistor failed short-circuit; this is actually very unlikely, and I have yet to come across a case of it. The unwanted result was that these amplifiers would not work on very low rail voltages because of the voltage drop across the resistor caused by the 6 mA tail current.

Safety When Working on Equipment

This section considers the safety of the designer and the service technician. The recommendations here are advisory only. Regulations bearing on the safety of the user are backed by law; they are considered in the next section.

There are some specific points that should be considered:

1. An amplifier may have supply rails of relatively low voltage, but the reservoir capacitors will still store a significant amount of energy. If they are shorted out by a metal finger-ring then a nasty burn is likely. If your bodily adornment is metallic then it should be removed before diving into an amplifier.
2. Any amplifier containing a mains power supply is potentially lethal. The risks involved in working for some time on the powered-up chassis must be considered. The metal chassis *must* be securely earthed to prevent it becoming live if a mains connection falls off, but this presents the snag that if one of your hands touches live, there is a good chance that the other is leaning on chassis ground, so your well-insulated rubber-soled shoes will not save you. All mains connections (neutral as well as live, in the case of miswired mains) must therefore be properly insulated so they cannot be accidentally touched by finger or screwdriver. My own preference is for double insulation; for example, the mains inlet connector not only has its push-on terminals sleeved, but there is also an overall plastic boot fitted over the rear of the connector, and secured with a tie-wrap.

Note that this is a more severe requirement than BS415, which only requires that mains should be inaccessible until you remove the cover. This assumes a tool is required to remove the cover, rather than it being instantly removable. In this context a coin counts as a tool if it is used to undo giant screwheads.

If you are working on equipment with exposed mains voltages, taking the time to improvise some temporary insulation with plastic sheet and tape might just save your life.

3. A Class-A amplifier runs *hot* and the heat-sinks may well rise above 70°C. This is not likely to cause serious burns, but it is painful to touch. You might consider this point when arranging the mechanical design. Safety standards on permissible temperature rise of external parts will be the dominant factor.
4. Readers of hi-fi magazines are frequently advised to leave amplifiers permanently powered for optimal performance. Unless your equipment is afflicted with truly doubtful control over its own internal workings, this is quite unnecessary. (And if it *is* so afflicted, personally I would turn it

off right now.) While there should be no real safety risk in leaving a soundly constructed power amplifier powered permanently, I see no point and some potential risk in leaving unattended equipment powered; in Class-A mode there may of course be an impact on your electricity bill.

Warning

This section of the book is intended to provide a starting point in considering safety issues. Its main purpose is to alert you to the various areas that must be considered. For reasons of space it cannot be a comprehensive manual that guarantees equipment compliance; it cannot give a full and complete account of the various safety requirements that a piece of electronic equipment must meet before it can be legally sold. If you plan to manufacture amplifiers and sell them, then it is your responsibility to inform yourself of the regulations. The regulations change, always in the direction of greater safety and hence greater severity, and you must keep up to date. All the information here is given in good faith and is believed to be correct at the time of writing, but I accept no responsibility for its use.

Safety Regulations

The overall safety record of audio equipment is very good, but no cause for complacency. The price of safety, like that of liberty, is eternal vigilance. Safety regulations are not in general hard to meet so long as they are taken into account at the start of the mechanical design phase.

European safety standards are defined in a document known as BS EN 60065:2002 'Audio, video and similar electronic apparatus – Safety requirements'. The BS EN classification means it is a European standard (EN) having the force of a British Standard (BS). The latest edition was published in May 2002. It is produced by CENELEC, the European Committee for Electrotechnical Standardization.

In the USA, the safety requirements are set by the Underwriters Laboratories, commonly known simply as 'UL'. The relevant standards document is UL6500 'Audio/Video and Musical Instrument Apparatus for Household, Commercial, and Similar General Use', ISBN 0-7629-0412-7. The name 'Underwriters Laboratories' indicates that this institution had its start in the insurance business, allegedly because American houses tend to be wood-framed and are therefore more combustible than their brick counterparts.

The requirements for Asian countries are essentially the same, but it is essential to decide at the start which countries your product will be sold in, so that all the necessary approvals can be obtained at the same time. Changing your mind on this, so things have to be retested, is very expensive.

Electrical Safety

This is safety against electrical shocks. There must be no 'hazard live' parts accessible on the outside of the unit, and precautions must be taken in the internal construction so that parts do not become live due to a fault.

A part is defined as 'hazardous live' if under normal operating conditions it is at 35 V AC peak or 60 V DC with respect to earth. Under fault conditions 70 V AC peak or 120 V DC is permitted. Professional equipment, defined as that not sold to the general public, is permitted 120 V rms; there are also special provisions for audio signals. You are strongly advised to consult page 50 of BS EN 60065:2002 for more detailed information.

To determine if a 'hazardous live' part is accessible, the jointed test finger called 'Test Probe B' (to IEC 61032) is used; it is pushed against the enclosure or inserted through any openings. This is repeated with 'small finger probes' (see IEC 61032 again). Openings are also tested by inserting a 4-mm-diameter metal test pin by up to 100 mm.

Connectors are also tested by inserting a 1-mm-diameter metal test pin by up to 100 mm. No hazardous voltages may be touched by it. This may be a serious problem with phono (RCA) connectors and some female XLR connectors, which allow the test pin to pass right through the rear of the connector and into the equipment. The force used on the 1 mm test pin is 20 newtons.

Mains connections are always well insulated and protected where they enter the unit, normally by IEC socket or a captive lead, so the likeliest place where such voltages may appear is on the loudspeaker terminals of a power amplifier. An amplifier capable of 80 W into 8 Ω will have 35 V AC peak on the output terminals when at full power.

This seems like a tricky situation, but the current interpretation seems to be that the contacts of loudspeaker terminals, if they are inaccessible when they are fastened down, may be 'hazardous live', provided they are marked with the lightning symbol on the adjacent panel. Strictly speaking one should consider the operation of connecting speaker cables to be 'by hand' and therefore the contacts should be inaccessible at all times, i.e. closed or open. However, the general view seems to be that the connection of speaker terminals is a rare event and that adequate user instructions will be sufficient for the 'by hand' clause to be disregarded. The instructions would be of the form: 'Hazardous live voltages may be present on the contacts of the loudspeaker terminals . . . before connecting speaker cables disconnect the amplifier from the mains supply . . . if in doubt consult a qualified electrician.' I would remind readers at this point that such an interpretation appears to be the current status quo, but things can change and it is their responsibility to ensure that their equipment complies with the regulations.

Loudspeaker terminals that can accept a 4 mm banana socket from the front have been outlawed for some time. Existing parts can be legally used if an insulating bung is used to block the hole; indeed this is the preferable way, because a lot of people (myself included) like to use banana plugs. It is therefore wise to configure the bung so it can be removed by the customer on his own responsibility without too much of a struggle.

Unless the equipment is double-insulated, also known as Class-II, an essential safety requirement is a solid connection between mains ground and chassis, to ensure that the mains fuse blows if live contacts the metalwork. The differences between Class-I (grounded) and Class-II (double-insulated) are described in Chapter 18. British Standards on safety require the mains earth to chassis connection to be a *protected earth*, clearly labeled and with its own separate fixing.

A typical implementation has a welded ground stud onto which the mains-earth ring terminal is held by a nut and locking washer; all other internal grounds are installed on top of this and secured with a second nut/washer combination. This discourages service personnel from removing the chassis ground in the unlikely event of other grounds requiring disconnection for servicing. A label warning against 'lifting the ground' should be clearly displayed.

The ground wire from the chassis to the rear of the IEC socket, if it is soldered, must be wrapped around the terminal to make a sound mechanical joint before soldering, and *not* just poked through the hole and soldered. If push-on connectors instead of solder are used no further restraint of the ground wire is required, but adding a cable-tie close to the IEC connector to keep live, neutral, and ground together is a sensible extra precaution.

In the internal construction, two of the most important requirements to be observed are known as 'creepage and clearance'.

Creepage is the distance between two conductors along the surface of an insulating material. This is set to provide protection against surface contamination that might be sufficiently conductive to create a hazard. While the provisions of BS EN 60065:2002 are complex, taking into account the degree of atmospheric pollution and the insulating material involved, the usual distances employed in domestic equipment are as in Table 19.2.

Table 19.2: Creepage distances between conductors

Conductors	Creepage distance (mm)
Live to earth	3
Neutral to earth	3
Live to neutral	6
Live to low-voltage circuitry	6

Different sorts of live tracks (e.g. before and after a mains switch, or a fuse) must have a minimum creepage distance of 2.5 mm between them for standard 230 V mains. More information can be found on page 74 of BS EN 60065:2002.

Clearance is the air gap between two conductors, set to prevent any possibility of arcing; obviously the spacing between live conductors and earthed metalwork is the most important. The minimum air spacing is 2 mm. More information can be found on page 70 of BS EN 60065:2002.

Live cables must be fixed so that they cannot become disconnected, and then move about creating a hazard. This is important where cables are connected directly into a PCB. If the solder joint to the PCB breaks, they must still be restrained. The two most common ways are:

1. Fixing the cable to an adjacent cable with a cable-tie or similar restraint. The tie must be close enough to the PCB to prevent the detached cable moving far enough to cause a hazard. Obviously there is an assumption here that two solder joints will not fail at the same time (see Figure 19.1).

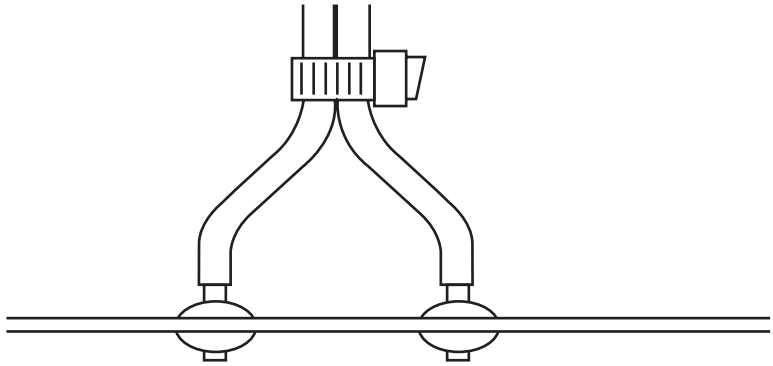


Figure 19.1: Cable restraint by fixing it to an adjacent cable

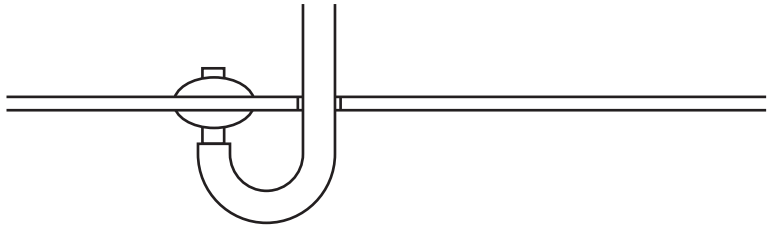


Figure 19.2: Cable restraint by hooking the cable through the PCB

2. Passing the cable through a plain hole in the PCB, and then bending it round through 180° to meet the pad and solder joint, as shown in Figure 19.2. This is often called 'hooking' or 'looping'.

Shocks from the Mains Plug

The need for EMC approval has resulted in X-capacitors being connected between live and neutral, and if the equipment mains switch is open these can hold a charge when the equipment is unplugged, depending on the mains voltage at the instant of breaking contact. To prevent shocks from mains plug pins, a drain resistor should be connected across the X-capacitor. An ordinary 270 k 1/4 W resistor can be used. A 220 k resistor is often seen in this position, but at 230 V it dissipates 240 mW, which is not exactly a large safety margin; 270 k only dissipates 196 mW and is to be preferred; 270 k with an X-cap of 470 nF gives a time-constant of 0.12 seconds, so in the worst case the voltage on the plug pins will have dropped from 230 to 32 V in a quarter of a second, and it is going to be difficult to get your fingers on the pins faster than that when you're unplugging something.

This is one time when the voltage rating of resistors matters: 1/4 W parts are usually rated at 500 or 700 V and this is fine for 230 V mains; 1/8 W resistors are often only rated at 200 V and are not suitable for this application, even if their value is such that they could withstand the power dissipation.

Touch Current

As mentioned in Chapter 18 when describing Class-I and Class-II equipment, the amount of current that can flow to ground via a human being when they touch the casework is an important issue. A Class-I (grounded) piece of equipment in normal use should have no touch current at all, as even a tenuous metallic connection to ground (and hopefully it is not tenuous) will have a negligible resistance compared with the body and no current will flow. For this reason Class-I equipment is tested for touch current with the protective earthing connection disconnected.

Class-II equipment has no ground connection, and the primary-to-secondary capacitance of the mains transformer can allow enough current to flow through to the casework for it to be perceptible in normal use. Clearly, if the current was big enough it would be hazardous.

Touch current is measured using a special network that connects the equipment to ground via resistors and capacitors, and is expressed in terms of the voltage that results; this is then compared with the voltages that make a part 'hazardous live'. The special network is defined in Annexe C of BS EN 60065:2002.

Here are a few more miscellaneous safety requirements, not necessarily enshrined in BS EN 60065:2002:

- Mains fuse ratings must be permanently marked, and a legend of the form 'WARNING: replace with rated fuse only' must be marked on the PCB.
- Internal wiring does not have to be color-coded (e.g. brown for live, blue for neutral) except for ground wiring, which must be green with a yellow trace.
- Crimp terminals on mains switches do not require color-coding of their plastic shrouds.

It is essential to keep an eye on mains transformer construction. With increasing globalization, transformers are now being made in parts of the world that do not have a long history of technological manufacturing, and mistakes are sometimes made, for example not using adequate insulation between primary and secondary.

Case Openings

As remarked elsewhere, in the section on mechanical design, case openings are subject to strict dimensional limits. A width of 3 mm is the maximum permitted. The old 'gold-chain' test has been removed from the latest edition of the standard, and is replaced by a narrow rigid test probe.

Equipment Temperature and Safety

There are limits on the permissible temperature rise of electronic apparatus, with the simple motivation of preventing people from burning themselves on their cherished hi-fi equipment. The temperature allowed is quoted as a rise above ambient temperature under specified test conditions. These conditions are detailed below. There are two regimes of ambient considered: 'moderate

climate', where the maximum ambient temperature does not exceed 35°C; and 'tropical climate', where the maximum ambient temperature does not exceed 45°C. In the tropical regime, the permitted temperature rises are reduced by 10°C. The temperature rise regulations are specified on pages 37–41 of BS EN 60065:2002 (Section 7).

The permitted temperature rise also depends on the material of which the relevant part is made. This is because metal at a high temperature causes much more severe burns than non-metallic or insulating material, as its higher thermal conductivity allows more heat to flow into the tissue of the questing finger.

The external parts of a piece of equipment are divided into three categories:

1. *Accessible, and likely to be touched often.*

This includes parts that are specifically intended to be touched, such as control knobs and lifting handles.

Metallic, normal operation – temperature rise 30°C above ambient

Metallic, fault condition – temperature rise 65°C above ambient

Non-metallic, normal operation – temperature rise 50°C above ambient

Non-metallic, fault condition – temperature rise 65°C above ambient.

This is usually an easy condition to meet, as knobs and switches are only connected to the internals of the amplifier via a shaft and a component such as a potentiometer or rotary switch that does not have good heat-conducting paths. Handles can be more difficult as they are likely to be secured to the front panel through a substantial area of metal, in order to have the requisite strength.

2. *Accessible, and unlikely to be touched often.*

This embraces the front, top, and sides of the equipment enclosure.

Metallic, normal operation – temperature rise 40°C above ambient

Metallic, fault condition – temperature rise 65°C above ambient

Non-metallic, normal operation – temperature rise 60°C above ambient

Non-metallic, fault condition – temperature rise 65°C above ambient.

This is the part of the temperature regulations that usually causes the most grief. To work effectively internal heat-sinks have vents in the top panel above them, allowing convective heat flow. The escaping air heats the top panel and this can get very hot. Some amplifier designs have a plastic grille over the heat-sink. This has several advantages. Since plastic is more economical to form than metal, the grille can have a structure that is more open and gives a larger exit area, while still complying with the 3 mm width limit for apertures. The grille itself is also allowed to get 20°C hotter because it is non-metallic, and for the same reason it conducts less heat to the surrounding metal top panel.

3. *Not likely to be touched.*

This includes rear and bottom panels, unless they carry switches or other controls that are likely to be touched in normal use, external heat-sinks and heat-sink covers, and any parts of the top enclosure surface that are more than 30 mm below the general level.

Normal operating conditions – temperature rise 65°C above ambient.

The permitted temperature under fault conditions is not specified, but it is probably safe to assume that a rise of 65°C is applicable.

The bottom panel is not likely to get very hot unless heat-sinks are directly mounted on it, as it gets the full benefit of the incoming cool air. The rear panel can be a problem as its upper section will be heated by convection, and is typically at much the same temperature as the top of the unit; it also often carries a mains switch, which takes it out of this category.

The test conditions under which these temperatures are measured are as follows:

- One-eighth of the rated output power into the rated load.
- All channels driven and with rated load attached.
- The signal source is pink noise, which is passed through an IEC filter to define the bandwidth to about 30 Hz–20 kHz. The details of the filter are given in Annexe C of BS EN 60065:2002.
- The mains voltage applied is 10% above the nominal mains voltage, so in Europe it is 230 V + 23 V = 253 V.

More information on the test conditions can be found on pages 24–27 of BS EN 60065:2002.

The introduction of temperature rise regulations caused external heat-sinks to become a rarity, despite the recognition that heat-sinks are rarely going to be touched in normal operation. It is usually much more cost-effective to have the heat-sinks completely enclosed by the casework, with suitable vents at top and bottom to allow convection. The heat-sinks can then be run much hotter, so they can be smaller, cheaper and lighter, obviously assuming that the semiconductor temperature limits are observed; the limit for power transistors is usually 150°C and for rectifiers 200°C. This usually allows the heat-sinks to be safely run at 90°C or more, depending on the details of transistor mounting and the amount of power dissipated by each device. Hot heat-sinks are more effective at dissipating heat by convection, but on the downside the restriction caused by the top and bottom vents, which must be of limited width, impairs the rate of air flow.

An exception to this is the use of massive heat-sinks to form part of the case, to make an aesthetic statement. In this case the heat-sinks are likely to be much larger for structural reasons than required for heat dissipation, and meeting the temperature-rise requirements is easy. Since aluminum extrusions are relatively expensive, this approach is restricted to ‘high-end’ equipment.

The importance of determining the temperature rise of the amplifier is such that it must be done as soon as possible in the product development process. If there are problems with EMC approval,

they can usually be fixed by relatively minor internal modifications. Heat problems are much more intractable, because fixing them may entail major mechanical redesign and rethinking of the aesthetics, or as a last resort reducing the amplifier power rating. It is therefore essential to test as early as possible, even if you haven't procured all the parts yet. Get a mains transformer that is roughly right, and use a variable transformer to get it to the right voltage. Mechanical parts not ready? Block up the holes with cardboard and tape. It is important, however, to use the correct heat-sink and get the ventilation arrangements as accurate as possible. The height off the test bench must be correct as it has a strong effect on air flow through vents in the bottom of the chassis.

Touching Hot Parts

It was described above how heat-sinks are often mounted internally, with air circulation through protective grilles. The air holes in these grilles must be small enough to prevent parts that exceed the temperature-rise regulations from being touched. The holes permitted are somewhat larger than those allowed near electrically hazardous parts. Testing is done with two probes. The 'toddler-finger probe', also known as Test-probe 19, simulates the finger of a child of 36 months or younger. It has a diameter of 5.6 mm and is articulated, with a hemispherical end. The other probe, Test-probe 18, covers persons from 36 months to 14 years; the diameter is 8.6 mm, and it is also articulated with a hemispherical end. The test force for both finger-type probes is 20 newtons. On the earth's surface a mass of 2.04 kg exerts a force of 20 newtons on its support. It is unlikely you will be doing the testing anywhere else.

Instruction Manuals

The instruction manual is very often written in a hurry at the end of a design project. However, it must not be overlooked that it is part of the product package, and must be submitted for examination when the equipment itself is submitted for safety testing. There are rules about its contents; certain safety instructions are compulsory, such as warnings about keeping water away from the equipment.

CHAPTER 20

Power Amplifier Input Systems

Most of this book deals with the actual power amplifier itself, and its immediate ancillary circuitry such as power supplies, overload and DC offset protection, control of fan cooling, and so on. It is quite feasible to take one or more such amplifiers and put them in a box with no further complications. Such a product was the much-loved Quad 303 power amplifier, which had no controls at all unless you count the mains voltage selector. However, amplifiers are very often also provided with input circuitry that gives a balanced input, gain control, filtering, level indication, and so on. This wholly new chapter deals with these subsystems, defining the amplifier input system as any part of the signal path before the actual power amplifier stage.

Balanced interconnections are seeing increasing use in hi-fi applications, and they have always been used in the world of professional audio. Their importance is that they can render ground loops and other connection imperfections harmless. Since there is no point in making a wonderful power amplifier and then feeding it with a mangled signal, making an effective balanced input is of the utmost importance and I make no apology for devoting a large part of this chapter to it.

Figure 20.1 shows a typical input system; many variations on this are of course possible. Firstly, RF filtering is applied at the very front end to prevent noise breakthrough and other EMC problems. The filtering must be done before the incoming signal encounters any semiconductors where RF demodulation could occur, and can be regarded as a 'roofing filter'. At the same time, the bandwidth at the low end is given an early limit by the use of DC-blocking capacitors, and overvoltage spikes are clamped by diodes. The input amplifier presents a reasonably high impedance to the outside world and almost invariably in professional amplifiers and, increasingly in hi-fi amplifiers, it is balanced so noise produced by ground loops and the like can be rejected. Sometimes the input is connected directly to a line output so that amplifiers can be daisy-chained together; this is much more economical of cable than having multiple fan-out cables each running from the source to one amplifier. The downside is that the failure of one amplifier in the chain can affect the feed to all of them. While the bandwidth of the amplifier has been very roughly circumscribed by the RF filtering and DC-blocking, it is usual to define it more precisely by the use of a subsonic filter and, less commonly, an ultrasonic filter. These very often come next in the signal path so they remove unwanted signals as soon as possible, and can benefit from a low-impedance drive from the input amplifier. After this comes the gain control, if there is one, and whatever buffering arrangements may be needed to drive the actual power amplifier stage.

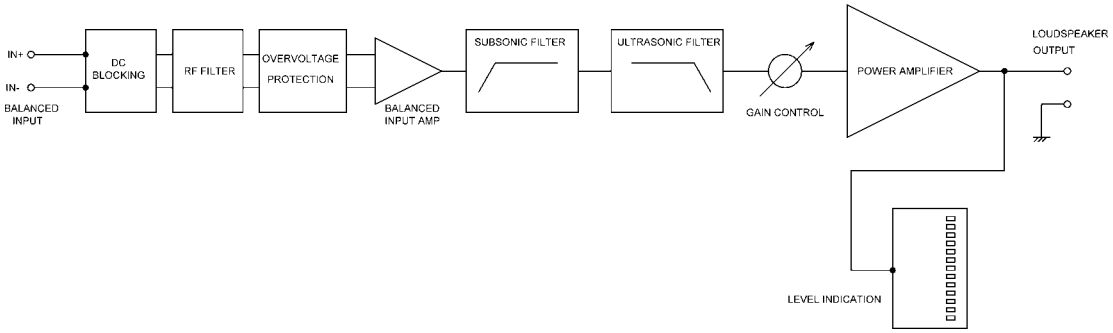


Figure 20.1: Block diagram of a comprehensive power amplifier input system

Table 20.1: Nominal signal levels

	V rms	dBu	dBv
Semi-professional	0.316	−7.78	−2.10
Professional	1.228	+4.0	+1.78
German ARD	1.55	+6.0	+3.78

If only a limited gain range is required, which is often the case, it is sometimes possible to combine it with the input amplifier in an active-gain-control format, which has advantages in minimizing noise and maximizing headroom in the input system. If a wide gain range is desirable, the gain control is almost always after the balanced input amplifier because making balanced inputs that retain good common-mode rejection when their gain is varied over a wide range is not so simple. More on that later.

External Signal Levels

There are several standards for line signal levels. The −10 dBv standard is used for a lot of semi-professional recording equipment as it gives more headroom with unbalanced connections – the professional levels of +4 and +6 dBu assume a balanced output that inherently gives twice the output level for the same supply rails as it is measured between two pins with signals of opposite phase on them (see Table 20.1).

Signal levels in dBu are expressed with reference to 0 dBu = 775 mV rms; the origin of this odd value is that it gives a power of 1 mW in a purely historical 600 Ω load. The unit of dBm refers to the same level but takes the power rather the voltage as the reference – a distinction of little interest nowadays. Signals in dBv (or dBV) are expressed with reference to 0 dB = 1.000 V rms.

These standards are well established, but that does not mean all amplifiers adhere to them and will give full output for a +4 or +6 dBu output. To take just one current example, the Yamaha P7000S requires +8 dBu (1.95 V rms) to give its full output of 750 W into 8 Ω.

There is not a great deal of consensus when it comes to the input sensitivity of hi-fi power amplifiers. The input required for full output can range from to 0 to +10 dBu or more.

Internal Signal Levels

It is necessary to select a suitable nominal level for the signal passing through the input system. It is always a compromise – the signal level should be high so it suffers the minimum of degradation by the addition of noise as it passes through the circuitry, but not so high that it is likely to suffer clipping before the gain control. It has to be considered that the gain control may be maladjusted by setting it too low and turning up the input level from the source equipment, making input clipping more likely. The levels chosen are usually in the range 388 mV–775 V rms (−6 to 0 dBu). Since the maximum output swing of an op-amp is around 9 V rms, this gives from 27 to 19 dB of headroom before clipping. If there is no gain control, then headroom in the input circuitry is not an issue. The power amplifier will always clip before the input circuitry, even though the latter is running off supply rails of only ±15 V.

In deciding the gain structure of the overall amplifier, it is wise to consider the needs of the power amplifier stage first. Making a linear Class-B power amplifier is always a challenge, and the only sensible way to meet it is to use as much negative feedback as can be safely applied without risking instability. This tends to result in a power amplifier stage gain in the region of 20–30 times. Thus a powerful amplifier delivering 500 W into 8 Ω, which is an output voltage of 63 V rms, will need an input voltage of between 2.1 V rms (+8.7 dBu) and 3.15 V rms (+12.2 dBu). These levels are well above any standard for equipment interfacing and at least 6 dB of gain will have to be provided by the input system, the exact amount depending on what you want the sensitivity of your amplifier input to be. This sort of gain can be provided by op-amps without adding significant distortion, and this is the best way to do it. The gain of a power amplifier stage can be increased by reducing the negative feedback ratio, and reducing the size of the compensation capacitor in proportion, to keep the feedback factor the same, but this process does seem to degrade the distortion performance by a certain amount. This is probably because the smaller compensation capacitor means there is less local feedback around the VAS.

If the incoming signal does have to be amplified, this should be done as early as possible in the signal path, to get the signal well above the noise floor as quickly as possible. If the gain is implemented in the first stage (i.e. the input amplifier, balanced or otherwise) the signal will be able to pass through later stages, such as filters, at a high level and so their noise contribution will be less significant.

The Choice of Op-Amps

It is the near-universal choice to make the amplifier input system out of op-amps, though there are notable exceptions; a few 'high-end' power amplifiers have quite complex discrete circuitry for implementing balanced inputs.

Until recently, the 5532 would have been the automatic choice for op-amps in the signal path, despite its great age (it was introduced in 1978). It has very low distortion and bipolar input devices that are quiet at the kind of impedances met with in most audio circuitry, with an input noise density of $4\,\text{nV}/\sqrt{\text{Hz}}$ (typical at 1 kHz).

The AD797 has remarkably low voltage noise at $0.9\,nV/\sqrt{Hz}$ (typical at 1 kHz) but it remains a specialized and very expensive part, costing approximately 25 times as much as a 5532 at the time of writing. It is only available in single versions, while the 5532 is a dual, so the cost factor per amplifier is actually 50 times. It has the reputation of being difficult to stabilize at HF, but in my experience it is not too hard.

The new chip on the block is the LM4562, a bipolar op-amp that has finally surpassed the 5532 in performance. It was just becoming freely available at the beginning of 2007. The LM4562 is a dual op-amp – there is no single or quad version. The input noise voltage is typically $2.7\,nV/\sqrt{Hz}$, which is substantially lower than the $4\,nV/\sqrt{Hz}$ of the 5532. For suitable applications with low source impedances, this gives a noise advantage of 3 dB or more. With a high source impedance (such as a moving-magnet cartridge) it is noisier than the 5532, because its current noise is higher. It is not fussy about decoupling and, as with the 5532, 100 nF across the supply rails usually ensures stability. Whether decoupling from rails to ground is required depends on the application. Slew rate is typically $\pm 20\,V/\mu s$, but the minimum is a bit lower at $\pm 15\,V/\mu s$. The LM4562 is a National Semiconductor product. I have had some difficulty in finding a representative cost, but at the time of writing it is something like 15 times more expensive than the 5532.

Unbalanced Inputs

The simplest unbalanced input feeds the incoming signal directly to the input of the power amplifier. This is not normal practice as at least some RF filtering will be required for EMC purposes. In addition, the power amplifier input impedance may be deliberately kept low, to $2\,k\Omega$ or less, so that offset voltages and noise generated in the feedback network are minimized (this is described in detail in Chapter 4); in this case a buffer amplifier will be needed. Figure 20.2 shows an unbalanced input amplifier, with the added components needed for interfacing to the real world. The op-amp U1:A acts as a unity-gain buffer or voltage follower; if you need gain just add two feedback resistors. A 5532 bipolar type is used here to keep the noise down; with the relatively low source impedances that are most likely to be encountered here, an FET-input op-amp would be 10 dB or more noisier. R1 and C1 make a first-order low-pass filter to remove incoming RF before it has a chance to reach the op-amp and demodulate into the audio band; once this has occurred any further attempts at RF filtering are of course useless. R1 and C1 must be as close to the input socket as physically possible to prevent RF from being radiated into the equipment enclosure before it is shunted to ground.

There is of course an inherent problem in selecting component values for input filters of this sort: we do not know what the output impedance of the source equipment is. If the source is an active pre-amp stage, then the output impedance will probably be around $50\,\Omega$, but it could be as high as $200\,\Omega$ or more. If a so-called 'passive preamplifier' is in use – i.e. just an input selector switch and a volume control potentiometer – then the output impedance will almost certainly be a good deal higher. (At least one passive preamplifier uses a transformer with switched taps for volume control, but I think analyzing that might be too much of a digression just at the moment.) If you really feel you want to use such a doubtful piece of equipment then a sensible potentiometer value is $10\,k\Omega$,

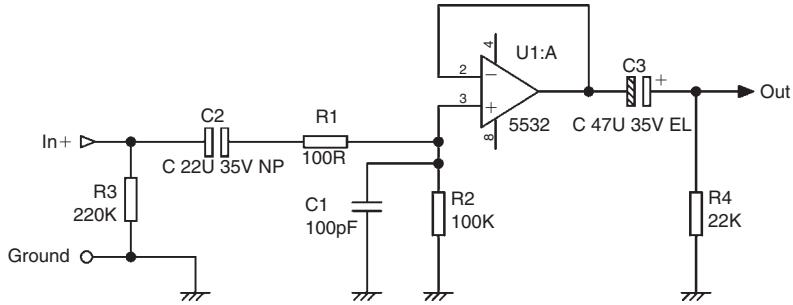


Figure 20.2: A typical unbalanced input amplifier with associated components

and its maximum output impedance (when it is set for 6 dB of attenuation) will be 2.5 kΩ, which is very different from 50 Ω. This resistance is in series with R1 and affects the turnover frequency of the RF filter. While it is very desirable to have effective RF filtering, it is also important to avoid a frequency response that sags significantly at 20 kHz.

If we take 2.5 kΩ as a worst-case source impedance and add the 100 Ω of R1, then 2.6 kΩ and 100 pF together give us −3 dB at 612 kHz ; this gives a loss at 20 kHz of only 0.0046 dB, so possibly C1 could be usefully increased; for example, if we made it 220 pF then the 20 kHz loss is still only 0.022 dB. If we stick with C1 at 100 pF and assume an active output with a 50 Ω impedance in the source equipment, then together with the 100 Ω resistance of R1 we have a total of 150 Ω, and 150 Ω with 100 pF gives −3 dB at 10.6 MHz.

This all seems very sensible and satisfactory, until you take a quick look at the sort of potentiometer values that passive preamplifiers really employ. I did a rapid survey, and while 10 kΩ seems to be a popular value, I quickly found one model with a 20 kΩ potentiometer, and another that had a value as high as 100 kΩ. The latter would have a maximum output impedance of 25 kΩ, and would give very different results with a C1 value of 100 pF – the worst-case frequency response would now be −3 dB at 63.4 kHz and −0.41 dB at 20 kHz, which is not good.

To put this into perspective, C1 is almost certainly going to be smaller than the capacitance of the interconnecting cable. Audio cable capacitance is usually in the range 50–150 pF/meter, so with a 2.5 kΩ source impedance you can only permit yourself a rather short run before there are significant effects on the frequency response; with a 25 kΩ source impedance you can hardly afford to have any cable at all. This is just one reason why 'passive preamplifiers' are not a good idea.

Another important constraint is that the series resistance R1 must be kept as low as practicable to minimize the generation of Johnson noise, but lowering this resistance means increasing the value of shunt capacitor C1, and if it becomes too big then its impedance at high audio frequencies will become too low. This can have two bad effects: too low a roll-off frequency if the input is connected to a source with a high output impedance, and a possible increase in distortion at high audio frequencies because of excessive loading on the source output stage.

Replacing R1 with a small inductor will give significantly better filtering but at increased cost. This is usually justifiable in professional audio equipment, but it is much less common in hi-fi. If you do

use inductors then it is essential to check the frequency response to make sure the LC circuit is well damped and not peaking at the turnover frequency.

C2 is a DC-blocking capacitor to prevent voltages from ill-thought-out source equipment from getting into the circuitry. Note that it is a nonpolarized type as voltages from the outside world are of unpredictable polarity. It is rated at 35 V so that even if it gets connected to defective equipment with an op-amp output jammed against one of the supply rails, no harm will come to our input circuit. R3 is a DC drain resistor that prevents the charge put on C1 by the aforesaid doubtful external equipment from remaining there for a long time and causing a thud when the connections are altered; as with all such drain resistors, its value is a compromise between draining the charge off the capacitor reasonably quickly and keeping the input impedance suitably high. The input impedance of the input circuit is R3 in parallel with R2, i.e. 220 k in parallel with 100 k, which comes to 68 k. This is a nice high value and should work well with just about any source equipment you can find.

Because of the presence of C2, R2 is required to provide biasing for the op-amp input; it needs to be quite a high value to keep the input impedance up, and bipolar input op-amps draw significant input bias current. The Fairchild 5532 data sheet quotes 200 nA typical and 800 nA maximum, and these currents would give a voltage drop across R2 of 20 and 80 mV respectively. This offset voltage will be faithfully reproduced at the output of the buffer, with the op-amp input offset voltage added on; this is only 4 mV maximum and so will not affect the final voltage much, whatever its polarity. The 5532 has NPN input transistors, and so the bias current flows into the input pins, and the voltage at Pin 3, and hence the output, will therefore be negative with respect to ground.

Such DC voltages are big enough to generate unpleasant thumps and bumps if the input stage is followed by any sort of switching, so further DC-blocking is required in the shape of C3; R4 is another DC drain resistor to keep the output at zero volts. It can be made rather lower in value than the input drain resistor R3 as the only requirement is that it should not significantly load the op-amp output. FET-input op-amps have much lower input bias currents, so that the offsets they generate as they flow through biasing resistors are usually negligible, but they still have input offsets of a few millivolts, so DC-blocking is still required if switches or relays downstream are to act silently.

With appropriate use of blocking capacitors, DC offsets should not cause trouble upstream or downstream, but you need to make sure that the offset voltages are not so great that the output voltage swing of the op-amp is significantly affected. The effect of our worst-case 80 mV offset here is trivial.

Balanced Interconnections

Balanced inputs on power amplifiers are used to prevent noise and crosstalk from affecting the input signal, especially in applications where long interconnections are used. They are standard on professional amplification equipment, and are steadily becoming more common in the world of hi-fi. A balanced input amplifier is sometimes called a line receiver. The basic principle of balanced interconnection is to get the signal you want by subtraction, using a three-wire connection. In some cases a balanced input is driven by a balanced output, with two anti-phase output signals, one

signal wire (the hot or in-phase) sensing the in-phase output of the sending unit, while the other senses the anti-phase output.

In other cases, when a balanced input is driven by an unbalanced output, as shown in Figure 20.3, one signal wire (the hot or in-phase) senses the single output of the sending unit, while the other (the cold or phase-inverted) senses the unit's output-socket ground, and once again the difference between them gives the wanted signal. In either of these two cases, any noise voltages that appear identically on both lines (i.e. common-mode signals) are in theory completely canceled by the subtraction. In real life the subtraction falls short of perfection, as the gains via the hot and cold inputs will not be precisely the same, and the degree of discrimination actually achieved is called the common-mode rejection ratio (CMRR), of which more later.

It is tedious to keep referring to non-inverting and inverting inputs, and so these are usually abbreviated to 'hot' and 'cold' respectively, though this does not necessarily mean that the hot terminal carries more signal voltage than the cold one. For a true balanced connection, the voltages will be equal. The 'hot' and 'cold' terminals are also often referred to as In+ and In−, and this latter convention has been followed in the diagrams here.

The subject of balanced interconnections is a large and subtle one, and a big fat book could be written on this topic alone. A classic paper on the subject is by Muncy[1]. To keep it to a reasonable length, this section has to concentrate on the areas most relevant to power amplifier interconnection.

Advantages

- Balanced interconnections discriminate against noise and crosstalk, whether they result from ground currents, or electrostatic or magnetic coupling to signal conductors.
- Balanced connections make ground loops much less intrusive, and usually inaudible, so people are less tempted to start 'lifting grounds' to break the loop. This tactic is only acceptable if the equipment has a dedicated ground-lift switch that leaves the external metalwork firmly connected to mains safety earth. In the absence of this facility, the optimistic will remove the mains earth (not quite so easy now that molded mains plugs are standard) and this practice is of course dangerous, as a short-circuit from mains to the equipment chassis will result in live metalwork.

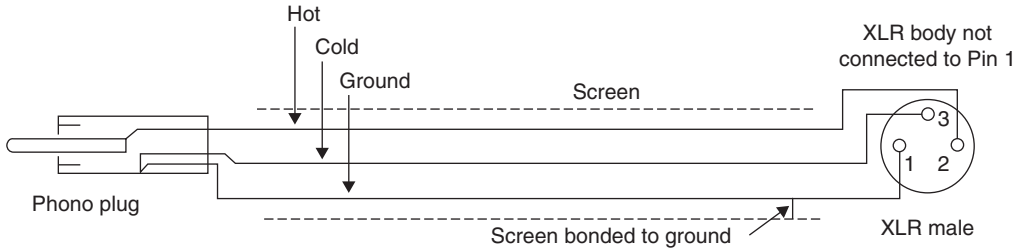


Figure 20.3: Unbalanced output to balanced input interconnection

- A balanced interconnection incorporating a true balanced output gives 6 dB more signal level on the line, which should give 6 dB more dynamic range. However, this is true only with respect to external noise – as the section below describes, the electronics of a standard balanced input is more than 6 dB noisier than the electronics of an unbalanced input.
- Balanced connections are usually made with XLR connectors. These are a professional three-pin format, and are a much superior connector to the phono (RCA) type normally used for unbalanced connections (more on this below).

Disadvantages

- Balanced inputs are inherently noisier than unbalanced inputs by a large margin, in terms of the noise generated by the input circuitry itself rather than external noise. This may appear paradoxical but it is all too true, and the reasons will be fully explained in this chapter.
- More hardware means more cost. Small-signal electronics is relatively cheap; unless you are using a sophisticated low-noise input stage (of which more later), most of the extra cost is likely to be in the balanced input connectors.
- Balanced connections may not provide much protection against RF intrusion – both legs of the balanced input would have to demodulate the RF in equal measure for common-mode cancelation to occur. This is not very likely, and it is important to provide the usual input RF filtering to avoid EMC difficulties.
- There are more possibilities for error when wiring up. For example, it is easy to introduce an unwanted phase inversion by confusing hot and cold in a connector, and this can go undiscovered for some time. The same mistake on an unbalanced system interrupts the audio completely.

While two wires carry the signal, the third is the ground wire, which has the dual duty of both joining the grounds of the interconnected equipment and electrostatically screening the two signal wires by being in some way wrapped around them. The ‘wrapping around’ bit can mean:

1. A lapped screen, with wires laid parallel to the central signal conductor. The screening coverage is not total, and can be badly degraded as the screen tends to open up on the outside of cable bends.
2. A braided screen around the central signal wires. This is more expensive, as it is harder to make, but opens up less on bends. Screening is not 100%, but certainly better than lapped screen.
3. An overlapping foil screen, with the ground wire (called the drain wire in this context for some reason) running down the inside of the foil and in electrical contact with it. This is usually the most effective as the foil is a solid sheet and cannot open up on the outside of bends. It should give perfect electrostatic screening. However, the higher resistance of aluminum foil compared with copper braid means that RF screening may be worse.

There are three ways in which any interconnection is vulnerable to hum and noise:

1. Electrostatic coupling. An interfering signal with significant voltage amplitude couples directly to the inner signal line, through stray capacitance. The stray capacitance between imperfectly screened conductors will be a fraction of a pF in most circumstances, as electrostatic coupling falls off with the square of distance. This form of coupling can be serious in studio installations with unrelated signals going down the same ducting.

The two main lines of defense against electrostatic coupling are effective screening and low impedance drive. An overlapping foil screen (such as used on Belden microphone cable) provides complete protection. Driving the line from a low impedance, of the order of 100 Ω or less, is also helpful because the interfering signal, having passed through a very small capacitance, is a very small current and cannot develop much voltage across such a low impedance. This is highly convenient because there are lots of other reasons for using a low output impedance, such as optimizing CMRR and avoiding HF loss due to cable signal-ground capacitance. For the best effectiveness the output impedance must remain low up to as high a frequency as possible; this can be a problem as op-amps invariably have a feedback factor that begins to fall from a low and possibly sub-audio frequency, and this makes the output impedance rise with frequency. From the point of view of electrostatic screening alone, the screen does not need to be grounded at both ends, or form part of a circuit[2]. It must of course be grounded at some point. Rearranging the cable-run away from the source of interference and getting some properly screened cable is more practical and more effective than trying to rely on very good common-mode rejection.

Stereo hi-fi balanced interconnections almost invariably use XLR connectors. Since an XLR can only handle one balanced channel, two separate cables are almost invariably used and interchannel capacitive crosstalk is not an issue. Professional systems, on the other hand, use multi-way connectors that usually do not have screening between the pins and there is an opportunity for capacitive crosstalk here.

2. Magnetic coupling. If a cable runs through an AC magnetic field, an EMF is induced in both signal conductors and the screen, and according to some writers the screen current must be allowed to flow freely or its magnetic field will not cancel out the field acting on the signal conductors, and therefore the screen should be grounded at both ends, to form a circuit[3]. In practice the magnetic field cancelation will be very imperfect and most reliance is placed on the common-mode rejection of the balanced system to cancel out the hopefully equal voltages V_m induced in the two signal wires. The need to ground both ends for magnetic rejection is not a restriction, as there are other good reasons why the screens should be grounded at both ends of a cable.

In critical situations the equality of these voltages is maximized by minimizing the loop area between the two signal wires, usually by twisting them tightly together in manufacture. In practice most audio cables have parallel rather than twisted signal conductors, and this seems adequate almost all of the time. Magnetic coupling falls off with the square of distance, so rearranging the cable-run away from the source of magnetic field is usually all that is required. It is unusual for it to present serious difficulties in a domestic environment.

3. Ground voltages coupled in through the common ground impedance, often called 'common-impedance coupling' in the literature[1]. This is the root of most ground-loop problems. In Figure 18.3 the equipment safety grounds cause a loop BCDGF; the mere existence of a loop in itself does no harm, but it is invariably immersed in a 50 Hz magnetic field that will induce mains-frequency current plus odd harmonics into it. This current produces a voltage drop down the non-negligible ground-wire resistance, and this once again effectively appears as a voltage source in each of the two signal lines. Since the CMRR is finite a proportion of this voltage will appear to be differential signal, and will be reproduced as such.

Common-Mode Rejection Ratio

Figure 20.4 shows a balanced interconnection reduced to its bare essentials: two source resistances and a standard differential amplifier. The balanced output in the source equipment is assumed to have two exactly equal output resistances R_{out+}, R_{out-}, and the balanced input in the receiving equipment has two exactly equal input resistances $R1$, $R2$. The balanced input amplifier senses the voltage difference between the points marked In+ (hot) and In− (cold) and ideally completely ignores common-mode voltages that are present on both. The amount by which it discriminates is called the common-mode rejection ratio or CMRR, and is usually measured in dB. Suppose a differential voltage input between In+ and In− gives an output voltage of 0 dB; then reconnect the input so that In+ and In− are joined together and the same voltage is applied between the two of them and ground. Ideally the result would be zero output, but in this imperfect world it won't be, and in real life the output could be anywhere between −20 dB (for a bad balanced interconnection) and −140 dB (for a very good one). The CMRR when plotted may have a flat section at low frequencies, but it commonly degrades at high audio frequencies (more on this later).

In one respect balanced audio connections have it easy. The common-mode signal is normally well below the level of the unwanted signal, and so the common-mode range of the input is not an issue.

The extremely simplified circuit of Figure 20.4, with a little SPICE simulation, demonstrates the need to get these resistor values right for good CMRR, before you even consider the rest of the

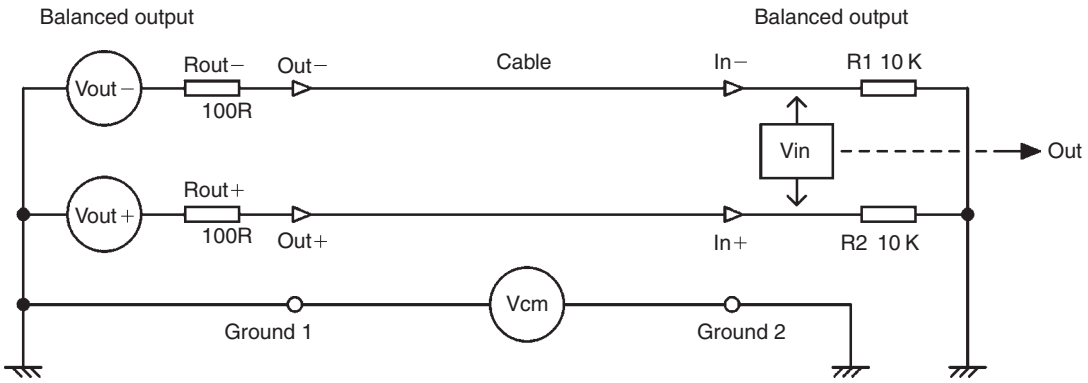


Figure 20.4: Balanced interconnection showing influences on CMRR

circuitry. The differential voltage sources V_{out+}, V_{out-} that represent the actual balanced output are set to zero, and V_{cm}, which represents the common-mode voltage drop down the cable ground, is set to 1 V to give a convenient result in dBv. The output resulting from the presence of this voltage source is measured by a mathematical subtraction of the voltages at In+ and In− so there is no actual input amplifier to confuse the results with its non-ideal performance.

Let us start out with R_{out+}, $R_{out-} = 100\,\Omega$ and $R1$, $R2 = 10\,k$, which are plausible values and nice round figures. When all four resistances are exactly at their nominal value, the CMRR is infinite, which on my simulator rather worryingly appears to be exactly −400 dB (presumably that is the mathematical 'noise floor'). If one of the output resistors or one of the input resistors is then altered in value by 1%, then the CMRR drops like a stone to −80 dB. If the deviation is 10%, things are predictably worse and the CMRR degrades to −60 dB, as shown in Table 20.2. That would be quite a good figure in real use, but since we have not begun to consider op-amp imperfections or other circuit imbalances, and have only altered one resistance out of the four that will in real circuitry all have their own tolerances, it's a bit unsettling. Clearly we need to understand how to improve things at this theoretical level before we start to complicate the circuitry.

The essence of the problem is that we have two resistive dividers, and we want them to have exactly the same attenuation. If we increase the ratio between the output and input resistors, by reducing the former or increasing the latter, the attenuation gets closer to unity and variations in either resistor have less effect on it. If we increase the input impedance to 100 k, putting aside the technical implications of doing this for the moment, things get 10 times better, as the R_{in}/R_{out} ratio has improved from 100 to 1000 times. We now get −100 dB for a 1% resistance deviation and −80 dB for a 10% deviation. An even higher input impedance of 1 MΩ, if it can be managed,

Table 20.2: How resistor tolerances affect the theoretical CMRR

R_{out+}	R_{out-}	R_{out} deviation	$R1$	$R2$	$R1$, $R2$ deviation	R_{in}/R_{out} ratio	CMRR (dB)
100	100	0	10 k	10 k	0	100	Infinity
100	101	1%	10 k	10 k	0	100	−80.2
100	110	10%	10 k	10 k	0	100	−60.2
100	100	0	10 k	10.1 k	1%	100	−80.3
100	100	0	10 k	11 k	10%	100	−61.0
100	100	0	100 k	101 k	1%	1000	−100.1
100	100	0	100 k	110 k	10%	1000	−80.8
100	100	0	1 M	1.01 M	1%	10,000	−120.1
100	100	0	1 M	1.1 M	10%	10,000	−100.8
68	68	0	20 k	20.2 k	1%	294	−89.5
68	68	0	20 k	22 k	10%	294	−70.3

raises R_{in}/R_{out} to 10,000, and gives −120 dB for a 1% resistance deviation and −100 dB for a 10% deviation.

As another angle of attack, we can reduce the output impedances to 10 Ω, ignoring for the moment the need to secure against HF instability caused by line capacitance, and return to an input impedance of 100 k. This again yields, as you have probably guessed, −120 dB for a 1% deviation and −100 dB for a 10% deviation.

In practical circuits, the combination of 68 Ω output resistors and a 20 kΩ input impedance is often encountered; the 68 Ω resistors are about as low as you want to go with conventional circuitry, to avoid HF instability. The 20 kΩ input impedance is what you get if you make a basic balanced input amplifier with four 10 kΩ resistors. I strongly suspect that this value is chosen because it looks as if it gives standard 10 kΩ input impedances – in fact it does nothing of the sort, and the common-mode input impedance, which is what matters here, is 20 kΩ on each leg (more on this later). It turns out that 68 Ω output resistors and a 20 kΩ input impedance give a CMRR of −89.5 dB for a 1% deviation, which is not at all bad. All these results are summarized in Table 20.2.

The conclusion is simple: we want to have the lowest possible output impedances and the highest possible input impedances to get the maximum common-mode rejection. This is highly convenient because low output impedances are already needed to drive multiple amplifier inputs and cable capacitance, and high input impedances are needed to minimize loading and maximize the number of amplifiers that can be driven.

Balanced Connectors

Balanced connections are most commonly made with XLR connectors. These are a professional three-pin format, and are a much better connector in every way than the usual phono (RCA) type. Phono connectors have the great disadvantage that if you are connecting them with the system active (inadvisable, but people are always doing inadvisable things) the signal contacts meet before the grounds and thunderous noises result. XLRs are wired with Pin 2 as hot, Pin 3 as cold, and Pin 1 as ground. The main alternative to the XLR is the stereo jack plug. These are often used for line-level signals in a recording environment, and are frequently found on the rear of professional power amplifiers as an alternative to an adjacent XLR connector. Jack sockets can be obtained with switching contacts that can be used to disable the XLR input to prevent the intrusion of noise. Balanced jacks are wired with the tip as hot, the ring as cold, and the sleeve as ground. Big sound-reinforcement systems often use large multi-way connectors that carry dozens of three-wire balanced connections.

Balanced Signal Levels

Many pieces of equipment, including preamplifiers and power amplifiers designed to work together, have both unbalanced and balanced inputs and outputs. The consensus in the hi-fi world appears to be that if the unbalanced output is say 1 V rms, then the balanced output will be created

simply by feeding the in-phase output to the hot output pin, and also to a unity-gain inverting stage, which drives the cold output pin with 1 V rms phase-inverted. The total balanced output voltage is therefore 2 V rms, and so the balanced input must have a gain of 0.5 or −6 dB relative to the unbalanced input to maintain consistent signal levels.

Balanced Inputs: Electronic versus Transformer

Balanced interconnections can be made using either transformer or electronic balancing. Electronic balancing has many advantages, such as low cost, low size and weight, superior frequency and transient response, and no problems with low-frequency linearity. While it is still sometimes regarded as a second-best solution, the performance is more than adequate for most professional applications. Transformer balancing has some advantages of its own, particularly for work in very hostile RF/EMC environments, but serious drawbacks. The advantages are that transformers are electrically bulletproof, retain their high CMRR performance forever, and consume no power even at high signal levels. Unfortunately they also generate LF distortion, particularly if they have been made as small as possible to save weight and cost. They tend to have HF response problems due to leakage reactance and distributed capacitance, and inevitably they are heavy and expensive compared with any electronic input. The first two objections can be surmounted with extra electronic circuitry, but the last two cannot. Transformer balancing is therefore relatively rare, even in professional audio applications, and most of this chapter deals with electronically balanced inputs.

The Basic Balanced Input

Figure 20.5 shows the basic balanced input amplifier using a single op-amp. To make it balanced R1 must be equal to R3 and R2 equal to R4. The amplifier in Figure 20.5 has a gain of R2/R1 (= R4/R3). The standard one-op-amp balanced input or differential amplifier is a very familiar circuit block, but its operation often appears somewhat mysterious. Its input impedances are *not* equal when it is driven from a balanced output; this has often been commented on[4] and some confusion has resulted.

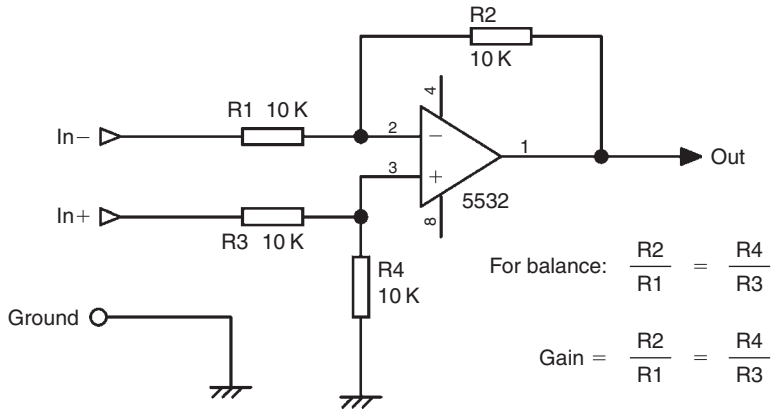


Figure 20.5: The basic balanced input amplifier

The root of the problem is that a simple differential amplifier has interaction between the two inputs, so that the input impedance on the cold input depends strongly on the signal applied to the hot input. Since the only way to measure input impedance is to apply a signal and see how much current flows into the input, it follows that the apparent input impedance on each leg varies according to the way the inputs are driven. If the amplifier is made with four 10k resistors, then the input impedances *Z* are as in Table 20.3.

Some of these impedances are not exactly what you would expect, and require some explanation. In Case 1, the balanced input is being used as an unbalanced input by grounding the cold input and driving the hot input only. The input impedance is therefore simply R3 + R4. R3 and R4 reduce the signal by a factor of 0.5, but this loss is undone as R1 and R2 set the amplifier gain to two times, and the overall gain is unity. If the cold input is not grounded then the gain is 0.5 times.

In Case 2, the balanced input is again being used as an unbalanced input, but this time by grounding the hot input and driving the cold input only. This gives a phase inversion and it is unlikely you would want to do it except in an emergency, to correct a phase error somewhere else. The instructive thing about this case is that the input impedance is now only 10k, the value of R1, because negative feedback through R2 creates a virtual earth at Pin 2 of the op-amp. This indicates that this simple circuit is not as symmetrical as it looks. The gain is unity, whether or not the hot input is grounded; however, if it is left floating the noise performance will be worse.

Case 3 is the standard balanced interconnection. Here the input is driven from a balanced output with the same signal levels on hot and cold, as if from a transformer with its center-tap grounded. The input impedance on the hot input is what you would expect: R3 + R4 add up to 20kΩ. However, on the cold input there is a much lower input impedance of 6.66kΩ, which at first sounds impossible as the first thing the signal encounters is a 10k series resistor, but the crucial point is that the hot input is being driven simultaneously with a signal of the opposite phase, so the inverting op-amp input is moving in the opposite direction to the cold input due to negative feedback, a sort of anti-bootstrapping that reduces the effective value of the 10k resistor to 6.66k. The vital point here is that these are the differential input impedances we are looking at, the impedances experienced by the balanced output driving them. Common-mode signals see a common-mode impedance of 20kΩ, as in Case 4 below. You will sometimes see the misguided statement that these unequal differential input impedances 'unbalance the line'. From the point of view of CMRR, this is not the case. The line is, however, unbalanced in the sense that the cold

Table 20.3: The input impedances for different input drive conditions

Case	Pins driven	Hot input res.	Cold input res.
1	Hot only	20k	Grounded
2	Cold only	Grounded	10k
3	Both (balanced)	20k	6.66k
4	Both common-mode	20k	20k
5	Both floating	10k	10k

input draws three times the current from the output than the hot one does. This imbalance might conceivably lead to inductive crosstalk in some multi-way cable situations, but I have never seen it. The differential input impedances can be made equal by increasing the R1 and R2 resistor values by a factor of 3, but this makes the noise performance worse and makes the common-mode impedances to ground unequal, which is usually much more undesirable.

In Case 4, both inputs are driven from the same signal, representing the existence of a common-mode signal. Now both inputs show an impedance of 20 kΩ. It is the symmetry of the common-mode input impedances that determines how effectively the balanced input rejects the common-mode signal.

In Case 5, where the input is driven as from a floating transformer with the center-tap (if any) unconnected, the impedances are nice and equal; they must be, because with a floating winding the same current must flow into each input. However, in this connection the line voltages are not equal and opposite: with a true floating transformer winding the hot input has all the signal voltage on it while the cold has none at all, due to the negative-feedback action of the balanced input amplifier. This seemed very strange when it emerged in SPICE simulation, but a reality check proved it true. The line has been completely unbalanced as regards talking to other lines, although its own common-mode rejection remains good. Even if perfectly matched resistors are assumed, the CMRR of this stage is not infinite; with a TL072 it is about −90 dB, degrading from 100 Hz upwards, due to the limited open-loop gain of the op-amp.

Common-Mode Rejection in the Basic Balanced Input

It is now time to look at the effect the op-amp itself has in the basic differential input circuit. Even if perfectly matched resistors are assumed, the CMRR of this stage is not infinite; the mathematics of the subtraction assume that the two op-amp inputs are at exactly the same voltage. This is an approximation because it depends on the efficacy of the negative feedback, and hence on the open-loop gain of the op-amp, and that is neither infinite nor flat with frequency up to the ultraviolet. Some more SPICE simulation is instructive. Figure 20.6 shows a simple balanced interconnection,

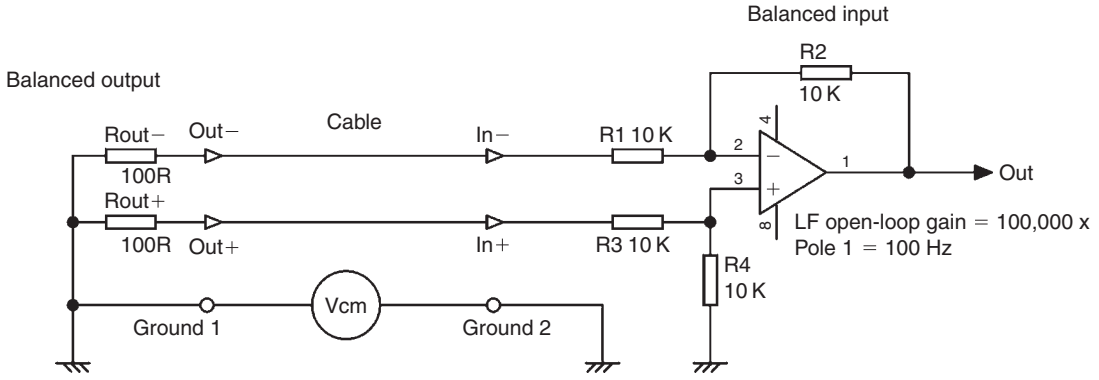


Figure 20.6: A simple balanced interconnection for SPICE simulation to show the effect that op-amp properties have on the CMRR

with the balanced output represented simply by two 100 Ω output resistances connected to the source equipment ground (Ground 1).

A common-mode voltage is injected between Ground 1 and Ground 2 by the voltage source V_{cm}, and the output is measured between the op-amp output and Ground 2. The balanced input amplifier is our standard differential amplifier with all resistances set to precisely 10 kΩ, and the op-amp represented by a simple model that has only two parameters: a low-frequency gain and a single-pole frequency that says where that gain begins to roll-off with a 6 dB/octave slope. The op-amp input impedances and the op-amp's own CMRR are infinite, as in the world of simulation they so easily can be. Its output impedance is zero. For the first experiments, even the pole frequency is made infinite, so now the only contact left with harsh reality is that the op-amp gain is finite. That is enough to give distinctly non-ideal CMRR figures, as Table 20.4 shows.

Thus for a low-frequency gain to 100,000, which happens to be the typical figure for a 5532 op-amp, even with perfect components the CMRR can never be better than −94 dB. The ratio is shown in the third column to emphasize that the common-mode error is inversely proportional to the gain.

Let us now set the low-frequency gain to 100,000, which gives a CMRR 'floor' of −94 dB. If we now introduce the pole frequency that determines where it begins to roll off, the result is that the CMRR now degrades at 6 dB/octave from a frequency that is set by the interaction of the low-frequency gain and the pole frequency. The results are summarized in Table 20.5, which shows that, as you might expect, the lower the open-loop bandwidth of the op-amp, the lower the frequency at which the CMRR begins to degrade. Figure 20.7 shows the situation diagrammatically.

Table 20.6 gives the gain and pole parameters for a few op-amps of interest. Both parameters, but especially the gain, are subject to considerable variation; the typical values from the manufacturers' data sheets are given here.

Table 20.4: The effect of finite op-amp gain on CMRR

Gain	CMRR (dB)	CMRR ratio
10,000	−74.0	19.9×10^{-5}
30,000	−83.6	66.4×10^{-6}
100,000	−94.0	19.9×10^{-6}
300,000	−103.6	6.64×10^{-6}
1,000,000	−114.1	1.97×10^{-6}

Table 20.5: The effect of op-amp open-loop pole frequency on CMRR

Pole frequency	CMRR breakpoint
10 kHz	10.2 kHz
1 kHz	1.02 kHz
100 Hz	102 Hz
10 Hz	10.2 Hz

Some of these op-amps have very high gains, but only at very low frequencies. This is good for DC applications, but in audio line input applications, where the lowest frequency of CMRR interest is 50 Hz, they will be operating above the pole frequency and so the gain available will be proportionately less.

Using a TL072 the CMRR is about −90 dB, degrading from 100 Hz upwards, due to the limited open-loop gain of the op-amp.

We have seen earlier in this chapter that the output and input impedances on a balanced interconnection must be both correctly proportioned and accurate in value for a good CMRR. That made no other assumptions about the actual input amplifier at the receiving end. We now need to know how deviations in resistor values in the amplifier itself compromise the CMRR. This is quite simple to calculate with a perfect op-amp, but if a finite op-amp gain is taken into account, SPICE simulation is much easier. In Table 20.7, LF gains of 100,000 and 1,000,000 are assumed, but the effect of finite op-amp bandwidth has not been taken into account. R1 is varied while R2, R3, and R4 are kept at exactly 10 kΩ. The balance output impedance is taken as exactly 100 Ω in each leg. The results give a good illustration of how resistor accuracy affects CMRR, but in real life – a phrase that keeps creeping in here, and shows how many factors affect a practical balanced interconnection – all four resistors will of course be subject to a tolerance, and a more realistic calculation would produce a statistical distribution of CMRR rather than a single figure. There is, however, not a great deal of point in attempting this unless you know how the resistor values are distributed within their tolerance window – usually you don't.

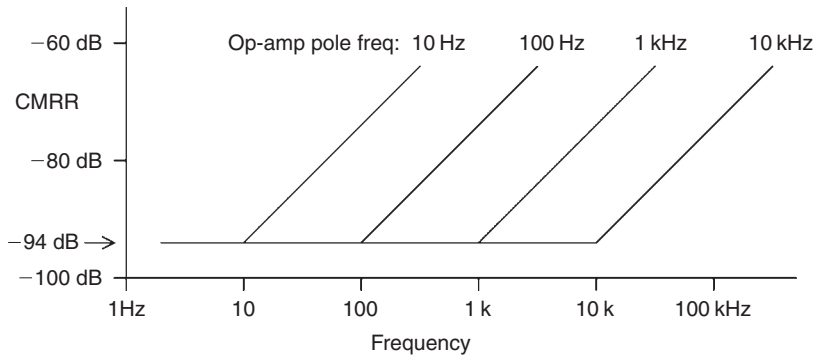


Figure 20.7: How the CMRR degrades with frequency for different op-amp pole frequencies. Resistors are assumed to be perfectly matched

Table 20.6: LF gain and open-loop pole frequency for some useful op-amps

Name	Type	LF gain	Pole
NE5532	Bipolar	100,000	100 Hz
LM4562	Bipolar	10,000,000	Below 10 Hz
TL072	FET	200,000	20 Hz
OPA2134	FET	1,000,000	3 Hz
OPA627	FET	1,000,000	20 Hz

Table 20.7: How resistor tolerances affect the CMRR with realistic op-amp O/I gain

R1	R1 deviation (%)	Gain (×)	CMRR (dB)
10 k	0	100,000	−94.0
10.0001 k	0.001	100,000	−96.5
10.001 k	0.01	100,000	−90.6
10.01 k	0.1	100,000	−66.5
10.1 k	1	100,000	−46.2
11 k	10	100,000	−26.6
10 k	0	1,000,000	−114.1
10.0001 k	0.001	1,000,000	−110.5
10.001 k	0.01	1,000,000	−86.5
10.01 k	0.1	1,000,000	−66.2
10.1 k	1	1,000,000	−46.2
11 k	10	1,000,000	−26.6

The results show immediately that our previous calculations, which took only output and input impedances into account, and determined that 68 Ω output resistors and a 20 kΩ input impedances gave a CMRR of −89.5 dB for a 1% deviation in either, were actually highly optimistic. If a 1% tolerance resistor is used for R1 (and at the time of writing this fifth edition there really is no financial incentive to use anything sloppier) the CMRR is dragged down to −46 dB; the same results apply to varying any other one of the four resistances. If you can run to a 0.1% tolerance for these components, the CMRR is a rather better −66 dB. This shows that there really is no point in worrying about the gain of the op-amp you use in balanced inputs; the effect of mismatches in the resistors around that op-amp are far greater. There are eight-pin SIL packages that offer four resistors that should have good matching; be wary of these as they usually contain thick-film resistive elements that are not perfectly linear. In a test I did a 10 k resistor with 10 V rms across it generated 0.0010% distortion. In the search for perfect audio, resistors that do not stick to Ohm's law are not a good start.

To conclude this section, it is obvious that errors in the resistors around the input amplifier have much more effect on the CMRR than do imbalances in the output resistance/input impedance system that we looked at earlier; this is because in a well-designed interconnection the output resistances and input impedances are of a different order of magnitude, whereas the amplifier resistors are, if not of the same value, then of the same order. In practical use it is the errors in the amplifier resistors that determine the CMRR, though either unbalanced capacitances (C1, C2 in Figure 20.8) or the finite op-amp bandwidth are likely to cause further degradation at high audio frequencies. If you are designing both ends of a balanced interconnection and you are spending money on precision resistors, you should put them in the input amplifier, not the balanced output. The LF gain of the op-amp has virtually no effect. It has to be said at this point that simple balanced input amplifiers with four 1% resistors are used extensively in the audio business, and almost always prove to be up to the job in terms of their CMRR. When a better CMRR is required, one of the resistances is made trimmable with a preset.

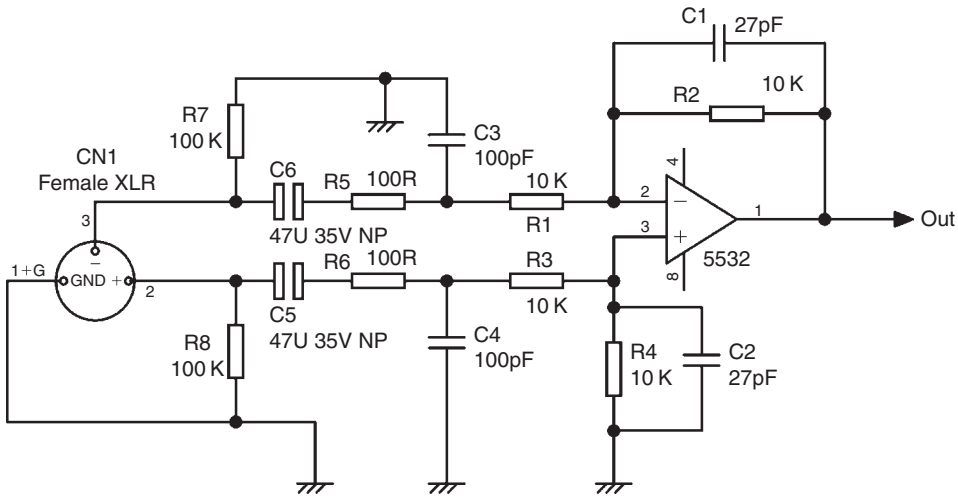


Figure 20.8: Balanced input amplifier with the extra components required for practical use

The Practical Balanced Input

The simple circuit shown in Figure 20.5 is not fit to face the world without some additional components. Figure 20.8 shows a more fully dressed version. Firstly, and most importantly, C1 has been added across the feedback resistor R2; this prevents stray capacitances from Pin 2 to ground causing extra phase shifts that lead to HF instability. The value required for stability is small, much less than that which would cause an HF roll-off anywhere near the top of the audio band. The values here of 10 k and 27 pF give −3 dB at 589 kHz, and such a roll-off is only down by 0.005 dB at 20 kHz. C2, of equal value, must be added across R4 to maintain the balance of the amplifier, and hence its CMRR, at high frequencies.

C1 and C2 are not particularly effective at RF rejection as C1 is not connected to ground, and there is every chance that RF will demodulate at the op-amp inputs. An RF filter is therefore added to each input, in the shape of R5, C3 and R6, C4. The capacitors here will shunt incoming RF to ground before it reaches the op-amp.

It was explained at some length in the previous section on unbalanced inputs that it is not at all easy to guess what the maximum source impedance will be, given the existence of so-called ‘passive preamplifiers’. Such a device clearly cannot give a balanced output (unless it is fitted with a transformer), but there is no reason why it could not be used to feed a balanced input, and so it needs consideration.

As before, circuit resistances must be kept as low as practicable to minimize the generation of Johnson noise. The situation here is, however, different from the unbalanced input as there have to be resistances around the op-amp, and they must be kept up to a certain value to give acceptably high input impedances; this is why a simple balanced input like this one is appreciably noisier than an unbalanced input. There is therefore more freedom in the selection of the values of R5 and R6. With the values shown, if we once more assume 50 Ω output impedances in both legs of the source

equipment output, then together with the 100 Ω resistances we have a total of 150 Ω; 150 Ω and 100 pF give −3 dB at 10.6 MHz.

Returning to a possible passive preamplifier with a 10 kΩ potentiometer, its maximum output impedance of 2.5 k plus 100 Ω with 100 pF gives −3 dB at 612 kHz, which remains well clear of the top of the audio band.

As with the unbalanced input, replacing R5 and R6 with small inductors will give much improved RF filtering but at increased component cost. Ideally a common-mode choke (usually two bifilar windings on a small toroid core) should be used as this improves performance. Once more, check the frequency response to make sure the LC circuits are well damped and not peaking at the turnover frequency.

C5 and C6 are DC-blocking capacitors. Once again they are rated at 35 V to protect the input circuit, and are nonpolarized types as voltages from outside may be of either polarity. The lowest input impedance that can occur with this circuit when using 10 kΩ resistors is, as described above, 6.66 kΩ when it is being driven in the balanced mode. The low-frequency roll-off is therefore −3 dB at 0.51 Hz. This may appear to be undesirably low, but the important point is not the roll-off but possible loss of CMRR at low frequencies due to imbalance in the values of C5 and C6; they are electrolytics and will have a significant tolerance. Therefore they should be made large so their impedance is a small part of the total circuit impedance; 47 μF is shown here but 100 or 220 μF can be used to advantage if there is the space to fit them in. The low-end frequency response must be defined somewhere in the input system, and the sooner the better, but this is not the place to do it.

R7, R8 are the DC drain resistors that prevent charges lingering on C5 and C6. These can be made lower in value than for the unbalanced input as the input impedances are lower, and a value of, say, 100 k rather than 220 k makes relatively little difference to the total input impedance.

There now follows a collection of balanced input circuits that offer advantages or extra features over the standard balanced input configuration that has just been described in remorseless detail. To make things clearer, the circuit diagrams mostly omit the stabilizing capacitors, input filters, and DC-blocking circuitry discussed above. They can be added in a straightforward manner; in particular bear in mind that a stabilizing capacitor like C1 is often needed between the op-amp output and the negative input to guarantee freedom from high-frequency oscillation.

Combined Unbalanced and Balanced Inputs

Very often both unbalanced and balanced inputs are required, and it is extremely convenient if it can be arranged so that no switching between them is required – switches cost money. A handy way to do this is shown in Figure 20.9, which for clarity omits most of the extra components required for practical use that are referred to above. For balanced use, simply connect to the balanced input and leave the unbalanced input unterminated. For an unbalanced input, simply connect to the unbalanced input and leave the balanced input unterminated. No mode switch is required. These unterminated inputs sound as though they would cause a lot of extra noise, but in fact the circuit works very well and I have used it with success in high-end equipment.

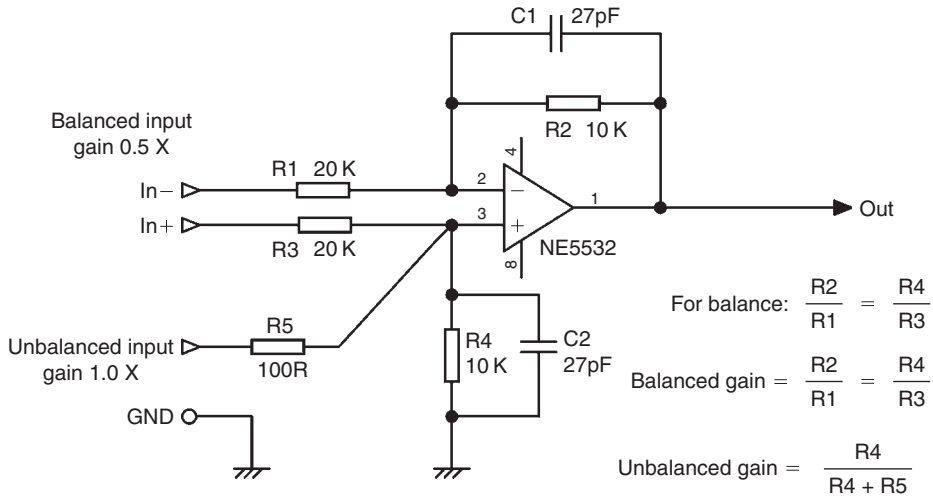


Figure 20.9: Combined balanced and unbalanced input amplifier with no switching required

As described above, in the world of hi-fi, balanced signals are at twice the level of the equivalent unbalanced signals, and so the balanced input must have a gain of 0.5 or −6 dB relative to the unbalanced input to keep the same system gain by either path. This is done here by increasing the value of R1 and R3 to 20 kΩ. The balanced gain of this circuit can be made either greater or less than unity, but the gain via the unbalanced input is always unity. The differential gain of the amplifier and the constraints on the component values for balanced operation are shown in Figure 20.5, and are not repeated in the text to save space. This applies to the rest of the balanced inputs in this chapter.

There are a few compromises in this scheme. The noise performance in the unbalanced input mode is worse than for the sort of dedicated unbalanced input circuitry described earlier in this chapter, because R2 remains effectively in the signal path in unbalanced mode. Also, the input impedance of the unbalanced input cannot be very high because it is determined by the value of R4, and if this is raised all the resistances around the op-amp must be increased proportionally and the noise performance is markedly worsened. A vital point is that only one input cable should be connected at a time. If an unterminated cable is left connected to an unused input, then the extra cable capacitance to ground will cause frequency-response anomalies and can in bad cases cause HF oscillation. A warning on the back panel is a very good idea.

Superbal Input

This version of the balanced input amplifier, shown in Figure 20.10, has been referred to as the 'Superbal' circuit because it gives equal impedances into the two inputs for differential signals. It was originated by David Birt of the BBC[5]. The configuration gives much better input symmetry than the standard differential amplifier, for with the circuit values shown the differential input impedance is exactly 10 k via both hot and cold inputs. Common-mode input impedance is 20 k for both inputs.

Because of the increased negative feedback, the gain with four equal resistors is −6 dB instead of the 0 dB given by the standard balanced input. The gain can be reduced below −6 dB by giving the inverter a gain of more than 1, but this is of limited use as the inverter U1:B will now clip before the forward amplifier U1:A, and the headroom will be impaired. If R1, R2, R3, and R4 are all equal, the gain is $1/(A + 1)$, where A is the gain of the inverter stage. If the gain of the inverter stage is gradually reduced from unity to zero, the stage slowly turns back into a standard balanced amplifier with the gain increasing from half to unity and the input impedances becoming less and less equal.

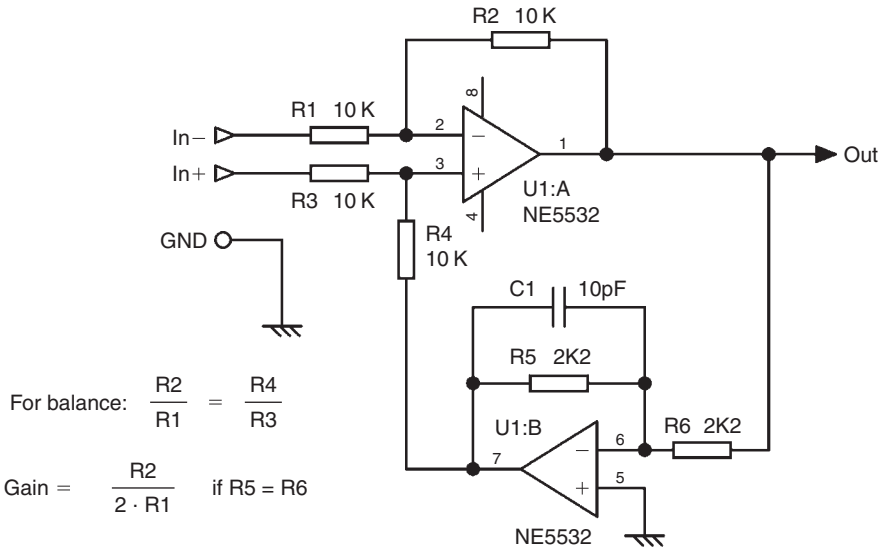


Figure 20.10: The Superbal balanced input amplifier

Note that R5 and R6 should be kept as low in value as possible to minimize their Johnson noise; they do not need to be equal in value to R1, etc. The only restriction is the ability of U1:A to drive R6 and U1:B to drive R5, both being effectively grounded at one end. The capacitor C1 will almost certainly be needed to ensure HF stability; the value in the figure is only a suggestion. It should be kept as small as possible because reducing the bandwidth of the U1:B stage impairs CMRR at high frequencies.

Switched-Gain Balanced Inputs

The need for a balanced input stage that can be switched to two different gains crops up frequently. Equipment often has to give an optimal performance with both semi-pro (−7.8 dBu) and

professional (+4 dBu) input levels. Depending on the nominal internal level of the amplifier input system, the input stage may have to be able to switch between amplifying and attenuating, while maintaining good CMRR in both modes.

The obvious way to change gain in a balanced input stage is to switch the values of either R1 and R3 or R2 and R4 in Figure 20.5, keeping the pairs equal in value to maintain the CMRR; this needs a double-pole switch for each input channel. A much more elegant technique is shown in Figure 20.11. Perhaps surprisingly, the gain of a differential amplifier can be manipulating by changing the drive to the feedback arm (R2, etc.) only, and leaving the other arm R4 unchanged, without affecting the CMRR. The crucial point is to keep the source resistance of the feedback arm the same, but drive it from a scaled version of the op-amp output. Figure 20.11 achieves this by the network R5, R6, which has a source resistance comprising 6k8 in parallel with 2k2, which comes to 1.662 kΩ. This is true whether R6 is switched to the op-amp output (low gain setting) or to ground (high gain setting), for both have effectively zero impedance. For low gain the negative feedback is not attenuated, but fed through to R2 and R7 via R5, R6 in parallel. For high gain R5 and R6 become a potential divider, so the amount of feedback is decreased and gain increased. The value of R2 + R7 is reduced from 7k5 by 1.662 kΩ to allow for the source impedance of the R5, R6 network; this requires the distinctly non-standard value of 5.838 kΩ, which is here approximated by R2 and R7, giving 5.6k + 240R = 5.840k. This value is the best that can be done with E24 resistors; it is obviously out by 2 Ω, but that is much less than a 1% tolerance on R2.

Note that this stage can attenuate as well as amplify if R1, R3 are set to be greater than R2, R4, as shown here. The nominal output level of the stage is assumed to be −2 dBu; with the values shown the two gains are −6.0 and +6.2 dB, so +4 and −7.8 dBu respectively will give −2 dBu at the output. Other pairs of gains can of course be obtained by changing the resistor values; the important thing is to stick to the principle that the value of R2 + R7 is reduced from the value of R4 by the source impedance of the R5, R6 network.

With the values shown the differential input impedance is 11.25 k via the cold and 22.5 k via the hot input. Common-mode input impedance is 22.5 k for both inputs.

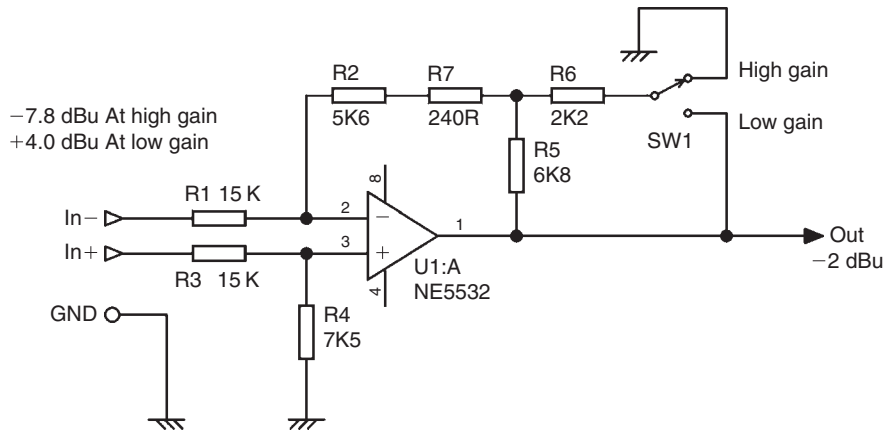


Figure 20.11: A balanced input amplifier with gain switching that maintains good CMRR

This circuit has the extra advantage that nothing bad happens when the switch is moved with the circuit operating. When the wiper is between contacts you simply get a gain that is intermediate between the high and low settings, which is pretty much the ideal situation. On the downside the CMRR may be slightly worse than usual because there are more resistor tolerances involved.

Variable-Gain Balanced Inputs

A variable-gain balanced input is advantageous because it gives the opportunity to get the incoming signal up to the desired internal level as soon as possible, exposing it to the minimum contamination from circuit noise. If the input stage can attenuate as well as amplify it also avoids the possibility of internal clipping that cannot be prevented because the stage doing the clipping is before the gain control. Unfortunately, making variable-gain differential stages is not so easy; the obvious method is to use dual-gang pots to vary two of the resistances, but this is clumsy and will give a CMRR that is both bad and highly variable due to the inevitable mismatches between pot sections. For a stereo input the resulting four-gang pot is not an attractive proposition.

There is, however, a way to get a variable gain with good CMRR and a single pot section. The principle is essentially the same as for the switched-gain amplifier above: to maintain constant the source impedance driving the feedback arm. The principle is shown in Figure 20.12. To the best of my knowledge I invented this circuit in 1982, but so often you eventually find out that you have re-invented rather than invented; any comments on this point are welcome. The feedback arm R2 is of constant resistance, and is driven by voltage-follower U1:B. This eliminates the variations in source impedance at the pot wiper, which would badly degrade the CMRR. R6 limits the gain range and R5 modifies the gain law to give it a more usable shape. Bear in mind that the center-detent gain may not be very accurate as it partly depends on the ratio of pot track (often no better than ±10%, and sometimes worse) to 1% fixed resistors.

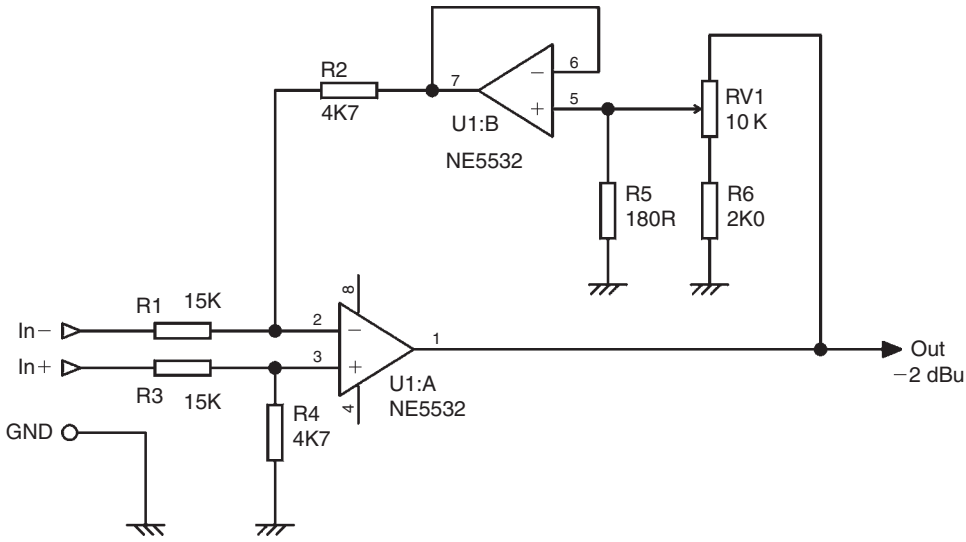


Figure 20.12: Variable-gain balanced input amplifier

This stage is very useful as a general line input with an input sensitivity range of −20 to +10 dBu. For a nominal output of 0 dBu, the gain of Figure 20.12 is +20 to −10 dB, with R5 chosen for 0 dB at the central wiper position. An op-amp in a feedback path appears a dubious proposition for stability, because of the extra phase shift it can introduce, but here it works as a voltage-follower, so its bandwidth is maximized and in practice the circuit is dependably stable.

High-Impedance Balanced Inputs

Since high input impedances are required to maximize CMRR, the circuit shown in Figure 20.13 has its uses because the input impedance is determined only by the input bias resistances R1 and R2. High-impedance balanced inputs are also useful for interfacing to valve equipment in the strange world of retro-hi-fi. Adding output cathode-followers to valve circuitry is expensive and consumes a lot of extra power, and so the output is often taken directly from a gain-stage anode; as a result even a so-called bridging loading of 10 k may seriously compromise the distortion performance and available output swing of the source equipment.

All of the balanced stages dealt with up to now have their input impedances determined by the values of input resistors, etc., and these cannot be raised without degrading noise performance. Figure 20.13 shows one answer to this. The op-amp inputs themselves pretty much have infinite impedance in audio terms, so the input impedance is determined by the need for R1, R2 to bias the non-inverting inputs. A remarkable and very useful property of this circuit is that adding the resistance R_g increases its gain, but preserves the circuit balance. This configuration cannot be set to attenuate because the gain of an op-amp with series feedback cannot be reduced below unity.

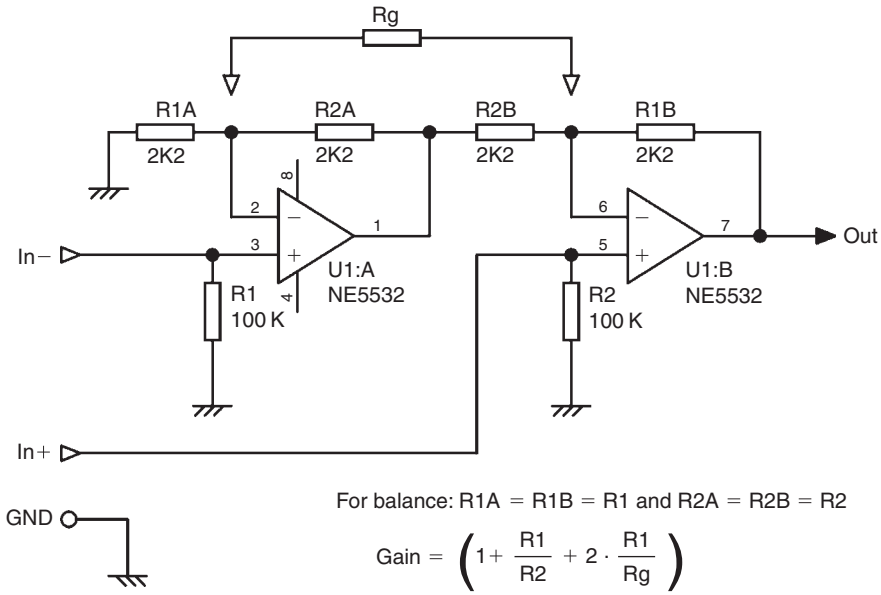


Figure 20.13: High-impedance balanced input

The Inverting Two-Op-Amp Input

This stage, depicted in Figure 20.14, is included here as it is sometimes advocated simply because the hot and cold inputs have the same impedances for differential signals, as well as for common-mode voltages[6]. It is not normally suited to high input impedances because high resistor values would have to be used that would generate excess Johnson noise, but the unique feature it does have is that it can cope with very high input levels if R1 and R3 are made high and R2, R4, and R5 are made low, the latter choice keeping the noise down. The CMRR may get worse at HF because the hot signal has gone through an extra op-amp and suffered phase shift; this can be compensated for by the network Rc-1, Rc-2, and C1; values depend on op-amp type.

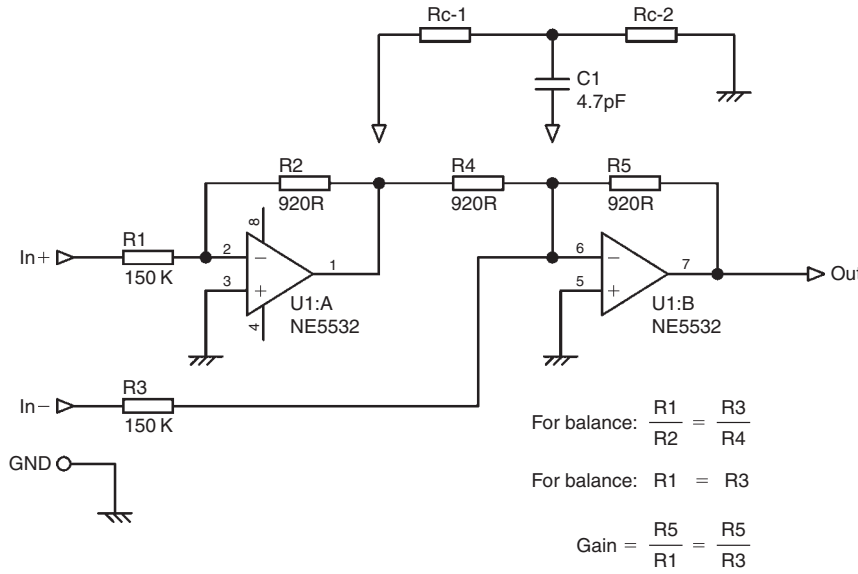


Figure 20.14: Inverting two-op-amp balanced input

The circuit values shown are suitable for connection to a 100 V loudspeaker line of the sort still in use for distributing audio over a wide area; it reduces the input voltage to a nominal −2 dBu (615 mV) internal level. If this circuit is used for that purpose then effective input overvoltage protection on the inputs is essential as large voltage transients are possible on a loudspeaker line if parts of it are unplugged or plugged with signal present, due to the inductance of the transformers connected to it. This protection can be conveniently provided by the usual diode clamps to the supply rails. Since the input resistors will be of high value in this application there is very little possibility of excessive inputs 'pumping up' the op-amp supply rails.

The Instrumentation Amplifier

Just about every book on balanced or differential inputs includes the three-op-amp circuit of Figure 20.15 and praises it as the highest expression of the differential amplifier. It is usually

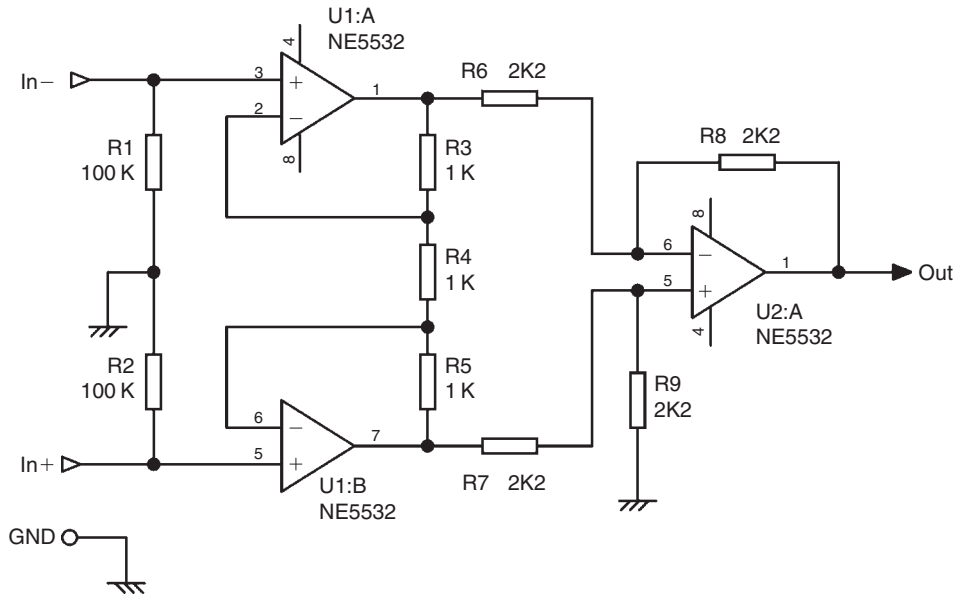


Figure 20.15: The instrumentation amplifier configuration

called the instrumentation amplifier configuration because of its undoubted superiority for data acquisition. (Note that specialized ICs do exist that are sometimes also called instrumentation amplifiers or in-amps; these are designed to give very high CMRR without external resistors. They are expensive and are not generally optimized for audio applications.)

The beauty of this configuration is that the dual input stage buffers the balanced line from the input impedances of the final differential amplifier; this means that the four resistances around the amplifier can be made much lower in value, reducing their Johnson noise by a significant amount, while retaining the CMRR benefits of presenting high input impedances to the balanced line. The other feature, which is usually much more emphasized because of its unquestionable elegance, is that the dual input stage can have a high differential gain, but the common-mode gain is always unity; this is not affected by mismatches in R3 and R5. The final amplifier then does its usual job of common-mode rejection, and the result can be a very high CMRR for high gains.

All well and good, but this stage as it stands is not very useful for audio balanced line inputs. A data-acquisition application may need a gain of 1000 times, which will allow a stunning CMRR to be achieved without using precision resistors, but the hard cruel fact is that in audio usage any gain at this point is very often simply not wanted. In a typical signal path comprised of op-amps, and powered from ±17 or ±15 V rails, the nominal internal level is usually between −6 and 0 dBu, and the level coming in is at the professional level of +4 dBu; what is needed is 6 dB of attenuation rather than gain. To introduce gain at this point and attenuation later would be to introduce what can only be described as a headroom bottleneck, if I may be permitted such a curdled metaphor. If the incoming level was the semi-pro −7.8 dBu a small amount of gain could be introduced, but then the CMRR advantage would be equally small, and certainly not worth the cost of two op-amp sections.

Transformer Balanced Inputs

When it is important that there is no galvanic connection (i.e. no electrical conductor) between two pieces of equipment, transformer inputs are indispensable. They are also useful if EMC conditions are severe. Figure 20.16 shows a typical transformer input. The transformer usually has a 1:1 ratio, and is enclosed in a metal shielding can that must be grounded. Good line transformers have an interwinding shield that must also be grounded or the high-frequency CMRR will be severely compromised. The transformer secondary must see a high impedance as this is reflected to the primary and represents the input impedance; here it is set by R2, and a buffer drives the circuitry downstream. In addition, if the loading is too heavy there will be increased transformer distortion at low frequencies. Line input transformers are built with small cores and are only intended to deliver very small amounts of power; they should not be used as line output transformers. An ingenious approach to solving the distortion problem by operating the transformer core at near-zero flux was published by Paul Zwicky in 1986[7].

There is a bit more to loading the transformer secondary correctly. If it is simply loaded with a high-value resistor there will be peaking of the frequency response due to resonance between the transformer leakage inductance and the winding capacitance. This is shown in Figure 20.17, where a Sowter 3276 line input transformer (a high-quality component) was given a basic resistive loading of 100 kΩ. The result was trace A, which has a 10 dB peak around 60 kHz. This is bad not only because it accentuates the effect of out-of-band noise, but because it also affects the audio frequency response, giving a lift of 1 dB at 20 kHz. Reducing the resistive load would damp the resonance, but it would also reduce the input impedance. The answer is to add a Zobel network, i.e. a resistor and capacitor in series, across the secondary; this has no effect except at high frequencies. The first attempt used R1 = 2k7 and C1 = 1 nF, giving trace B, where the peaking has been reduced to 4 dB around 40 kHz, but the 20 kHz level is actually slightly worse. R1 = 2k7 and C1 = 2 nF gave trace C, which is a bit better in that it only has a 2 dB peak. A bit more experimentation ended up with R1 = 3k3 and C1 = 4.3 nF (3n3 + 1 nF) and yielded trace D, which is pretty flat, though there is a small droop around 10 kHz. The Zobel values are fairly critical for the flattest possible response, and must certainly be adjusted if the transformer type is changed.

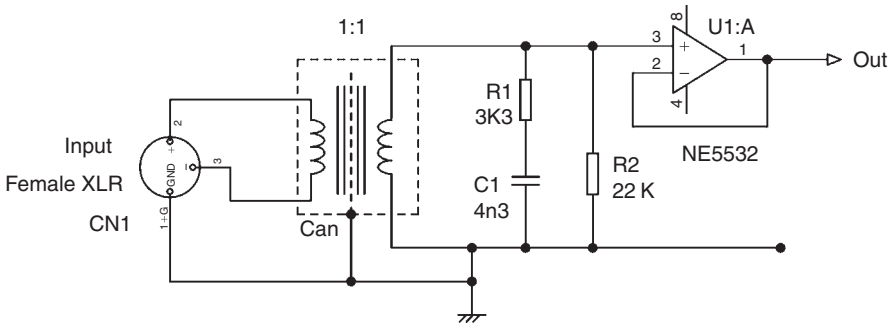


Figure 20.16: A transformer balanced input. R1 and C1 are the Zobel network that damps the secondary resonance

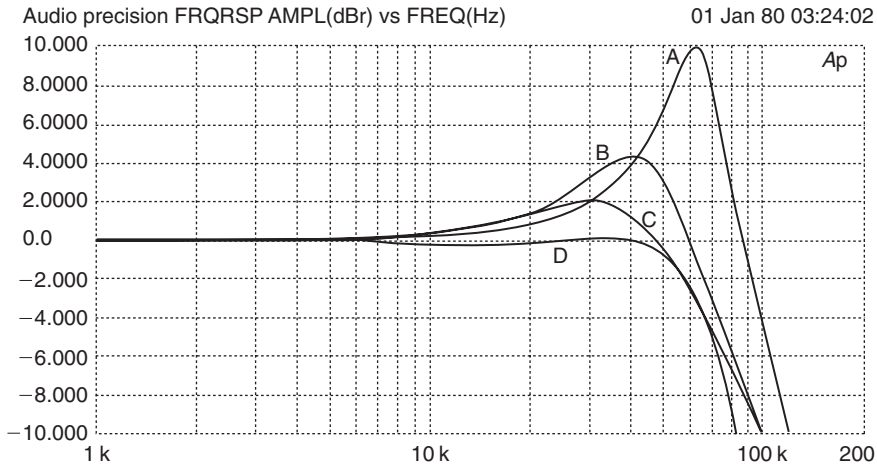


Figure 20.17: Optimizing the frequency response of a transformer balanced input with a Zobel network

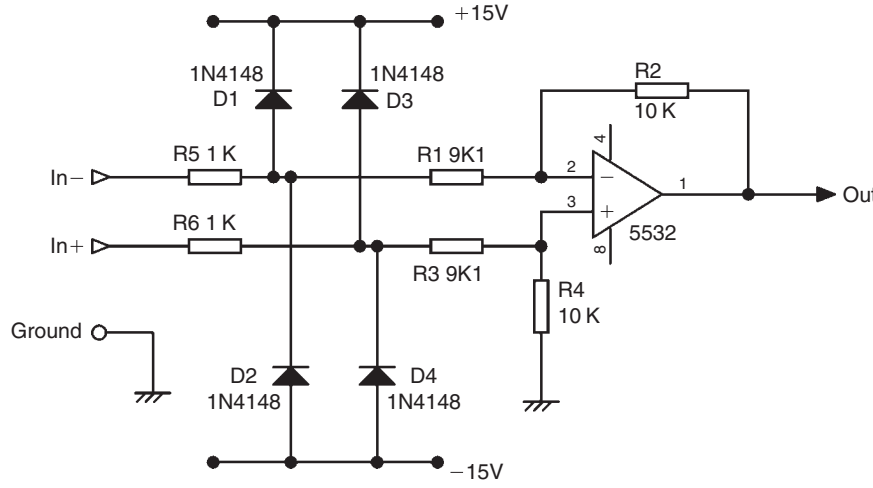


Figure 20.18: Input overvoltage protection for a balanced input amplifier

Input Overvoltage Protection

Input overvoltage protection is not common in hi-fi applications, but is regarded as essential in professional amplifier use. The normal method is to use clamping diodes, as shown in Figure 20.18, which prevent certain points in the input circuitry from moving outside the supply rails.

This is very straightforward, but there are two points to watch. Firstly, the ability of this circuit to withstand excessive input levels is not without limit. Sustained overvoltage may burn out R5 and R6, or pump unwanted amounts of current into the supply rails; this sort of protection is mainly aimed at transients. Secondly, diodes have a nonlinear junction capacitance when they are reverse-biased, so if the source impedance is significant the diodes will cause distortion at high frequencies. To quantify this problem here are a few figures. If the circuit of Figure 20.18 is being fed from

the usual kind of line output stage, the impedance at the diodes will be approximately 1 kΩ and the distortion introduced with an 11 V rms 20 kHz input will be below the noise floor. However, in a test I conducted where the impedance was increased to 10 kΩ with the same input, the THD at 20 kHz was degraded from 0.0030% to 0.0044% by adding the diodes. I have worked out a rather elegant way to eliminate this effect completely, but this is not the place to disclose it. As you might have guessed, I am rather hoping to sell the idea.

Noise and the Input System

In Chapter 4 the sources of noise inside the actual power amplifier circuitry were examined. As an example we looked at a real amplifier with a closed-loop gain of +30.6 dB and measured noise at the amplifier output of a pleasingly low −92 dBu over the standard 22–22 kHz bandwidth. This is a noise level referred to the amplifier input of only −122.6 dBu, and there is no obvious way to reduce it, so we will take it as fixed for the time being. What is immediately clear is that almost anything we connect to the input of this amplifier is going to compromise the noise performance, but by how much?

As a first example, consider that in some amplifier designs a first-order RC low-pass filter is placed immediately before the power amplifier input, in the pious hope of preventing slew-rate limiting. It was offered as a magical panacea against transient intermodulation distortion (TIM) before it became clear to everybody – and it took an unconscionably long time – that TIM was just another way of referring to slew-limiting, which rarely, if ever, occurred. This was always a dubious expedient if there was no buffering before the RC filter because the source resistance of the external equipment would affect the turnover frequency; the resistor was usually in the range 470 Ω–1 kΩ in the hope that this would be significantly greater than the external source resistance and so minimize the variation. If we take an 820 Ω resistance as typical (for some reason it does seem to have been a particularly popular value) then its Johnson noise is −123.5 dBu. If we rms sum this with the amplifier EIN of −122.6 dBu, then the result is −120.0 dBu, and we have degraded the noise performance by 2.6 dBu already with this one component, with no active circuitry upstream. (It might be as well to mention here that putting resistances directly in series with a power amplifier input is a bad plan for another reason – it induces input current distortion. This ticklish topic is dealt with in Chapter 4.)

Let us now see what the noise consequences are of putting active circuitry that actually does something useful in front of the power amplifier. I need to say at this point that the op-amp noise data quoted here is taken from extensive real-life measurements rather than theoretical calculations, but averaged over a relatively small number of samples. It is my experience that (providing you stick to reputable manufacturers) the noise performance of bipolar op-amps such as the 5532 shows relatively little variation. The aim here is to show the general principles of low-noise design rather than get too picky about the last decimal place.

Firstly, we put a simple unity-gain buffer in front of the power amplifier stage; this might be done to drive an input resistor that has been given a low value to minimize input offset voltages, so we can still present a high input impedance to the outside world, or to prevent input current distortion by

driving the amplifier from a very low impedance (the latter issue is described in Chapter 4). We will take the impedance seen by this buffer stage as 50Ω, which is about as low as we might hope for; the Johnson noise from this is only −135.2dBu. The noise output of a NE5532 unity-gain buffer with these input conditions is −119dBu; this is a very low value and is obtained only because there are no medium-value feedback resistances in the op-amp circuit and all we are seeing is the op-amp voltage noise. When this noise level is added to the amplifier EIN of −122.6dBu, it gives an rms total of −117.4dBu. We have added what at first sight is the quietest possible preceding stage but amplifier noise output has already been increased by 5.2dB. And now things are about to get worse.

It is much more useful to put a balanced input stage in front of the power amplifier itself, as it gives all the benefits of common-mode rejection. The standard one-op-amp unity-gain balanced input is very commonly made with four 10k resistors; this is a compromise that, as we have seen, gives a respectable if not stunning 20k common-mode impedance on each input, combined with levels of Johnson noise that are usually considered acceptable (see Figure 20.19a). The noise output of

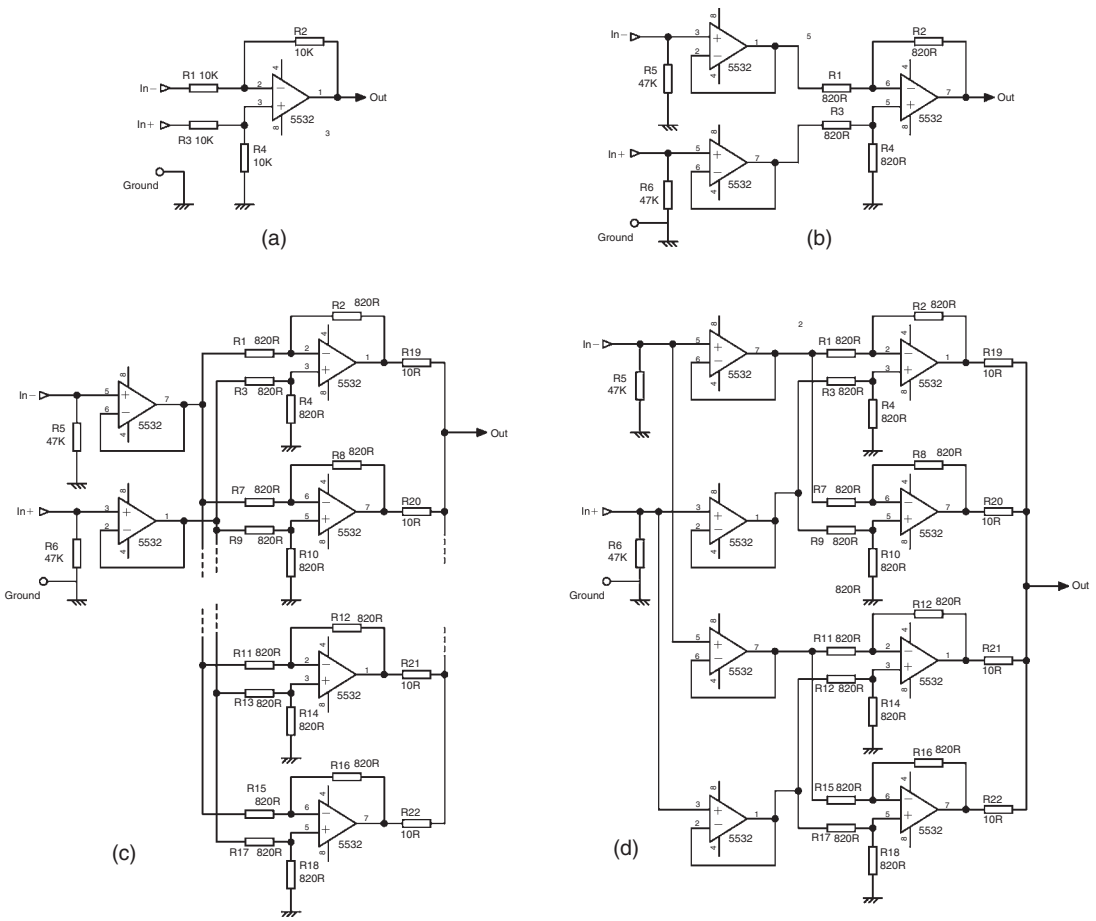


Figure 20.19: Low-noise unity-gain balanced inputs using multiple 5532 buffers and differential amplifiers

a 4 × 10k balanced input amp using an NE5532 is −105.1 dBu, which completely swamps the power amplifier EIN and degrades the overall noise performance by 17.5 dB. The noise at the amplifier output increases from −92 to −75 dBu. This is not a very happy outcome, but is the inevitable result of using conventional balanced input technology. The ideal would be an input stage that has negligible noise compared with the power amplifier; if we accept that the power amplifier noise should be increased by only 0.1 dB, a few seconds juggling with RMS-summation shows that the input stage noise output would have to be a very low −139 dBu. That is the Johnson noise of a 21 Ω resistor basking in a room temperature of 25°C, and is clearly a pretty tall order. If we reluctantly agree that the power amplifier noise can be worsened by 1.0 dB, which would be hard to detect even in ABX testing, we then need an input stage with a noise output of −128.5 dBu, which is still an ambitious target. We might decide, in a fleeting moment of realism, that we can live with equal noise contributions from the input stage and the power amplifier; in other words the input stage noise will degrade the overall noise output by 3 dB. That will require an input stage noise output that is the same as the power amplifier EIN, which is −122.6 dBu. That might be attainable, with a bit of thought. These noise requirements figures are summarized in Table 20.8, where the bottom line shows the effect of the standard balanced input with four 10k resistors.

So, we have a target: a balanced input stage with an output noise of −122.6 dBu. Let's see what can be done about it.

Low-Noise Balanced Inputs

We have seen that the noise output of the standard one-op-amp balanced input with four 10k resistors in Figure 20.19a is −105.1 dBu; we clearly need to apply some more serious electronics to the noise problem. In all the measurements that follow the source impedance was 50 Ω to ground on both inputs. We have seen that the instrumentation amplifier configuration is of limited use for audio work as it only gives an improvement in CMRR commensurate with the gain of its first stage. However, if we reduce it to a standard differential amplifier with a unity-gain buffer on each input, we can reduce the value of the four resistors around the final differential amplifier, reducing their Johnson noise, and at the same time increase the input impedance presented to the outside world, and so possibly improve the CMRR. This arrangement is shown in Figure 20.19b. There is a limit to how far the four resistors can be reduced, as the differential stage has to be driven by the input buffers, and it also has to drive its own feedback arm, but against this is the relatively small

Table 20.8: How noise from an input stage degrades power amplifier output noise performance

Input stage noise O/P (dBu)	Input noise summed with power amp EIN (dBu)	Power amp noise worsened by (dB)
−139.00	−122.50	0.10
−128.50	−121.61	0.99
−122.60	−119.59	3.01
−105.10	−105.02	17.58

part of the output swing that is used if the input stage is directly coupled to the power amplifier; the latter will always clip a long time before the op-amps get anywhere near their maximum output. If NE5532s are used a safe value that gives no measurable deterioration of the distortion performance is about 820R, and an NE5532 differential stage alone (without the buffers) and 4 × 820R resistors gives a noise output of −111.7 dBu, which is 6.6 dB lower than the standard 4 × 10k version. Adding the two input buffers degrades this only slightly to −110.2 dB, because we are adding only the voltage noise component of the two new op-amps, and we are still 5.1 dB quieter than the 4 × 10k version. The interesting point here is that we have three op-amps in the signal path instead of one, but we still get a significantly lower noise level. Power amp noise, however, is still degraded by 12.6 dB.

This might appear to be all we can do; it is not possible to reduce the value of the four resistors around the differential amplifier any further without compromising linearity. In fact, there is almost always some way to go further in the great game that is electronics, and here are three possibilities. A step-up transformer could be used to exploit the low source impedance (remember we are still assuming the source impedances are 50Ω) and it might well work superbly in terms of noise alone, but transformers are always heavy, expensive, susceptible to magnetic fields, and of doubtful low-frequency linearity. We could design a discrete-op-amp hybrid stage that uses discrete input transistors, which are quieter than those integrated into IC op-amps, coupled to an op-amp to provide raw loop gain; this can be effective but you need to be very careful about high-frequency stability. Thirdly, we could design our own op-amp using all discrete parts; this approach tends to have less stability problems but does require rather specialized skills, and the result takes up a lot of PCB area.

If those three expedients are rejected, now what? One of the most useful techniques in low-noise electronics is to use two identical amplifiers so that the gains add arithmetically, but the noise from the two separate amplifiers, being uncorrelated, partially cancels. Thus a 3 dB noise advantage is gained each time the number of amplifiers used is doubled. This technique works very well with multiple op-amps; let us apply it and see how far it may be usefully taken.

Since the noise of a single 5532 unity-gain buffer is only −119 dBu, and the noise from the 4 × 820R differential stage (without buffers) is a much higher −111.7 dBu, the differential stage is the place to start work. We will begin by using two identical 4 × 820R differential amplifiers as shown in Figure 20.19c, both driven from the existing pair of input buffers. This will give partial cancelation of both resistor and op-amp noise from the two stages when their outputs are summed. The main question is how to sum the two amplifier outputs; any active solution would introduce another op-amp, and hence more noise, and we would almost certainly wind up worse off than when we started. The answer is, however, beautifully simple. We just connect the two amplifier outputs together with 10Ω resistors; the gain does not change but the noise output drops. The signal output of both amplifiers is nominally the same, so no current should flow from one op-amp input to the other. In practice there will be slight gain differences due to resistor tolerances, but with 1% resistors I have never experienced any problems. The combining resistor values are so low at 10Ω that their Johnson noise contribution is negligible.

We therefore have the arrangement of Figure 20.19c, with single input buffers (i.e. one for each of the two inputs) and two differential amplifiers, and this drops the noise output by 2.3 dB to −112.5 dBu, which is quieter than the original 4 × 10 k version by a hefty 7.4 dB. We do not get the full 3 dB noise improvement because both differential amplifiers are handling the noise from the input buffers; this is correlated and so is not reduced by partial cancelation. Power amplifier output noise is now only worsened by 10.5 dB. The role of the input buffer noise is further emphasized if we take the next step of using four differential amplifiers. (There is nothing special about using amplifiers in powers of 2. It is perfectly possible to use three or five differential amplifiers in the array, which will give intermediate amounts of noise reduction.)

So, sticking with single input buffers, we try the effect of four differential amplifiers. These are added on at the dotted lines in Figure 20.19c. We get a further improvement, but only by 1.5 dB this time. The output noise is down to −114.0 dBu, quieter than the original 4 × 10 k version by 8.9 dB, but still making the power amplifier 9.2 dB noisier when connected. You can see that at this point we are proceeding by decreasing steps, as the input buffer noise begins to dominate, and there seems little point in doubling up the differential amplifiers again; the amount of hardware would be getting out of hand, as would the PCB area occupied. The increased loading on the input buffers is also a bit of a worry.

A more fruitful approach is to tackle the noise from the input buffers, by doubling them up as in Figure 20.19d, so that each buffer drives only two of the four differential amplifiers. This means that the buffer noise will also undergo partial cancelation, and will be reduced by 3 dB. There is, however, still the contribution from the differential amplifier noise, and so the actual improvement on the previous version is 2.2 dB, bringing the output noise down to −116.2 dBu. This is quieter than the original 4 × 10 k version by a thumping 11.1 dB, but still makes the power amplifier noisier by 7.3 dB, and we are some way short of our target of 3 dB. Using this input stage, if we rms-sum its noise output with the power amplifier EIN the effective input noise at the power amplifier is −115.3 dB, and the lesson is that the power amplifier noise is no longer negligible; now it increases the total noise at the output by 0.9 dB. Remember that there are two inputs, and 'double buffers' means two buffers per input, giving a total of four in the complete circuit.

Since doubling up the input buffers gave us a useful improvement, it's worth trying again, so we have a structure with quad buffers and four differential amplifiers, as shown in Figure 20.20, where each differential amplifier now has its very own buffer. This improves on the previous version by a rather less satisfying 0.8 dB, giving an output noise level of −117.0 dBu, quieter than the original 4 × 10 k version by 11.9 dB. Connecting this input stage to the power amplifier increases its noise output by 6.7 dB. The small improvement we have gained indicates that the focus of noise reduction needs to be returned to the differential amplifier array, but the next step would seem to be using eight amplifiers, which is not very appealing. Thoughts about ears of corn on chessboards tend to intrude at this point.

This is a good moment to pause and see what we have achieved. We have built a balanced input stage that is quieter than the standard circuit by 11.9 dB, using standard components of low cost. We have used increasing numbers of them, but the total cost is still small compared with power transistors,

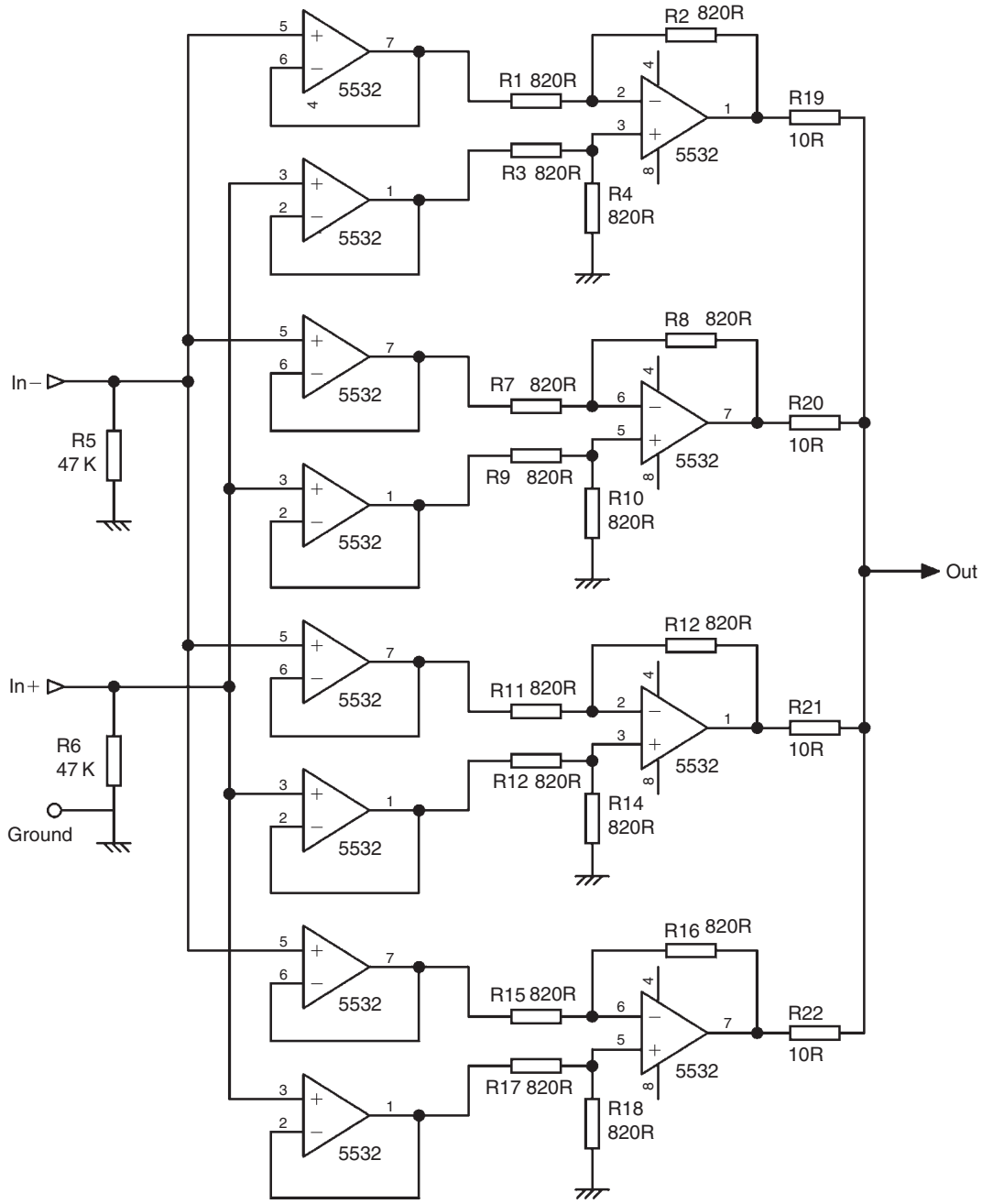


Figure 20.20: The final 5532 low-noise unity-gain balanced input stage, with quad input buffers and four differential amplifiers. The noise output is only −117.0 dBu

heat-sinks, and transformers. The power consumption is clearly greater, but trivial compared with the quiescent current of the average Class-B power amplifier. The technology is highly predictable and the noise reduction reliable, in fact bulletproof. The linearity is as good as that of a single op-amp of the same type, and in the same way there are no HF stability problems. Not bad.

. . . And Quieter Yet

In the last section we rested in our efforts, having achieved an economical balanced input stage with an output noise of −117.0 dBu. It would be wrong to conclude from this that the resources of electronic design are exhausted. We were aware that the dominant noise source was the differential amplifier array, and shrank from doubling it again to use eight amplifiers. What we will do now is put aside considerations of cost, and see how the removal of that constraint changes the game.

A certain way to make the differential amplifier array quieter is to use quieter (and more expensive) op-amps in it, the focus being on low-voltage noise rather than current noise because of the low resistor values. The LM4562 op-amp will certainly give a significant noise reduction, but the king of low-noise-with-low-source-resistance op-amps is the AD797. The AD797 (Analog Devices) is a single op-amp with no dual version, so more PCB area is required.

Firstly, we leave the quad 5532 input buffers as they are but replace all four op-amps in the differential amplifiers with expensive AD797s. The noise drops by an impressive 2.9 dB, giving an output noise level of −119.9 dBu, which is quieter than the original 4 × 10 k version by 14.8 dB. The power amplifier noise is now only degraded by 4.6 dB.

Our final flourish is to replace the quad 5532 input buffers with dual (not quad) AD797 buffers. This requires another four AD797s (two per input) and is once more not a cheap move. We retain the four AD797s in the differential amplifiers. The noise drops by another 0.7 dB, yielding an output noise level of −120.6 dBu, quieter than the original 4 × 10 k version by 15.5 dB. The power amplifier noise performance is now only degraded by 4.1 dB.

So, we have not quite reached the target of only degrading the power amplifier noise performance by 3 dB, but I suggest we have got commendably close. The small noise improvement in the last step we made tells us that the differential amplifier array is still the dominant noise source, and any further development would have to focus on this. A first step would be to see if the current noise of the AD797s is significant with respect to the surrounding resistor values, and if so to see if the resistor values can be further reduced.

These noise results are summarized in Table 20.9. The bottom three lines about 'notional input stages' relate to the choice of noise reduction target in Table 20.8.

Noise Reduction in Real Life

I don't want you to think that this noise-reduction exercise is simply a voyage off into pure theory. As an example of this technique in action, consider the Cambridge Audio 840 W power amplifier, which, I will modestly mention in passing, won a CES Innovation Award in January 2008. This unit has both unbalanced and balanced inputs, and for the reasons given above, conventional technology would have meant that the balanced inputs would have been the noisier of the two. Since the balanced input is the 'premium' input, many people would think that was the wrong way round. We therefore decided that the balanced input was required to be quieter than the unbalanced input. Using 5532s in a structure similar to those outlined above, this requirement proved quite

Table 20.9: A summary of the noise improvements made to the balanced input stage

Buffer	Differential amplifier	Input stage noise output (dBu)	Improvement on previous version (dB)	Improvement over 4 × 10k diff. amp (dB)	Power amp noise degraded by (dB)
None	Standard diff. amp 10k 5532	−105.10		0	17.58
None	Single diff. amp 820R 5532	−111.70	6.60	6.6	11.24
Single 5532	Single diff. amp 820R 5532	−110.20		5.1	12.64
Single 5532	Dual diff. amp 820R 5532	−112.50	2.30	7.4	10.50
Single 5532	Quad diff. amp 820R 5532	−114.00	1.50	8.9	9.16
Dual 5532	Quad diff. amp 820R 5532	−116.20	2.20	11.1	7.30
Quad 5532	Quad diff. amp 820R 5532	−117.00	0.80	11.9	6.66
Quad 5532	Quad diff. amp 820R AD797	−119.90	2.90	14.8	4.57
Dual AD797	Quad diff. amp 820R AD797	−120.60	0.70	15.5	4.12
	Notional input stage	−122.60		17.5	3.01
	Notional input stage	−128.50		23.4	0.99
	Notional input stage	−139.00		33.9	0.10

practical, and the finalized balanced input design was both economical and quieter than its unbalanced neighbor by a dependable 0.9 dB.

Two other versions were evaluated that made the balanced input quieter than the unbalanced one by 2.8 dB, and by 4.7 dB, at slightly greater cost and complexity. These were put away for possible future upgrades.

Unbalanced and Balanced Outputs

A balanced interconnection is very much a system – the balanced output, the interconnecting cable, and the balanced input and their interactions must all be considered to get the best performance. Therefore in the next section we will take a quick look at the types of unbalanced and balanced

outputs that may drive a power amplifier input. This is also relevant to amplifier design as such; separate balanced outputs, as opposed to just hard-wiring line input and output sockets together, are also useful if you want to daisy-chain amplifier inputs; if an amplifier downstream fails and is switched off it will not affect the signal integrity of the amplifiers upstream, because of the buffering inherent in the output stage driving it. If there is a failure upstream, however, you may be in trouble; if smoke is pouring from every orifice of an amplifier it is probably a good plan to switch it off, which will disable its line output stage.

Unbalanced Outputs

There are only two electrical output terminals for an unbalanced output – signal and ground. However, the unbalanced output stage in Figure 20.21a is fitted with a three-pin XLR connector, to emphasize that it is always possible to connect the cold wire in a balanced cable to the ground at the transmitting (output) end and still get the benefits of common-mode rejection. If a two-terminal connector is fitted, the link between the cold wire and ground has to be made inside the connector, as shown in Figure 20.3 at the start of this chapter. This is vitally important when an unbalanced output is being used with a balanced input. The output amplifier in Figure 20.21a is configured as a unity-gain buffer, though in some cases it will be connected as a series feedback amplifier to give gain. A nonpolarized DC-blocking capacitor C1 is included; 100 µF gives a −3 dB point with a 600 Ω load of 2.6 Hz, which is amply low. A drain resistor R1 ensures that no charge can be left on the output. The op-amp is isolated from the line shunt capacitance by a resistor R2, in the range 47–100 Ω, to ensure HF stability, and this unbalances the hot and cold line impedances.

If the output resistance R2 is taken as 100 Ω worst-case, and the cold line is simply grounded as in Figure 20.21a, then the presence of R2 degrades the CMRR of the interconnection to an uninspiring −40 dB even if the balanced input at the other end of the cable has infinite CMRR in itself and perfectly matched 10 k input impedances.

To fix this problem, Figure 20.21b shows what is called an impedance-balanced output. There are now three physical terminals: hot, cold, and ground. The cold terminal is neither an input nor an output, but a resistive termination R3 with the same resistance as the hot terminal output impedance R2. If an unbalanced input is being driven, the cold terminal on the transmitting (output) equipment

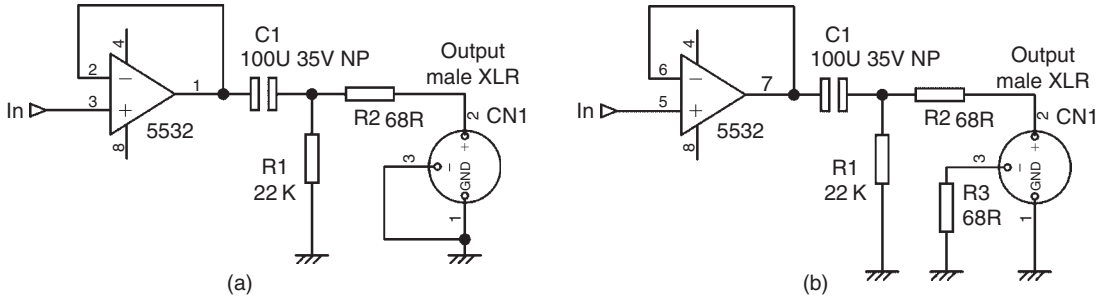


Figure 20.21: Unbalanced outputs: simple output (a) and impedance-balanced output (b) for improved CMRR when driving balanced inputs

can be either shorted to ground locally or left open-circuit. The use of the word 'balanced' is perhaps unfortunate, as when taken together with an XLR output connector it appears to imply a true balanced output with anti-phase outputs. The impedance-balanced approach is not particularly cost-effective. It requires significant extra money to be spent on a balanced-capable connector with three connections. Adding an op-amp inverter to that to make it a proper balanced output costs little more, especially if there happens to be a spare op-amp half available, and sounds much better in the advertising.

Ground-Canceling Outputs

This most ingenious technique, also called a ground-compensated output, surfaced in the early 1980s in mixing consoles (see Figure 20.22a). It allows ground voltages to be canceled out even if the receiving equipment has an unbalanced input; it prevents any possibility of creating a phase error by mis-wiring; and it costs virtually nothing except for the provision of a three-pin output connector.

Ground-canceling works by separating the wanted signal from the unwanted by addition at the output end of the link, rather than by subtraction at the input end. If the receiving equipment ground differs in voltage from the sending ground, then this difference is added to the output signal so that the signal reaching the receiving equipment has the same voltage superimposed upon it. Input and ground therefore move together and there is no net input signal, subject to the usual effects of component tolerances. This requires the connecting lead to be differently wired from the more common unbalanced-out balanced-in situation; now the cold line must be joined to ground at the *input* or receiving end.

The cold pin of the output socket is now an input, and has a unity-gain path summing into the main signal going to the hot output pin. This path R3, R4 usually has a very low input impedance equal to the hot terminal output impedance, so if it *is* used with a balanced input the line impedances will be balanced, and the combination will work well. The 6 dB of attenuation in the R3–R4 divider is undone by the gain of 2 set by R5, R6. It is unfamiliar to most people to have the cold pin of an

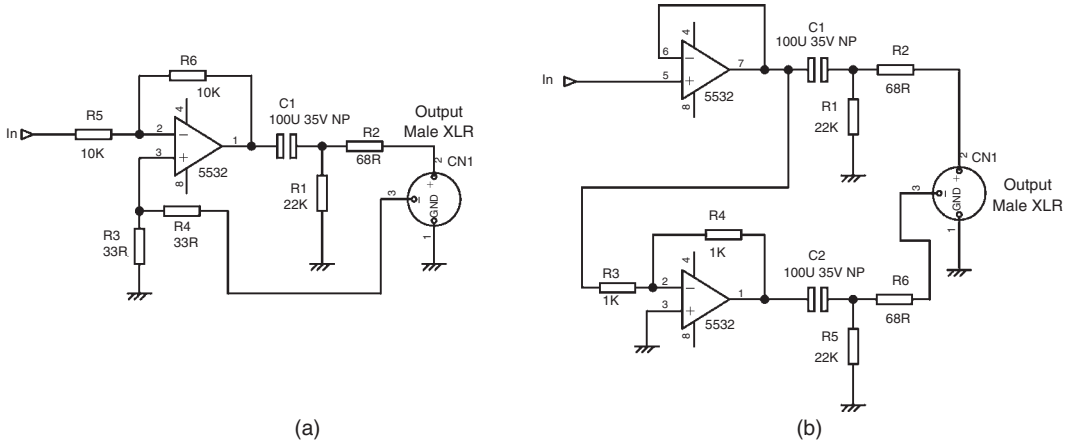


Figure 20.22: (a) A ground-canceling output. (b) A true balanced output

output socket as a low-impedance input, and its very low input impedance minimizes the problems caused by mis-wiring. Shorting it locally to ground merely converts the output to a standard unbalanced type. On the other hand, if the cold input is left unconnected then there should be only a very small degradation of noise due to the very low input resistance of R3. Note that this version of the ground-canceling output phase-inverts; this may or may not be convenient depending on the circuitry upstream. It is also possible to make non-inverting ground-canceling output stages.

Ground-canceling outputs are an economical way of making ground loops innocuous when there is no balanced input, and it is rather surprising they are not more popular. Perhaps people find the notion of an input pin on an output connector unsettling.

Balanced Outputs

Figure 20.22b shows a simple balanced output, where the cold terminal carries the same signal as the hot terminal but phase-inverted. This can be simply arranged by using an op-amp stage to invert the normal in-phase output. The resistors R3, R4 around the inverter should be as low in value as possible to minimize Johnson noise, but bear in mind that R3 is effectively grounded at one end and its loading, as well as the external load, must be driven by the first op-amp. A unity-gain follower is shown for the first amplifier, but this can instead be any other shunt or series feedback stage as required. The inverting output if not required can simply be ignored, but it must not be grounded, because the inverting op-amp will then spend most of its time in current-limiting, probably injecting unpleasant distortion into the pre-amp grounding system, and possibly suffering overheating. Both hot and cold outputs must have the same output impedances (R2, R6) to keep the line impedances balanced and the interconnection CMRR maximized.

A balanced output has the advantage that the total signal level on the line is increased by 6 dB, which if correctly handled can improve the signal-to-noise ratio. Another advantage is that it is less likely to crosstalk to other lines even if they are unbalanced, as the current injected via the stray capacitance from each crosstalking line will tend to cancel at the receiving end; how well this works depends on the physical layout of the conductors. All balanced outputs give the facility of correcting phase errors by deliberately swapping hot and cold outputs. This is, however, a two-edged sword, because it is probably how the phase became wrong in the first place.

There is no need to worry much about the exact symmetry of the output – that is, the equality in level of the two output signals – unless capacitive crosstalk really is a major concern. Even quite large gain errors only affect the signal-handling capacity of the interconnection by a small amount.

This form of balanced output is the norm in hi-fi balanced interconnection, but is less common in professional audio, where the quasi-floating output below gives more flexibility.

Quasi-Floating Outputs

This purely electronic output stage approximately simulates a floating transformer winding; if both hot and cold outputs are driving signal lines, then the outputs are balanced, as if a center-tapped output

transformer were being used, though clearly the output is not galvanically isolated from ground. If, however, the cold output is grounded, the hot output doubles in amplitude so the total level hot-to-cold is unchanged. This condition is detected by the current-sensing feedback taken from the outside of the 75Ω R10, and the current driven into the shorted cold output is automatically reduced to a low level that will not cause problems (see Figure 20.23a).

Similarly, if the hot output is grounded, the cold output doubles in amplitude and remains out of phase; the total hot-to-cold signal level is once more unchanged. This system has the advantage that it can give the same level into either a balanced or unbalanced input, given an appropriate connector at the input end; 6 dB of headroom is, however, lost when the output is used in unbalanced mode. It is most useful in recording studios, where various bits of equipment may be temporarily connected; it is not of value in a PA system with a fixed equipment line-up.

When an unbalanced output is being driven, the quasi-floating output can be wired to work as a ground-canceling connection, with rejection of ground noise no less effective than the true balanced mode. This requires the cold output to be grounded at the remote (input) end of the cable. Under adverse conditions this might cause HF instability, but in general the approach is sound. If you are using exceptionally long cable, then it is wise to check that all is well.

If the cold output is grounded locally, i.e. at the sending end of the cable, then it works as a simple unbalanced output, with no noise rejection. When a quasi-floating output stage is used unbalanced, the cold leg *must* be grounded, or common-mode noise will degrade the noise floor by at least 10 dB, and there may be other problems with high distortion.

Quasi-floating outputs use a rather subtle circuit with an intimate mixture of positive and negative feedback of current and voltage. This performs the required function quite well, but a serious drawback is that it accentuates the effect of resistor tolerances, and so a preset resistor is normally required to set the outputs for equal amplitude; the usual arrangement is shown in Figure 20.23a. If the balance preset is not correctly adjusted one side of the output will clip before the other and reduce the total output headroom.

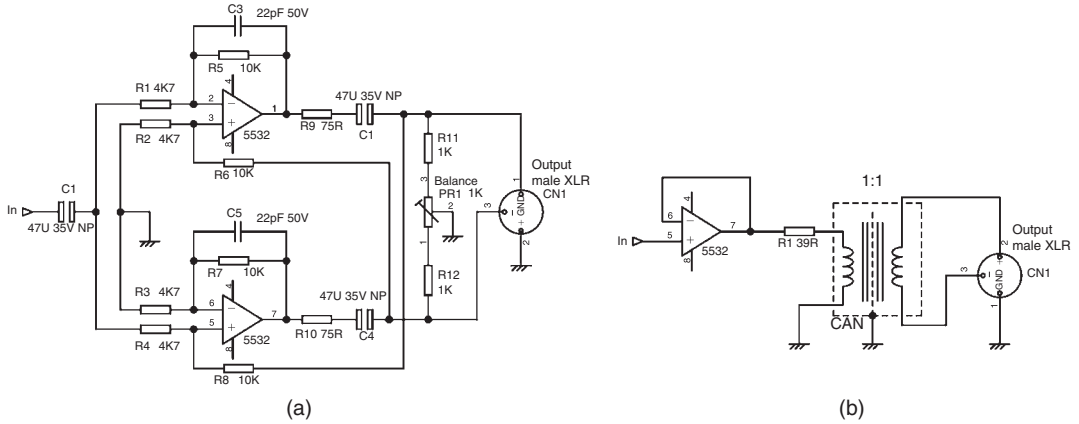


Figure 20.23: (a) Quasi-floating balanced output. (b) Transformer balanced output

Transformer Balanced Outputs

If true galvanic isolation between equipment grounds is required, this can only be achieved with a transformer. The technique is rarely used, because of the cost, weight, and performance problems of transformers, but is sometimes found in touring PA systems (usually only in the mic-splitter box on the stage) and broadcast environments where huge RF field strengths are encountered. A basic transformer balanced output is shown in Figure 20.23b.

Transformers have a well-known problem with linearity at low frequencies. This is because the voltage induced into the secondary winding depends on the rate of change of the magnetic field in the core, and so the lower the frequency, the greater the change in magnitude must be for transformer action. The current drawn by the primary winding to establish this field is nonlinear, because of the well-known nonlinearity of iron cores. If the primary had zero resistance and was fed from a zero source impedance, as much distorted current as was needed would be drawn and no one would ever know there was a problem. But there is always some primary resistance and the distorted current flowing through this introduces distortion into the magnetic field established, and so into the secondary output voltage. Sometimes there is also series resistance such as R1 deliberately inserted into the primary circuit, presumably with the intention of avoiding HF instability provoked by transformer winding and line capacitances; this definitely makes the LF distortion problem worse, and a better means of isolation is a low-value inductor of, say, 4 μH in parallel with a low-value resistor of around 39 Ω.

The LF distortion can be reduced by applying negative feedback via a tertiary transformer winding. Another cure is to cancel out the transformer primary resistance by an electronically generated negative resistance; comprehensive details on this approach can be found in Bruce Hofer's patent, which covers the transformer output of the Audio Precision System One[8].

DC flowing through the primary winding is also bad for linearity, and if necessary should be stopped by a DC-blocking capacitor.

Using a Balanced Power Amplifier Interface

The use of a balanced input to a power amplifier is a thoroughly good idea, canceling out ground voltages and rendering ground loops innocuous. This technology is useful even *inside* the power amplifier unit, where ground currents can otherwise put a limit on achievable hum performance; typically it may be necessary when passing the signal at line level from one PCB to another. The obvious approach is to put a differential amplifier in front of the power amplifier, but this rather smacks of overkill to deal with internal grounding problems, and adding an op-amp in such a position almost invariably degrades the overall noise performance. At this point the idea may strike – we have an amplifier here already, with a differential-pair input stage: can it be made into a differential power amplifier? The answer is yes, if you do it with a little care.

Now and again the circuit shown in Figure 20.24a has appeared in the literature. It suggests that all you have to do to make a differential power amplifier is to treat the bottom of the negative-feedback network as the cold input. There are several things wrong with this.

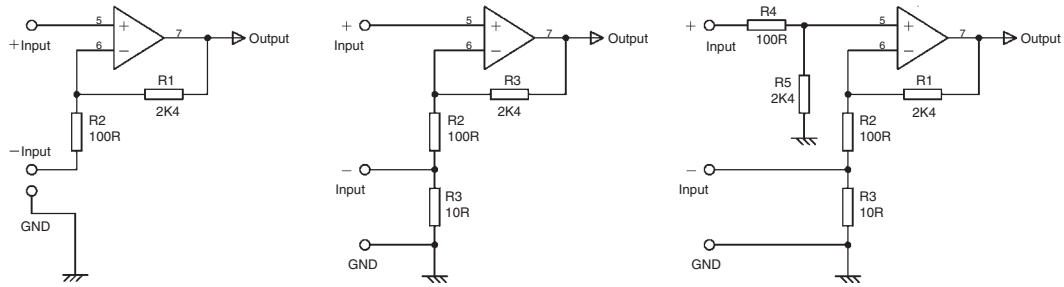


Figure 20.24: Balanced power amplifier interfaces: (a) is dangerous; (b) works very nicely; (c) improves the CMRR

Firstly, if the cold input becomes disconnected – say, by someone pulling out the input connector – the closed-loop gain of the amplifier is abruptly reduced to unity. This will almost certainly render it completely unstable, with massive high-frequency oscillation and an excellent chance of damaging both the amplifier and any attached loudspeakers. This problem can be prevented by adding a low-value resistor R3 at the bottom of the negative-feedback network as shown in Figure 20.24b. In this example the closed-loop gain is 25 times (+28.0 dB) with the cold input connected and 22.8 times if it goes open; a power amplifier with reasonable stability margins should not be worried by this. The gain change can be reduced by making R3 smaller, so long as it remains high in value compared with the ground resistance. For example, making R3 lower at 4R7 gives normal gains of 25 times and 23.9 times with cold disconnected, and I have found this value works well in practical use.

Secondly, it is not a proper differential circuit as the gains via the two inputs are not quite the same. The gain via the inverting input is R1/R2, while the gain via the non-inverting input is R1/R2 +1. The common-mode gain (by simulation) is +1.8 dB, so the CMRR is only 26.2 dB.

This can be easily corrected by using the circuit shown in Figure 20.24c, which adds a little attenuation to the non-inverting input in the form of R4 in conjunction with R5. R4 is very often already present as part of an RC input filter, and it is simple to scale the filter values so that R5/R4 = R1/R2. Likewise R5 is usually already there to define the DC level of the input stage. The input gains are now the same and good common-mode rejection is achieved.

The values shown reduce the normal gain very slightly to +27.6 dB and reduce the common-mode gain sharply to −12.4 dB, so the CMRR improves markedly to 40.0 dB. In a perfectly balanced circuit, and with an infinite open-loop amplifier gain, the CMRR would in theory be infinite, but here the CMRR achievable is determined by the realistic open-loop gain assumed for the amplifier when the simulations were run. Open loop gain was set to 35,000 with a single pole at 20 Hz. This gives a flat CMRR curve up to 10 kHz.

Finite open-loop gain can be compensated for by increasing R4 to 110 Ω, which improves the CMRR to 44.5 dB, but heading in this direction involves canceling two quite different effects and is not very respectable. If 40 dB is not enough then there is something seriously wrong with your grounding system and you need to seek out the root cause of the problem.

Finally, it is not a proper differential circuit in that the two inputs are not interchangeable. If you are trying to correct an absolute phase error by driving the cold input (which is not exactly good practice in itself), you will soon find it is not at all suitable for the task, having much too low an impedance for use as a signal input. This is not a problem if the cold input is used appropriately, to cancel the voltage drop on the ground connection. Since the resistance of any sensible ground is a small fraction of an ohm, a cold input impedance of 4.7 or 10 Ω will not cause significant errors.

While this approach could be used to interface to the outside world, it is much more useful for making internal amplifier connections between input circuitry and the power amplifier itself. In one instructive case there was a difference in ground potential between the two sections, which were on separate PCBs; both gave an exemplary hum-free performance alone, but when the signal connection between them was made, hum appeared. It was at a very low level and quite inaudible but it was uncomfortably visible on the amplified noise output. There was clearly a 50 Hz current flowing through the ground of the signal connection, for when the ground was reinforced with the traditional piece of 32/02 green-and-yellow wire, the hum level dropped significantly. It has to be said that the deep reason for this current remained obstinately obscure, despite intense investigation. The most obvious possible cause was that capacitive coupling between primary and secondary of the mains transformer was causing 50 Hz currents to flow – but the transformer was a high-quality item with a grounded interwinding screen to prevent just that happening, and disconnecting the screen rather disconcertingly made no difference to the problem.

The total noise level at the power amplifier output was −94.1 dBu over the 22–22 kHz bandwidth. When the PCB–PCB interconnection was changed to the balanced system, this dropped to −94.4 dB. That may sound like a pretty trivial improvement, but in fact it meant that the obtrusive 50 Hz component disappeared. To quantify this, a narrowband noise measurement, made with a 50 Hz bandpass filter, was reduced from −100.8 to −109.6 dBu. The coherent 50 Hz waveform disappeared completely, leaving just the random excitation of the bandpass filter by wideband noise. The balanced interface was working beautifully.

References

[1] N. Muncy, Noise susceptibility in analog and digital signal processing systems, JAES 3 (6) (June 1995) p. 447.

[2] T. Williams, EMC for Product Designers, Newnes (Butterworth-Heinemann), 1992, p. 176.

[3] T. Williams, EMC for Product Designers, Newnes (Butterworth-Heinemann), 1992, p. 173.

[4] S. Winder, What's the difference?, Electronics World (Nov 1995) p. 989.

[5] D. Birt, Electronically balanced analogue-line interfaces, Proceedings of the Institute of Acoustics Conference, Windermere, UK, Nov 1990.

[6] W. Jung (Ed.), Op Amp Applications Handbook, Newnes (Elsevier), 2006.

[7] P. Zwicky, Low-distortion audio amplifier circuit arrangement, US Patent No. 4,567,443, 1986.

[8] B. Hofer, Low-distortion transformer-coupled circuit, US Patent No. 4,614,914, 1986.

CHAPTER 21

Input Processing and Auxiliary Subsystems

This chapter deals with input processing functions such as gain control and filtering, and other power amplifier features that are not part of the signal path, but perform ancillary functions such as level indication or control of on/standby switching.

Ground-Lift Switches

If balanced inputs with high CMRR are used, they should deal effectively with the humming and buzzing ground currents that result from the existence of ground loops. However, in an emergency it may be useful to have a means of lifting the ground without creating a shock hazard. It has to be said that it is often not very effective, and is a last resort when the audience is starting the slow handclap.

The typical ground-lift switch breaks the direct connection between mains earth and the equipment signal earth, but leaves them connected by a resistor, usually in the range 10–100 Ω. This resistor is high enough to prevent significant ground-loop current passing, but low enough to prevent the signal earth from floating about. Without it the signal circuitry may have a 120 V AC voltage on it due to interwinding capacitance in the mains transformer, if it is not otherwise grounded. If the mains transformer has an interwinding screen then this must remain connected to mains earth and not the signal earth.

The resistor is usually shunted with a small capacitor (usually 10–100 nF), which makes the impedance low at radio frequencies but does not let a significant current at 50 Hz flow.

Phase Reversal Facility

A feature found on a small minority of professional amplifiers is a phase reverse switch that can be used to correct phase reversals elsewhere in the audio system. If only a balanced input is provided, phase reversal is easily implemented by a switch that swaps over the inputs. If there is also an unbalanced input, however, then things are a bit more difficult; it will be necessary to add a unity-gain inverting stage to create the anti-phase signal.

Gain Control

Gain controls on hi-fi power amplifiers are relatively rare as volume is almost always controlled from the preamplifier. They do, however, come in very handy when setting up a system, and this is

more true the more powerful the amplifier. Alternatively, a switch giving 20 or 30 dB of attenuation would be useful, but unless you are a hi-fi reviewer you are unlikely to be changing the system around often enough to justify the cost.

Professional amplifiers usually have gain controls; there are usually separate controls for each channel as such an amplifier is more likely to be handling two bands of a multi-amped loudspeaker system rather than two stereo channels. If stereo *is* being handled, ganged potentiometers are not very satisfactory as gain controls because all but the most expensive tend to have poor channel balance at large attenuations. For this reason a large number of professional amplifiers use switched attenuators based on rotary switches and resistor ladders.

At least one manufacturer (Harrison) has produced a range of professional amplifiers with VCA modules that allow remote control of output level by means of a DC voltage.

Subsonic Filtering: High-Pass

Subsonic filters are a common feature of professional amplifier input systems. Power amplifiers are not likely to be damaged by subsonic inputs, but the loudspeakers they drive are. It is essential to prevent large and uncontrolled excursions of the loudspeaker bass units. This is particularly important when using reflex (ported box) loudspeakers that have no restraint on cone movement at very low frequencies. In sound-reinforcement applications subsonic signals can be generated by microphones with insufficient blast protection, by direct-injected bass guitars, and in many other ways. Mixing desks almost always include effective low-frequency filters on microphone input channels, but these are usually configured for a fairly high roll-off such as 100 Hz to deal with microphone placement problems, so it is still good practice to incorporate true subsonic filters in the power amplifiers. In hi-fi applications subsonic filtering is usually done at the preamplifier end, but there is very often some last-ditch low-end bandwidth limitation in the power amplifier as well. In this case the most common source of subsonic signals are the warps on vinyl records.

While large subsonic signals are not likely to actually damage a power amplifier, they will consume valuable headroom by pushing the amplifier into clipping that would not happen if it was only reproducing the wanted signals in the audio band, and this is another reason for including an effective subsonic filter.

The high-pass filters used are typically of the second-order or third-order Butterworth (maximally flat) configuration, giving roll-off rates of 12 and 18 dB/octave respectively, the latter being preferable. Fourth-order 24 dB/octave filters are sometimes used but are less common, as people get worried (from all the evidence, unnecessarily) about the possible subjective effects of rapid phase changes at the very bottom of the audio spectrum. The Butterworth response (sometimes called the Butterworth alignment) is not the only one possible; the Bessel alignment gives a slower roll-off, but aims for linear phase, i.e. a constant delay versus frequency, and so reproduces the shape of transients better. There are other filter alignments such as the Chebyshev that give faster initial roll-offs than the Butterworth, but they do so at the expense of ripples in the pass-band or stop-band gain so they are not used in this sort of application.

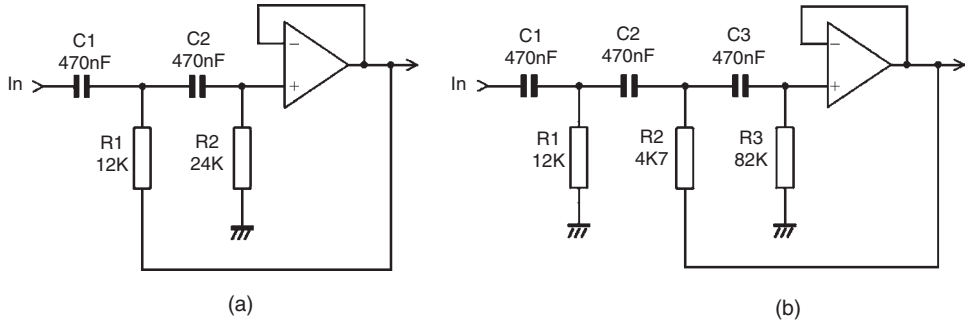


Figure 21.1: Subsonic filtering: second-order and third-order Butterworth high-pass filters, both −3 dB at 20 Hz

The most popular filter configuration is the well-known Sallen-and-Key type. The second-order version is very simple to design; the two series capacitors C1, C2 are made equal and R2 is made twice the value of R1. A second-order Butterworth filter with a −3 dB point at 20 Hz is shown in Figure 21.1a. The response is 12.3 dB down at 10 Hz and 24.0 dB down at 5 Hz, by which time the 12 dB/octave slope is well established. Above the −3 dB roll-off point the response is −0.78 dB down at 30 Hz, which is intruding a little into frequencies we want to keep. Other roll-off frequencies can be obtained simply by scaling the component values while keeping C1 equal to C2 and R2 twice R1.

Third-order filters are a bit more complicated, and many versions of them use two op-amps instead of the one required for a second-order filter. But it can be done with just one, as in Figure 21.1b, which is a third-order Butterworth filter also with a −3 dB point at 20 Hz. The resistor value ratios are now 2.53:1.00:17.55, and the circuit shown uses the nearest E24 values to this – which by happy chance come out as E12 values. These values are not quite mathematically correct, and there is a little gain peaking of 0.001 dB around 80 Hz, according to SPICE simulation, whereas the true Butterworth response has no peaking at all but stays flat until the roll-off. If you can't live with this, then you will have to go to E96 resistors; R1 = 12.1 k, R2 = 4.75 k and R3 = 84.5 k give a very accurate response with no peaking at all.

For the third-order filter the response is 18.6 dB down at 10 Hz and 36.0 dB down at 5 Hz, which is some pretty serious filtering, but the 30 Hz response is only −0.37 dB down, which demonstrates that a third-order filter is much more effective than a second-order one for this sort of job. As before, other roll-off frequencies can be had by scaling the component values while keeping the resistor ratios the same.

An important consideration with frequency-dependent networks like filters is the input impedance; this can sometimes drop to unexpectedly low values, putting excessive loading on the previous stage. In the high-pass case, the input impedance is high at low frequencies but falls as frequency increases. For the second-order version, it tends to the value of R2. R1 is bootstrapped and has no effect. In the third-order version, it tends to the value of R1 in parallel with R3, which here is 10.58 k. In neither case should the previous stage be embarrassed by the loading.

Because of the large capacitances, the noise generated by the passive elements in a high-pass filter of this sort is usually well below the op-amp noise. For the values used here, SPICE shows that the resistors in the second-order filter produce −132.4 dBu at the filter output, while in the third-order they produce −125.0 dBu at the output (22 kHz bandwidth, 25°C).

Ultrasonic Filtering: Low-Pass

Amplifier input systems are sometimes fitted with an ultrasonic filter (somehow a 'supersonic filter' sounds wrong) to define the upper limits of the working bandwidth, though they are rather less common than subsonic filters. This is *not* a duplication of the input RF filtering which, as described in Chapter 20, has to be designed very cautiously as the source resistance is unknown. If an active low-pass filter is used, driven from a low and known source impedance, the turnover frequency of the filter can be accurately defined, and therefore set much lower in frequency without the fear that it will ever encroach on the wanted audio bandwidth. One of the main uses of an ultrasonic filter is the protection of the power amplifier and loudspeakers against ultrasonic oscillation in the audio system.

The filters used are typically second-order with roll-off rates of 12 dB/octave; third-order 18 dB/octave filters are rather rarer at the high end of the audio spectrum, probably because there seems to be a general feeling that phase changes are more audible at the top end of the audio spectrum than the bottom. As with subsonic filters, either the Butterworth (maximally flat) or the Bessel type can be used. It is unlikely that there is any real audible difference between the two types of filter in this application, as most of the action occurs above 20 kHz, but using the Bessel alignment does require compromises in the effectiveness of the filtering because of its slow roll-off, as I will demonstrate.

Figure 21.2a shows a second-order Butterworth low-pass filter. This time the resistors are equal while the capacitors must have a ratio of 2. This creates problems as capacitors are available in a much more limited range of values than resistors – often in the E6 series, which runs 10, 15, 22, 33, 47, 68 – so filter values often have to be made up of two capacitors in parallel. (It is perfectly possible to make Sallen-and-Key filters where both the R and the C values are equal – you just have to replace the unity-gain buffer with an amplifier with a gain of 1.586 times[1]. However, this gain is often unwanted and inconvenient.)

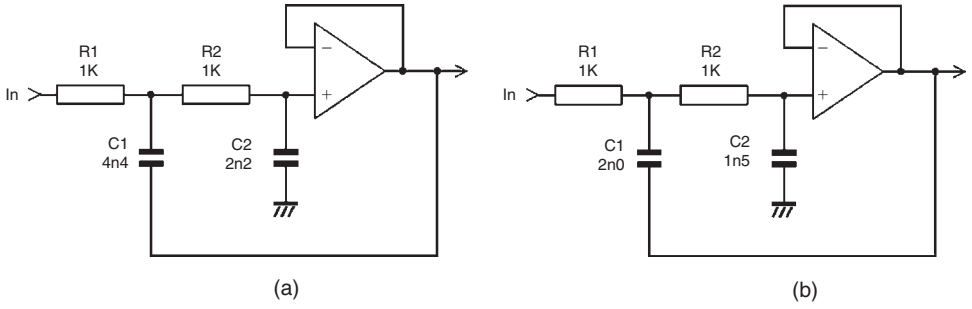


Figure 21.2: Ultrasonic filtering: Butterworth (a) and Bessel (b) second-order filters

The Butterworth filter in Figure 21.2a has a −3 dB point set at 50 kHz, and this gives almost exactly 0.0 dB at 20 kHz, so there is no intrusion into the audio band. The response is −12.6 dB at 100 kHz and a useful −24.9 dB at 200 kHz. C1 is made up of two 2n2 capacitors in parallel.

But supposing we are worried about linear phase and we want to use a Bessel filter. The only change is that C1 is 1.335 times as big as C2 instead of twice, but the response is very different. If we design for −3 dB at 50 kHz again, we find that the response is −0.39 dB at 20 kHz, which is not exactly a stunning figure to put in your spec sheet. If we decide we can live with about −0.1 dB at 20 kHz, then the Bessel filter has to be designed to be −3 dB at 72 kHz, and this is the Bessel filter shown in Figure 21.2b, with C1 made up of two 1 nF capacitors in parallel. Due to this change, and the inherently slower roll-off, the response is only down to −5.8 dB at 100 kHz and −15.8 dB at 200 kHz; the latter figure is almost 10 dB worse than for the Butterworth filter. Think hard before you decide to go for the Bessel option.

Once again we need to consider the way the filter input impedance loads the previous stage. In this case, the input impedance is high in the pass-band, but above the roll-off point it falls until it reaches the value of R1, which here is 1 kΩ. This is because at high frequencies C1 is not bootstrapped, and the input goes through R1 and C1 to the low-impedance op-amp output. Fortunately this low impedance only occurs at high frequencies, where one hopes the level of the signals to be filtered out will be low.

Another important consideration with low-pass filters is the balance between the R and C values in terms of noise performance. R1 and R2 are in series with the input and their Johnson noise will be added directly to the signal. Here the two 1 kΩ resistors generate −119.2 dBu of noise (22 kHz bandwidth, 25°C) while SPICE simulation of the complete filter gives −118.9 dBu, which is pretty close; this ignores the op-amp noise, which must be included to give the overall noise performance. The obvious conclusion is that R1 and R2 should be made as low in value as possible without causing excess loading (and 1 kΩ is not a bad compromise) with C1, C2 proportioned to maintain the same roll-off frequency.

Combined Filters

The subsonic and ultrasonic filters can be combined into one stage for convenient bandwidth definition. This is feasible only because the two turnover frequencies are widely separated. Figure 21.3 shows a second-order Butterworth high-pass filter combined with a second-order Butterworth low-pass filter; the response of the two filters is exactly the same as described above for each separately, with the slight proviso that the mid-band gain is now −0.088 dB rather than precisely unity. Hopefully your overall system design can cope with this. The loss occurs because the series combination of C3 and C4, together with C2, forms a capacitive potential divider with this loss figure, and this is one reason why the turnover frequencies need to be widely separated for the combining of the filters to work. If C3, C4 were smaller and C2 bigger the loss would be greater.

A third-order high-pass filter can be used instead of the second-order version, in exactly the same way. The only difference is that the mid-band loss is now −0.15 dB because there are three

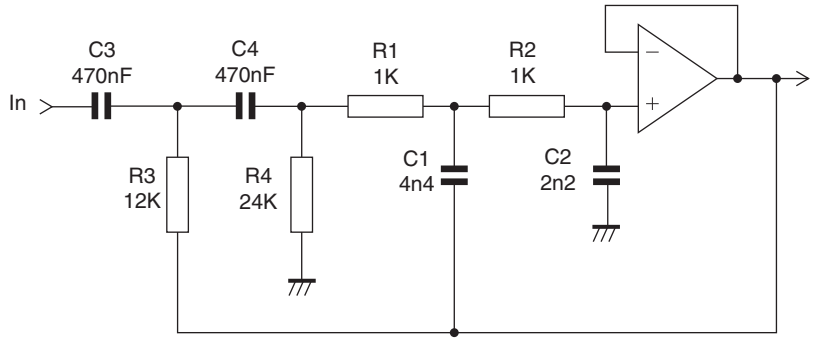


Figure 21.3: Subsonic and ultrasonic second-order Butterworth filters combined

capacitors in the high-pass filter rather than two. Combined filters have the advantage that the signal now has to only pass through one op-amp rather than two, and it can also be very useful if you only have one op-amp section available, and you would otherwise need another dual package, of which one half would just sit there twiddling its inputs and consuming quiescent current.

This is not the place to go any deeper into the vast subject of active filters; two good in-depth references are Van Valkenburg[2] and Williams and Taylor[3].

Electronic Crossovers

A few manufacturers (BGW is one example) have produced amplifiers with internal electronic crossovers to split the incoming signal into three or more bands, which are applied to separate power amplifiers and separate kinds of loudspeaker that specialize in a given frequency range. The subject of electronic crossover design is a large and complex area, and there is no room to go into it here.

Digital Signal Processing

A minority of professional power amplifiers include digital signal processing (DSP) facilities. The types of processing offered include filtering, equalization, delay, level control by limiting, and the implementation of electronic crossover systems. The delay option is particularly useful in compensating for sound delays resulting from speaker placement.

An LCD screen is usually provided to allow changes in configuration and parameter settings. A number of storable user-defined DSP presets are usually available so that the amplifier can quickly be reconfigured for different uses. DSP facilities are somewhat outside the scope of this book and will not be considered further here.

Signal-Present Indication

Professional amplifiers are often fitted with a ‘signal-present’ indicator that gives reassurance that an amplifier – possibly in a bank of 20 others – is receiving a signal and doing something with it. The level at which it triggers must be well above any likely noise level, but also well below the

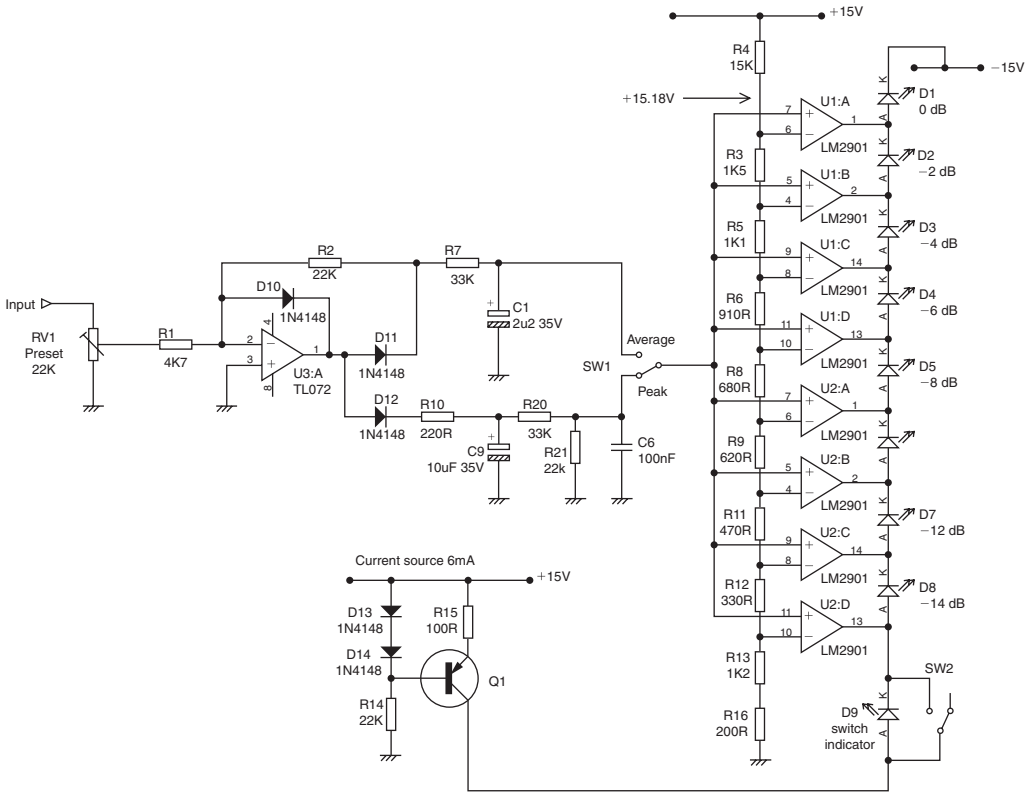


Figure 21.4: LED bar-graph meter with selectable peak/average response

maximum amplifier output. They are usually provided for each channel of a multichannel amplifier, and are commonly set up to illuminate when the channel output level exceeds 2 V rms, which is equivalent to 0.5 W into an 8 Ω load, or 1 W into a 4 Ω load. A trigger level of 4 V rms is also used.

A vital design consideration is that the operation of the circuit should not introduce distortion into the signal being monitored; this could easily occur by electrostatic coupling or imperfect grounding if there is a comparator switching on and off at signal frequency. This is not a problem with clipping detectors as when clipping occurs the signal is already distorted. A typical circuit would comprise just the bottom step of the LED bar-graph meter shown in Figure 21.4.

Output Level Indication

Many power amplifiers include some indication of the output level. This may be an LED bar-graph, analog VU-type meters, (though these are used mostly as a fashion statement on hi-fi amplifiers, being too fragile for on-the-road use) or simply a clipping indicator light on its own.

VU meters consist of a relatively low-resistance meter winding driven by rectifier diodes, sometimes with a series resistor. It is important to remember that this represents a horribly nonlinear load to an external circuit, and it must never be connected across a signal path unless it

has near-zero impedance. In practice a buffer stage is always used between a signal path and the meter to give complete isolation.

Bar-graph meters are commonly made up of an array of LEDs. Vacuum fluorescent displays are sometimes used but require hefty tooling charges if you want a custom display, and their high-voltage operation makes driving them more complicated.

An LED bar-graph meter can be made effectively with an active-rectifier circuit and a resistive divider chain that sets up the trip voltage of an array of comparators; this allows complete freedom in setting the trip level for each LED. A typical circuit which indicates from 0 to −14 dB in 2 dB steps with a selectable peak/average characteristic is shown in Figure 21.4 and illustrates some important points in bar-graph design.

U3 is a half-wave precision rectifier of a familiar type, where negative feedback servos out the forward drop of D11, and D10 prevents op-amp clipping when D11 is reverse-biased. The rectified signal appears at the cathode of D11, and is smoothed by R7 and C1 to give an average, sort-of-VU response. D12 gives a separate rectified output and drives the peak-storage network R10, C9, which has a fast attack and a slow decay through R21. Either average or peak outputs are selected by SW1, and applied to the non-inverting inputs of an array of comparators. The LM2901 quad voltage comparator is very handy in this application; it has low input offsets and the essential open-collector outputs.

The inverting comparator input are connected to a resistor divider chain that sets the trip level for each LED. With no signal input, the comparator outputs are all low and their open-collector outputs shunt the LED chain current from Q1 to −15 V, so all LEDs are out. As the input signal rises in level, the first comparator U2:D switches its output off and LED D8 illuminates. With more signal, U2:C also switches off and D7 comes on, and so on, until U1:A switches off and D1 illuminates. The important points about the LED chain are that the highest level LED is at the bottom of the chain, as it comes on last, and that the LED current flows from one supply rail down to the other, and is not passed into a ground. This prevents noise from getting into the audio path. The LED chain is driven with a constant-current source to keep LED brightness constant, despite varying numbers of them being in circuit; this uses much less current than giving each LED its own resistor to supply rail, and is universally used in mixing-console metering. Make sure you have enough voltage headroom in the LED chain, not forgetting that yellow-and-green LEDs have a larger forward drop than red ones. The circuit shown has plenty of spare voltage for its LED chain, and so it is possible to put other indicator LEDs in the same constant-current path; for example, D9 can be switched on and off completely independently of the bar-graph LEDs, and can be used to indicate the status of a ground-lift switch or whatever. An important point is that in use the voltage at the top of the LED chain is continually changing in 2 V steps, and this part of the circuit should be kept well away from the audio path to prevent horrible crunching noises from crosstalking into it.

This meter can of course be modified to have a different number of steps, and there is no need for the steps to be the same size. It is as accurate in its indications as the use of E24 values in the resistor divider chain allows.

If a lot of LED steps are required, there are some handy ICs which contain multiple open-collector comparators connected to an in-built divider chain. The National LM3914 has 10 comparators and a divider chain with equal steps, so they can be daisy-chained to make big displays, but some law bending is required if you want a logarithmic output. The National LM3915 also has 10 comparators, but with a logarithmic divider chain covering a 30 dB range in 3 dB steps.

Signal Activation

With increasing attention being paid to economy in the use of energy, it is now quite common for power amplifiers to have a standby mode where the main transformer is disconnected from the supply when the unit is not in use, but a small standby transformer remains energized to run a housekeeping microcontroller. This is particularly pertinent if large Class-A amplifiers are in use. It is therefore very convenient to have the amplifier wake up automatically when a signal is applied. It is now only necessary to pop in a CD or whatever, and start it playing; there is no need to push a button on the power amplifier. This is especially useful when large amplifiers are in use that are hidden away out of sight, or when monobloc power amplifiers are used, in which case two buttons in different locations have to be pushed.

Signal activation operates by detecting when a low-level signal appears at any of the amplifier inputs, low level in this case meaning a long way, such as −60 dB, below that which gives maximum output but hopefully well clear of the noise floor to avoid false triggering. The idea is that triggering on a low enough level will mean that you don't miss much of the start of the music, but you are always going to miss a fraction of a second. If the amplifier has a wake-up time of several seconds, which is quite common, because of the need for an inrush suppression delay, followed by a period in which muting is maintained while the internal voltages settle, you are going to miss rather more and you will probably find yourself restarting the CD. Signal activation is not without its limitations. In the case of multichannel amplifiers, seven or more inputs must all be monitored for the onset of a valid audio signal.

The principle of signal activation is very simple, and it appears straightforward to implement – you just have to sum all the amplifier inputs together (you cannot use the amplifier outputs, of course, because the amplifiers are not active yet), put them through a high-gain amplifier, and apply the result to a level detector, which in turn signals a microcontroller or otherwise wakes up the system. However, it may not come as a total surprise that in the real world things are a little more complicated than that. The challenges and their solutions are described with reference to the signal-activation system shown in Figure 21.5, which is loosely based on one of my commercial designs, is considered in some detail below to bring out the various important points. The design interfaces to a microcontroller, but the same design principles can be used to interface with discrete logic.

Firstly, if you have a high-gain amplifier connected to the amplifier inputs, it is going to be clipping hard all the time when normal signal levels are applied. This is not a good thing, as it takes sharp-edged gulps of current from the op-amp supply rails and dumps them into ground; even with good grounding practice it is possible for these ugly currents to contaminate a signal that you are trying to reproduce

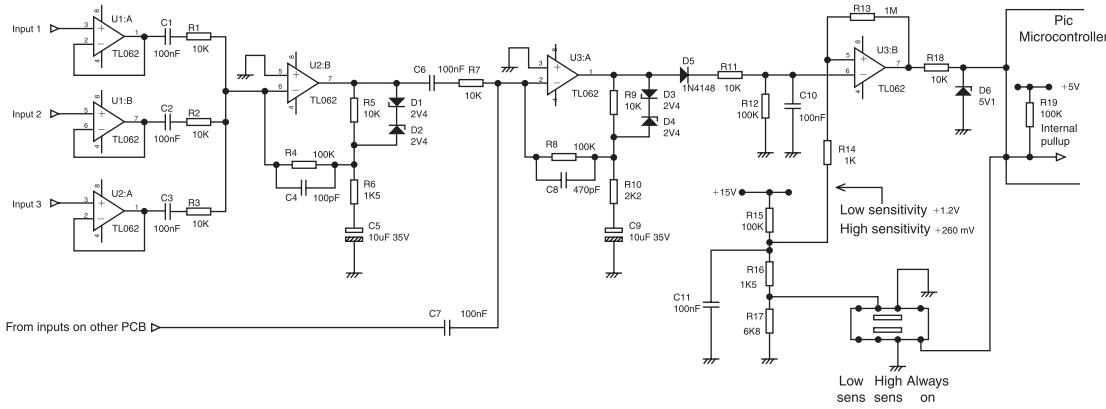


Figure 21.5: Circuit diagram of a signal-activation system for a multichannel amplifier

with less than 0.001% THD. In addition, the high-gain amplifier output will consist of a series of sharp edges that may get into the audio path by simple capacitive crosstalk. It is wise to put the activation circuitry as far away physically from susceptible audio paths, but in modern equipment that is not usually very far. Screening plates would obviously help with this but extra bits of metalwork cost money and are to be regarded as a last resort.

The answer to the problem is to prevent the high-gain amplifier from clipping by clamping its output, applying increased negative feedback when the limits of the desired output excursion are reached. In this way the output of the stage always remains under negative-feedback control. There are many ways to do this, but the simplest is adding a couple of Zener diodes to the feedback path. This approach greatly reduces the potential problems, but there is of course still a distorted signal on the high-gain amplifier output, and this must still be kept away from sensitive audio circuitry.

Secondly, it is important to avoid false triggering. If an amplifier switches off the supply to the main transformer with relays, these will usually be quite hefty and will make audible clicking noises when switching from standby to the active state and vice versa. If mains noise, etc. is continuously initiating these cycles of clicking, the paying customer is soon going to get irritated. This is bad enough, but there is also the point that most inrush-protection circuitry will not take kindly to continuous on/off cycling, and ultimately the inrush resistor might burn out – not good.

Ideally, the activation system should not trigger even when all the inputs of the amplifier are left open-circuit. This is a rather severe requirement, and not everyone would agree it was necessary, but it can be met by informed design and thorough testing of prototypes. The design shown here includes a switchable high/low sensitivity control so that signal activation is still usable in non-optimal conditions, and also an ‘always on’ switch position to cater for those who feel that amplifiers have to be powered constantly if they are to give of their best.

The main defense against false triggering is the restriction of the bandwidth the activation system responds to. Curtailing the frequency response at the low end prevents hum from giving a false trigger; likewise, limiting the high-frequency response discriminates against noise and transients. In the case of an open-circuit input, the typical hum waveform from electrical pickup is a severely

distorted travesty of a 50 Hz sine wave, with strong harmonics going up to at least 500 Hz, and so the low-frequency response must be curtailed quite dramatically.

When the amplifier is activated, it is important that it remains on for some time after the signal inputs cease – say 30 seconds or longer. This will prevent the amplifier going to standby between tracks of music.

Thirdly, since all the inputs meet at the signal-activation system, it is important that this is not a source of crosstalk between them.

The signal activation system of Figure 21.5 was designed for a seven-channel home theater amplifier, which has happily proved to have a long life in the marketplace. The electronics were divided into two big PCBs, one carrying three channels of power amplifier and the other four. Here is the description.

The seven inputs are of the usual unbalanced type with phono (RCA) connectors. The activation circuitry is on the PCB carrying the three power amplifiers, and each input is connected to a unity-gain buffer (U1:A, U1:B, and U2:A), which eliminates the possibility of crosstalk between them. The three buffered outputs are DC-blocked by C1, C2, and C3 and then fed through R1, R2, and R3 into the virtual-earth summing point of amplifier U2:B. The small capacitor value of 100 nF together with a 10 k input resistance gives an LF roll-off of −3 dB at 160 Hz to discriminate against hum. It might be objected that the buffers are superfluous as the virtual earth will stop signals coming in one input and then sidling out of the other, but in real life a virtual earth is not a perfect earth, and carries enough residual signal to compromise the crosstalk performance. The other point is that R1, R2, and R3 are low enough in value to excessively load the amplifier inputs. They have to be relatively low to give high gain with a feedback network that uses practicable resistor values.

It should be explained here that one of the reasons why the high-gain amplifier is divided into two stages is that the large amount of closed-loop gain required (about +70 dB) would mean a shortage of open-loop gain at high audio frequencies if only one op-amp was used.

The first amplifier stage U2:B has a mid-band gain of +38 dB, and uses shunt feedback in the usual way to generate a virtual earth at Pin 6. The feedback network R4, R5, R6 is in the form of a T-network C4 across R4 providing an HF roll-off of −3 dB at 16 kHz to discriminate against HF noise. C5 reduces the gain to unity at DC, and gives an LF roll-off at 11 Hz. D1 and D2 are 2V4 Zener diodes that provide output clamping by increasing the negative feedback when the output exceeds about 3 V peak in either direction.

The second amplifier stage U3:A is a similar shunt-feedback stage with a virtual earth at Pin 2. It is fed from the first stage just described via capacitor C6 and input resistor R7. Once again the capacitor value of 100 nF together with R7 gives an LF roll-off of −3 dB at 160 Hz. The other PCB, which carries the four power amplifiers, also has four more identical unity-gain buffers feeding another first amplifier stage. On this PCB there also reside DC-blocking and input components equivalent in function to C6 and R7, and there is an important point relating to why they are there

and not on the same PCB as the second amplifier stage. At first sight it is risky to send a signal from one PCB to another in current mode (i.e. at virtual earth) because such signals are vulnerable to capacitive crosstalk unless they are screened. For cost reasons, no screening was used here; inter-PCB signals were carried by a ribbon cable and adding a screened wiring assembly to this was not a tempting proposition. However, the first amplifier stages raise the signal level sufficiently so that there is no possibility of crosstalk from other signals causing false triggering of the activation system.

The real benefit of this philosophy is that the signal, being in current mode and at a negligible voltage, cannot itself crosstalk to parts of the main audio paths. This approach works very well.

The second gain stage U3:A has a mid-band gain of +35 dB, using shunt feedback through T-network R8, R9, R10 to generate the virtual earth at Pin 2. C8 across R8 provides an HF roll-off of −3 dB at 3.4 kHz, which discriminates heavily against HF noise. C9 reduces the stage gain to unity at DC, rolling off at 7.2 Hz, and D3 and D4 are once more 2V4 Zener diodes that clamp the output to about 3 V peak.

The output of the second gain stage drives a simple peak detector made up of D5 and C10. This only gives half-wave rectification but seems to be perfectly satisfactory in practice and the extra expense of full-wave rectification appears to be pointless. R11 gives a slow attack time to further discriminate against isolated noise pulses, and R12 defines the decay time. The stored voltage on C10 is applied to comparator U3:B, which has a threshold voltage generated by the divider R15, R16, R17 and switched by SW1 to give the high- and low-sensitivity settings. For the low-sensitivity mode, one section of SW1 removes the short across R17 and increases the threshold voltage from +260 mV to +1.2 V, reducing sensitivity by 13 dB.

R13 and R14 provide a small amount of positive feedback to introduce a little hysteresis and give clean comparator switching. The comparator output is clamped to +5 V by R18 and Zener D6, and applied to the input port of a microcontroller – in this case one of the PIC family. Another input port is used in conjunction with the remaining half of SW1 to sense when sensitivity switch SW1 is in the 'always on' position; the internal pull-up facility of the PIC is used to simplify this bit of the circuit. Note that TL062 op-amps, rather than the more familiar TL072, are used because of their lower input offset voltage. This is particularly important in the comparator stage where the voltages are low.

The long turn-off delay before the unit returns itself to standby is implemented in software in the microcontroller, as it is inconveniently long to be done in hardware. It is also possible to incorporate further discrimination against false triggering in the software, for example by disregarding single input pulses that are not followed by further input signals within a specified time interval.

The circuitry described above fully met the demanding requirement that it should not false-trigger when all the amplifier inputs were left open-circuit. In other applications this immunity will depend on many details of the design, such as the input impedances, the type of input connectors used, their proximity to mains wiring, and so on.

Twelve-Volt Trigger Activation

Another method of activating a power amplifier from standby is the use of a 12V trigger. Typically a preamplifier (which will be at hand for access to the controls) is connected to a remote power amplifier by a cable with 3.5 mm jack plugs at each end. When the preamplifier is turned on, it sends out a +12V DC signal that tells the power amplifier to come out of standby and become active. A vital point is that the connection must be opto-isolated to prevent the formation of a ground loop; this is done at the input (power amplifier) end. A typical +12V trigger input and output system is shown in Figure 21.6.

It is important to remember that a 12V trigger line might be connected to almost anything, as it is not an audio in/out and there are a lot of people out there who are pretty vague about how it works. Having once seen a cable with a 13A mains plug on one end and a phono connector on the other, I believe anything is possible. (I also saw the result of plugging this lead into an expensive Revox reel-to-reel recorder; the owner, who was lucky to be alive, did not seem to like the sound of 'Beyond economic repair, guv.')

The 12V trigger output must therefore be protected against short-circuits and against being connected to reverse voltages. In Figure 21.6 the PIC microcontroller switches on Q3, which switches on Q2, which in turn applies +15V to the 100 mA 78L12 regulator U2. This regulator not only provides a regulated +12V output, but is also current-limiting. D3 protects against intrusive reverse polarities.

The trigger input must be protected against excessive input voltages and reverse polarity; it must also be designed to reject quite high levels of electrical noise, at both low and high frequencies.

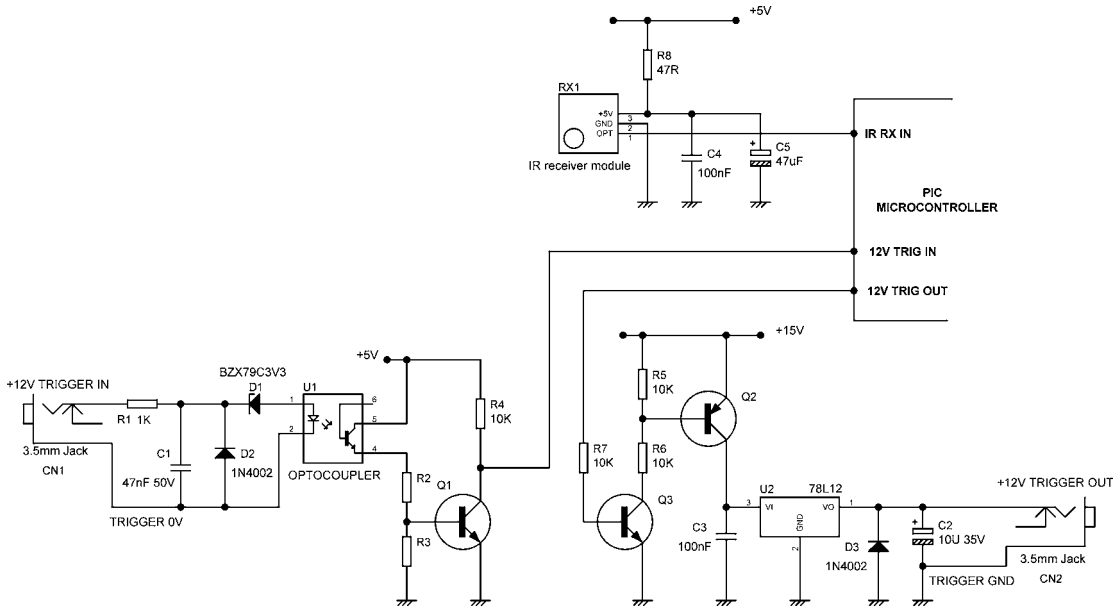


Figure 21.6: Typical 12V trigger in/out system and IR receiver facility for a power amplifier

R1 limits the current drawn from the transmitting unit, C1 filters out noise, and D2 protects against reverse polarity. The 3V3 Zener D1 ensures that incoming voltages have to exceed a threshold of about 5 V before the opto-isolator is activated. The opto output turns on Q1, which sends a low to the PIC when an incoming trigger occurs; the values of R2 and R3 depend on the opto characteristics.

Infrared Remote Control

An infrared remote control facility is now very common on preamplifiers for source selection, volume and balance control, muting, and on/standby control. Very often there are other control functions as well. The application of IR control to power amplifiers is rather rarer, but it is sometimes used for on/standby switching. Commands are transmitted using the Philips/Sony RC5 code modulated onto a carrier, typically at 37 kHz; this in turn is modulated on the IR emitted by the hand controller. The receiving circuitry required is very simple to arrange, most of the complexity being contained in a small transistor-sized component such as the Toshiba TSOP348XX series. The IR sensor, amplifier, AGC loop, band-pass filter, and demodulator circuitry are all integrated; carrier frequencies between 30 and 56 kHz are available. The only real precaution required with these devices is to make sure you have effective supply-rail decoupling close to the module. This is carried out by R8 and C4, C5 in Figure 21.6. After demodulation from the carrier the RC5 decoding is carried out in software by the microcontroller.

Other Amplifier Facilities

There are several other facilities that may appear on a professional amplifier, but are much less likely to be found in the hi-fi world:

- *Temperature indication*. Some amplifiers go beyond a simple 'over-temperature' indicator, and have a bar-graph display that reads the heat-sink temperature. This can be useful as a rise in temperature due to obstructed ventilation or whatever can be detected before it puts the amplifier into shutdown. It does of course assume that someone has the time to keep an eye on a dozen or more temperature displays.
- *Fan running indicator*. An LED that illuminates when a thermostatic fan-control system turns the fan on. This gives confidence that the cooling system is working, and can also give advanced warning of imminent overheating before shutdown becomes necessary.
- *Fuse indicators*. A few amplifiers are fitted with LEDs that indicate when internal fuses have blown.

References

[1] H.M. Berlin, Design of Active Filters with Experiments, Blacksburg, 1978, p. 85.

[2] M. Van Valkenburg, Analog Filter Design, Holt-Saunders International Editions, 1982.

[3] A. Williams, F. Taylor, Electronic Filter Design Handbook, fourth ed., McGraw-Hill, 2006.

Index

D

E

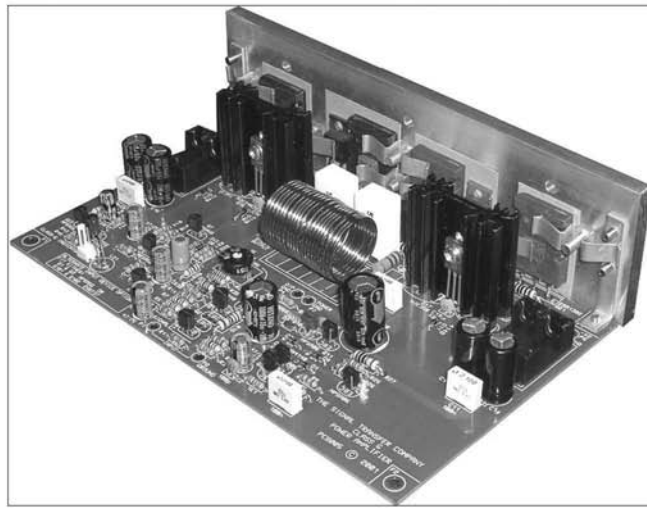
F

G

This page intentionally left blank

The Signal Transfer Company

The Signal Transfer Company is the only source for PCBs guaranteed to comply with the preamp and power amplifier design philosophies pioneered by Douglas Self



Shown above is the Class-G power amplifier, combining improved efficiency with first-class performance. The design is described in detail by Doug Self in Chapter 10.

The following PCBs are available:

- **The Precision Preamplifier** – active gain-control and variable-frequency tone controls
- **The Load-Invariant Power Amplifier** – very low distortion into heavy loads
- **The Trimodal Power amplifier** – ultra-low distortion Class A, switchable to Class B
- **The Class-G power amplifier** – with driver circuit for Class-G indicator LED

Our PCBs have been designed with meticulous care at every point. The power amplifier board layouts are precisely the same as those approved by Douglas Self when his famous series of articles on power amplifier distortion appeared in Electronics World. You can therefore be confident that proper operation is built-in.

We supply the finest quality fibreglass PCBs, single or double-sided as appropriate. All boards have a full solder mask, tinned pads, and a silk-screen component layout. Each PCB is supplied with extensive constructional notes, previously unpublished information about the design, and a detailed parts list to make ordering components simple.

Kits of parts to build the above PCBs are also available. These contain all PCB-mounted parts, including machined heatsink coupling plates for the power amplifiers, as shown in the illustration above.

For prices and more information go to **http://www.signaltransfer.freeuk.com/** or contact:

The Signal Transfer Company
35 Hirst Grove
Dodd Naze
Hebden Bridge
West Yorkshire
HX7 8DN
United Kingdom

Tel: 01422 885196