Radio Hobbyist's Designbook

by
Leonard H. Anderson
K6LHA

A Dedication
To
Hugo John Anderson (1900 - 1975)
and
Stuart Kemsley Golding (1900 - 1977)

My father and father-in-law,
respectively, both born the year before
Guglielmo Marconi first sent the letter S across the Atlantic
and three years before the Wright Brothers flew the first
heavier-than-air aircraft at Kittihawk, NC...yet both
lived to see the first humans set foot on the Moon, live,
after a quarter million mile rocket flight,
from the comfort of their homes, via television.

What a marvelous century was the 1900s!

Radio: Radio is a subset of the much larger technology field called *electronics*. Radio communications seems to be where much of electronics began as publicly demonstrated in 1896. Radio works by the **same** laws of physics as everything else in electronics.

Hobbyist: A person engaging in an avocational pursuit for personal enjoyment and interest. For **fun.** Great personal pleasure can be obtained by creating something from a concept in the mind, then on paper, finally in hardware...and making it work. It is a very satisfactory feeling to ***design*** something, be it a whole radio, an equipment for testing, modifying some older radio, or just learning about how electronics and radio work. The author has been a hobbyist in electronics since before he became a professional in the industry...and continues the avocation into retirement.

Design: *To contrive, to plan to do, intend to do* according to many dictionaries. As a part of that is *creativity*, the ability to make something different, unique, and to have it useful and performing an intended function. This book cannot teach ***creativity*** but it can help the designer by providing some information on the natural laws, circuits, systems, a myriad of things that can all be used in making a radio work and perform.

Designbook: An invented name. This is not a college-level textbook for assigned reading in a university course. It doesn't cover first principles or engage in multitudes of mathematical explanations that are more arcane than informative. It does include mathematics of the high school level algebra and trigonometry, enough needed for the designer to calculate parts values and analyze results...along with numerical examples to help the reader understand each process. This isn't a *handbook* either, rather a more advanced book...without getting too deep.

Radio Hobbyist's Designbook is there to explain essential basics, to describe circuits and systems, to show a bit of history, describe some available components. Essential math is given, but not much further. A formal academic degree is not needed...but it may take some work to learn how to apply components. As hobbyists we can experiment, try out, make different versions of those wonders, explore the technology. It is truly fascinating with much more to come.

This is NOT a simple work. It assumes the reader knows SOMETHING about electricity and can handle simple workshop tools. It handles subjects based on that. Chapters are arranged in an order thought best for more advanced understanding. Readers can skip over familiar Chapters as desired. Arrangement is hoped to induce personal thinking about Chapter subjects.

How RHdb Was Slowly Born...

The author got interested in *radio* as a middle-school teenager in 1947 while flying model airplanes as a hobby. *Radio control* of models was desired to enhance that flying. About the only source of information on either radio or electronics back then were monthlies like *Popular Science* and *Radio Craft* from newsstands. Understandable textbooks were almost nil; available textbooks were rooted in pre-WWII radio technology. The first *transistor* had been born in a laboratory but had not reached the electronics market. This all changed for the author beginning military service in 1952 in the U. S. Army. The first posting was to Army station ADA in Tokyo, Japan, in January

1953. ADA served the (then) entire Far East Command, operated 24/7 on HF with 40-plus transmitters from 1 to 40 KW output with a teleprinter message relay throughput of 220,000 per month average in 1955. Multi-channel microwave radio relay was installed there in 1954 and the author became an operations and maintenance supervisor on microwave relay in his last duty assignment. There was a wealth of information to be gleaned in Signal Corps manuals and those were studied intently to learn more. Information obviously existed but little was written up.

Moving to southern California and into aerospace electronics 1956, there was little to be found in newsstand literature other than broadcast radio, high-fidelity sound, or TV. *Detailed* information on how to calculate simple amplifier stage circuits had to be gleaned from various sources; hardly any text had any complete how-to-do-it procedures.

The opening of the Internet to the public in 1991 and the availability of electronic publishing tools to individuals has changed information access enormously. One of the major drawbacks is some website descriptions components that have become obsolete. It may be interesting to dissect some component of *unobtainium*[1] but it is of little practical use.

A godsend to hobbyists is the ease of getting *manufacturer's specifications* and *datasheets* free over the Internet. In nearly a half century of semiconductor industry growth and thousands of new devices introduced, perhaps half or more are now obsolete. They are all made of *unobtainium*.

Radio as a communications medium was first demonstrated in 1896 by Guglielmo Marconi in Italy and Aleksandr Popov in Russia. The first triode vacuum tube wasn't invented until 1906 and it took perhaps a decade more to make *tubes* practical. The age of semiconductors couldn't become practical until the electronics industry invented new things just to make semiconductors. Then the Integrated Circuit or *IC* was invented. With the IC practical the semiconductor era exploded into an exponential growth curve that doesn't seem to stop. Nostalgia of times gone by and *the way things were* can be an emotional stopper to analytical thought and a barrier to new design. One can't be stuck with old technology. One has to be open-minded about applying old technology using new components.

Gestation of RHdb

The author wrote a number of various articles, first for himself, then for publications that paid money, sight-unseen, based on the work. That was refined until it resulted in ***RHdb***, the informal title for this work.

The two gentlemen to whom this work is dedicated, the author's father and father-in-law, were both born in the year 1900. Radio as a communications medium had just begun but it would not be until a year later that Marconi received the first trans-Atlantic radio signal. It would be three years until the brothers Wright flew the first successful heavier-than-air flying machine. Both men lived long and experienced the wonders of electronics and they both lived long enough to see the first humans set foot on the moon on ***live*** television from a quarter of a million miles away.

They had seen and heard the first *talking* motion pictures, heard radio broadcasts from *overseas* in the years before the United States was engaged in World War II, saw television from the living rooms of their respective homes. While they still lived, telephoning had replaced the dial with push-buttons, telephone calls coast-to-coast were as clear as across town, *color* television would become reality, the *communications satellites* would relay television programs from

[1] *Unobtainium:* noun; rare substances hard to get and expensive in scarcity; used only for exact duplicates of old electronics; may be replaced by newer components if esthetics allow.

anywhere in the globe.

The *computer* had also become reality but the *personal computer* would not exist until just after their passing. They would not have experienced web-surfing on a PC, using a *cell phone* small enough to carry in a shirt pocket, watch a movie recorded on a *DVD*, nor see *High-Definition Digital Television* on a thin flat-panel screen. They would not have had a *radio watch* that could listen to WWVB on 60 KHz and have it set itself to the correct time automatically every night. They didn't know about the *GPS* nor the existence of display screens showing how to get to a destination. They knew *wireless* as an old term for radio.

What a marvelous past century enabling all of us to improve our quality of life! We have more active and passive devices on the market, for lower prices than a half century ago, and can place orders for parts electronically and transact business the same way. ***RHdb*** was created to fill in some of the how-to-do-it gaps of knowledge for the hobbyist. ***RHdb*** is more of an ***advanced*** hobbyist reference. But, as said, it isn't a college textbook nor does it concentrate on *old radios* of the tube type. If anything, it concentrates on receivers...how to get signals of many kinds from both near and far. Solid-state technology is prominent and digital devices are a good part of that. Radio is no longer just analogue. Neither is it confined to one type of semiconductor per stage as in the old vacuum tube era. There are *systems on a chip* that can encapsulate most circuits in a single package.

The author is both a professional and a hobbyist in electronics, has been there and done that, is continuing to experiment, tries to innovate, and is enjoying life.

Always remember: ***Electrons, fields and waves obey THEIR laws.*** They aren't affected by emotions, nostalgia, glossy advertisements, personal braggadocio, magazine promises nor of some paper award given out by a few humans in an organization. *Learn those basic physical laws in order to make electrons, fields and waves do* ***your*** *bidding.*

Leonard H. Anderson
Retired Electronics Design Engineer, but retired only from Regular Office Hours.
Life Member, Institute of Electrical and Electronic Engineers (IEEE)
First Class Radiotelephone License since 1956, changed by FCC to lifetime GROL in 1985.
United States Amateur Radio Licensee K6LHA (Extra), first licensed as an amateur in 2007.
12 April 2014

Table of Contents

Subsystems

Projects

* Indicates chapter contains program listings

Totals

948 pages, 540 Illustrations

All drawings done by the author. Chapter 38 has 13 halftone illustrations from USA government sources. Chapter 39 has 9 photographs from the author's personal collection.

A few Chapters have blank last pages to make 2-sided printing possible.

Chapter 1

An Introduction To The Design Process

Design is a creative endeavor. In radio and electronics a designer must use available parts and arrange them with certain values to follow physical laws, beginning with a certain task/function, then arranging circuitry and components to accomplish the task. The arrangement may be a whole radio, modifications to a specific radio, or just an interface connection between radios or electronic boxes. The arrangement cannot be haphazard. Physics and the availability of components are ruling factors. ***Hobbyists*** are perhaps more free to be creative than professional radio designers...they are not ruled by market goals, are not required to be part of a company bureaucracy having excess paperwork, can adapt used, working components rather than sticking to corporate stock parts and don't have to attend staff meetings. On the other hand, hobbyists may not be specialists in certain areas and are required to *learn enough* about everything of their project in order for the entirety to work as desired. ***RHdb*** was created to assist the hobbyist in this learning...to present tutorials on radio and electronics fundamentals, circuits commonly used and necessary formulas for calculation of component values in those circuits...compiled by an author who is both a professional and hobbyist in radio and electronics.

What You Will Need As A Designer

1. Knowledge of electronics. This book and other materials are your references.
2. Paper and pen or pencil to write, print, sketch, and keep notes on everything.
3. Manufacturer's data on components, direct, from catalogs or from the Internet.
4. A safe, orderly place to keep all the information and all the parts.
5. The ability to do simple algebra and trigonometry.
6. A scientific calculator to crunch the numbers of algebra and trigonometry.
7. Some familiarity with electronics construction.
8. A computer is not imperative but computer circuit analysis programs can act as a quick trial-breadboard; use depends on the complexity of certain circuits..
9. Curiosity and some skepticism plus the desire to know more.

Let's take all those items in reverse order.

Curiosity, Skepticism, Desire To Know More

Curiosity is a good attribute. There may be more than one way to accomplish a circuit block's function and relation to other circuit blocks. There may be a better way that can use up some of those parts in your junk box. Or maybe not. You have to judge and you can't be a good judge without "evidence."

Anyone can make a mistake. Learn to double-check lengthy calculations. If some new circuit appears in a publication, try to check it out, learn from it. It may fit your need or maybe not. Be skeptical about "needing" some new feature in a manufactured radio advertisement, think about whether or not it is to your advantage. Remember that the purpose of an advertisement is to get you to buy the product. Ads aren't technical forums nor are they the judge of what you "should" do.

Never stop learning...and don't restrict your knowledge input to any one group in radio. Radio began as such in 1896. Its technology hasn't stopped advancing since that time.[1] The vacuum tube revolutionized all electronics when it became commercially available just before World War 1. The transistor was invented in 1947 and the first integrated circuits appeared on the market in the 1960s. The solid-state era was a "revolution squared" in all of electronics. Vacuum tube use was relegated to very specialized niches, no longer the active-device workhorse of radio. By the turn of the century in the new millennium there was a ***new era*** in all electronics.

Computers Aren't Imperative...

But personal computers can be a godsend in some aspects of designing. The *SPICE* derivative programs can be tools for breadboarding a circuit or subsystem without physically making it or analyzing it in either time- or frequency-domains.[2] A SPICE program user simply types in the component description and its location in the circuit as *node* numbers, then selects the analysis mode (time or frequency) and whether the result is presented as voltage, current, power and so forth. All of these analysis programs are mathematically correct and the results are as accurate as the typed-in circuit *model*.

Still other computer programs can calculate all the component values in filters, take the calculation drudgery out of working with transmission lines and their complex quantities. Some SPICE programs are available as free Internet downloads.

Some Familiarity With Construction

You can't complete an effective design without knowing something about how to physically build it. Vacuum tube radios used sheet-metal chassis having tube socket holes made by special punches. Wiring those was point-to-point using component leads soldered to socket or terminal strip solder lugs.

Solid-state circuitry construction convention uses printed circuit boards (PCBs) with wiring being the copper *trace* of the etched board. At frequencies above about 30 MHz, the lead lengths of both construction types becomes a part of the circuit itself. Etching PCB stock and drilling wire

[1] The first publicly-demonstrated radio communication took place in Italy (Guglielmo Marconi) or Russia (Aleksandr Stepanovich Popov) in 1896. Which one was "first" is relative to your particular ethnic view, definition, or historical reference. The fact is that radio had become a part of our lives less than a half century after that year.

[2] The word *breadboard* came from early radio of the 1920s and 1930s, referring to using a flat piece of wood as the "chassis" to hold all the parts. Since most kitchens of that time had such cheap pull-out wood flats for making bread, they became an easy-to-use prototyping platform. The word became synonymous with electronic prototyping while the kitchen breadboard (for bread making) is slowly becoming extinct.

holes has become the equivalent of point-to-point manual wiring.

With electronics miniaturization has come *Surface Mount Technology* or SMT. Lead spacing has become very small and manual soldering requires a new set of manual skills. If you don't have these new skills, you have to work around components that are SMT-only or get a partner to do the assembly. Whatever construction mode is used, good soldering technique is a must. That can't be taught by reading a book. It is a skill acquired by manual practice.

A Scientific Calculator as a Primary Number Cruncher

Most of the common "four-function" pocket calculators have the necessary number of digits to do accurate calculations. Most of those also include a fifth function in finding the square-root of numbers. That is fine for things like money amounts, but the ubiquitous palm-sized four-function pocket calculator doesn't have the *range* of number values needed in electronics and does not come with the *transcendental* functions needed for logarithms, exponentials, or trigonometric functions. The numeric range in electronics goes from picoFarads in capacitors to GigaHertz in frequency, a numeric range of 10^{21}:1. That can only be handled by a calculator with a *decimal exponent* capability. The *scientific* calculator has all of those plus extra calculation capability such as conversion between "English" and metric units and an ability to handle *complex quantity* calculations as easy as conventional scalar quantities.

This scientific calculator type can become your primary tool in design calculations...and you will be doing a great deal of calculations as part of the design process. The author's choice for a scientific calculator is, presently, a Hewlett-Packard 32SII and 35S.[3] The author is familiar with, and prefers, the *RPN* (Reverse Polish Notation) form built into most H-P calculators. Texas Instruments makes equally-good, versatile scientific calculators, as does Casio, both using the algebraic notation common to the inexpensive four-function calculators..

The *graphing scientific calculators* are fine for students first learning mathematics, but the extra graphing feature has low resolution with little practical value in design.

The Need For Simple Algebra and Trigonometry

That need is inescapable. Every component in a circuit needs calculation for the proper value. Most of those calculations involve simple algebra solutions, the remainder need simple trigonometry. Necessary skill level is that of a 12th grade graduate who has completed high school algebra and trigonometry classes and knows what a *logarithm* and *power of a number* are, no more.

There is little advantage in knowing more advanced mathematics rules beyond algebra and trigonometry. While that may seem sacrilegious to college instructors, it is nonetheless a truism in the practical design world. In the author's entire career in electronic design he had only three occasions to use calculus at work and those were for reports describing already-designed-and-tested

---

[3] The author bought an HP-35 in 1972 for $395 (fair-trade price then), an HP-67 programable calculator in 1977 for about the same price. The HP 32SII cost $60 in the year 2000, has some programmability but the distinguishing feature of that model is the ability to handle complex quantities in addition, subtraction, multiplication, and division by a single extra keystroke. By 2004 the HP 33S replaced the 32S II, same price. Both the 32 and 33 models have programming capability; their small batteries last for a long time thanks to internal very low power CMOS active devices and low-power LCD screens.

systems. Calculus and other advanced math are fine for a deep understanding and for developing solutions for techniques not yet discovered. This book contains systems and circuitry that already have developed solutions. There is some math greater than simple algebra presented (such as in Chapter 4) but only as a reference for later chapters on specific modulation and demodulation circuitry.

Chapter 2 has a review of simple algebra and simple trigonometry plus an introduction and small tutorial on *complex quantities.* Complex quantities are necessary for proper handling of design tasks involving resonance, matching, and transmission lines.

If you are comfortable with simple algebra, trigonometry, and complex quantity calculations, feel free to skip that chapter. If not, please review that section and the tutorials. Either way, this book will show worked-out numerical solutions for most formulas as examples.

A Safe, Orderly Information Storage Place

That may seem redundant since a home workshop seems like a safe, secure place to store notes, sketches, data sheets and so forth. But, a hobbyist does not always enjoy an organized environment. Weeks, perhaps months may go by due to necessary interruptions for family, travel, vacation, or other needs. Organizing information, dating notes and calculations, keeping component data in order are all useful in keeping the creative mind reminded.

Anyone who has done a lot of computer programming will understand the need for note-making in complex tasks. Few can remember in detail exactly what they had done a few months or years ago. It may be an irritation to make such notes but it is a greater irritation to come back after a month's absence and not remember *why* certain values were jotted down on a diagram.

Loose leaf notebooks are good organizers. It is easy to organize things by subject and date. The notebook can stand by itself, less subject to damage from dirt and spills.[4]

Data Sheet References

The Internet may be one of the best all-around sources for component information. Use it when available, print out those data sheets on selected parts. By the year 2004, all major US active and passive component manufacturers had detailed data on their products available through downloads. Much data was also available on components *no longer made*.

If you have only one original printed copy of often-used component data, a copy for personal archival storage is a safe bet. Copies of things like standard tables of 1, 2, 5 percent values or some often-used formulas can be useful.

Catalogs from parts distributors can be reference texts of a sort, letting you know what is available and how much it costs. In the year 2004 some of the major distributors had *links* to manufacturers' data on their Internet sites, valuable for quick judgement on component possibilities.

Application Notes from manufacturers are useful in showing typical applications and tips on

[4] Pets and little children roaming around hobby quarters can be accidental mess-makers despite their adorable existence. The author is allowed to coexist in his house by his wife's cats. Cats are very independent creatures, not given to much discipline training. While they have free run in the home office, they are physically barred from the workshop. There are dangerous things in workshops for small, inquisitive critters. It is best to exercise prudence to protect them...as well as your own valuable notes.

choices. While Appnotes (a common term) were designed primarily to interest others in buying manufacturer's products, many are mini-tutorials valuable as much as other textbooks. A few Appnotes **are** tutorials. Most are available as free downloads over the Internet.

Writing To Yourself

Sound absurd? Taking detailed notes for your own records is a good reminder of what you *were* thinking or considering for your project a few days ago or last week. Yes, your design project is one of a kind and not intended for production but hobby projects are subject to all sorts of interruptions at home, everything from family members to incoming telephone calls to shooing away pets. All of those can break your concentration and you might lose some creative spark just firing up. ***Date your notes and sketches.*** Dated notes jog your mind. Those are *very* useful for the future; you might want to change and improve your original project later.

If you have to test some component or sub-system on the bench to determine how it will fit into your design project, keep those notes. Such notes become the equivalent of manufacturer's data sheets.

Quadrille paper, cross-ruled in quarter or eighth inch increments is useful in roughing out a PCB layout or its position within a housing or cabinet, even some necessary mechanical item's sketches. You don't need to make it look pretty on paper, only to make it accurate to build, accurate to fit where it is supposed to be.

For manual PCB layouts *vellum* paper is available in conventional paper sizes. Vellum is translucent, can be used to trace package outlines, and is tough enough to take lots of erasures. Translucency allows manual layouts on both sides of the page to simulate double-sided PCBs. Vellum is even tough and stable enough to be used as a punch guide for drilling holes with 1:1 scale trace layouts..

Above all, keep all the paper information together. That saves time in hunting up some scrap of data that meandered into some other pile of data.[5]

Electronics Knowledge

You need to have sufficient knowledge to know a circuit, sub-system, or method exists and where to find the detailed information. This book is one starting point**.** Other books are fine but there are many of them, expensive, and may not explain in detail what you want to know. Selectivity in choosing the most useful ones *to you* is difficult. About the only way (in the author's opinion) is to spend some browsing time at technical book sellers.

Previous radio kit building is fine. It self-teaches assembly skills, soldering, and use of hand tools. Radio and electronic kits don't cover much of the *theory* behind the circuitry and few cover such proprietary items as coding of embedded processors that are part of the kit. Someone else did the design work and the design was oriented towards making the kit a profit-making item.

On the other hand, a thorough knowledge of radio and electronic device advertisements is not necessarily a good reference for what *should* be used as part of your radio. Ready-built, working

[5] Data seems to exhibit some strange phenomenon of "transporting" itself to another location without telling you. Disciplining data will keep it home as part of your project "family."

boxes of electronics are competitive items in a marketplace, each vying with other manufacturer's boxes for a sale...and thus a profit for the manufacturer and the dealer. Touted extra features may not be optimally designed and may not even be necessary.

Another area of acquiring more knowledge is electronics industry *trade publications*. These are *controlled circulation* free magazines targeting a specific electronics discipline.[6] Some examples are ***EDN, Electronic Design, RF Design, Microwaves & RF,*** and ***Microwave Journal***, all covering electronics disciplines. Be aware that such publications are targeting commercial electronics designers, may contain some confusing math in descriptions. But, they are also *free.*

Methodology In Hobby Radio Design

The design process is very interactive. It is a complex game of mental ping-pong trading off goals versus practicality versus parts availability versus everything else. There's no set of hard and fast rules but the following are good generalities:

1. Establish project goals, technical specifications.
2. Know or have the circuit knowledge available.
3. Have test equipment available, owned or borrowed.
4. Know component availability, where it is available and at what price.
5. Be prepared to compromise on any of the preceding..

Item (1) needs to be done for purposes of practicality. A project from an article is one thing, starting from scratch is quite another. Both need test equipment (3) and both require enough knowledge (2) to give you confidence you can complete it.

You must be able to get the parts from stock or a workshop *junk box.* Articles in magazines, especially very old ones, have parts made of ***unobtainium,*** a very rare material. You must be able to either duplicate that or figure out suitable replacements that are electronically equal..

In some cases you may be able to complete much of your project but one section is unfamiliar or uses parts that aren't fully specified. Junk box parts are a good example of the latter. You may have to make that a separate sub-project of measuring it thoroughly before proceeding. There is no point to continue trying to use something unknown despite some nostalgic emotional appeal..

The Block Concept and Architecture

The term "architecture" may seem strange for electronics but it applies there as well as to buildings. Just as any concrete-block building is made of blocks cemented together, a complete radio or electronic system is made of *circuit blocks*. The smallest blocks are individual components. Next come the individual circuits. Each individual circuit block is characterized by function, input-output conditions, power supply demand, and so forth. Several blocks together may perform a *sub-system* function, have its own input-output condition and power demand. All the blocks together

---

[6] The term *controlled circulation* refers to subscribers submitting their qualifications to a publisher. Subscription is free and the publication relies on advertising revenue for sole income. The qualification sent in by subscribers is used as proof that the publication reaches an audience that advertisers want to target.

make the *system* function. Such a system could be a complete radio.

Radio design during the vacuum tube era used a *stage* concept based on one particular tube for each stage. Stages were cascaded to form a complete radio receiver, for example. Radio receivers were often characterized by the number of tubes, presumably the greater the number of tubes the "better" the receiver. The stage concept of design persisted into the beginning of the solid-state era.

The advent of the integrated circuit brought about a change in design thinking. Each IC could contain a whole group of internal blocks performing a specific function. Individual designers were seldom able to design their own ICs so the available IC designs were used as whole circuit blocks. Function, input-output, power demand were all characterized by the IC manufacturer so designers could arrange ICs as easily as using individual transistors or resistors or capacitors. The *SOC* or *System On a Chip* term was born, each small solid-state package containing its own sub-system.

A transmitter with phase-locked loop (PLL) frequency control might have its frequency control section as a sub-system block. That block could be described as containing separate smaller blocks such as the variable (for tuning) oscillator, a crystal controlled reference oscillator, a phase detector and selectable count-down divider. The divider block could itself have cascaded binary flip-flops, each flip-flop a smaller circuit block. All the smaller blocks join together to function as the larger block to generate and control frequency. As to the transmitter, its overall architecture considers the frequency control sub-system as a functioning block. Inside the frequency control block all the smaller blocks must function together to form the whole.

Overall design can begin by the largest blocks arranged by function. These blocks can then be further detailed as to how they can interact. That leads to detailing the smaller blocks and their internal interaction. Eventually the smallest blocks are reached and characterized, then the process begins outward. Smaller blocks can now be examined in more detail to form larger blocks and so on until the overall system is reached. Along the way you have visibility into each block and possibly revising a few things to insure that smaller blocks can interact to form larger blocks.

Setting up the design in the block concept is sort of a ping-pong game. It goes outside-in then inside-out as needed to complete the whole design concept. While it is a game of sorts, its only score is that of knowing that the whole seems possible using practical means and available parts.

Organization Of This Book

RHdb is designed to be both a reference work and a tutorial on electronic circuits and systems focusing on radio. Readers can range from beginning radio hobbyist to those who are professional technicians and engineers. Information and subjects are arranged somewhat differently than conventional textbooks since this is not intended as a college-level text. The sections and subjects are grouped according to complexity, proceeding from passive components through active devices to functional circuits and finally whole radio systems. Readers may skip sections they are familiar with and are encouraged to study unfamiliar subject sections.

RHdb can be thought of as *filling in some gaps* in popular articles and texts. While some texts go to great lengths in explaining the *insides* of small ready-built devices ranging from diodes to ICs, the general tone is to refer to the ***exterior conditions*** necessary to make them work. With the increasing microscopic size of ICs, it does little good to explain something that *cannot be changed by a hobbyist.* At one time, with the aid of some fancy vacuum-drawing equipment and glass-blowing equipment plus a spot-welder, it was possible to actually make a vacuum tube. By the 1960 time it was already beyond the capabilities of even a rich hobbyist to get into a transistor, and certainly not an integrated circuit.

What exists is simply the interconnections between the parts and how to make them. In itself that is not an impossible task. It simply takes some skull-sweat to form and calculate what is needed. Paper and all the scribbles on it are good enough. If done correctly, it will show you what ***can be done.***

Book Construction

Each chapter contains a main subject body and some chapters have individual chapter appendices. Page numbering, formula numbers are all given by chapter number first, a dash, then the individual item number. This was felt to afford the easiest reference tool to reach an appropriate detail or formula. Footnotes are numerous and range from more detail on a particular item, reference notes to other texts and information sources, or just the author's comments.

Formulas abound in this book. Those are unavoidable, but are restricted as much as possible to simple forms of algebra or trigonometry. Most have numeric examples to further illustrate their use in step-by-step calculation. Formulas are identified by sequential numbers in parenthesis plus reference to their Chapter origin.

Some references give Internet websites. These were valid when the manuscript was prepared and the author doesn't guarantee that such will remain. The Internet environment is ephemeral, fluid and volatile. Printed, bound books remain intact, unchangeable as long as the paper remains readable. References to literature are in [square brackets, by number] with such references listed in numerical appearance order at the end of each chapter. References to preceding formulas is given in the **(**familiar) form with the first number denoting the Chapter.

There exist some lines of computer *pseudo-code;* i.e., higher-level computer language statements not in conventional syntax but still readable. Some of the operations are short enough to be translated to pocket size programmable calculators.[7] A few Assembler code examples are found, those following manufacturer's syntax...but with more extensive commentary included.

There exist references to particular vacuum tubes, transistors, integrated circuits given in this book; where possible those components were available or still being manufactured when this book was completed. Most vacuum tube types have already been discontinued before the new millennium. EIA-registered transistor types, once listed in the thousands of different numbers have dropped a hundred-fold by the new millennium only to be inflated by semiconductor manufacturer's own numbers. IC part numbers may change or be modified, some even becoming extinct due to decreased market demand. Electronics technology keeps changing, has changed even while the manuscript was prepared. The hobbyist designer must retain current references as to what is available, when, and for how much.

A few subjects have been deliberately omitted. There will be no details on how to spoof subscription services. Piracy is illegal. Neither are there specific *mods* or modification information for transmitters to operate at unauthorized frequencies. High power transmitter final amplifier

[7] Many programmers insist that the **C** languages are better for computer programming. While that may be true from the programmer's viewpoint, this book is not about programming but about hardware. BASIC is an *interpreter* language based on the syntax and expressions of FORTRAN whose acronym is derived from *FORmula TRAnslation*. Either language better illustrates the relationship between a formula and its computer equivalent.

information is scarce here; the ARRL has many publications on that. *DSP* or Digital Signal Processing techniques are mentioned but not detailed. The main focus is on the frequency range below about 40 MHz. A reason for that is the relative low cost of test equipment for the hobbyist below VHF.

A Word on Illustrations

All the schematics in ***RHdb*** were drawn by the author. All. No exceptions. For quite some time the author has developed a personal style that is partly illustrative. Figure 1-1 shows that.

The two examples at left in Figure 1-1 are the most used. A connection is very definite; it is shown with a connecting dot. Rather than use two lines just crossing (but no connection), there are some slight spaces on the vertical conductor that ***appears as if it went under*** the horizontal conductor. This was felt to be more of the quick-glance sort of appearance. In the older method of just having two conductor lines crossing, that appeared confusing.

A schematic can have several sets of conductor lines both vertical and horizontal that are crossing. In all those cases, actual connections will have the connecting dots while all others will be continuous but with small spaces.[8]

The small ***x*** has been used by others, particularly for IC pins which are not connected to anything. The ***Broken line*** technique borrows from several technological disciplines for graphics. The jagged line at right angles to the line being broken has no significance except to indicate there is a space between each.

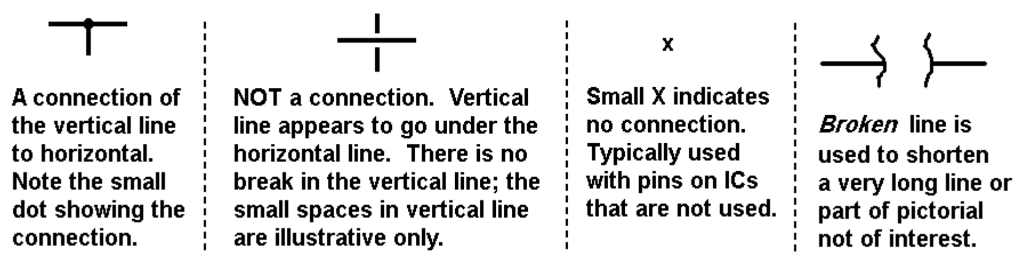

Figure 1-1 Particular drawing scheme for illustrations in *RHdb*.

A Final Word...

RHdb shies away from featuring the *latest and greatest* techniques and hardware, just for the sake of *marketing a book.* This book is a reference book and will remain so for many years. The author's work experience over the last half-century showed that some techniques and components have entered the electronics market, found wanting, and quietly disappeared as techniques became more standardized...yet were lionized in press releases *for the future* when first introduced.

This book is part textbook. Textbooks are associated with classes that have tests and homework. Is there homework in this book? Absolutely. You are doing work at home. Does anyone grade your homework *tests*? Yes, **you** do. Your ***grade*** is how well you accomplished a task for yourself and how much fun you've had. Let's hope your ***grades*** are excellent!

---

[8] All schematics were composed with MS Paint. For the passing-under conductor line, the small break is the same width as the conductor line itself. This borrows from illustration use, particularly with arrows pointing to component parts in photographs.

Chapter 2

Mathematics Needed In Design

Simple algebra and trigonometry is essential to calculate circuit component values. Transcendental functions are used in phase and power level calculations. Complex quantities let you do interstage matching and determine network components. All are invited to review this chapter but those familiar with it can skip it or keep it as a reference. Those who left all the "math stuff" after their last school class should study it carefully. Simple mathematics is absolutely essential in design. You must feel comfortable with it.

Powers of Ten and Exponents

Radio and electronic numeric values can range from several billion (1,000,000,000) to one-trillionth (1 / 1,000,000,000,000). To avoid needless writing of zeroes, *scientific* notation is much more convenient. This is a system of value multipliers having the following values and names:

$$\mathbf{K}\text{ (Kilo)} = \text{x } 1000 = \text{x } 10^{3} \qquad \mathbf{m}\text{ (milli)} = \text{x}\left(\frac{1}{1000}\right) = \text{x } 10^{-3}$$

$$\mathbf{M}\text{ (Mega)} = \text{x } 1{,}000{,}000 = \text{x } 10^{6} \qquad \mathbf{u}\text{ or }\mu\text{ (micro)} = \text{x } 10^{-6}$$

$$\mathbf{G}\text{ (Giga)} = \text{x } 10^{9} \qquad \mathbf{n}\text{ (nano)} = \text{x } 10^{-9}$$

$$\mathbf{T}\text{ (Tera)} = \text{x } 10^{12} \qquad \mathbf{p}\text{ (pico)} = \text{x } 10^{-12}$$

$$\mathbf{P}\text{ (Peta)} = \text{x } 10^{15} \qquad \mathbf{f}\text{ (femto)} = \text{x } 10^{-15}$$

The little **x** indicates *multiplied by*. The multiplier letter immediately precedes a value name. For example, a resistor of 2,200,000 Ohms is written ***2.2 MegOhms***. Since resistors are always valued in Ohms, it is common to note it as ***2.2 Meg*** or just ***2.2 M***. Resistors less than 1.0 Ohm are rare and usually described as a decimal fraction of an Ohm. For another example, a capacitor of 0.00018 Farad value is written as ***180 µFd*** with Farad abbreviated.. A frequency of 121,500,000 cycles per second is written ***121.5 MHz*** using *Hz* for Hertz. An inductance of 0.010 Henry is written ***10 mHy***. The convention throughout the ***RHdb*** is to capitalize the first letter of abbreviated value multipliers if the value is above unity. Value multipliers below 1.0 are not capitalized. This is a small help in differentiating the large from the small.[1].

---

[1] This is not an absolute standard but the change in *style* avoids some common mistakes in notation.

Value names and their abbreviations used in following chapters are:

Frequency: Hertz = **Hz** Resistance: **Ohms** or **Ω**

Capacitance: Farads = **Fd** Inductance: Henrys = Henries = **Hy**

Voltage: Volts = **V** Current: Amperes = **Amp** or **A**

Time: Seconds = Sec or **s**

Metric Length: Meter = **m** Metric Mass: Kilogram = **kg**

English Length: Feet = **ft** English Mass: Pound = **lb**

Some other texts use the Greek letter Omega (Ω) for the Ohms value name. There is some doubt as to the plural of the name of inductance. While the singular is *Henry* there are two forms of the plural; this book stays with the common-text plural using *ies* as the ending.

Wavelengths for frequencies greater than 300 MHz are commonly described in centimeters or ***cm***, a centimeter being 0.01 meter, or in millimeters or ***mm***, a millimeter being 0.001 meter. Note that the metric mass already has the multiplier of kilo but the abbreviation ***k*** is in lower-case.[2]

Note also multipliers less than unity: Their power-of-10 exponent is *negative*. A negative exponent always refers to the *reciprocal* of a number whose value was a positive exponent before being divided into unity. The multiplier itself is not negative unless a negative sign *precedes* the whole multiplier.

Laws of Exponents

Any variable name's trailing superscript is that variable's *exponent*. That exponent denotes the variable is *raised to the power of that exponent value.* If an exponent is, for example, 2, it means the variable is *squared* or multiplied by itself. But, suppose you encounter a variable raised to a *negative power* or to a fraction of the power. What then? The following applies to such situations:

$$A^{(-X)} = \frac{1}{\left(A^{X}\right)} \qquad B^{(1\,/\,Y)} = \sqrt[Y]{B} \ = \ \text{Yth root of B}$$

Those follow from the following *Laws of Exponents*:

[2] The author can understand some confusion over styles here, but ***kg*** is the internationally-accepted standard for mass. Acceptable style for value names is to capitalize the first letter if the name is in honor of a person such as Hertz, Volta, Faraday, Henry, and Ampere (NIST Special Publication 811, 1995).

$$A^X \cdot A^Y = A^{(X+Y)} \qquad A^X \cdot B^X = (A \cdot B)^X$$

$$\left(A^X\right)^Y = A^{(X \cdot Y)} \qquad \frac{A^X}{A^Y} = A^{(X-Y)}$$

$$A^{(X/Y)} = \text{Yth root of}\left(A^X\right) = \left(A^X\right)^{(1/Y)} = \sqrt[Y]{A^X} \qquad \mathbf{A^0 = 1}$$

The little *dot* is a common convention to denote multiplication. This replaces the small *x* often used in arithmetic as a multiplier; that replacement reduces confusion with variables using x or X. Do not confuse the multiply dot with a decimal point. Decimal values less than unity will be written with a numeric 0 preceding the decimal point such as *0.2* meaning two-tenths.

Note that any variable raised to a zero power always results in unity.

Raising a variable to a power is simple if the value of the power is an *integer*. If the power is 2, the variable is *squared* or multiplied by itself. If the power is exactly ½ the *square root* of the variable is required. Most scientific calculators have an intrinsic square root function capability. Non-integer powers of variables can be found through *logarithms*. There's more on logarithms in a few pages.

Algebra Review

As an example for review, following formula needs to be solved for B, C, or D:

$$A = \frac{2B\,(C+1)}{D^2 - 1}$$

Note the parentheses. Those denote the entire quantity of (C + 1) is multiplied by 2B in the numerator. In the denominator, the quantity of D is squared first before 1 is subtracted from it. No parentheses are used for the denominator but those could be used without altering the identity. A, B, C, and D are *variables* since their numerical value is related to both other variables as well as *constants* of 1 and 2.[3]

In solving for other values, let's suppose you know the values of variables A, B, and D, but need to solve the numeric value of C. The manipulation can begin with multiplying both sides of the equation by the whole denominator:

$$A\left(D^2 - 1\right) = 2B\,(C+1)$$

The next step would be to divide both sides of the equation by 2B in order to "isolate" C on one side:

$$\frac{A\left(D^2 - 1\right)}{2B} = C + 1$$

Note: Multiplying or dividing *both* sides of the equation is absolutely necessary to maintain

[3] Terminology of *variable* and *constant* are from computer programming convention rather than mathematics.

equality..

The final step would be to subtract 1 from both sides of the equation. We can turn the expression left for right about the equal sign for the convention of showing the unknown at the left:

$$C = \frac{A\left(D^2 - 1\right)}{2B} - 1 = \frac{A\left(D^2 - 1\right) - 2B}{2B}$$

We can use parentheses, square brackets, curly brackets to *nest* sub-expressions in a formula. That is sometimes useful in ordering operations with a handheld calculator. We can rewrite the solution for C in any of the following forms without changing the *identities:*

$$C = \frac{A\left(D^2 - 1\right)}{2B} - 1 = \left[\frac{A\left(D^2 - 1\right)}{2B}\right] - 1 = \left[\frac{A\left(D^2 - 1\right)}{2B}\right] - \left[\frac{2B}{2B}\right] =$$

$$\frac{\left[A\left(D^2 - 1\right)\right] - 2B}{2B} = \frac{\left[A\left(D^2 - 1\right) - 2B\right]}{2B}$$

It should be obvious that 2B divided by 2B will result in *unity* or the value of 1. Such manipulations are useful in obtaining a *common denominator* such as in the lower two identities.[4]

The equation format used is largely one for clarity. In some computations, especially in computer programs, two nearly-equal sub-expressions will have strange results. If the following were true...

$A\left(D^2 - 1\right) \approx 2B$ [note the *approximation* sign rather than an equal sign]

...then the numerator of the entire expression for C would approach zero and the value of C itself would approach zero. A similar thing would happen if D came very close to the numeric value of 1; the numerator would go to a negative value.

If the value of B was very low the value of C would approach *infinity*. The numerator value would be dependent only on the values of A and D but the denominator contains only B. Some *high level language* computer programs may crash if a denominator value reaches zero.[5]

$$B = \frac{A\left(D^2 - 1\right)}{2(C+1)} \quad D^2 = \frac{2B(C+1) + A}{A} \quad D = \sqrt{\frac{2B(C+1) + A}{A}}$$

Factoring: $D^2 - 1 = (D+1)(D-1)$

---

[4] There is sometimes a confusion between *equality* and *identity*. An equality is always denoted by the equal sign and indicates the expression on the left of the equal sign is equal in value to the expression on the right. An identity is using or manipulating one of the expressions, perhaps using different functions, to maintain the equality even though the appearance might be quite different than the original.

[5] It is a good idea in computer programming to add a *trap* to detect such things ahead of time.

The Quadratic Equation and Something On Roots

If an equation can be reduced to the form of:

$$A X^2 + B X + C = 0$$

...and provided that X does not have a value of zero, then there are two possible *roots* of X:

$$X = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}$$

The plus-minus sign indicates *two* solutions

The following rules apply on quadratic roots (note the *greater-than* and *less-than* symbols):

If $B^2 \rangle 4AC$, roots are real and unequal

If $B^2 = 4AC$, there is only one solution with $X = \frac{-B}{2A}$

If $B^2 \langle 4AC$, roots are imaginary (trying to take square root of -1)

For an explanation of *real* and *imaginary* values, skip ahead to *complex numbers* in this Chapter. Hint: An imaginary value would be an attempt to get a square root of a negative number.

In most cases of design calculations, the roots will be real and unequal. The selection of applicable root value for two real solutions depends on what is being calculated. While such a numeric value test is not good form in mathematics, it is nonetheless quick and easy, as accurate as with a formal method. In the following Chapters, ***RHdb*** formulas have the proper root already selected for a practical solution.

Logarithms

"The *logarithm X* of the number *N* to the base *B* is the exponent of the power to which *B* must be raised to give *N*." That is expressed in formulas as:

$$\mathrm{Log}_B\, N = X \qquad \text{- or -} \qquad B^X = N$$

When base *B* is 10, the logarithm is known as a *common logarithm*. In that specific situation we can write the mathematical definition as:

$$\mathrm{Log}_{10}\, N = X \qquad \text{- or -} \qquad 10^X = N$$

In even more common terms, the left-hand expressions are referred to simply as the *log of N* (to find *X*) and the right-hand expressions are the *antilogarithm of X* (to find *N*). Since *common logarithms* are so common in practice, the subscript 10 to indicate the base is usually omitted.

The *natural* or *Naperian logarithm* uses ***e*** (2.718281...) as the base. Natural logarithms can be written as either log_e or *Ln*. The definitions of natural logarithms can be written as:

$$\mathrm{Ln}\, N = X \qquad \text{- or -} \qquad \varepsilon^X = N \qquad [\varepsilon = 2.71828.......]$$

As an example, suppose N equals 4. The *common logarithm of 4* would be 0.60206. Raising 10 to the power of 0.60206 would result in 4.0000. In older *BC* times (Before Calculators) tables of logarithms would be used; tables precalculated usually to only 5 decimal figures. Pocket scientific calculators calculate the exact (to the stated number of digits accuracy) logarithm or anti-logarithm to at least 10 decimals.

Most high-level computer language intrinsic logarithm functions are *natural*. Common engineering design convention is to use *common* logarithms or logs to the base 10. To convert from one to the other we can use rules on *change of base of logarithms*:

$$\mathrm{Log}_B\, M = \mathrm{Log}_C\, M \cdot \mathrm{Log}_B\, C = \frac{\mathrm{Log}_C\, M}{\mathrm{Log}_C\, B}$$

If *base B* is ε and *base C* is 10 then the conversion can be noted as:

$$\mathrm{Ln}\,(M) = [\mathrm{Log}\ (M)]\cdot[\mathrm{Ln}\,(10)] = \frac{\mathrm{Log}\,(M)}{\mathrm{Ln}\,(\varepsilon)}$$

That allows some easier conversion for calculator and computer programs since:

Ln (10) = 2.3025 85092 99404 56840

Log (ε) = 0.43429 44819 03251 82765

Twenty-digit accuracy is given here for completeness. The style of separating long number strings to five numerals with a space is common in large numerical value tables. The parentheses around the logarithm *arguments* is common convention.

For seven digit accuracy the base conversion formulas become:

Ln (M) = 2.302585 Log (M)

Log (M) = 0.4342945 Ln (M)

Pocket calculator accuracy is about ten digits and high-level computer languages have accuracies from 7 to 15 digits. It should be noted that the numeric value of *Ln 10* is the reciprocal of the numeric value of *Log **e***; something that may be handy in writing calculations for computer programs.

The power of logarithms lies in the law of exponents: To multiply two numbers, add their logarithms, then use that sum to find the antilogarithm to find the product. To divide two numbers, subtract the divisor's logarithm from the dividend's logarithm, then use the difference to find the antilogarithm and quotient. That logarithm addition and subtraction was the basis for the old, non-electric, highly-portable *slide rule*.

Slide rules have two primary scales marked with a *logarithmic progression* of decimal numbers. That is, the logarithms of the decimal numbers are in a linear progression, not the decimal numbers. The decimal numbers are far apart at low values but compressed at high values. The "C" and "D" slide rule scales cover a 1.0 to 10 range. To multiply two numbers one aligns the *index* or edge mark of the sliding scale above the position of the first number on the fixed scale. A sliding *hairline* marker (separately-moveable glass or plastic) is positioned over the other number on the sliding scale. The product value is then read off the fixed scale under the hairline.

Slide rule division is done similarly except the sliding scale is moved in the opposite direction. What is happening is the addition (moving to the right) or subtraction (moving to the left) of the logarithms. The antilogarithm is found under the hairline marker. Slide rules are very handy for quick calculations but suffer from an inherent lack of accuracy. The advent of the semiconductor-based pocket calculator at the start of the 1970s marked the end of slide rule use.

Decibels

Magnitudes in radio and electronics can vary over a billion to one range or more. A 10 KW transmitter may have an RF RMS value of 707 Volts at its 50 Ohm feedline...yet a distant receiver may have the same signal at its antenna input of only 7.07 microVolts. That is a 100,000,000 to 1 ratio in voltage. If the distant receiver's feedline is also 50 Ohms, the received power is only 1 picoWatt. The power ratio between those two extremes is 10,000,000,000,000,000:1 or 10^{16}:1! Rather than writing all those zeroes, it is better to use logarithms. The ***bel***, named for Alexander Graham Bell, is such a value, the common logarithm of the power ratio. One-tenth of a *bel* or ***decibel*** (***db*** in common notation) is expressed by:

$$db_{power} = 10 \cdot \mathrm{Log}\left(\frac{Watts_{out}}{Watts_{in}}\right) \qquad \text{- or -} \qquad Watts_{ratio} = 10^{\left(\frac{db_{power}}{10}\right)}$$

A ***db*** is dimensionless. That is, it expresses only the ratio of two values of the same type. In the example given, the large power ratio would be written as 160 db. In general, a negative db value denotes a power loss, a positive db value a power gain. The sign of a db value depends on how one is defining the ratio. If the text says specifically that there is so much *db loss* or that a signal is *db down*, then the given decibel value is written positive while the actual value is negative.[6]

Voltage and current ratios can also be written as db values. These are:

$$db_{E} = 20 \cdot \mathrm{Log}\left(\frac{E_2}{E_1}\right) \qquad \text{- or -} \qquad E_{ratio} = 10^{\left(\frac{db_E}{20}\right)}$$

$$db_{I} = 20 \cdot \mathrm{Log}\left(\frac{I_2}{I_1}\right) \qquad \text{- or -} \qquad I_{ratio} = 10^{\left(\frac{db_I}{20}\right)}$$

Note: The subscripts of db are seldom used in practice, are shown here for illustration. A circuit or system response is given in un-subscripted db values regardless of power, voltage, or current ratios. One must exercise caution to use the right ratio type when finding a new power, voltage, or current value as a result of db loss or gain from the original reference value.

The reason for the difference of 10 versus 20 multiplier of the log of the ratio is from the definitions of power. Given a fixed resistance R, power P is related to voltage E and current I as:

$$P = E \cdot I \qquad P = \frac{E^2}{R} \qquad P = I^2 \cdot R$$

Voltage squared or current squared is responsible for doubling the multiplier of the log of a voltage

[6] Bad practice but common.

or current ratio.

In the original example the power loss between transmitter and receiver was 160 db. In determining the receiver voltage (in a 50 Ohm system where R = 50), the voltage ratio is 100,000,000:1 or 10^8:1 (exponent of power of 10 is 160/20 = 8). The current ratio is also the same 10^8:1, from about 14 Amps at the transmitter to 0.14 µA at the receiver. Since the receiver input voltage is 7.07 µV, the received power is thus 1.0 pW.

Some Decibel Values Referenced

While decibels are themselves dimensionless, the radio and electronics industry has adopted some common suffix conventions, principally for power levels. These are:

dbm: 0 dbm = 1 mW in a 50 Ohm system

dbw: 0 dbw = 1.0 Watt

dbc: 0 dbc = Carrier power as specified

dbµ: 0 dbµ = 1.0 µV

The most common is **dbm**. *In a 50 Ohm system* refers to the characteristic impedance of a system of coaxial cable joined components. A power level expressed as 13 dbm would mean 20 mW in a 50 Ohm system. A power level of **-**13 dbm would be 50 µW (negative exponent, 1 mW divided by 20). Microwave region powers are also expressed in dbm even though they are in waveguide rather than in coaxial structures.

A lesser-used level is **dbw**, still found on a few component specifications. It would have a power level 30 db higher than 1.0 dbm. The **dbµ** level is principally used in receiver designs to express the low signal levels received..

The referenced level of **dbc** is most often used with sideband components of a modulated carrier wave or with possible *intermodulation* products or *phase noise* effects. There the carrier level is used as the reference of 0 dbc with sidebands, products, or effects being so many db *down from the carrier level*.

Some older texts refer to audio levels in terms of **dbu** or **VU** (Volume Units). Those have a 0 dbu equivalent to a 1.0 mW level in a 600 Ohm system.

Simple Trigonometry

AC and RF waveforms have two major components. The amplitude of the waveform is the *magnitude* while the time position of the waveform (referred to some starting time) is the *phase angle* (familiarly just *phase*). The amplitude at any point in the waveform is a function of both magnitude and phase angle. The amplitude of a sinusoidal waveform at a particular point can be determined by simple trigonometry.

Figure 2-1 shows one point in time of an AC or RF waveform as a triangle on a rectangular plot. The waveform magnitude is represented by the triangle hypotenuse extending from plot center. Waveform phase angle is represented by angle alpha. A zero angle would mean the hypotenuse lies on the X axis in the direction to the right. Increasing angle value is a counter-clockwise rotation.[7]

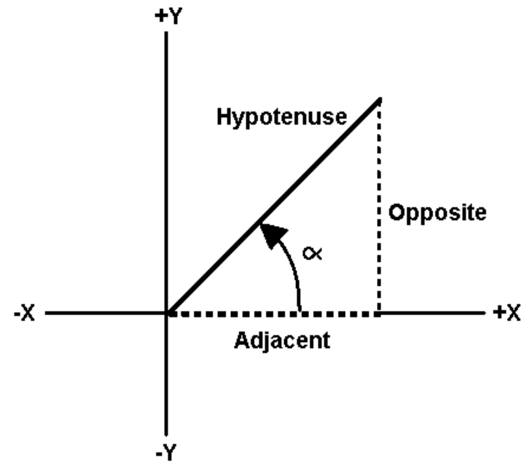

Figure 2-1 Triangle formed by magnitude (length of hypotenuse) and angle α (alpha) with side-opposite and side-adjacent shown by dotted lines. The X (horizontal) and Y (vertical) axes' signs are also shown.

The outermost end of the hypotenuse (away from plot center) represents the waveform amplitude at any one point in time. The hypotenuse is often referred to as the ***vector*** or *radial line*. A plot of magnitude and phase angle is called a ***polar form***. The center of a polar form plot is the *point of origin*.

The triangle can also be described in a ***rectangular form*** using the side opposite the angle and the side adjacent to the angle. A rectangular form plot is also called *Cartesian* in mathematics texts.[8] The *coordinates* in rectangular form are the **x** (horizontal) and **y** (vertical) ***axes***. Dimensions in rectangular form refer to the distance from each of the axes.[9]

Angle α can be expressed through ratios of the sides. Length of hypotenuse is ***M*** while adjacent, opposite sides expressed in terms of ***X*** and ***Y*** axis dimensions:

$$Sine\ \alpha = \frac{\mathrm{Y}}{\mathrm{M}} = \frac{side\ \mathrm{opposite}}{hypotenuse} = \mathrm{Sin}(\alpha)$$

$$Co\sin e\ \alpha = \frac{\mathrm{X}}{\mathrm{M}} = \frac{\mathrm{side\ adjacent}}{\mathrm{hypotenuse}} = \mathrm{Cos}(\alpha)$$

$$Tangent\ \alpha = \frac{\mathrm{Y}}{\mathrm{X}} = \frac{\mathrm{side\ opposite}}{\mathrm{side\ adjacent}} = \mathrm{Tan}(\alpha)$$

The sine or cosine of an angle will always have a ratio between 0 and ±1.0. Neither the opposite nor adjacent sides can be longer than the hypotenuse. The tangent of an angle can be any ratio between 0 and ± infinity. It is convention to abbreviate notations of trigonometric functions to ***Sin***, ***Cos***, and

[7] Hopefully the future will still have some clocks with "hand" indicators so that everyone will know what rotation is described by *clockwise* and *counter-clockwise* (or *anti-clockwise* as the British say).

[8] After French mathematician and philosopher Rene Descartes.

[9] Math texts usually call the horizontal coordinate location as the *abscissa* and the vertical as the *ordinate*. There are enough new names and values in this book that we don't have to worry about all of the "proper" names for mathematical terms.

Tan rather than writing out the entire function name. Angle α is the *argument* for a Sin, Cos, or Tan function, usually shown within parentheses. Knowing the angle and one side allows solution of an unknown side:

$$Y = M \cdot \mathrm{Sin}(\alpha) = X \cdot \mathrm{Tan}(\alpha) = \sqrt{M^2 - X^2}$$

$$X = M - \mathrm{Cos}(\alpha) = \frac{Y}{\mathrm{Tan}(\alpha)} = \sqrt{M^2 - Y^2}$$

$$M = \frac{Y}{\mathrm{Sin}(\alpha)} = \frac{X}{\mathrm{Cos}(\alpha)} = \sqrt{X^2 + Y^2}$$

The triangle's angle can be found by the *inverse* trigonometric function using any two known sides. The inverse is denoted by the ***Arc*** prefix but is commonly noted as the superscript of ***-1*** meaning a mathematical ***inverse.***

$$\alpha = \mathrm{Arcsin}\left(\frac{Y}{M}\right) = \mathrm{Sin}^{-1}\left(\frac{Y}{M}\right)$$

$$\alpha = \mathrm{Arccos}\left(\frac{X}{M}\right) = \mathrm{Cos}^{-1}\left(\frac{X}{M}\right)$$

$$\alpha = \mathrm{Arctan}\left(\frac{Y}{X}\right) = \mathrm{Tan}^{-1}\left(\frac{Y}{X}\right)$$

Any value or variable between natural limits can be the *argument* of an inverse trigonometric function.[10] The limits of an arcsine or arccosine argument are between plus and minus unity. There are, essentially, no limits on the range of an arctangent argument since the maximum magnitude is infinity.

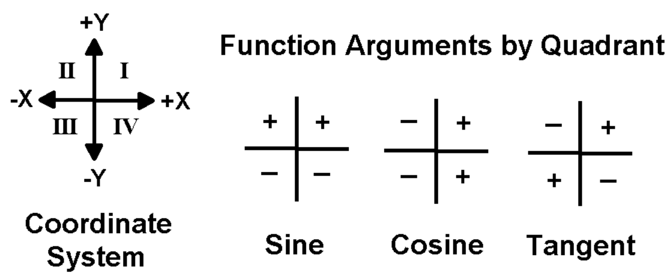

Figure 2-2 Guide to angle function arguments.

Note: Some high-level computer languages and calculators may not accept angular arguments greater than a full circle (360°). If so, you must subtract exactly a full circle until the argument will be in range of the computer language. If a negative value results, that is the argument.

Polarization of Angles in Trigonometric Functions

Figure 2-2 is a guide to the resulting polarity for Sine, Cosine, and Tangent functions dependent on the argument angle's sign. Note that an increasing positive angle turns counter-clockwise and

[10] An *argument* of a function refers to the variable to be acted upon in a function. For example, if the tangent of angle α is to be found, α is the *argument* of a tangent function. If α is to be found from the arctangent of x, x is the *argument* of the arctangent function.

an increasing negative angle turns clockwise. The Sine of $\pm 90°$ is always unity while the Cosine of $\pm 90°$ is always zero. The Sine of $0°$ and $\pm 180°$ is also always zero while the Cosine of $0°$ and $\pm 180°$ is always unity.[11] The Tangent of $0°$ and $\pm 180°$ is always zero while the Tangent of $\pm 90°$ is infinity.

Mathematical texts denote *quadrants* by Roman numerals, beginning with quadrant I as the upper right and moving counter-clockwise to quadrant IV at lower right.[12]

Complex Numbers

All numbers, variables, and constants presented so far have been ***scalar***. Scalar numbers have only one dimension. Complex numbers have two dimensions, defining a particular point in two-dimensional space or representing a *vector* having a radius and angle. In the polar form we can express a complex number in this manner:

$$C_X = \text{Magnitude} \angle \text{Angle}$$

In radio and electronics the Magnitude could be an RF voltage or current while the Angle would be the phase angle of that RF voltage or current. The same complex number can be expressed in rectangular form as:

$$C_X = \text{Real} + j\,\text{Imaginary}$$

The notation of the little ***j*** prefix is to firmly establish the Imaginary part.[13] The plus sign indicates the sign of the Imaginary part; it does not indicate that the Real part is added to the Imaginary part. A minus sign preceding the ***j*** would indicate that the Imaginary part is signed negative. The Real part may have either a plus or minus sign preceding it to sign only the Real part. By convention, no Real part sign is taken to be positive.

Polar form notation may have negative signs for either Magnitude or Angle or both. As in rectangular form notation, either sign pertains only to that particular part. Again, by convention, no sign is taken to be positive. The relationships are:

$$\text{Real} = \text{Magnitude} \cdot \text{Cos}(\text{Angle})$$

$$\text{Imaginary} = \text{Magnitude} \cdot \text{Sin}(\text{Angle})$$

$$\text{Magnitude} = \sqrt{\text{Real}^2 + \text{Imaginary}^2}$$

$$\text{Angle} = \text{Tan}^{-1}\left(\frac{\text{Imaginary}}{\text{Real}}\right)$$

---

[11] The relationship of Sine to Cosine will have a direct bearing on modulation, demodulation, and some forms of mixing.

[12] Defining a quadrant is seldom required in electronics design work. The quadrant ***location*** for an angle or vector is extremely important for proper polarities. Try to visualize the four quadrants.

[13] Mathematicians use ***i*** to denote the Imaginary part. Since a lower-case i is often used for an AC current, electronics and radio notation uses the lower-case j as a substitute.

Figure 2-4 graphically depicts the relationships. Note the similarity to the simple trigonometric functions of Figure 2-2. The polar form magnitude value is the same as the triangle's hypotenuse, sometimes referred to as a *vector*. The rectangular form Real part is the same as the X values and the Imaginary part is the same as the Y values.

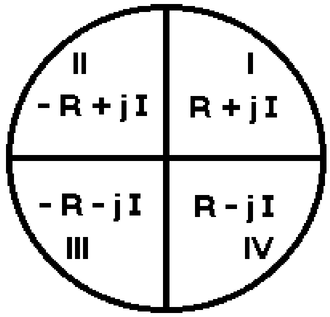

Figure 2-4 Rectangular form part polarities relative to quadrant.

In rectangular form, the Imaginary part is always signed either positive or negative as a notation that it "belongs with" the real part. Complex quantities will always have two parts even if one of them has a value of zero. Note: Sometimes only one part of a complex quantity is shown; if so, the part not shown is considered zero.

A vector's angular position in a quadrant will determine the signs of the rectangular-form parts, as shown in Figure 2-5. The value of the angle can never exceed ±360 degrees (a full circle rotation) but it can also be given within the limits of ±180 degrees. In limiting angles to a half-circle, subtract 360° from a positive angle or add 360° to a negative angle.

Note that rotating a vector's angle by exactly 180 degrees is the same as changing the signs of the entire complex quantity in rectangular form. That can be seen by rotating a vector from quadrant I to III or another vector from quadrant IV to II.

In polar form notation the magnitude is positive but the angle changes depending on the quadrant. The angle sign will be inverted in quadrants II or IV after a rotation from quadrants I or III.

Can a polar form have a *negative magnitude*? Yes.
Signing a magnitude negative is the same as a quadrant rotation of 180 degrees. That's also the same as putting a signal through a *phase inverter* circuit.[14] Rotation inverts the signs of both parts in its rectangular form. Converting a rotated rectangular form to polar form will show the effect.

Magnitude polarity can be changed back to positive by the following rules regarding the angle's polarity and sign:

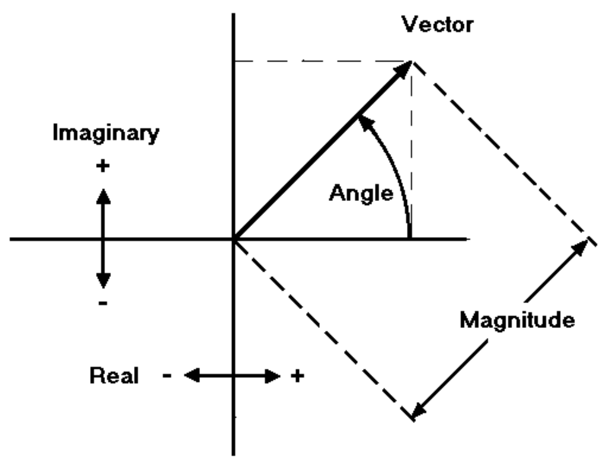

Figure 2-3 Graphical relationship of complex quantity in both rectangular and polar forms.

If the angle was positive, subtract 180° (or π radians) from it.

If the angle was negative, add 180° (or π radians) to it.

---

[14] More on that much later in chapters on active devices. Keep that in mind because such things will occur in the practical world and can show up as the result of complex quantity math calculations.

Complex Quantity Arithmetic

Complex arithmetic has notation the same as scalar values. Complex quantities may be noted the same as scalar quantities in expressions. But, because each complex quantity has two parts, the actual arithmetic operations are more involved than with scalar quantities.

Adding or subtracting two complex quantities is done in the rectangular form. Real parts are added or subtracted only with other Real parts; Imaginary parts are added or subtracted only with other Imaginary parts. Given two complex quantities, **Cx** and **Dx**, addition or subtraction would be:

If: $C_X = R_C + j I_C$ and $D_X = R_D + j I_D$

Addition: $C_X + D_X = (R_C + R_D) + j(I_C + I_D)$

Subtraction: $C_X - D_X = (R_C - R_D) + j(I_C - I_D)$

By notation convention the sign between the real and imaginary parts can take the sign of the imaginary part and vice-versa:

$$R + j(-I) = R - jI$$

A *scalar* (or "real" quantity in computer language terminology) can be added to or subtracted from a complex quantity. A scalar quantity would have a zero Imaginary part so it transforms to the complex form shown below, K as the example scalar quantity:

$$K_{(\text{complex form})} = K + j\,0$$

Mechanization of adding a scalar to, or subtracting a scalar from a complex quantity is:

$$C_X + K = (R_C + K) + j(I_C + 0) = (R_C + K) + j I_C$$

$$C_X - K = (R_C - K) + j(I_C - 0) = (R_C - K) + j I_C$$

If a complex quantity is subtracted *from* a scalar, then something new happens:

$$K - C_X = (K - R_C) + j(0 - I_C) = (K - R_C) - j I_C$$

Note that the sign of the Imaginary part has reversed. It is important to keep track of signs and the order of operations.

If a complex quantity is signed in an expression, the sign change must apply to *both* real and imaginary parts.

If: $C_X = R_C + jI_C$ Then $-(C_X) = -R_C - jI_C$ -or -

If: $C_X = R_C - jI_C$ Then $-(C_X) = -R_C + jI_C$

Multiplication and Division of Complex Quantities

Multiplication and division of complex quantities is easiest to do in polar form. In polar form, the Magnitudes are multiplied or divided, but the Angles are added or subtracted. Those rules are illustrated by:

If: $C_X = M_C \angle A_C$ and $D_X = M_D \angle A_D$

Multiply: $C_X \cdot D_X = (M_C \cdot M_D) \angle (A_C + A_D)$

Divide: $\dfrac{C_X}{D_X} = \left(\dfrac{M_C}{M_D}\right) \angle (A_C - A_D)$

Multiplying or dividing a Polar form complex quantity by a scalar quantity K must consider K the same as a complex having a magnitude equal to its scalar value and an angle equal to zero.

If: $C_X = M_C \angle A_C$ and $K = (K) \angle 0$

$C_X \cdot K = (M_C \cdot K) \angle (A_C + 0) = (M_C \cdot K) \angle A_C$

$\dfrac{C_X}{K} = \left(\dfrac{M_C}{K}\right) \angle (A_C - 0) = \left(\dfrac{M_C}{K}\right) \angle A_C$

If the complex quantity is multiplied or divided by a scalar, there is no change in the Angle sign. However, if the scalar K is divided by complex C_X, then the Angle sign changes:

$\dfrac{K}{C_X} = \left(\dfrac{K}{M_C}\right) \angle (0 - A_C) = \left(\dfrac{K}{M_C}\right) \angle -A_C$

Inversion of a complex quantity is that same as dividing a complex quantity into a scalar unity:

$\dfrac{1}{C_X} = \left(\dfrac{1}{M_C}\right) \angle -A_C$

The reciprocal of the Magnitude is taken and the Angle sign reverses.

Multiplying and dividing with two complex quantities in rectangular form is more complicated:

If: $C_X = R_C + jI_C$ and $D_X = R_D + jI_D$

Multiply: $C_X \cdot D_X = (R_C R_D - I_C I_D) + j(I_C R_D + I_D R_C)$

Divide: $\dfrac{C_X}{D_X} = \left(\dfrac{R_C R_D + I_C I_D}{R_D^2 + I_D^2}\right) + j\left(\dfrac{I_C R_D - I_D R_C}{R_D^2 + I_D^2}\right)$

Invert: $\dfrac{1}{D_X} = \left(\dfrac{R_D}{R_D^2 + I_D^2}\right) - j\left(\dfrac{I_D}{R_D^2 + I_D^2}\right)$

Inverting a complex quantity D_X in Rectangular form can assume that $C_X = 1 + j\ 0$ and substitute unity for R_C, zero for I_C.

Multiplying a rectangular form complex quantity D by a scalar K is a relatively easy process:

$$K \cdot D_X = (K \cdot R_C) + j\,(K \cdot I_C)$$

Note that D is written as if it were scalar but it has been *defined* as complex. Similarly, if D is divided by scalar K then it would be:

$$\frac{D_X}{K} = \left(\frac{R_D}{K}\right) + j\left(\frac{I_D}{K}\right)$$

In both multiplication and division of a complex quantity by a scalar, the sign of the scalar would apply to each part of the rectangular form complex result.

Dividing a scalar by a complex quantity would be the same as multiplying the inverse of the complex by the scalar:

$$\frac{K}{D_X} = \left(\frac{K \cdot R_D}{R_D{}^2 + I_D{}^2}\right) - j\left(\frac{K \cdot I_D}{R_D{}^2 + I_D{}^2}\right)$$

Some Other Operations with Complex Quantities

The square of a complex quantity D_X in both rectangular and polar forms is:

$$D_X{}^2 = \left(R_D{}^2 - I_D{}^2\right) + j\,(2\,R_D\,I_D) = M_D{}^2 \angle 2A_D$$

To raise a polar form complex quantity to the Nth power, the magnitude is raised to the Nth power while the angle is multiplied by N:

$$C_X{}^N = M_C{}^N \angle (N \cdot A_C)$$

A square root takes the root of the Magnitude while the Angle is divided by 2. That and Nth roots of complex quantity D_X are:

$$\sqrt{D_X} = \sqrt{M_D} \angle \left(\frac{A_D}{2}\right) \qquad \sqrt[N]{D_X} = \sqrt[N]{M_D} \angle \left(\frac{A_D}{N}\right)$$

Conjugates of Complex Quantities

A *conjugate* of a complex quantity is another complex quantity whose Real parts are equal and whose Imaginary parts have the same value but are of opposite signs. The phrase ***conjugate match*** may have been seen in regards to impedance or admittance matching of a source to its load. A conjugate match occurs when the Imaginary part of the load is equal to, but of opposite sign to the Imaginary part of the source and the Real parts of both have the same value. The effect of a conjugate match is to reduce the Imaginary parts to zero, leaving only the equal Real parts, thus insuring a maximum transfer of power from source to load.

If: $C_X = 4 + j\,3$ and $D_X = 4 - j\,3$

Then: C_X is the **complex conjugate** of D_X

This is an important feature of complex quantities that permits many solutions for circuit design in regards to *impedance matching* and *admittance matching*..

Radian Angle Handling

Radian angle measure is based on the value of Pi, specifically 2 Pi radians is equal to a full circle. A one-radian angle is approximately equal to about 57 degrees of common angle measure. Most high-level computer program languages require angle calculation in radians rather than degrees. Pocket scientific calculators have built-in degree-to-radian and radian-to-degree conversion. Programmers must supply the constants to convert input degree data to radians, then convert the radian angle result back to degrees for output after calculation. See the Appendices of this chapter for these and other useful conversion constants.

Important Intrinsic Functions of Scientific Pocket Calculators

Polar-to-Rectangular and *Rectangular-to-Polar* conversion may be the most important intrinsic function of a pocket calculator used in radio design.[15] While addition and subtraction is straight-forward in real and imaginary parts, conversion is required if the data is available only in polar form. Similarly, rectangular form data must be converted to polar form first for convenient multiplication and division. Proper handling of impedance and admittance data requires that conversion funtion.

A built-in constant of Pi is very convenient, especially in calculating *radian frequency* (2 Pi f) that occurs in many formulas. A degree to radian angle conversion is handy but not an absolute necessity. To convert degrees to radians, multiply by Pi and divide by 180.

An intrinsic square-root function is almost a necessity; square-roots also occur often in formulas. Programmability is convenient for frequently-used formulas. A down side is that no calculators have a common programming language between manufacturers. Each programmable calculator requires a short learning period.

Root Mean Square or *RMS*

[15] The later Hewlett-Packard scientific pocket calculators, such as the Model 32S II and 35S, handle complex quantities as a number-pair and can do the four functions of addition, subtraction, multiplication and division directly as a pair of pairs. This relieves users of separate part handling.

This refers to the ***square root of the sums of the squares*** of samples of voltages or currents measured within equidistant time intervals of some period. Mathematically it is shown in Figure 5-7 and Chapter 5 has a more complete description.

All waveforms are not nice, neat sinewaves. In particular, digital signals can have varying *on* and *off* times. The peak values may be constant but most simple measurement devices will show variation of voltage or current due to simpler circuitry within an electronic meter. This was solved by a *thermoelectric* technique converting the varying waveshape to ***heat*** and then measuring the heat rise relative to ambient temperature.

Personal Computer Programs

Personal computers having high-level languages can be a great aid in repetitive calculation. BASIC and FORTRAN have a syntax related to mathematical formulas. Both languages allow program storage and, with some experience, "reading" of the formulas being calculated. The various dialects of BASIC are most suited for casual computing since they are *interpreters* and can be run immediately to verify that a program is running correctly.[16]

Since Pi and degree-to-radian conversion multipliers occur often, computer program variables used as constants can be *declared* at program start. Most FORTRAN requires all angle data to be in radians so the following statement can declare variables TWOPI or ONEPI for use as constants:

```
TWOPI = 8.0 * ATAN(1.0)

ONEPI = 4.0 * ATAN(1.0)
```

Note: Be sure to follow the language convention on all floating-point variables.

Programs written in FORTRAN or BASIC are directly readable as equations. Later high-level languages have different program syntax and not easily read as such. Most FORTRAN compilers support complex quantities declared as a single variable name and can handle statements of complex quantity operations directly. Source code is converted into the correct mathematical operations. Other languages require statements of separate real and imaginary or magnitude and angle part operations.

---

[16] An *interpreter* does not have to be *compiled* separately (as with FORTRAN or C), can run from written program source code. As a consequence a BASIC interpreter program runs slower. "Slow" is relative considering that most personal computers have clock rates higher than 100 MHz.

Appendix 2-1

Mathematical Constants

π = 3.1415 92353 58979 32385 2π = 6.2831 85307 17958 64789

$\frac{1}{\pi}$ = 0.3183 09886 18379 06715 $\frac{1}{2\pi}$ = 0.1591 54943 09189 53358

π^2 = 9.8696 04401 08935 86188

$\frac{1}{\pi^2}$ = 0.1013 21183 64233 77715 $\frac{1}{4\pi^2}$ = $2.5330\ 29591\ 05844\ 42861 \bullet 10^{-2}$.

1 radian = 57.295 77951 30823 20877 degrees = $\frac{180}{\pi}$

1° = 0.0174 53292 51994 32957 69 radians

e = 2.7182 81828 45904 52354 (base of natural logarithms)

Log_{10} *e* = 0.43429 44819 03251 82765 **Ln 10** = 2.3025 85092 99404 56840

Log_{10} 2 = 0.30102 99956 63981 19521 **Ln 2** = 0.69314 71805 59945 30942

Log_2 *e* = 1.6514 96129 47231 87981 **Log_2 10** = 3.3219 28094 88736 23478

Physical Constants

Speed of Light in a Vacuum

299,792,458 m / sec (international standard) 299,792.458 km / sec

983,571,056 ft / sec 186,282.397 mi / sec (statute miles)

161,874.977 mi / sec (nautical miles) [1 nautical mile = 1853.000 meters]

Boltzmann Constant = 1.38054×10^{-23} joules / °K

Conversion Factors

Celsius	Fahrenheit	Kelvin
°C = (°F - 32) / 1.8	**°F =** 1.8 °C + 32	**°K** = °C + 273.15

1 inch (in) = 2.54 centimeters (cm) [exact] 1 cm = 0.39970 07874 in.

1 mil (0.001 in) = 0.0254 mm 1 mm = 39.970079 mils

1 foot (ft) = 0.3048 meters (m) = 304.8 cm = 12 in 1 m = 3.2808 39895 ft

1 micron (μ) = 1 • 10^{-6} m 1 nanometer (nm) = 1 • 10^{-9} m 1 ångstrom = 0.1 nm

1 statute mile = 5280 ft = 1609.344 m 1 nautical mile = 6076.11549 ft = 1852 m

1 gallon (gal) [liquid] = 3.785512 liter = 0.003785412 cubic meter (m^3) = 4 quarts = 8 pints

1 quart (qt) [liquid] = 0.9463529 liter (L) 1 liter [liquid] = 1.056688 qt

1 pound (lb) = 0.4535924 kilogram (kg) 1 kg = 2.2046225 lb = 35.27397 oz

1 foot pound-force = 1.355818 joules (J)

1 PSI (pound-force/in^2) = 6894.757 pascal (Pa)

1 Horsepower(electric) = 1 HP_e = 746 Watts 1 Watt hour (W • h) = 3600 joules

1 Ampere hour (A • h) = 3600 coulombs (C) 1 Watt second (W • s) = 1 joule

1 standard atmosphere = 101,325 pascal (Pa) ≈ 29.92 inches of mercury (inH_g)

1 footcandle = 10.76391 lux (lx) 1 footlambert = 3.426259 candela / sq.meter (cd/m^2)

Appendix 2-2

Greek Letters and Common Use in Formulas

Common use refers to repeated convention of using Greek letters to denote certain quantities. This is not absolute (no definite standard) but has been found common by the author.

Α	α	Alpha, phase angle in lower case	Ν	η	Nu
Β	β	Beta, phase angle in lower case	Ξ	ξ	Xi
Γ	γ	Gamma	Ο	ο	Omicron
Δ	δ	Delta, variation in upper case	Π	π	Pi, mathematical constant
Ε	ε	Epsilon, dielectric constant in lower case	Ρ	ρ	Rho
Ζ	ζ	Zeta	Σ	σ	Sigma, summation upper case
Η	η	Eta	Τ	τ	Tau, time
Θ	θ	Theta, angle in lower case	Υ	υ	Upsilon
Ι	ι	Iota	Φ	φ	Phi, angle in lower case
Κ	κ	Kappa	Χ	χ	Chi
Λ	λ	Lambda, wavelength in lower case	Ψ	ψ	Psi
Μ	μ	Mu, permeability in lower case	Ω	ω	Omega, Ohms in upper case radian frequency in lower case

Notes: Radian Frequency $\omega = 2\pi f$ where f is frequency and π is the mathematical constant

Ω is sometimes found elsewhere on schematics and there refers to Ohms.

Lower case Greek letters are generally used to avoid conflict with English upper case.

References for Chapter 2

[1] “Handbook of Mathematical Tables and Formulas,” by Richard Stevens Burrington, Ph.D, Handbook Publishers, Inc., Sandusky, Ohio, Third Edition 1953, is the source for all basic mathematical operations in this chapter, including the definition of logarithms. Any mathematics handbook covering algebra, trigonometry, and transcendental functions will have the same information. The presentation of basic information has been re-arranged to show the essentials for the hobbyist designer.

[2] “Guide for the Use of the International System of Units (SI),” by Barry N. Taylor, NIST Special Publication 811, 1995 Edition, was used for the value prefix multiplier values and names plus the conversion constants in Appendix 2. NIST is the National Institute of Standards and Technology of the United States Department of Commerce. This publication was obtained free over the Internet in February, 2000, but could be purchased in hard-copy form from the United States Government Printing Office (US GPO).

[3] “Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables,” edited by Milton Abramowitz and Irene A. Stegun, National Bureau of Standards Applied Mathematics Series 55, June 1964, Chapters 1 through 4, was the source for the remaining mathematical constants and conversion values plus a check on the basic mathematical operations. The US National Bureau of Standards became a part of NIST after this handbook edition was published.

Chapter 3

Waveforms, Heterodynes, Mixing

Waveforms are important in all disciplines of electronics and these are explained insofar as elementary waveshapes and their harmonic content, ways to calculate those harmonics. Heterodyning is introduced and those lead to mixers, a basic component of frequency control/tuning systems.

Elementary Waveforms

A Pure Sinewave

The term *sinewave* denotes an AC or RF waveform that has **no** harmonics. Voltage or current amplitude over the *period* of a sinewave follows the mathematical sine function, zero amplitude at start, maximum positive amplitude at 90 degrees, zero amplitude at 180 degrees, maximum negative amplitude at 270 degrees, finally zero again at 360 degrees.

The amplitude over time can also be described by a mathematical *cosine* function but that requires the amplitude to be maximum at the period start. The cosine waveform will ***lead*** a sine waveform of the same frequency by 90 degrees. Conversely, a sine waveform will ***lag*** a cosine waveform by 90 degrees.

For general convenience, the sine function can describe the waveform having no harmonics. Such a non-harmonic-content signal is referred to simply as a *sinewave*. A signal having a sinewave appearance is also called a *sinusoid.*

Analysis of Repetitive Waveforms

A truly pure sinewave exists only as a mathematical abstraction. **All** waveforms contain some harmonics of the fundamental frequency, regardless of how pure they appear to be. The inverted time of the period of one cycle of a sinewave is its ***fundamental frequency,*** sometimes referred to as just ***fundamental. Harmonics*** are multiples of the fundamental frequency. The *second harmonic* is twice the fundamental frequency, the *third harmonic* is triple the fundamental frequency, and so forth. Any deviation from the pure sinusoidal shape will have some harmonic content.[1]

The total content of fundamental and harmonics can be shown on a frequency spectrum chart with frequency the horizontal axis, and amplitude the vertical axis. The amplitude can be expressed in any form: Power, voltage, or current. Voltage relative to the fundamental amplitude is the usual representation, often in terms of *decibels relative to fundamental amplitude.* Such a representation is called the *spectral content* of any waveform.

---

[1] It is possible to generate sinewaves with extremely small harmonic content, below 1% of the power of the fundamental, but there is always some amount of harmonics in it.

What was vexing to all in very, very early days of electronics was a way of expressing the spectral content of any waveform. By assuming the waveform was *continuously repetitive*, this could be solved by using a ***Fourier transform.***[2] That mathematical technique converted equal-time amplitudes of a repetitive waveform into equal-frequency amplitudes, resulting in a set of *Fourier coefficients* that could describe the harmonic frequency content. That transform assumed that each resulting harmonic was itself a sinewave.

It should be understood from the beginning that a Fourier transform is based on a ***repetitive waveform***, repetitive in both time and amplitude for a very long time. Obviously, if the waveform ceases, there is nothing to transform, and if the waveform changes amplitude the repetition also changes and the elementary transform will not handle that.[3] The coefficients of a Fourier transform carry with it the relative amplitude and phase, referred to the fundamental frequency. Zero phase of the fundamental is considered the starting point, the beginning of a period of the fundamental frequency.

There is also a Fourier transform of the spectral content from frequency back to time that will, in effect, reconstruct the original waveform solely from frequency-domain spectral content. It is possible to limit the number of harmonics of the frequency spectral content and reconstruct a repetitive waveform from that. How faithfully the waveform is reconstructed relative to its original wave shape will determine the ***necessary bandwidth*** in the frequency-domain to communicate or transfer the waveform to another place through a bandwidth-limited medium such as radio.[4] The following table shows energy distribution over the spectrum for several waveforms:

Table 3-1 Comparison of 9 Waveforms' Energy Content

Waveform Type	Percentage Energy in Fundamental	Percentage Energy in all Higher Harmonics
Pure Sine Wave	100	0
20-Segment Sinusoid Approximation	99.4	0.6
10-Segment Sinusoid Approximation	98.5	1.5
Square Wave, 25% rise, fall times	81.1	18.9
Square Wave, 10% rise, fall times	70.0	30.0
Square Wave, 4% rise, fall times	15.9	84.1
Square Wave, 2% rise, fall times	8.0	92.0
Square Wave, 1% rise, fall times	4.0	96.0
Half-sine pulse, 20-segment approximation	32	68

The *segment sinusoid approximation* was generated by using a limited number of sample amplitudes

[2] It is not necessary for the hobbyist to know Fourier Analysis or its transforms, only that such mathematical tools exist. For language purists, it is a French surname with the approximate English pronunciation of *"foo-re-ay."*

[3] There are extensions of the basic Fourier transforms that will work with continuously-changing waveforms. One example seen by many is the more-expensive home music system showing the changing amplitude content of music and speech on an LED or LCD bargraph display.

[4] This is not the sole determination of bandwidth of a radio or other electronic medium. In radio it is common to assume the necessary bandwidth is the same as what is transmitted. However, in a radio circuit the receiver may deliberately limit the bandwidth and that may or may not change the content of the received signal.

along one period of a pure sinewave; had more, shorter-time samples been used the fundamental energy content would approach 100% and the total harmonic energy content approach zero. The square waves used rise times relative to the period of the square wave. It shows that faster rise-time, faster fall-time square waves need a large bandwidth than slower rise- and fall-times. If the frequency-to-time reconstruct of the spectral content of fast transition harmonics, the rise and fall times would have become slower and distorted..

A question can now be asked, what happens to a reproduced waveform if the number of harmonics is limited or distorted? The fewer the higher harmonics, the slower the rise and fall times. A great reduction of higher harmonics can result in a slope at the top and decay at the bottom, *ringing* (appearance of a superimposed, higher-frequency sine wave at transitions) and *overshoot* (ringing exceeding the amplitude of the top). As the harmonics are limited more and more towards the fundamental, the waveform approaches a sinusoidal shape with lesser and lesser peak amplitude (loss of energy contained in higher harmonics). A very rough value of bandwith necessary for a waveform is (3/t) where t is the fastest transition time of the waveform such as rise or fall time of a pulse. If a pulse has a 0.1 μs rise time then the rough transmission bandwidth would be about 30 MHz.

Fourier Coefficients of Basic Waveforms

Repetitive waveforms' harmonics can be calculated in both amplitude and phase (relative to the fundamental) through Fourier coefficient formulas using the three basic waveshapes shown in Figure 3-1. The coefficients will be complex quantities.

Any arbitrary-shape waveform can be described by ***pieces*** of its shape, the piece lengths or *segments* being parts of its repetition time. Combinations of two of the three basic waveshapes can make up one piece of time of the arbitrary waveform to be analyzed. The coefficients of each waveshape can be added for the one piece of the *piecewise approximation* of one part of the waveform. But, the pieces must be *time-shifted* relative to the rectangular waveform for making up segments of the original wave shape.

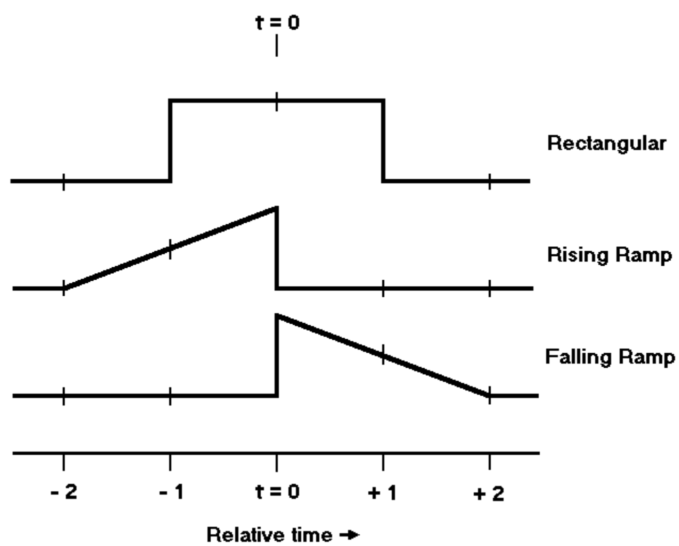

Figure 3-1 Comparison of the three basic waveshapes used in synthesizing segments of repetitive waveforms. Note alignment at time = zero. All are assumed to have zero rise and fall times.

All coefficients are calculated over a preselected number of harmonics. From the sheer amount of mathematical calculations involved, this is only practical on a personal computer. What follows can be adapted to such a personal computer program using the formulas described.

For a Rectangular Waveshape

$$C_N = \left(\frac{2\ A_S}{\pi\ n}\right)\left[\mathrm{Sin}\left(\frac{\omega_n\ t_W}{2}\right) - j\,0\right] \qquad (3-1)$$

Where: C_N = Complex coefficient at harmonic n

A_S = Peak amplitude or rectangular waveshape

t_W = Width of rectangle, seconds

ω_N = Radian frequency of harmonic n $= \dfrac{2\,\pi\ n}{\mathrm{Period}_{SECONDS}}$

Note that each coefficient is a complex quantity but the imaginary part of the rectangular coefficient is always zero. Time and frequency are seconds and Hertz, respectively. Amplitude **A** may be in Volts or Amperes. Rectangle width equals the segment time. The Period is the repetition time of the waveform to be approximated.

For a quick calculation of a perfect square wave using sine arguments in degrees:

$$C_N = \left(\frac{2\ A_S}{p\ n}\right)\left[\mathrm{Sin}\,(90\ n) + j\,0\right] \qquad (3-2)$$

One thing not apparent is that the C_N polarity is positive for n of 1, 5, 9, 13, 17......but negative for n of 3, 7, 11, 15, 19 and so on. Since these are complex quantities, a negative real part with zero imaginary part is the same as a phase reversal relative to a positive real part with zero imaginary part.

A perfect square wave will not have any even harmonics. If n is 2, then the real part in brackets would require a sine of 180° which is zero. In the practical world square waves have finite rise-times and fall-times, limited by the generator of them and the transmission medium.

Start of Rising Ramp and Falling Ramp Waveshapes

For a rising ramp waveshape:

$$C_n = \left(\frac{A_R}{p\ n\,w_n\ t_R}\right)\left\{\left[1 - \mathrm{Cos}\,(w_n\ t_R)\right] + j\left[(w_n\,t_R) - \mathrm{Sin}(w_n\ t_R)\right]\right\} \qquad (3-3)$$

Where: C_n = Coefficient of rising ramp waveshape at harmonic n

A_R = Peak amplitude of rising ramp waveshape

t_R = Time of rising ramp, seconds

w_n = Radian frequency of harmonic $= \dfrac{2\,p\ n}{\mathrm{Period}_{seconds}}$

For equations (3-1), (3-3) through (3-5) all angles are in radians, all times in seconds. Note the term $(\omega_n\,t_R)$ which occurs four times in equation (3-3). That product can be calculated once for each

n, equated to a variable, then the variable used in four places for the final calculation statement. Peak amplitude refers only to the rising ramp waveshape. Again, the Period is the repetition time of the waveform to be approximated. For a falling ramp waveshape:

$$C_N = \left(\frac{A_F}{\pi\, n\, \omega_N\, t_F}\right)\left\{\left[\,1 - \mathrm{Cos}\,(\omega_N\, t_F)\,\right] + j\left[\,\mathrm{Sin}\,(\omega_N\, t_F) - (\omega_N\, t_F)\,\right]\right\} \qquad (3\text{-}4)$$

Where: C_N = Coefficient of falling ramp waveshape at harmonic n

A_F = Peak amplitude of falling ramp waveshape

t_F = Length of time of falling ramp waveshape, seconds

ω_N = Radian frequency of harmonic = $\dfrac{2\,\pi\, n}{\mathrm{Period}_{\mathrm{SECONDS}}}$

Note the two term groups in the imaginary part. Those are essentially the same as the imaginary part term groups except polarity is reversed for a rising ramp waveshape. This can be used to advantage in computer program writing where a temporary variable can hold common data.

Alignment of Basic Waveshapes to Create a Segment

The basic waveshapes in Figure 3-1 all have different *time-zero* positions. In order to calculate coefficients accurately, the rectangular shape must be delayed by half of a segment time while the falling ramp must be delayed by a whole segment time. Since the falling ramp time-zero is at the beginning of its shape, it is already positioned equivalent to the start of a segment. A ***delay multiplier*** must be created [equation (3-5)] to modify the rectangular shape and the rising ramp shape.

All of the waveshape coefficient formulas assume a *positive going* waveform with a DC baseline of zero. This is not a problem for bandwidth analyses since the coefficient magnitudes and phase angles of the fundamental and all harmonics will be the same regardless of the DC baseline. For the purposes of finding coefficients, the peak negative swing can be taken as zero voltage and all other amplitudes as excursions from that value. That will make all **A** amplitudes positive.

$$D_N = \mathrm{Cos}\left(\frac{2\,p\,n\,t_D}{t_P}\right) - j\,\mathrm{Sin}\left(\frac{2\,p\,n\,t_D}{t_P}\right) \qquad (3\text{-}5)$$

Where: D_N = Delay multiplier, a complex number quantity

n = Harmonic number

t_D = Time delay required, seconds

t_P = Period of approximated waveform, seconds

The two ramp formulas have similar overall multipliers and real parts; but the imaginary parts have reversed polarity. A plot of both ramps' harmonic magnitudes will be the same (with equal rise and fall times) but the phase angle reversal accounts for the shape difference on a reverse transform. Again, as with the square wave, only odd harmonics have a real part. Unlike square waves, the even harmonics do have a magnitude but only from non-zero imaginary parts.

The start and stop amplitudes of a particular straight-line segment piece will determine which wave-shape is used for a particular segment. The six possible combinations are shown in Figure 3-2. Segment amplitude entries should begin with zero as the first start point as a convenience. The first segment will then have a choice of no amplitude for the stop time (no waveshape) or a rising ramp.

A segment's stop amplitude will always become the start amplitude of the next segment. Successive segments' stop time amplitude must be compared to the start time amplitude in order to determine the proper waveshape selection.

Figure 3-2 shows voltage entry for start and stop amplitudes. If voltage entries are done, a reconstruction of the approximated waveform will have voltage output. Current could also be entered with a reconstruction being done with current output.

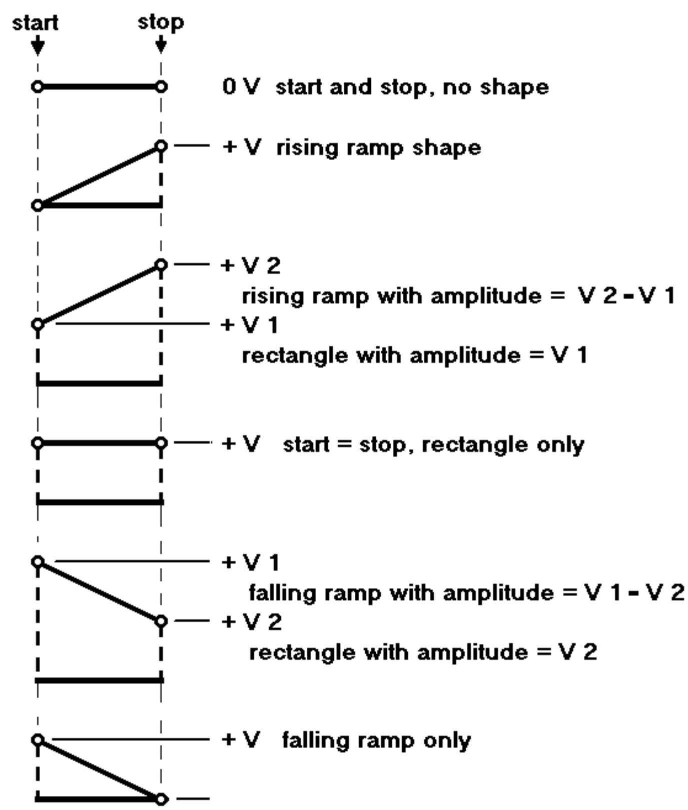

Figure 3-2 The six possible combinations of basic wave-shapes to create a piece or segment of an arbitrary-shape repetitive waveform. All are shown for positive-going segment voltages.

Amplitude entry requires a decision structure in a computer program to select one or two of the three basic waveshapes. Figure 3-4 can be used to organize the decision structure. A set of amplitude entries will also have to incorporate start and stop times in order to determine the waveshape's time length. The stop time of a preceding segment becomes the start time of the next segment.

The Delay Multiplier's Use

To make each segment, it must be delayed in time. For each segment's coefficients:

$$C_{DN} = C_N \cdot D_N \quad \text{[as complex number quantities]} \qquad (3\text{-}6)$$

where: C_N is the Fourier coefficient at harmonic n

D_N is the delay multiplier at harmonic n

Depending on who does the calculating or computer program writing, the polar form of complex number quantities might be more convenient to use. In polar form, the delay multiplier becomes the following:

$$D_n = 1 \angle \left(\frac{2\ p\ n\ t_d}{t_p} \right) \qquad \text{[radian angle]} \qquad (3-7)$$

All a polar form multiplier does is to change the angle. Recall that polar form multiplication arithmetically multiplies magnitudes but adds angles. Since the delay multiplier magnitude is unity, a multiplied coefficient's magnitude is not changed in value.

Setting Up Coefficient Calculation of a Piece-wise Waveform Approximation

Calculation involves so many variables that it must be done in a computer program. Several items of information are required by such a program before starting. The first is the number of harmonics desired. That will drive requirements for internal memory storage. Typical values for accuracy of reconstruction would be 200 to 500. Using *double precision* variables (8 Bytes or 64 Bits each variable) a complex coefficient array would need 8 x 2 x 502 = 8032 Bytes for 500 harmonics plus fundamental plus the DC component.[5] The *DC Component* is optional and represents the average DC baseline of a single waveform period.

The next item is the segment time interval. Manual entries will be in amplitude at some time within the approximated waveform's period for each segment. By storing the previous segment's stop time, that can be used as the start time for the next segment. Entries can be done at any time position within a period. Identical time increments are neat and require the least overhead tasks in the program but may be inconvenient for irregular waveforms or those with relatively long equal-amplitude sections.

A suggestion on the main portion of a program to handle amplitude entries is to assume the approximated waveform's starting amplitude and ending amplitude is zero. The first amplitude entry would correspond to a time position after the waveform start; the first segment's start time position would be zero. The next segment's start position would use the previous segment's stop position and the amplitude entry would be at that segment's stop time. The last segment's stop position would equal or be extremely close to the end of the waveform period. That last segment amplitude entry would be zero to correspond with the waveform's period start amplitude.

The time-and-amplitude entry routine must include a decision tree to compare the previous segment stop amplitude with the current segment stop amplitude. See Figure 3-4. That will determine which waveshapes are included in the current segment. This requires temporary storage of the previous segment amplitude.

Waveform Approximation DC Baseline Calculation

Amplitude entries for the coefficient formulas can be assumed to be positive-going. If a non-zero DC baseline is needed, that can be stored in a single temporary variable. Positive and negative-going amplitude entries of each segment can add the negative of that DC baseline to each entry for the coefficient calculations; addition of the negative baseline value will make the coefficients' **A** values positive-going.

The DC value is still required for basic waveshapes and those are accumulated in a temporary

---

[5] Double precision or "Real 8" has an accuracy of 15 decimal digits in most high-level computer languages.

variable. That DC value is simply the average of the area of a waveshape relative to the entire waveform period. For the rectangular waveshape it is:

$$E_{DC} = A_S \left(\frac{t_W}{t_P} \right) \qquad (3\text{-}8)$$

Where:

A_S = Peak amplitude of rectangular waveshape

t_W = Width of rectangle in seconds

t_P = Period time, seconds

Note: The **A** amplitude is only for the rectangle waveshape. If the rectangle and a ramp are combined for a segment, different **A** values are used with (3-9) and (3-10) but the DC values from each are added for the segment's DC value.

For the ramp waveshapes they are:

$$E_{DC} = A_R \left(\frac{t_R}{2\ t_P} \right) = A_F \left(\frac{t_F}{2\ t_P} \right) \qquad (3\text{-}9)$$

Where: A_R = Peak amplitude of rising ramp waveshape

A_F = Peak amplitude of falling ramp waveshape

t_P = Approximated waveform period, seconds

t_F = Time length of falling ramp, seconds

t_R = Time length of rising ramp, seconds

Note: The **A** amplitudes are only for the ramp waveshapes. If a segment requires both a rectangular and ramp waveshape the two DC values are added.

The total of all segments' DC values are added or *accumulated* in a temporary variable. That variable is not affected by any delay multiplier.

Array Storage Required, Choice of Complex or Scalar Variables

Coefficient calculation requires only one complex variable *array*.[6] The complex array is required for accumulation of waveshape coefficients. It is the property of Fourier coefficients that individual time-section coefficients can all be added to form the final waveform approximation's coefficient set.

Most dialects of FORTRAN or BASIC allow complex variables to exist as an internally-ordered pair of scalar variables. The language allows statements of direct arithmetic, handling all the rules of complex arithmetic internally. A few BASIC language dialects allow complex variables or statements. All arithmetic statements can be expressed as separate real and imaginary part

---

[6] An *array* is a group of contiguous variables in memory *declared* to be such at the beginning of a program. Individual variables in an array are selected by their *subscript* numbers in parentheses following the name of the array variable. See the individual language manual for syntax and available functions..

variables, obeying the rules of complex quantities given in Chapter 2. In that case, there will be a need of four variable arrays declared but the memory storage will be about the same as with two complex arrays.

What to Do With a Coefficient Set After Calculation is Finished

Fourier coefficients of a repetitive waveform represent the magnitude and phase angle of all harmonics relative to the fundamental frequency that exist in the frequency domain. The obvious application would be to use the coefficients as a voltage or current source for a circuit. Since the circuit will probably alter magnitudes and phase angles, the output of the circuit could be reconstructed to show the effect of the circuit on a waveform.

Examination of the magnitudes of harmonics will give some indication of the bandwidth necessary to retain waveform fidelity. For a true picture of the effect of changing phase angles of the harmonics, a reconstruction of the waveform is required.

Reconstruction of the Approximated Waveform from Fourier Coefficients

Reconstruction involves calculation and summation of all waveform Fourier coefficients for each point of time desired.[7] That is the reverse Fourier transform, from frequency domain to time domain. For coefficients having amplitudes in voltage it is:

$$e(t) = E_{DC} + \sum_{n=1}^{n=L} M_n \left[\text{Cos}\left(w_n t + q_n \right) \right] \qquad (3\text{-}10)$$

Where:
$e(t)$ = Instantaneous voltage e at time t, t in seconds
E_{DC} = DC voltage of entire approximated waveform
L = Total number of harmonics used in coefficients
M_n = Magnitude of coefficient at harmonic n
q_n = Phase angle of coefficient at harmonic n, radians [Cosine argument
w_n = Radian frequency of harmonic n = $\dfrac{2\ p\ n}{t_p}$ is in Radians]
t_p = Waveform period in seconds

The strange-appearing "Σ" in (3-11) is an upper-case Greek letter Sigma denoting that all terms to the right are to be summed together between harmonics **1** (*n = 1 at* the Sigma undersign) up to a maximum of **L** (*n = L* at the Sigma oversign). The undersign and oversign mathematically signify *lower* and *upper* limits, respectively, of the summation. If there were 500 harmonics in the coefficient set, **L** would be set to 500. There would be 500 calculations of the magnitude times the cosine at each harmonic, accumulating each one, finally adding the DC baseline voltage. The result would be the reconstructed waveform's voltage at the selected time position, **t**.

For the next time position (a new **t**) the summation would have to be repeated all over again.

---

[7] Many points must be calculated in order to form a graphical picture of the reconstruction, itself becoming a piecewise approximation of "connecting the dots."

The reconstruction routine could be two DO loops. The inner loop would perform the (3-11) calculation at time position **t** while the outer loop would increment **t** and store each **e(t)** in a scalar array for later display purposes.

Amplitudes of Fourier coefficients and the reconstruction can also use current. In that case the **e(t)** becomes **i(t)** or current as a function of time. The use of lower-case letters for voltage or current is common notation to separate instantaneous values who are functions of another variable rather than steady-state voltage or current. Steady-state conditions are noted by upper-case characters.

Notation of variable **e** with an immediate following variable name in parenthesis (the **(t)** in this case) means that **e** ***is a function of*** **t**. Such notation is found on a few other calculations in following chapters. Without such short forms of notation (3-10) would have to be written as:

$$e(t) = E_{DC} + \left[M_1 \operatorname{Cos}(\omega_1 t + \theta_1)\right] + \left[M_2 \operatorname{Cos}(\omega_2 t + \theta_2)\right] +$$
$$\left[M_3 \operatorname{Cos}(\omega_3 t + \theta_3)\right] + \left[M_4 \operatorname{Cos}(\omega_4 t + \theta_4)\right] +$$
$$\cdots$$
$$+ \left[M_{L-1} \operatorname{Cos}(\omega_{L-1} t + \theta_{L-1})\right] + \left[M_L \operatorname{Cos}(\omega_L t + \theta_L)\right]$$

Gibbs Phenomenon and Hamming Weighting

Square pulses that are transformed into Fourier coefficients then reconstructed again will be perfect in their reconstructed form if and only if the number of harmonics are a very high value. Limiting the number of harmonics will show a "ringing" effect on the edges of the wave shapes with certain maximum harmonic limits. This ringing or "spiking" is called the *Gibbs Phenomenon* but it is only a mathematical condition due to a finite number of harmonics. A way to avoid the appearance of this phenomenon is ***weighting***, a modification of the coefficient amplitudes or phases. One such is *Hamming Weighting* or *Hamming Windowing*.[8] To apply such weighting, multiply each magnitude (except the DC component) of a Fourier coefficient by:

$$H_{WN} = 0.54 + 0.46 \operatorname{Cos}\left(\frac{n\ \pi}{L}\right) \qquad \text{Where:} \qquad (3-12)$$

n = Harmonic number [Cosine argument in radians]

L = Maximum number of harmonics

The Hamming Weight result of (3-12) is a multiplier value between 1.0 at the first harmonic and near-zero at the highest harmonic. The result of such weighting would make the system handling the waveform appear to have a low-pass filter with a long, slow cut-off. The effect on a sharp-transition pulse is to cause a very slight rounding of the transition corners. Use of such weighting is preferrable in circuit analysis programs having a few (under 200) harmonics. It will remove the Gibbs phenomenon which might cause the user to think the circuit itself caused the ringing.

Weighting can also be applied to the entire Fourier coefficient set, such as in the complex coefficient variable array. Such Hamming Weighting can simulate the effects of finite bandwidths

---

[8] After Richard W. Hamming. The particular Hamming Weight or Window is given in an Internet free-distribution PC program, **Fourier2.EXE** written by Harry Garland. A well-done program, it illustrates waveform reconstruction in quick graphics at the DOS level.

in amplifiers and oscilloscopes.

Two Waveforms Having Less Harmonic Content

Two more elementary waveforms can be described, the *cosine* and the *cosine squared*. Their names correspond to the mathematical function describing their shape. The cosine pulse is the "half-sinusoid" waveform included in Table 3-1. The cosine squared waveform in amplitude resembles a full cycle of a sinusoid from negative peak to negative peak.

The important part of the cosine squared pulse is that its harmonic content is the least of all pulses. If the object is to transmit *bits* of information, a cosine-squared pulse is ideal for transmitter modulation. The RF output spectrum would be narrower than that required for rectangular waveforms. A relatively narrow receiving filter can be used without unduly distorting the cosine-squared waveform.

The cosine pulse has about 32% of the total waveform energy in the fundamental. The first ten *even* harmonics contain about 50% of the total waveform energy. The cosine-squared waveform has about 97% of the total energy in the fundamental; highest-amplitude harmonics are at least 60 db down from the fundamental amplitude.[9]

Creation of a cosine-squared waveform can begin with a rectangular pulse. Passing that pulse through a low-pass filter whose cutoff frequency is equal to the reciprocal of the width time will attenuate all harmonics from the 3rd on upward. Filter output will closely resemble the cosine-squared waveform. A *Gaussian* waveshape is very similar to a cosine-squared, having the shape of a *Gaussian distribution.*[10] In practical applications their shapes are nearly identical.

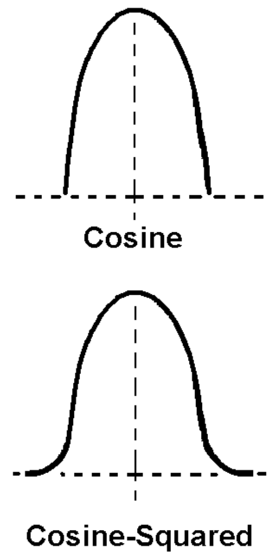

Figure 3-3 Two waveforms having few harmonics.

Comparison of Repetitive Versus Non-Repetitive Waveforms

It is mathematically possible to calculate the spectral components of non-repetitive waveforms. That is not worth the effort for most hobbyists. The waveform *shape* determines the spectral content and the bandwidth, content and bandwidth will be the same for non-repetitive and repetitive waveforms. Digital Signal Processing, or *DSP*, can tell more about an unknown waveform's content in the frequency domain. DSP is a subject by itself and there is an excellent book (at a reasonable price) available through the Internet on that subject.[11]

Use of Fourier transforms of repetitive signals to calculate necessary bandwidths is both a

---

[9] 20-segment approximations were used to confirm energy distribution.

[10] The *bell-shaped curve* is found in nearly every beginning book on statistics.. The actual waveshape looks prettier than Figure 3-3 but I think you get the idea.

[11] Those who wish to go further on DSP can find answers in *The Scientist and Engineer's Guide to Digital Signal Processing* by Steven W. Smith, PhD, California Technical Publishing, San Diego, CA 92150-2407. The ISBN is 0-0660176-7-6 for hardback version ($64 + S&H) or 0-9660176-6-8 for the free electronic version.available http://www.DSPguide.com on the Internet as of mid-2007. 630 pages, 500 illustrations.

calculation convenience and a real-world tool to see the spectral components. Spectrum analyzers can "tune in on" each and every harmonic component of a repetitive waveform. If a *Digital Sampling Oscilloscope* (DSO) is used, built-in time-to-frequency domain calculation functions on such instruments will show the same spectral components.[12]

Non-repetitive waveforms have a different *average* energy even though their peak amplitudes are the same as repetitive waveforms of the same shape. While the average energy may be important to the output stages of a transmitter, the peak waveform energy is more important insofar as intelligence to be transmitted.

Heterodynes

The definition of *heterodyne* means "to mix, combine, or join different powers." Over the years the term has been applied to many different things and phenomena. For example, two or more AM carriers present within a receiver's passband and differing slightly in frequency will produce an audio tone or squeal that many call a heterodyne. What is happening there is simply the detector or demodulator is producing the difference of the two steady-state carrier frequencies, no different than it is reconstructing the various RF components (such as sidebands of an AM signal) in comparison with that signals' carrier.

True *heterodyning* is found in a ***mixer circuit*** where an incoming signal is actual shifted in frequency, split and duplicated, moved in frequency to the sum or difference of the signal input relative to the frequency of a ***local oscillator*** injecting its power into the mixer.

Mixers and Mixing

A *mixer* is basically a switch operated by the *local oscillator* or ***LO***, switching at the rate of the LO frequency. The output of a mixer contains two *new* frequencies:

$$F_{OUT} = F_{LO} + F_{IN} \quad \text{and} \quad F_{OUT} = F_{LO} - F_{IN} \qquad (3\text{-}13)$$

------ Sum ----- --- Difference ---

where: F_{LO} is the frequency of the Local Oscillator
F_{IN} is the frequency of the Input RF
all frequencies are in the same units

The end result is that each frequency output is approximately at the same power level and a passive mixer always has a 6 db power loss relative to the input. Active mixers will have a gain defined on manufacturer's datasheets.

The mixer was a godsend in terms of frequency control, both receiving and transmitting. It made possible the *superheterodyne* receiver, the common configuration for most of the tunable receivers ever designed and made. It made possible the transmitters that could be manually tuned

---

[12] These exist at the beginning of the millennium but were still too expensive for most hobbyists. Such functions are built into higher-end DSOs (Digital Sampling Oscilloscopes) made by Tektronix and LeCroy. Audio range sprectrum analyzers using embedded microprocessors have been built into consumer-grade stereo music systems, commonly seen as the bar graph display of amplitude over sections of the audio spectrum.

to any frequency within its VFO[13] range and did not force them to use a different quartz crystal for each transmit frequency.[14]

In the case of *balanced mixers*, most of the passive ones or the active mixers involving double differential amplifiers such as the *Gilbert* topology, the RF input and LO injection frequencies might leak through to the output but those would be at least 30 db below their respective inputs. At the output port the sum or difference component would be selected by resonant circuits or filters; the unwanted output component would either be reflected back to the mixer or absorbed in the resonant circuit or filter. Most of the *unbalanced mixers* were the vacuum tube type, many designed to contain both the mixer and the Local Oscillator tube active devices. Unbalanced mixers would also pass their input and LO frequencies to the output port so proper selection of the wanted frequency components was a very important design task.[15]

Spurious Outputs from Mixers

Regardless of the purity of a sinewave from an LO, there are always some harmonics of its fundamental frequency present in the mixing process. Depending on the mixer type, those harmonics may be generated by unbalances or distortion in the mixer itself. The results in a new mathematical representation:

$$F_{OUT} = m\,F_{LO} \pm n\,F_{IN} \qquad \text{where:} \qquad (3-14)$$

F_{LO} is the LO frequency as in equation (3 - 13)

F_{IN} is the RF input frequency also as in (3 - 13)

m is the integer harmonic of F_{LO}, above zero

n is the integer harmonic of F_{IN}, above zero

Anything that is not the wanted frequency output of equation (3-13) is termed ***spurious***, or more colloquially, just a ***spur.***[16]

Spurious mixing products can be either inside the wanted RF output range or outside it; if outside it is usually dismissed in design. Spurious mixer outputs are responsible for some of the false continuous wave non-signals commonly called *birdies.* Identifying the spurious mixing products falling within the desired mixer output range can be identified ahead of time by the following method.

---

[13] Variable Frequency Oscillator, manually tunable by a user.

[14] The first stable-frequency transmitters required a quartz crystal controlled oscillator. That was fine for broadcast transmitters assigned to one specific frequency but other transmitters such as radio amateur or civil aviation radios allowed to use any part of an allocated band had to have more flexibility in frequency control.

[15] More on mixer circuits' details are given in the chapters on Modulation/Demodulation and on Superhet receivers.

[16] If both m and n equal unity, equation (3-14) is equal to equation (3-13).

Low / High Method of Identifying Spurious Mixing Products[17]

The procedure is to divide the *lower* mixer input frequency by the *higher* mixer input frequency. It doesn't matter if either is the RF input or the LO injection. The end result is a fraction above zero but less than unity. Either of two tables following are consulted for the mixer output being the *difference* of the two inputs (Table 3-2) or the *sum* of the two inputs (Table 3-3).

The *order* of the spurious output equation is the sum of ***m*** and ***n*** in equation (3-14). In general the lowest-order spur product will have the greater output, the highest-order spur product the least output power.

Table 3-2
Low / High Fractions and Spurs for Difference Frequency Mixing

F_L/F_H Fraction	Lowest-Order Spur Product	Spur Out Due To Low Freq.	Spur Out Due To High Freq.	Order
0.091	$10\ F_L$	$10\ F_L$	$10\ F_H / 11$	10
0.100	$9\ F_L$	$9\ F_L$	$9\ F_H / 10$	9
0.111	$8\ F_L$	$8\ F_L$	$8\ F_H / 9$	8
0.125	$7\ F_L$	$7\ F_L$	$7\ F_H / 8$	7
0.143	$6\ F_L$	$6\ F_L$	$6\ F_H / 7$	6
0.167	$5\ F_L$	$5\ F_L$	$5\ F_H / 6$	5
0.222	$8\ F_L - F_H$	$7\ F_L / 2$	$7\ F_H / 9$	9
0.250	$3\ F_L$	$3\ F_L$	$2\ F_H / 4$	3
0.286	$6\ F_L - F_H$	$5\ F_L / 2$	$5\ F_H / 7$	7
0.333	$2\ F_L$	$2\ F_L$	$2\ F_H / 3$	2
0.375	$7\ F_L - 2\ F_H$	$5\ F_L / 3$	$5\ F_H / 8$	9
0.400	$4\ F_L - F_H$	$3\ F_L / 2$	$3\ F_H / 5$	5
0.429	$6\ F_L - 2\ F_H$	$4\ F_L / 3$	$4\ F_H / 7$	8
0.500	F_L	F_L	$F_H / 2$	1
0.571	$6\ F_L - 3\ F_H$	$3\ F_L / 4$	$3\ F_H / 7$	9
0.600	$4\ F_L - 2\ F_H$	$2\ F_L / 3$	$2\ F_H / 5$	6
0.667	$2\ F_L - F_H$	$F_L / 2$	$F_H / 3$	3
0.714	$6\ F_L - 4\ F_H$	$2\ F_L / 5$	$2\ F_H / 7$	10
0.750	$3\ F_L - 2\ F_H$	$F_L / 3$	$F_H / 4$	5
0.800	$4\ F_L - 3\ F_H$	$F_L / 4$	$F_H / 5$	7
0.833	$5\ F_L - 4\ F_H$	$F_L / 5$	$F_H / 6$	9

To use Table 3-2, assume an example the first mixer of a receiver that has several input frequency ranges with a fixed, crystal-controlled first LO for each range.[18] The output of the first mixer goes to a broadband first IF with a frequency range of 8.4 to 8.9 MHz. The first step is to find the low/high frequency fraction for each band, from the lowest to highest input frequency; 1st LO is fixed so the identification of any spur found is easier. Calculations, all frequencies in MHz:

---

[17] The method is based on the January 10, 1964, article by Roger T. Stevens published in *Electronics* magazine, once a McGraw-Hill biweekly subscription magazine. This was expanded and refined by the author in 1972, then published by RCA Corporation Electromagnetic and Aviation Systems Division as *L/H Method of Spurious Signal Identification*, RCA document number VNES-73-EM-001, January 1973.

[18] Described in more detail on Superheterodyne Receivers. While this is skipping ahead many chapters, the reader may already be aware of the existence of spurious mixing problems and that example is a real-world one to better illustrate the method.

BAND	1^{st} LO	Low/High Fraction	Possible Spurs Found
3.5 - 4.0	12.395	0.282 to 0.323	7^{th} Order at 0.286
5.7 - 6.2	14.595	0.391 to 0.425	5^{th} Order at 0.400
7.0 - 7.5	15.895	0.440 to 0.472	-none-
9.5 - 10,0	18.395	0.516 to 0.544	-none-
11.5 - 12.0	20.395	0.564 to 0.588	9^{th} Order at 0.571
14.0 - 14.5	22.895	0.611 to 0.633	-none-
15.0 - 15.5	23.895	0.628 to 0.649	-none-
17.5 - 18.0	26.395	0.663 to 0.682	3^{rd} Order at 0.667
21.3 - 21.8	30.195	0.705 to 0.722	-none-
26.9 - 27.4	35.795	0.752 to 0.765	-none-

BAND	Spur Product	Spur Calculation	Spur @ Tuning
3.5 - 4.0	$6F_L - F_H$	12.395 x (5/7) = 8.853571	3.541429
5.7 - 6.2	$4F_L - F_H$	14.595 x (3/5) = 8.757000	5.838000
11.5 - 12.0	$6F_L - 3F_H$	20.395 x (3/7) = 8.740714	11.654286
17.5 - 18.0	$2F_L - F_H$	26.395 x (1/3) = 8.798333	17.596667

In actual testing only two spurs were heard, the one at 5.838 MHz and the one at 17.597 MHz, both faint, the higher frequency spur the most prominent. It makes a difference on which mixer input requires the higher multiplier on the spur product column: At 5.838 MHz, it must be the 4th harmonic of the antenna input while the 17.597 MHz spur needs only the 2nd harmonic of the input.

The 1st IF is one input to the 2nd mixer with the tunable 2nd LO (5.5 to 5.0 MHz) its other input. In checking them, with both mixer inputs variable, the extremes of the Low/High fractions have to be used: (5.0/8.9) = 0.562 and (5.5/8.4) = 0.655. Only two spur orders show up in that fraction's range, 0.571 (9th order) and 0.600 (6th order). To identify just *where* the spur is located in frequency, an identity can be used from the *Spur Out Due To Low Freq.* column of Table 3-2:

$$\text{For Low / High fraction of 0.571:} \quad \frac{3\ F_L}{4} = 3.4 \text{ [IF frequency] and } F_L = 4.533333$$

That isn't in the tuning range of the 2nd LO and is disregarded

$$\text{For Low / High fraction of 0.600:} \quad \frac{2\ F_L}{3} = 3.4 \text{ and } F_L = 5.100000$$

The 0.600 Low/High fraction is within the 5.5 to 5.0 2nd LO tuning range (its frequency change direction is inverse of antenna input) and that spur might fall on all bands 400 KHz up from the bottom of each band. That is a 6th order spur product and would require a very strong signal input to create a 4th harmonic to produce the spur. It was not observed in normal tuning towards the top of each band.[19]

The *Spur Out Due To High Freq.* column of Table 3-2 could have been used as well. All it requires is another identity to equate to the IF and the 2nd mixer input frequency could be found.

[19] Tests on the SB-310 were conducted in 1972. Heath produced the SB-310 kit from 1972 to 1974.

Table 3-3
Low / High Fractions and Spurs for Summation Frequency Mixing

F_L/F_H Fraction	Lowest-Order Spur Product	Spur Out Due To Low Freq.	Spur Out Due To High Freq.	Order
0.111	$10\ F_L$	$10\ F_L$	$10\ F_H / 9$	10
0.125	$9\ F_L$	$9\ F_L$	$9\ F_H / 8$	9
0.143	$8\ F_L$	$8\ F_L$	$8\ F_H / 7$	8
0.167	$7\ F_L$	$7\ F_L$	$7\ F_H / 6$	7
0.200	$2\ F_H - 5\ F_L$	$6\ F_L$	$6\ F_H / 5$	6
0.250	$5\ F_L$	$5\ F_L$	$5\ F_H / 4$	5
0.286	$8\ F_L - F_H$	$9\ F_L / 2$	$9\ F_H / 7$	9
0.333	$4\ F_L$	$4\ F_L$	$4\ F_H / 3$	4
0.400	$6\ F_L - F_H$	$7\ F_L / 2$	$7\ F_H / 5$	7
0.429	$8\ F_L - 2\ F_H$	$10\ F_L / 3$	$10\ F_H / 7$	10
0.500	$3\ F_L$	$3\ F_L$	$3\ F_H / 2$	3
0.600	$6\ F_L - 2\ F_H$	$8\ F_L / 3$	$8\ F_H / 5$	8
0.667	$4\ F_L - F_H$	$5\ F_L / 2$	$5\ F_H / 3$	5
0.750	$5\ F_L - 2\ F_H$	$7\ F_L / 3$	$7\ F_H / 4$	7
0.800	$6\ F_L - 3\ F_H$	$9\ F_L / 4$	$9\ F_H / 5$	9

Table 3-3 is used much the same way as Table 3-2 except it applies to *summation* of frequencies in the mixer.

Tables 3-2 and 3-3 were both limited to a maximum of 10 orders of spurious products. The original and revised versions use maximum orders of 16. There is no real limit on the number of orders except that the harmonics get much larger. A maximum order of 10 was felt to be practical in nearly all hobbyist applications.

Intermodulation Distortion

Linear circuits have limits on their range of linear input levels. Exceeding that results in distortion and creation of harmonics that weren't in the source. A severe case of that results in a square wave. That harmonic content can be measured using a *spectrum analyzer* or specialized ***THD*** (Total Harmonic Distortion) instruments. A THD is measured as the sum of energies of the harmonics relative to total energy of fundamental and all harmonics. A typical THD maximum value of 2 percent or less indicates linear operation.

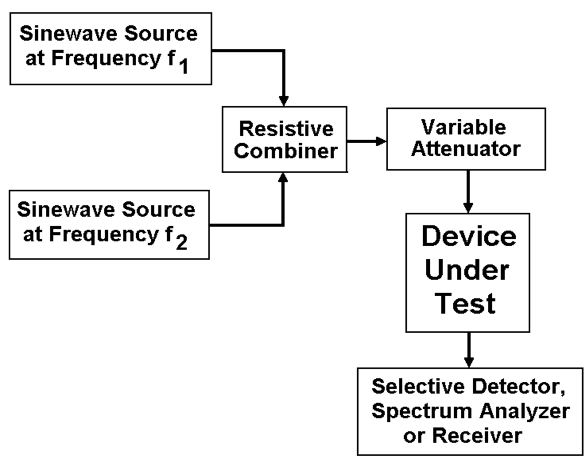

Figure 3-4 Equipment set-up to do two-frequency intermodulation distortion testing.

An easier and less-expensive test of linearity is to do a ***two-tone intermodulation test***.[20] This test setup is shown in Figure 3-4 and requires two sinewave sources of equal level, a resistive combiner, an attenuator for total input level adjustment, and a selective detector (or spectrum analyzer or wide-tuning receiver). The

[20] "Less expensive" is a relative term and refers to the cost of a spectrum analyzer instrument, usually a thousand or more dollars, as compared to a simpler setup of the two-generator arrangement.

difference frequency mixing product due to non-linear range operation is referred to as ***intermodulation*** or more familiarly, ***intermod***.

When the Device Under Test begins distorting, it will act as a mixer and generate sum and difference frequencies of the inputs in addition to passing both input signals. The selective detector will see only the fundamental or difference frequency component due to tuning. Both frequency sources should be set to the same output power. The intermodulation output level is relative to the input level of both equal-level sine sources.

Two-tone tests owe their name to early landline telephony when the frequencies were in the audio range, referred to as *tones*. In actual practice, the input frequencies can be at any frequency from sub-audible to the limit of microwaves. With a large dynamic range detector, this test is sensitive in determining the input maximum of the DUT's linear range.

Conversion Gain and Intercept Point

Passive mixers have definite limits to their signal input limits, usually given by manufacturers, along with their LO input levels. When input levels begin exceeding a linear input level, it is characterized as the *1 db compression* from a straight-line plot of output level versus input. While that shows device linearity, it is difficult measure.

An easier method of determining compression and non-linearity is the ***3rd Order Intercept Point***, also called ***IP3*** or just ***intercept point.*** The *order* in the name refers to the multiplier total in the mixer output products.[21] 1st order refers to only one input. 2nd order is the sum or difference of the inputs (multiplier of 1 for each frequency). The 3rd order has the sum and difference of one frequency and the 2nd harmonic of the other frequency (multipliers add up to 3, hence 3rd order). Figure 3-5 illustrates a log-log graph of input level versus output for an IP3 test result.

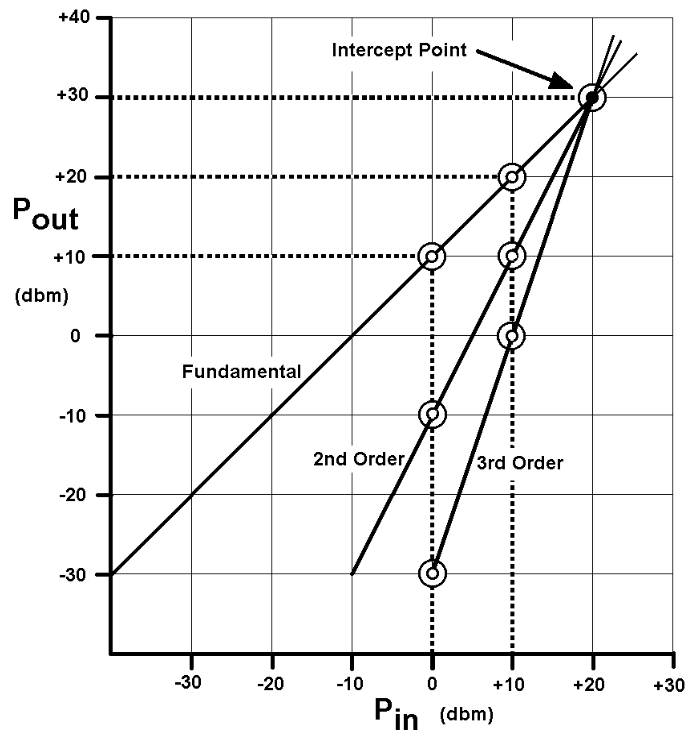

Figure 3-5 Graphical plot of fundamental (1st order), 2nd and 3rd order input versus output level responses to determine Intercept Point.

The *IP3* test can be done for amplifiers with the Figure 3-4 setup; for mixers there would be only one signal input generator while the other generator might be the mixer's LO injection. A spectrum analyzer is preferred due to the multiplicity of mixing product frequencies and the changing product magnitudes.

A plot would begin for amplifiers with just one signal input. Low-level signals would be

---

[21] EDN magazine, 3 May 2001, pp 53-61, "RF Transistors Meet Wireless Changes" by Bill Travis.

applied, noted, and output levels measures. This forms the 1:1 slope on Figure 3-5 called *fundamental.* As the signal level is increased the selective detector is tuned to the 2nd or 3rd order and output level measured while input level is noted. Using *log-log* graph paper, the input and output levels are plotted and a line drawn through fundamental, 2nd order, and 3rd order data. The intercept point itself is the intersection of the three slopes in Figure 3-2. Power levels are commonly done in dbm (0 dbm = 1 mW in a 50 Ohm system).

The Power Budget

The term *power budget* refers to the level of an RF signal along stages (usually in block diagram form) from its origin to a destination. It is in wide use in the microwave region, almost always at the common *characteristic of impedance* of 50 Ohms. It can refer to any power, voltage, or current level diagram. Radio amateur publications have mentioned similar budgets in transmitters but generally neglect any sort of RF level in receivers.

A lot of attention has been paid to receiver input stages in the 1990s and 2000s as to their ability to handle large input signal levels. While that is a noble effort for those who must use receivers in the vicinity of high-power RF sources, they also seem to be neglecting the following stages in any receiver. The later stages can also be prone to overloading and thus causing distortion.

Receiver design should include a *power budget* chart, especially with AGC or Automatic Gain Control chart rows beginning with the antenna input power levels that result in the AGC control voltages or currents. Along each row would be each stage's input signal level, whether it is AGCed or not. Each input can then be compared to the individual circuit and its linear amplification characteristics. It may be desirable to *reduce* a stage's gain, perhaps by negative feedback, in order to keep it more linear. Linearity of the entire chain is desirable to prevent distortion that can result in IMD or, in extreme cases, spurious frequency products from mixer-like input overdrive.

Chapter 4

Bandwidth, Modulation, Noise

Radio conveys information. In doing so, the radio signal ***carries*** information via some form of ***modulation*** on that signal. The signal center frequency is called the ***carrier*** and is called the ***carrier wave***. That total radio signal requires a certain ***bandwidth*** for faithful reproduction of the modulation waveform. Bandwidth or spectral occupancy depends on the rate of the waveform and the rate of change in time of the wave shape plus the type of modulation. The total radio signal occupies spectrum that others cannot use without interference. It determines the minimum signal level at a receiver, and establishes a maximum rate of information transfer. ***Thermal noise*** is ever present and must be calculated to determine the minimum possible receiver input level.

Shannon's Limit

The maximum rate of information transfer in any communications channel was mathematically analyzed by Claude E. Shannon in 1948:

$$C = W \operatorname{Log}_2 \left(\frac{P + N}{B} \right) \qquad (4\text{-}1)$$

Where:

C = Channel or circuit information rate in bits per second (BPS)

W = Bandwidth of channel or circuit, Hertz

P = Average power of signal, Watts

N = Average power of random noise in channel, Watts

This deceptively simple formula became known as *Shannon's Limit*. or sometimes *Shannon's Law.* Base-2 logarithms were used since Shannon was working with binary teleprinter codings having *bit* states of on or off. The value of bits per second is further defined as the rate having an *arbitrarily small frequency of errors*.

Shannon used a very generic rate description and did not elaborate on the type and kind of the waveform carrying the information. He did show that there is a relationship between signal power and noise power (rate can be increased if the signal-to-noise ratio is higher) and that bandwidth is proportional to the information rate. For conventional base-10 logarithms, Shannon's Limit can be expressed as:

$$C_{BPS} = BW_{Hertz} \bullet 3.3 \bullet \operatorname{Log}_{10} \left(\frac{\text{Signal} + \text{Noise}}{\text{Noise}} \right) \qquad (4\text{-}2)$$

If the signal to noise ratio is 10:1, a 10 db ratio, then the channel capacity C at a bandwidth of 3 KHz is approximately 3000 times 3.3 times the logarithm of 11 or about 10,400 bits per second (10.4 KBPS). Increasing the signal-to-noise power ratio by 10 (to 20 db) will increase the rate to about 20 KBPS. A 30 db signal-to-noise ratio would allow a rate of 30 KBPS and so on. However, if the signal-to-noise power ratio is only 3 db, then the channel capacity rate has decreased to 4.7 KBPS. If the signal power is equal to noise power then the rate is about equal to the bandwidth.

Information theory and Coding theory has advanced since 1948. By compressing redundant data and using modulation combinations, it is possible to send binary information over a 3 KHz bandwidth telephone line at a rate of 56 KBPS. Shannon's Limit has not been compromised but the use of information theory and coding has allowed a virtual increase in the signal-to-noise ratio (SNR).[1]

Bandwidth itself is generally defined as a frequency range of minimum attenuation; i.e., within the frequencies between attenuation increases of 3 to 6 db. However compact the information coding or the signal-to-noise ratio, a waveform's shape and the type of modulation will still determine the necessary radio bandwidth.

Basic Modulation Types

There are three basic types: Amplitude Modulation (**AM**), Frequency Modulation (**FM**), and Phase Modulation (**PM**). **AM** increases or decreases the apparent RF amplitude of an RF carrier wave without changing the carrier frequency. **FM** moves the RF carrier frequency above and below the RF carrier wave frequency but does not affect carrier amplitude. **PM** advances or retards the phase of the RF carrier wave from the modulation, but does not change amplitude of the carrier.

While most radio modulation involves only the basic types by themselves, it is possible to combine AM and FM and AM and PM. The latter is commonly done in the telephone modem used with personal computers. That AM-PM combination makes it possible to maximize binary data throughput (rate of information transfer) close to the Shannon Limit.

A unique property of AM enables removal of one sideband and the carrier, re-inserting the carrier at the receiver, to both reduce transmitter power requirements and bandwidth occupancy. That mode is common on HF and known by its colloquial name, *sideband.*

[1] A modulator-demodulator or *modem* in year 2001 in a computer-telephone connection can transfer 56 KBPS on a "clean" household subscriber line. The actual rates may be limited to 33 to 45 KBPS due to extraneous, random noise impulses or the frequency response of all telephone circuits. The virtual SNR is still high even though the signal may be periodically corrupted.

Amplitude Modulation

The instantaneous voltage at any point in time of an amplitude modulated RF waveform is:

$$e(t) = E\left[1 + M\cos\left(\omega_m t\right)\right] \sin\left(\omega_c t\right) \qquad (4-3)$$

Where:

$e(t)$ = Instantaneous RF voltage e at time t

t = Time position, seconds

E = Peak voltage of unmodulated RF carrier

ω_m = Radian frequency of modulation $= 2\pi f_m$

f_m = Modulation frequency, Hertz

ω_c = Radian frequency of RF carrier $= 2\pi f_c$

f_c = Carrier frequency, Hertz

M = Modulation factor in range of 0 to 1

Sine and cosine arguments are in radians

When modulation factor **M** is zero, there is no modulation, only the carrier. When **M** is unity, the peak carrier amplitude with modulation is twice the peak value of **E**, the peak carrier amplitude without modulation. The modulation factor is sometimes expressed as a percentage so that an **M** of unity would represent 100 percent modulation.

AM creates ***sidebands*** on either side of the carrier. If equation (4-3) is expanded and a trigonometric identity applied, it can have a new form:

$$e(t) = \underbrace{E\sin\left(\omega_c t\right)}_{\text{- Carrier -}} + \underbrace{\left(\frac{EM}{2}\right)\sin\left[\left(\omega_c + \omega_m\right)t\right]}_{\text{- -Upper Sideband - -}} + \underbrace{\left(\frac{EM}{2}\right)\sin\left[\left(\omega_c - \omega_m\right)t\right]}_{\text{- -Lower Sideband - -}}$$

Where: (4 - 4)

All terms are as in equation (4 - 3)

Both equation (4-3) and (4-4) assume the modulation is a single frequency. Note that the carrier remains unchanged by modulation. Note also that the two sidebands differ only in frequency; but are identical as to content. The bandwidth of an AM signal is twice that of the modulation frequency.

What is disturbing to some is that the classical oscilloscope pictures of a 100% modulated AM signal shows the “carrier” apparently going to a zero RF value during a negative modulation swing. Does the carrier actually go to zero? No. RF voltage **e(t)** is a linear summation of the carrier and both sideband terms. During a negative modulation voltage swing the sum of both sidebands is equal to the carrier but their phase angles are opposite to the carrier. A negative modulation peak at 100% modulation results in a zero-value **e(t)**. The carrier and sidebands still exist but a linear addition causes them to cancel each other out. At a positive modulation peak the carrier and both sideband terms are in phase so that all add to produce twice the **e(t)** value.

The spectral content of AM with a single tone for 100% modulation is shown in Figure 4-1. If the modulation is 50% then the amplitude of the sidebands drop to 0.25 **E** but the carrier amplitude remains the same. As long as the modulation remains between 0 and 100 percent there will be *no* change in carrier amplitude.[2]

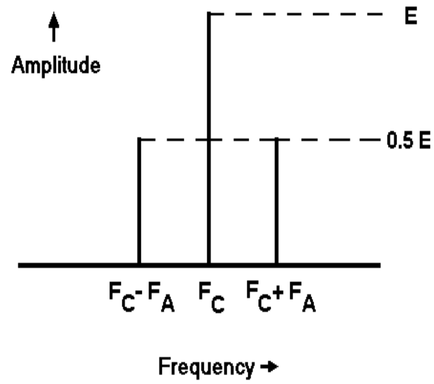

Figure 4-1 Ideal AM spectrum with 100% modulation by a single sinusoid.

At 100% modulation the carrier contains half the total power of the signal. Each sideband contains a quarter of the total power. That is fine for the simple example of a sine wave as the modulation source but voice (as one example) contains many frequencies. It is useful to look at conditions where two sinusoids of different strengths AM the carrier.

Equation (4-3) can be manipulated as in (4-4) with the resulting form for two sine waves amplitude modulating the carrier shown in equation (4-5):

$$e(t) = E\,\mathrm{Sin}\left(\omega_C\, t\right) + \left(\frac{E\,M_A}{2}\right)\mathrm{Sin}\left[\left(\omega_C + \omega_{FA}\right) t\right] + \left(\frac{E\,M_B}{2}\right)\mathrm{Sin}\left[\left(\omega_C + \omega_{FB}\right)\right]$$

$$+ \left(\frac{E\,M_A}{2}\right)\mathrm{Sin}\left[\left(\omega_C - \omega_{FA}\right) t\right] + \left(\frac{E\,M_B}{2}\right)\mathrm{Sin}\left[\left(\omega_C - \omega_{FB}\right) t\right] \qquad (4\text{-}5)$$

Where:

e, t, ω_C, E are is with (4 - 4) and

ω_{FA}, ω_{FB} are the radian frequencies of modulation frequencies F_A, F_B

M_A, M_B are the modulation factors for F_A, F_B, respectively

All sine arguments are in radians

Figure 4-2 shows the spectral content of AM with two modulating sinusoids for a total modulation factor of 1.0 or 100% modulation. Modulation factor for the *A* sine is about 0.83 and that for the *B* sine is about 0.55. The total power of the lower sidebands are still only a quarter of the entire AM signal. Equation (4-3) can be expanded to include all modulation frequencies but it is useful to use sine waves as modulation in order to compare all basic types of modulation.

A question to ask now is, if each sideband set has the same information, why transmit both of them? One of the reasons lies in early radio technology and the ease of both generating and

[2] The classic AM 100% modulation RF envelope has been deliberately omitted to avoid confusion with the oscilloscope picture. The oscilloscope is linearly adding carrier and sidebands and giving the *appearance* of null power at most-negative swing due to sideband phases adding and canceling the carrier phase.

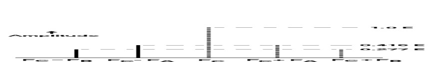

Figure 4-2 Spectral content of

detecting an AM signal. A second question is that, if the carrier maintains a steady power, why not eliminate that at the transmitter and reinsert it at the receiver? The answer to both questions existed in landline telephony just prior to World War One.

Single Sideband Amplitude Modulation

Note the solid and dotted lines in Figure 4-2. If the carrier and upper sideband components (dotted lines) were eliminated, the lower sideband components would carry all the modulation information. The necessary bandwidth is now no wider than the difference between lowest and highest modulation frequencies.

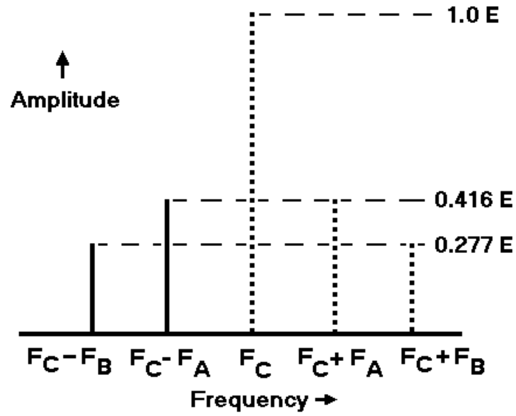

Figure 4-2 Spectra of an AM signal with carrier suppressed completely and only one sideband set.

The lower sideband components (solid lines in Figure 4-2) are still at radio frequencies even though they carry modulation information. To restore them, a steady carrier frequency must be *reinserted* or generated at the receiver. The receiver's detector will *mix* the received sideband components and the reinserted carrier and the result is the original modulation signal.[3] ***Sideband*** or ***SSB*** receivers need more frequency stability than AM receivers to maintain the reinserted carriers' frequency equal to that of the transmitter.

There are several advantages of SSB over conventional AM. The most obvious is bandwidth reduction. Second, the required transmitter power output is reduced, achieving more electrical mains efficiency and reduced local transmitter heating. Third, there is an increase in the *Signal to Noise Ratio* or SNR at the receiving end..

Early landline long-distance telephony was hampered by the cost of miles of wire pairs, one pair per circuit. While necessary voice bandwidth on any circuit was only about 3 KHz, each wire pair's bandwidth was at least five times wider. Long-distance telephone providers devised terminal equipment that *multiplexed* up to four voice-bandwidth telephone circuits on one wire pair. This *carrier* equipment (its colloquial name) used SSB techniques to put three more voice circuits occupying 3 to 6, 6 to 9, and 9 to 12 KHz bands on the same pair that carried a single 0 to 3 KHz voice circuit. Long-distance providers could quadruple their circuit capacity using *Frequency Division Multiplexing* via SSB.

It took nearly another generation before radio technology was developed enough to utilize SSB techniques effectively. Chief among that development was high-power linear RF amplifiers and frequency stability at each end. Commercial and military SSB communications began in the early 1930s adapting the old *carrier* terminal equipment to allow four voice-bandwidth circuits on

[3] Detector circuits are basically non-linear. They act as mixers in the demodulation process. Actual mixers can and are used as "detectors" to mix sidebands and a reinserted carrier to recover the original modulation signal as the difference between carrier and sidebands.

each SSB communications path.[4] It was typical between the 1930s to the 1960s to have similar teleprinter carrier equipment that multiplexed several teleprinter signals in one voice-bandwidth circuit.

Single-channel SSB came into popularity in the early 1950s, chiefly for voice communications by single users. SSB had become the voice communications mode of choice on HF by nearly all radio services at the start of the new millennium.

Carrier Suppression in AM or SSB

What would happen if the steady carrier is removed entirely from an AM signal? At the transmitter the spectral content would have only the lower and upper sidebands. The information is still there but it sounds terrible, barely intelligible at a standard AM receiver.

An obvious advantage of AM carrier suppression is that the total transmitted energy drops considerably. At 100% modulation the carrier energy represented 2/3 of the total RF power transmitted and did not contribute anything at all to the information being sent. A double-sideband, suppressed carrier AM signal would require only half the RF power of conventional AM to transmit the *same* information at the *same* amplitude. Another advantage of suppressing the carrier is the elimination of a source of heterodyne interference to anyone listening at adjacent frequencies.

A compromise can be reached between full AM and SSB by reducing the carrier power rather than eliminating it. This allows simpler demodulation in receivers since there is some carrier power received; carrier reinsertion is not needed.. Television broadcasting uses a *vestigial sideband* technique with both analog and digital video modes. This sends a small part or *vestige* of the lower video sideband along with about half of the carrier power needed for full video AM.[5]. A driving reason for that system was hardware. A bandpass filter following a conventional modulator could shape the response to partially attenuate the carrier but could not completely eliminate one sideband. A lowered carrier power allowed greater life for the final video power amplifier and reduced down-time at a broadcast station. TV receivers could retain simple and less-costly circuitry.[6]

While some experimentation has been done with AM other than in TV for suppressing the carrier power, few have shown any practical advantage. *Short wave* AM receivers (HF broadcast bands) appearing first in the early 1990s used *synchronous detection* to improve reception quality on HF paths having propagation perturbations. Synchronous detection techniques use a local carrier reinsertion in detection, the local carrier *phase-locked* to the received carrier. The advantage there is to maintain a power-steady carrier unaffected by propagation-changed carrier power levels.[7]

---

[4] The landline-originated carrier equipment created the 12 KHz wide audio in the first commercial and military SSB transmitters and also separated the received audio into separate voice bandwidth channels. Carrier equipment is largely unknown or unfamiliar to single-channel SSB users.

[5] An average of worldwide TV broadcast standards as of 2001. Either video sideband could be sent, the other suppressed.

[6] The advent of large-scale integrated circuitry obviated such dependency on receiver circuitry but the standards were developed when semiconductors were in their infancy. USA black and white TV standards were formalized just before World War Two, color TV standards a decade after that ending; vacuum tube architecture ruled then.

[7] The term is ***selective-fading*** due to unequal sideband frequency component shifts due to ionospheric acts.

Pulsed AM

Radar was the first widespread use of short-duration, high peak power pulse transmission. A pulsed carrier is an extreme form of AM. Pulse transmission has been used in other applications: Aircraft DME and ATC transponders, early multi-channel radio relay, but most often in very low power automobile keyless entry systems. On-off keyed telegraphy transmitters can be considered and analyzed as pulsed AM.

A typical radar spectral bandwidth is 10 to 30 MHz. If the radar transmitter pulse is assumed to have zero rise and fall times, its carrier and sideband spectral content can be approximated by:

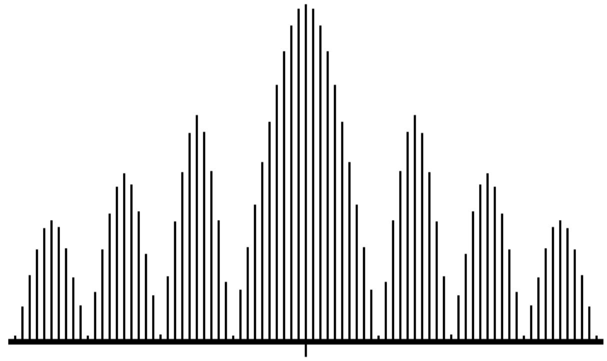

Figure 4-3 Spectra of a radar pulse.

$$C_O = 2 \pi D \qquad \text{[Carrier]}$$

$$C_N = 2 A D \left[\frac{\mathrm{Sin}\left(N \pi D\right)}{N \pi D} \right] \qquad \text{[Sidebands]} \qquad (4 - 6)$$

Where:

A = Peak amplitude of pulse   Sine argument in radians

N = Integer harmonic number

$D = \text{Duty Cycle} = \dfrac{\text{Pulse Width}}{\text{Repetition Period}}$

Sidebands are spaced at integral multiples of the pulse repetition rate on both sides of the carrier as shown in Figure 4-3. The "dips" in the sidebands occur at increments of the reciprocal of the pulse width time.

Figure 4-3 is an ideal case for illustration only. The exact amplitude and phase of sidebands would follow the rectangular pulse harmonic equation, with the lower sidebands being positioned as the difference between carrier frequency and the harmonic frequency. See Chapter 3 for rectangular pulses having finite rise and fall times.

While a radar transmitter may use a 1.0 µSec pulse width repeating every 2.5 mSec, a manual telegraphy transmitter might have on-times of 200 mSec with 30 mSec rise and fall times. An on-off keyed radiotelegraph transmitter still generates sidebands, albeit in a relatively very narrow bandwidth. Radiotelegraph receivers use correspondingly narrow bandwidths. Control over the rise and fall time of telegraphy transmitters' RF output is important in reducing interference to near-frequency radio users.[8]

[8] Too rapid an on-off switching time produces momentary sidebands that can exceed the average 400 Hz bandwidths of radiotelegraphy. Those sidebands sound like "clicks" in adjacent frequencies. The *keyclicks* can be disturbing if the clicker is geographically close and running high power. Too slow a transition will reduce the

The bandwidth required by an AM pulse signal depends on the fidelity required in an application. Radar needs very wide bandwidths for target discrimination. Pulse code train transmission can get by with reduced band-widths and some pulse shape distortion; the information in the *pulse group* is important while the individual pulse shape needs depend on keeping a good SNR. A few rare cases can have bandwidths equal to the reciprocal of the pulse width.[9]

Frequency Modulation

While AM keeps the carrier frequency constant and varies the amplitude with modulation, FM varies the frequency and keeps the amplitude constant. The instantaneous voltage of an FM signal having a single sine wave as modulation is:

$$e(t) = E\,\mathrm{Sin}\left[(\omega_c\, t) + \left(\frac{\Delta f}{f_m}\right)\mathrm{Sin}(\omega_m\, t)\right] \qquad (4\text{-}7)$$

Where:

$e(t)$ = Instantaneous amplitude e at time t, t in seconds

E = Peak amplitude of carrier

ω_c = Radian frequency of unmodulated carrier $= 2\pi f_c$

ω_m = Radian frequency of modulating sine wave $= 2\pi f_m$

Δf = Peak frequency deviation due to modulation

Sine argument in radians, frequency in Hz

Delta-f is the maximum excursion of the modulated carrier frequency referenced to the unmodulated carrier frequency. Delta-f over modulation frequency forms the ***modulation index*** or:

$$M_i = \frac{\Delta f}{f_m} \qquad (4\text{-}8)$$

The modulation index of FM is not at all like the modulation factor or percentage modulation of AM.. The value of the modulation index has no real bounds, only practical ones of anything from 0.2 to 5.0. The modulation factor of AM is limited to between zero and unity.

Determining delta-f may be done if the FM modulator can respond to DC equally well as to an AC signal. A DC test signal can be applied and varied until it matches the AC modulation signal, then the modulation-moved carrier frequency is measured directly. It is quite another thing if the FM modulator has a low-frequency limit, requiring *carrier null* measurements shown later in this section.

---

intelligibility of short on-times.

[9] Having a BW = (1 / width) is sometimes called a *matched filter* situation. A rectangular pulse through such a narrow filter will have an RF envelope at the output of a cosine-squared pulse envelope shape. If the filter center frequency is moved to encompass the two sidebands nearest the carrier, the resulting filter output RF envelope will have a shape similar to a wide bow tie with a peak amplitude less than a quarter of the on-carrier-frequency filter.

It will be shown that FM sidebands mathematically extend to an infinite bandwidth on either side of the carrier frequency. In practical applications the FM bandwidth is finite and can be referenced to the modulation index, modulating frequency, and a desired signal-to-noise ratio.

Through some substitution of identities, equation (4-7) can be transformed into a series equation which describes the FM sidebands:[10]

$$\begin{aligned} e(t) = E\,\{ & J_0(M_i)\,\mathrm{Sin}\left[\omega_c t\right] \\ & + J_1(M_i)\,\mathrm{Sin}\left[(\omega_c + \omega_m)\,t\right] - J_1(M_i)\,\mathrm{Sin}\left[(\omega_c - \omega_m)\,t\right] \\ & + J_2(M_i)\,\mathrm{Sin}\left[(\omega_c + 2\omega_m)\,t\right] - J_2(M_i)\,\mathrm{Sin}\left[(\omega_c - 2\omega_m)\,t\right] \\ & + J_3(M_i)\,\mathrm{Sin}\left[(\omega_c + 3\omega_m)\,t\right] - J_3(M_i)\,\mathrm{Sin}\left[(\omega_c - 3\omega_m)\,t\right] \\ & + J_4(M_i)\,\mathrm{Sin}\left[(\omega_c + 4\omega_m)\,t\right] - J_4(M_i)\,\mathrm{Sin}\left[(\omega_c - 4\omega_m)\,t\right] \\ & + J_5(M_i)\,\mathrm{Sin}\left[(\omega_c + 5\omega_m)\,t\right] - J_5(M_i)\,\mathrm{Sin}\left[(\omega_c - 5\omega_m)\,t\right] \\ & + \ldots\} \end{aligned} \qquad (4\text{-}9)$$

Where:

$e(t)$, E, M_i, ω_c, ω_m are as with equation (4 - 7)

$J_n(M_i)$ = Bessel Function to order n with argument M_i

All sine arguments are in radians

Bessel Functions are explained in more detail in Appendix 4-2. Based on the order (**n**) and argument (modulation index) they return a number to be used as a multiplier of the **Sine [ ]** terms.[11]

Note that the Sine argument in parenthesis has both summation and difference of the radian frequencies of the unmodulated carrier and the modulating frequency. This indicates that FM sidebands exist in frequency and

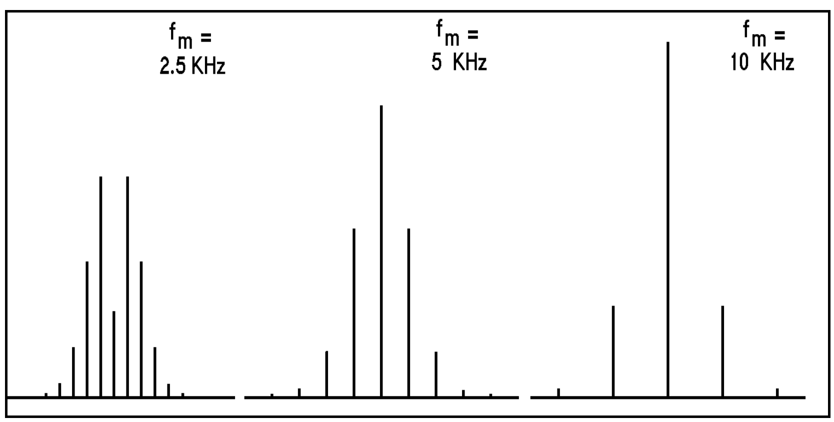

Figure 4-4 Spectral content of an FM signal having 5 **KHz deviation but with three modulation frequencies of 2.5, 5.0, and 10 KHz. All are to same scale. Note that carrier amplitude varies with modulation frequency.**

[10] Shown in Appendix 3-1 for the benefit of those who actually like all the math...

[11] Other texts usually reprint a graph of overlapping almost-sine curves which was handy back in slide rule days. With small, powerful calculating tools of the start of the millennium it is easier to just calculate the numerical value rather than puzzle over which curve is which on those charts.

amplitude symmetry about the carrier frequency. Even with a pure sine wave for modulation, there are more than two sidebands created. Mathematically, there are an infinite number of sidebands (the series does not end as indicated by the ellipsis notation) but the practical case is that higher-order Bessel Functions have a numeric value decreasing with each order. The successively higher order terms will eventually become too small to be of any use.

With the same peak deviation (delta-f), differing modulation frequencies will have different sideband amplitudes. That is illustrated in Figure 4-4. If the modulation frequency remained the same, say 5 KHz, the spectra of Figure 4-4 would correspond to varying modulation indexes of 2, 1, and 0.5, left to right. That inverse relationship comes from equation (4-8).

FM spectral content is not intuitive as with AM. Sidebands and the carrier change amplitude, sidebands change spectral position depending on the peak deviation and the modulation frequency. It should be noted that, if one took the sum of the powers of all spectral components, they would always equal the power of the unmodulated carrier, regardless of modulating frequency or peak deviation. It remains that FM signal power is always constant. However, since spectral components change amplitude individually, those can be used to calibrate either peak deviation or modulation index by looking for a *null amplitude* of the carrier frequency or one of the sideband pairs. This method of deviation calibration applies to both FM and PM, is covered later in this chapter.

Frequency Multiplication of an FM Signal

Since the total power of an FM signal remains constant regardless of modulation amplitude, it follows intuitively that there is no problem if the FM signal is changed in amplitude for any reason. FM receivers have *limiter* stages prior to detection. Limiters will limit all amplitudes over a certain limiting threshold level. The combination of amplification and limiting in an FM receiver act to hold the received FM signal at the detector the same over a very wide range of antenna input levels. Limiting acts to keep **E** in equation (4-7) constant. If **E** is constant then the FM sidebands are preserved. Modulation remains constant at the demodulator input.

Frequency multiplier circuits exist whose tuned outputs have an integer multiple of their input frequency. What happens if an FM signal is multiplied in frequency? Let's say it is put through a special circuit which multiplies the input frequency **n** times, the **n** being an integer. Going back to equation 3-7, an FM signal at the output of the multiplier will be:

$$e(t) = E\,\mathrm{Sin}\left[\left(\mathbf{n}\,\omega_c\,t\right) + \left(\frac{\mathbf{n}\ \Delta f}{f_m}\right)\mathrm{Sin}\left(\omega_m\,t\right)\right] \qquad (4\text{-}10)$$

Where:

$e(t)$, t, E, ω_m, f_m, Δf are all as in equation (4 - 7)

n is an integer of 1 or greater

Sine arguments are in radians

Two things have happened. From the first term within the square brackets, the carrier frequency has been multiplied by **n** as intuitively expected. The second thing is a multiplication of the *modulation index* by **n**. Recall that the modulation index is delta-f over the modulating frequency. This is the same thing as multiplying the entire square bracket term by **n**. The multiplication increases the carrier and frequency and the frequency deviation by the same value of **n**. Sideband components (the inner sine term argument) are not changed in frequency or occurrence so the modulating signal still has the same information.

Integer multiplication cannot be done with an AM signal nor can limiting be used in an AM receiver. Doing so will remove the amplitude variations carrying the modulation information. That should be intuitive from the AM definition but it can be proven mathematically by multiplication of the radian frequencies in equation (4-3).

Bandwidth, Distortion, and Noise in FM

An approximate necessary bandwidth equation for FM signals is:

$$BW = 2\left(\Delta f + f_{max}\right) = 2 f_{max}\,(M_i + 1) \tag{4-11}$$

Where:

BW = Necessary bandwidth, Hz

Δf = Peak frequency deviation, Hz

f_{max} = Maximum modulation frequency, Hz

M_i = Modulation index as in equation (4-8)

This approximation takes into account all "significant" sidebands (approximately 97% or better) and is familiarly known as *Carson's Rule.*[12] It is obvious that FM will take more bandwidth than a SSB for the same maximum modulation frequency. FM bandwidth approaches the bandwidth of an AM signal only when the modulation index is small.

So-called *Narrow Band FM* or NBFM using very low modulation indexes has been tried on HF bands for voice modulation. While it works there is no significant advantage other than in simpler transmitter hardware compared to SSB; NBFM receiver hardware is made slightly more complex. FM on HF has been largely restricted to *Frequency Shift Keying* or FSK of teleprinter mark-space information at rates usually under 300 Hz and with peak frequency deviations of 170 to 850 Hz. A moderate bandwidth FM is used at VHF and above, voice modulation with peak deviation within 25 KHz. A much higher modulation index is used in FM broadcasting, modulation frequencies up to 15 KHz and peak deviations up to 75 KHz.

The advantage of FM lies in the ability of the FM receiver's limiters to remove any stray signal variations, impulse noise spikes less than the signal's level, and even suppressing other radio signals in adjacent frequencies. The "volume" of a demodulated FM signal remains quite constant without any *AGC* or Automatic Gain Control circuitry required for AM and SSB.

Phase Modulation

PM is almost the same as FM. Rather than changing the carrier frequency, PM changes the carrier ***phase***. Carrier power amplitude is not changed by modulation. Given a sine wave modulating voltage, a PM signal's instantaneous voltage can be defined as:

---

[12] After John Renshaw Carson of ATT. who did the paper analysis in 1915 and was awarded a USA patent on it in 1922.

$$e(t) = E \operatorname{Sin}\left[\omega_c t + \Delta\theta \operatorname{Cos}\left(\omega_m t\right)\right] \qquad (4\text{-}12)$$

Where:

$e(t)$, t, E, ω_c, ω_m are the same as in equation (4 - 7)

$\Delta\theta$ = peak value of phase deviation due to modulation, radians

Sine and cosine arguments are in radians

The modulation spectra of a phase-modulated signal can be obtained expanding equation 3-11 until there is an identity with Bessel Functions of the First Kind. Then a trigonometric identity is used to get the following series:

$$\begin{aligned} e(t) = E \{ & J_0(\Delta\theta) \operatorname{Sin}[\omega_c t] \qquad (4\text{-}13) \\ & + J_1(\Delta\theta) \operatorname{Cos}[(\omega_c + \omega_m) t] + J_1(\Delta\theta) \operatorname{Cos}[(\omega_c - \omega_m) t] \\ & - J_2(\Delta\theta) \operatorname{Sin}[(\omega_c + 2\omega_m) t] - J_2(\Delta\theta) \operatorname{Sin}[(\omega_c - 2\omega_m) t] \\ & - J_3(\Delta\theta) \operatorname{Cos}[(\omega_c + 3\omega_m) t] - J_3(\Delta\theta) \operatorname{Cos}[(\omega_c - 3\omega_m) t] \\ & + J_4(\Delta\theta) \operatorname{Sin}[(\omega_c + 4\omega_m) t] + J_4(\Delta\theta) \operatorname{Sin}[(\omega_c - 4\omega_m) t] \\ & + J_5(\Delta\theta) \operatorname{Cos}[(\omega_c + 5\omega_m) t] + J_5(\Delta\theta) \operatorname{Cos}[(\omega_c - 5\omega_m) t] \\ & - J_6(\Delta\theta) \operatorname{Sin}[(\omega_c + 6\omega_m) t] - J_6(\Delta\theta) \operatorname{Sin}[(\omega_c - 6\omega_m) t] \\ & - J_7(\Delta\theta) \operatorname{Cos}[(\omega_c + 7\omega_m) t] - J_7(\Delta\theta) \operatorname{Cos}[(\omega_c - 7\omega_m) t] \\ & + J_8(\Delta\theta) \operatorname{Sin}[(\omega_c + 8\omega_m) t] + J_8(\Delta\theta) \operatorname{Sin}[(\omega_c - 8\omega_m) t] \\ & + \ldots \} \end{aligned}$$

Where:

$e(t)$, t, E, ω_c, ω_m are as in equation (4 -12)

$J_n(\Delta\theta)$ = Bessel Function of order n, argument $\Delta\theta$

All angles, sine and cosine arguments in radians

Note the ordering of signs and the alternation of **Cos** and **Sin** in the orders. Other than that, equation (4-13) is remarkably similar to FM equation (4-8). The sine and cosine arguments in parenthesis indicates that PM sidebands occur at the same frequencies as with FM but their amplitudes and phases may change in comparison to FM.

Note also the argument of the Bessel Functions in PM. If the value of the radian phase deviation equals the FM modulation index then equations (4-7) and (4-12) are the same in amplitude but **e(t)** will differ in phase by one-half pi radians or 90 degrees.

Bandwidth, Distortion, and Noise in PM

Minimum bandwidth for a PM signal is the transposed Carson's Rule:

$$BW = 2 f_{max} (1 + \Delta\theta) \qquad (4\text{-}14)$$

Delta-theta is in radians. Note that bandwidth of (4-14) is identical to that of equation (4-11) if the radian phase deviation equals the FM modulation index.

A PM signal is essentially the same as an FM signal insofar as receivers are concerned.

Limiting in PM receivers may be applied the same as with FM receivers; limiting affects only **E** in equations (4-12) and (4-13) which does not change relative sideband values.

Intuitively, a phase change in a carrier results approximately in a frequency change and vice versa at higher modulating frequencies. The major difference between FM and PM is at low frequencies, especially with DC modulation. In FM a DC modulation shift results in a definite frequency shift, measurable in frequency. In PM there is no easy way to sense a DC modulation shift unless there is some definite preceding period of no modulation.

Frequency Multiplication of the PM Signal

If a PM signal is multiplied in frequency, the basic e(t) equation of (4-12) will appear approximately the same as in equation (4-10) except a cosine replaces the sine term and delta-theta replaces the FM modulation index. The multiplied PM signal has the form:

$$e(t) = E \operatorname{Sin}\left[\mathbf{n}\,\omega_c\, t + \mathbf{n}\,\Delta\theta \operatorname{Cos}\left(\omega_m\, t\right)\right] \qquad (4\text{-}15)$$

Where:

$e(t)$, E, t, ω_c, ω_m are as in equation (4-12), trig arguments in radians

$\mathbf{n}$ is an integer of 1 or greater

The carrier frequency and phase deviation are both multiplied by **n** but the modulation information itself remains the same as before multiplication.

Setting An Accurate Modulation Input

The amplitude of any modulator input must be related to the amount of modulation of the RF regardless of type of modulation. This is required to reference the modulating signal input to some value, usually to a maximum of modulation.

Setting the modulation percentage of AM involves observing the modulated RF envelope on an oscilloscope and noting the modulation level where the envelope peak is twice that of unmodulated RF envelope or where the envelope dip appears to go to zero. That is 100% modulation. With a linear response on the oscilloscope, lesser modulation percentages can be set via observing the RF envelope amplitude.[13]

The frequency deviation of an FM signal may be set by direct measurement if the modulator response is linear down to DC. A frequency counter can measure the RF directly and the DC level into the modulator easily adjusted, measured by a DC voltmeter. It is quite another thing when the FM modulator input does not respond to DC. It is even more complicated with PM since it is very difficult to ascertain the modulated RF phase relative to the unmodulated RF phase. The way out of these dilemmas is the ***null observation*** of the FM or PM carrier or one of its sidebands. A spectrum analyzer can view the carrier or selected sidebands to see the null or a narrow bandwidth receiver can be used to separate spectral components and observe the null.

[13] Coupling a high-level RF signal directly to a cathode ray tube's vertical deflection plates is an old, easy way to get a linear representation on the oscilloscope. The horizontal plates can be driven by conventional sawtooth sweep waveforms. The linearity will be as good as the electron geometry of the cathode ray tube.

Basis for the Null Observation Technique

The spectral components of FM are mathematically shown in equation (4-9) on page 4-10, those of PM shown in equation (4-12) on page 4-13. The carrier center frequency is represented by the first term group. The sidebands are represented by the pairs of term groups following the carrier. The sidebands appear symmetric about the carrier center frequency indicated by the integer multiplier of the radian modulating frequency. If the magnitudes of each term group are calculated with varying modulation indices and modulating frequencies, it will be seen that they vary considerably, even to reaching very low levels. The very low level condition is the **null**.

Bessel Functions are cyclic in value. That is, depending on their **order** and **argument**, the function's result value (which is what is directly used in calculation) go positive through zero to negative and swing back up through zero to a positive value, repeating forever. The zero value Bessel Function result value is the key to this technique. When that Bessel Function returns zero, that FM spectral component amplitude will also be zero or nulled.

The carrier frequency component is usually used in setting an FM modulation index so that part can be taken from equation (4-9) and equated to zero using the modulation index from (4-8):

If $J_0(M_i)\mathrm{Sin}[\omega_c\, t] = 0$ then $J_0(M_i)$ can equal zero and...

If $M_i = \left(\frac{\Delta f}{f_m}\right)$ and B_{zn} is the argument of $J_0(\ )$ that returns zero:

Then

$$B_{zn} = \left(\frac{\Delta f}{f_m}\right) \qquad f_m = \left(\frac{\Delta f}{B_{zb}}\right) \qquad \Delta f = f_m\, B_{zn} \tag{4-16}$$

By solving for the **argument** of a Bessel Function of **order 0** that returns zero, an identity has been created that is useful in setting FM by observing the carrier frequency null. Values of B_{zn} can be found from Table 4-1.

Nulling of the sideband components can be found the same way except that there are pairs of sidebands created with a single sine wave of modulation. The Bessel Function order can be taken as the "harmonic" of the modulating frequency, order 1 being the fundamental, order 2 being the "2nd harmonic," and so on.

Table 4-1
Bessel Function Arguments Returning a Zero Function Value

Order	Bz1	Bz2	Bz3	Bz4	Bz5
JO	2.404826	5.520078	8.653728	11.79153	14.93092
J1	0	3.831706	7.015587	10.17347	13.32369
J2	0	5.135622	8.417244	11.61984	14.79595
J3	0	6.380161	9.761023	13.01520	16.22347
J4	0	7.588342	11.06471	14.37254	
J5	0	8.771484	12.33860	15.70017	
J6	0	9.936110	13.58929	17.00382	
J7	0	11.08637	14.82127	18.28758	
J8	0	12.22509	16.03777		
J9	0	13.35430	17.24122		

To use the Table 4-1 zero-return values with FM carrier or sidebands, select either the frequency deviation (delta-f) or the modulating frequency (a single sine wave) as fixed. If the frequency deviation is set, divide it by the zero-return value to find the modulating frequency. As an example, assume the carrier is to be nulled and the frequency deviation is 75 KHz. The Bz1 column for order 0 is 2.404826. Divide 75 KHz by 2.404826 to find the modulating frequency of 31.18729 KHz. Using Bz2 through Bz5 columns, the modulating frequencies would be 13.58676, 8.666784, 6.360498, 5.023133 KHz, respectively. Adjust the modulator's sine wave input amplitude until a carrier null occurred with any of those frequencies.

The same procedure can be followed by selecting one of the sideband pairs. This is slightly more difficult (not overly so) since the sideband observing device must be set to the proper sum or difference of modulating frequency from carrier center frequency.

With PM signals the identity of (4-16) becomes:

$$B_{zn} = \Delta\theta \qquad (4-17)$$

The zero-return argument value is directly equal to the phase deviation in radians per second. Assuming the phase deviation directly proportional to modulator input amplitude, that input amplitude could be adjusted for a null, then another phase deviation interpolated from a proportional modulator input amplitude.

Secrecy in Modulation

Most of the *trunk lines* between telephone exchanges in the U.S. are digital in nature. Voice and data information is digitized by sampling the audio-range waveforms, converting the sampled amplitude to a group of digital bits that represent the sample. The digital bits, in digital *words* or bytes (eight bits per byte) can be stored in temporary digital memory, then combined with other sampling so that many telephone circuits can be carried digitally over one coaxial or fiber optic cable.

Once in digital form, the digital information can be *scrambled* to order. That is, the bits within a digital information group are re-arranged from their expected order or position. If the receiving end can unscramble the re-arrangement, the digital sampling information can be recovered correctly. This is one kind of *encryption*, a general term for secret communications. Such encryption needs a form of synchronizing of the unscrambling, of having the *key* which unlocks the scrambling. Without the key such a digital encryption circuit sounds like so much random noise, quite unintelligible.

Encryption and decryption is regularly used in military communications. An example is the COMSEC (Communications Security) mode of the SINCGARS (Single Channel Ground-Air Radio System) family of radios in the U.S. military.[14] There exist scramblers that can work over telephone lines to keep both government and civilian communications *secure* from eavesdropping. There exist scrambling systems for cable and satellite-downlink subscription television so that only subscribers

[14] SINCGARS manpack and vehicular radios came into U.S. military operational use about 1989 and operate in the 30 to 88 MHz range. Both voice and data can be encrypted with the additional feature of frequency hopping the carrier frequency every tenth of a second. As of October, 2006, over 300,000 SINCGARS radios have been produced and fielded, making it the most-manufactured military radio set in history.

can see their TV.[15] Scrambling is not necessarily confined to secret communications but there are methods of scrambling that allow for co-existence of many radio users on the same frequencies without mutual interference. The co-existence lies in a general technique known as ***spread spectrum***.

Direct Sequence Spread Spectrum or DSSS

DSSS involves digital data or digitized analog signals processed through a digital shift register prior to modulation. The end result of the processing is a **spreading** of the spectral content evenly over a particular bandwidth. Modulation usually involves a combination of amplitude and phase modulation of the digital signals. The receiving end reverses the processing to recover the original digital data.

Compared to conventional modulation, DSSS bandwidth is large. The advantage of DSSS is that the energy of a portion of the bandwidth is small and resembles natural non-coherent noise. If different ***spreading codings*** are used among different circuits, none will interfere with one another even if the carrier frequencies are the same.[16]

DSSS does not directly seek to encipher communications but rather to approximate natural noise in the total modulation spectrum bandwidth. If the demodulation coding is synchronized with modulation coding, communicated intelligence can be extracted coherently from a background of incoherent noise.

Frequency Hopping Spread Spectrum or FHSS

FHSS performs spectrum spreading by deliberately jumping to new carrier frequencies. The modulation type and characteristics must accommodate the jump time interruption and resulting new carrier frequency dwell time. Such modulation is typically digitized data or voice. FHSS systems need extraordinary time-frequency stability and a method of synchronizing receivers with transmitters. The U.S. military SINCGARS radio family of field and vehicular radios has an optional FHSS mode.

If the hop frequencies are many and done with a pseudo-random selection, an FHSS signal is quite difficult to detect, let alone intercept. It will sound like an occasional random click to anyone listening with a single-frequency receiver. The chief use of FHSS is for communications security.

NOISE

Thermal Noise

All resistances generate a wideband (up to about 100 GHz) *noise voltage* due to random

---

[15] Encryption and decryption details will <u>not</u> be covered in this book.

[16] The spreader-coding is known by various names, *polynomial shift-register* being the most common. These belong to the *PRSG* or Pseudo-Random Signal Generator family of subcircuits mechanized in digital shift registers having binary feedback. More in later chapters on details.

electron movement above *absolute zero* (-273.15 °C). The *thermal noise* is evenly distributed over the wide bandwidth. All conductors have some finite resistance so thermal noise is present in them as well. Average noise power can be calculated from average noise voltage and the resistance. Since thermal noise is truly random, noise powers *always* add when combined with other thermal noise sources or a signal source. Average noise voltage is dependent on bandwidth and temperature and is calculated by:

$$e_n = \sqrt{4\ k\ T\ B_W\ R} \qquad (4-18)$$

Where:

k = Boltzman Constant $= 1.38 \cdot 10^{-23}$ joules per degree Kelvin

T = Temperature, degrees Kelvin

B_W = *Bandwidth*, Hertz

R = Resistance, Ohms

Electronic industry standard noise measurement temperature is 290 K (16.85 °C or 62.35 °F) so equation (4-18) can simplify to:

$$e_n = 1.265 \cdot 10^{-10} \sqrt{B_W\ R} \qquad (4-19)$$

While that temperature is fine for controlled, colder laboratories, a slightly higher temperature of 300 Kelvin (roughly 81 °F or about 27 °C) can take into account ambient temperature on the bench:

$$e_N = 1.287 \cdot 10^{-10} \sqrt{B_W\ R} \qquad (4-20)$$

At a resistance of 75 Ohms and 6 MHz bandwidth, equation (4-20) would calculate as:

$$e_n = 1.287 \cdot 10^{-10} \sqrt{75 \cdot 6 \cdot 10^6} = 1.287 \cdot 10^{-10} \sqrt{4.5 \cdot 10^8} = 2.73\ \mu V$$

The average noise power in the 6 MHz bandwidth would be equal to the average noise voltage squared divided by resistance of 75 Ohms:

$$P_n = \frac{\left(2.73 \cdot 10^{-6}\right)^2}{75} = 99.38 \cdot 10^{-15} = 99.38\ fW \approx -130\ dbW = -100\ dbm$$

Thermal noise power is sometimes expressed as *PSD* or *Power Spectral Density* which is often in terms of power per unit bandwidth or power in Watts per Hertz. 100 femtoWatts over a 6 MHz bandwidth would become equal to about -198 dbW per Hz.

Noise Figure or NF

Noise Figure, expressed in positive decibels, refers to the ***inside-the-device*** generated noise of a device. It is important in evaluating Signal to Noise ratios in very low-level amplifiers. NF determines an ***equivalent input noise*** by multiplying input thermal noise by the NF in db equal to the value 10 raised to the quantity (NF / 20).

As an example, suppose the device has a NF of 5 db and the calculated thermal noise at the input is 290 nV. That 5 db NF would be the equivalent of raising it 1.778 times, 10 raised to the power of (NF / 20), or 515.7 nV. The ratio of 515.7 to 290 in Volts is equal to 5 db.

The calculated ***equivalent*** noise at the input is all noises multiplied by the Noise Figure. For voltage it would be a multiplier value. For power, usually in dbm, it is the noise power in dbm plus the db of the NF. In the above example in a 50 Ohm system, the thermal noise at the input would be 1.682 fW or -147.7 dbm. The equivalent noise (due to NF) would be 5.319 fW or -142.7 dbm.

Noise Temperature

Device noise is sometimes quantified as ***Noise Temperature*** relative to ambient noise temperature:

$$T_N = (NF - 1) \cdot T_A \qquad (4-21)$$

Where:

T_N = Device Noise Temperature, °K

T_A = Ambient noise temperature, standard temperature of 290 °K

NF = Noise Figure of equation (4 - 21)

Noise Temperature does not refer to the physical temperature of the device but rather an equivalent temperature to which the device would be raised to generate that noise power. If the ambient noise temperature and device noise temperature are known, Noise Figure may be calculated from a rearranged (4-22):

$$NF = \frac{T_N - T_A}{T_A} \qquad (4-22)$$

Noise Floor

This refers to the noise power of the environment without any coherent signals. It derives partly from the jagged, jittering horizontal baseline on spectrum analyzers due to the analyzer's own internal noise; external signals appearing to "stand on that noise floor."

White Noise and Pink Noise

White noise is uniformly distributed over a wide bandwidth. *Pink* (or other color) noise is not uniformly distributed over a wide bandwidth. Color-designated noises are used for various specialized testing purposes. The color descriptor probably comes from *white light* where the total light is made up of an even distribution of light wavelengths. Light filters would make the distribution of wavelengths unequal over the spectrum and thus appear to have some specific color.

Other Noises

Galactic noise is random over a wide but finite bandwidth originating outside our solar system. ***Atmospheric noise*** refers mainly to sudden but broadband impulse noise such as from lightning or auroras over regions closer to the earth's polar regions. ***Ignition noise*** refers to periodic, short-pulse noise from older automobile spark plugs. ***Shot noise*** is caused by the quantized and random nature of current flow. In general, it is device specific to active devices, unequal in spectral distribution, higher levels at low frequencies. ***Phase noise*** refers to minor variations in phase of continuous wave oscillators resulting in additional random frequencies close to the carrier. Phase noise is measured in *db per Hz bandwidth* (usually 1 Hz bandwidth) relative to the carrier power as ***dbc***, stronger near the carrier and becoming weaker going away from the carrier frequency. Phase noise is important to higher-rate digital communications such as cellular telephony and most oscillator components have phase noise (or lack of it) touted in the market competition. In general the phase noise of LOs for voice-bandwidth radios is of so little effect on a circuit that it can be ignored.

Mixing or Combining Noise

Noise is random or *incoherent*. If noise is the input to a mixer, the LO does not change the noise spectral content. Specifically band-limited noise, such as running it through a filter first, can be shifted in the spectrum by the mixer's LO. Wideband noise through a mixer is not affected by the LO, comes out as noise at a power level shifted by the mixer's conversion gain or loss.

Combining noise and a coherent signal will result in a total noise-plus-signal power that is the sum of each powers. Random noise power ***always*** statistically adds.

Appendix 4-1

Computer Calculation of Bessel Functions for FM and PM

Calibration of modulation indices in FM and PM is not helped by the usual curve presentation of Bessel Functions of the first kind. At best, two digit accuracy is about the best available from a large graphics image. There exist a few tables of Bessel Functions among several math texts but they seldom carry arguments beyond two decimal digits. Reference [5] carries arguments to three digits. For precision, a personal computer can be used with a short program. A quick review of the Bessel Function notation:

$J_N(X)$ Where:

J is the descriptor letter for a Bessel Function of the first kind.

N is the order of the Function, beginning with 0.

X is the argument of the Function, the value supplied.

Reference [6] defines a Bessel Function of order **n** with argument **x** as:

$$J_N(X) = \left(\frac{X^N}{2^N G(N+1)}\right)\left[1 - \frac{X^2}{2(2N+2)} + \frac{X^4}{2 \bullet 4(2N+2)(2N+4)} - \;\ldots\ldots\right]$$

The letter T with one arm missing is the Greek letter gamma. That indicates a Gamma Function , once described as "a factorial but it is missing something."[17] The Gamma Function is a factorial equal to its argument minus 1 with the proviso that the function does not go lower than unity. The Gamma Function has been precalculated in the subroutine example following.

The series under the square brackets may be re-arranged as:

$$\left(1 - \frac{X^2}{2(2N+2)}\left(1 - \frac{X^2}{4(2N+4)}\left(1 - \frac{X^2}{6(2N+6)}\left(1 - \frac{X^2}{8(2N+8)} \;\ldots\right)\right)\right)\right)$$

The factorization of the square bracket series allows fairly simple conversion to an iterative function in a computer program or a programmable calculator. A FORTRAN subroutine is illustrated by the BESSEL routine following. It iterates 40 times beginning with the right-most (or innermost) term and working to the left. At finish it multiplies the iteration result with the term group in large parentheses in the function series.

Ten iterations were found good enough for Bessel Function arguments less than unity. Forty iterations are good for argument values up to 20. Accuracy was equal to those in reference [5] for argument values up to 17.5 and 15 decimal digit accuracy.

Subroutine BESSEL's listing is only a partial one. The original was in FORTRAN using Double Precision floating-point variables. The calling routine has the Data array GAMMA(9), supplies ARGUMENT (Bessel argument) and NORDER (Bessel order). The subroutine returns BESS as the Bessel Function value.

---

[17] Description by a calculus instructor, far too long ago...

```
* SUBROUTINE BESSEL
*
* Calling routine supplies ARGUMENT, NORDER, receives BESS
*
* Data array declared in Main program:
*     GAMMA(1) = 2.0          GAMMA(2) = 8.0
*     GAMMA(3) = 48.0         GAMMA(4) = 384.0
*     GAMMA(5) = 3840.0       GAMMA(6) = 46080.0
*     GAMMA(7) = 645120.0       GAMMA(8) = 10320920.0
*     GAMMA(9) = 185794560.0
*
* Initialize conditions
*
      BRACKET = 1.0
      ARGSQUAR = ARGUMENT * ARGUMENT
*
* Do the "bracket loop" 40 times counting down
*
      FOR ITER = 40,1,-1
         NDIV = ITER + ITER
         NDIV = NDIV *(NORDER + NORDER + NDIV)
         BRACKET = 1.0 - (BRACKET*ARGSQUAR / NDIV)
      NEXT ITER
*
* Finish up with outermost term
*
      BESS = BRACKET*(ARGUMENT^NORDER)/GAMMA(NORDER)
*
      RETURN
*
* Note: If NORDER is an integer variable it may have to have
* a FLOAT(NORDER) intrinsic conversion statement in order to
* operate with a floating-point variable.  Consult language
* being used for proper syntax.
```

The GAMMA( ) data array was initialized in the main program to avoid needless calculations. It is rather easy to hand-calculate having successive integer multipliers in the progression of 4, 6, 8, 10, etc. Since this subroutine was intended for a maximum order of 9, and the lowest argument of the Gamma function is the order-plus-one, the value of Gamma(1) was omitted and all other values "moved down one." Subroutine operations are still exact as to the equation statements.

Temporary variable NDIV is used for convenience and the reduction of variable use. Reduction of the number of steps, statements, and variable use reduces the total execution time. Another temporary, ARGSQUAR, reduces the number of multiplications of ARGUMENT within the loop (it must be squared to obtain BRACKET). A raise-to-a-power statement almost always invokes an internal series polynomial approximation subroutine that involves many more arithmetic operations than multiplying a variable by itself. While it isn't noticeable in a few tens of loop iterations as here, a few thousands of iterations in another routine will be noticed.

References for Chapter 4

[4] "Electronic Designers' Handbook," by Robert W. Landee, Donovan C. Davis, Albert P. Albrecht, 1957, McGraw-Hill Book Company, Inc. In particular Chapter 5 on modulation.

[5} "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables," edited by Milton Abramowitz and Irene A. Stegun, National Bureau of Standards Applied Mathematics Series 55, 1964, U.S. Government Printing Office. For checking Bessel Function accuracy.

[6] "Theory and Problems of Laplace Transforms," by Murray R. Spiegel, Schaum's Outline Series, 1965, Schaum Publishing Co., New York. For definitions and identities to Bessel Function calculation series.

[7] "Reference Data for Radio Engineers," Fourth Edition, H. P. Westman, editor, 1956, International Telephone and Telegraph Corporation. In particular Chapter 35 on Fourier Waveform Analysis.

[8] Application Note 150-1, "Spectrum Analysis," April 1996, Hewlett-Packard Test and Measurement Division (now Agilent), Hewlett-Packard Company. While aimed at measurement of AM and FM modulation measurement with spectrum analyzers, it is also a good tutorial on modulation. Animation good only for older PCs.

[9] U.S. Army Field Manual FM 24-24 "Signal Data References: Signal Equipment," HQ Department of the Army, Washington, DC, 29 December 1994. This an old document but available. A better one is FM 6-02.53 "Tactical Radio Operations" dated 5 August 2009, more modern and up-to-date. Due to Internet restrictions as a result of the "9-11" attack on the USA, it is not available from the U.S. Army Training and Doctrine command at http://www.train.army.mil without proper computer authenticators. Some have been leaked and may be found by a Search feature on most browsers (as the author did).

[10] Advanced Television Systems Committee [ATSC], 1776 K Street NW, 8th Floor, Washington, DC, 20006-2304 (as of 2013). The ATSC was formed in 1982, has 200 members that represent the television industry of the USA, Canada, South Korea, Taiwan, and Argentina. The U.S. Federal Communications Commission adopted the ATSC's DTV characteristic in 1996. Canada and South Korea adopted the same in 1997 followed by Taiwan and Argentina in 1998. ATSC standards website is (as of 2013) http://www.atsc.org/cms/index.php/standards/standards?layer=default for all standards (PDF).