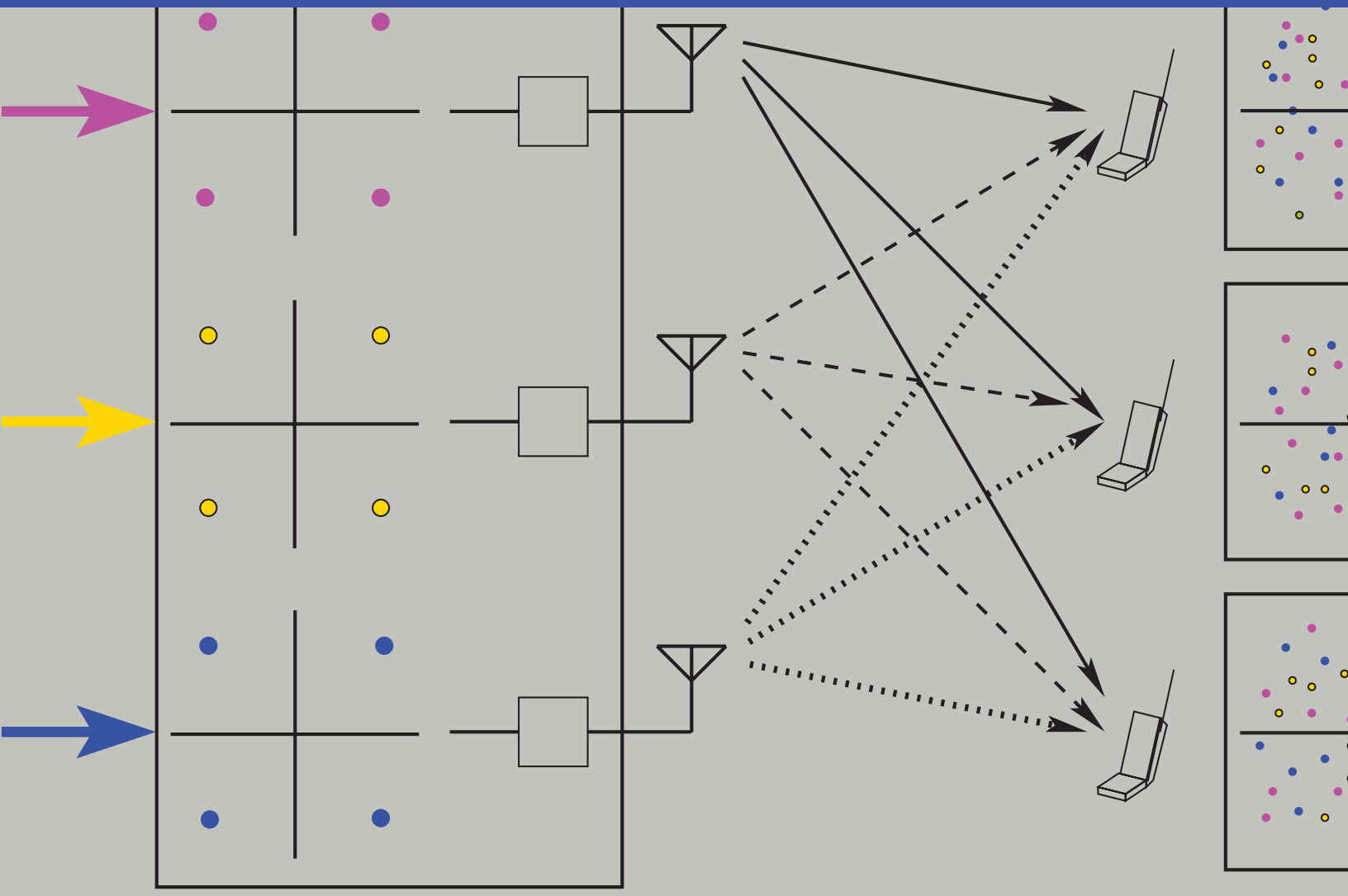
Volume I

MICROWAVE AND RF DESIGN

RADIO SYSTEMS



Michael Steer

Third Edition

Microwave and RF Design
Radio Systems

Volume 1

Third Edition

Michael Steer

Microwave and RF Design

Radio Systems

Volume 1

Third Edition

Michael Steer

Citation: Steer, Michael. *Microwave and RF Design: Radio Systems*. Volume 1. (Third Edition), NC State University, 2019. doi: https//doi.org/10.5149/9781469656915_Steer

ISBN 978-1-4696-5690-8 (paperback)
ISBN 978-1-4696-5691-5 (open access ebook)

Published by NC State University

NC STATE UNIVERSITY

Distributed by the University of North Carolina Press
www.uncpress.org

Printing: 1

To

Ross Lampe, in honor of his dedication to our profession

Preface

The book series *Microwave and RF Design* is a comprehensive treatment of radio frequency (RF) and microwave design with a modern "systems-first" approach. A strong emphasis on design permeates the series with extensive case studies and design examples. Design is oriented towards cellular communications and microstrip design so that lessons learned can be applied to real-world design tasks. The books in the Microwave and RF Design series are:

- Microwave and RF Design: Radio Systems, Volume 1
- Microwave and RF Design: Transmission Lines, Volume 2
- Microwave and RF Design: Networks, Volume 3
- Microwave and RF Design: Modules, Volume 4
- Microwave and RF Design: Amplifiers and Oscillators, Volume 5

The length and format of each is suitable for automatic printing and binding.

Rationale

The central philosophy behind this series's popular approach is that the student or practicing engineer will develop a full appreciation for RF and microwave engineering and gain the practical skills to perform system-level design decisions. Now more than ever companies need engineers with an ingrained appreciation of systems and armed with the skills to make system decisions. One of the greatest challenges facing RF and microwave engineering is the increasing level of abstraction needed to create innovative microwave and RF systems. This book series is organized in such a way that the reader comes to understand the impact that system-level decisions have on component and subsystem design. At the same time, the capabilities of technologies, components, and subsystems impact system design. The book series is meticulously crafted to intertwine these themes.

Audience

The book series was originally developed for three courses at North Carolina State University. One is a final-year undergraduate class, another an introductory graduate class, and the third an advanced graduate class. Books in the series are used as supplementary texts in two other classes. There are extensive case studies, examples, and end of chapter problems ranging from straight-forward to in-depth problems requiring hours to solve. A companion book, *Fundamentals of Microwave and RF Design*, is more suitable for an undergraduate class yet there is a direct linkage between the material in this book and the series which can then be used as a career-long reference text. I believe it is completely understandable for senior-level students where a microwave/RF engineering course is offered. The book series is a comprehensive RF and microwave text and reference, with detailed index, appendices, and cross-references throughout. Practicing engineers will find the book series a valuable systems primer, a refresher as needed, and a

reference tool in the field. Additionally, it can serve as a valuable, accessible resource for those outside RF circuit engineering who need to understand how they can work with RF hardware engineers.

Organization

This book is a volume in a five volume series on RF and microwave design. The first volume in the series, *Microwave and RF Design: Radio Systems*, addresses radio systems mainly following the evolution of cellular radio. A central aspect of microwave engineering is distributed effects considered in the second volume of this book series, *Microwave and RF Design: Transmission Lines*. Here transmission lines are treated as supporting forward- and backward-traveling voltage and current waves and these are related to electromagnetic effects. The third volume, *Microwave and RF Design: Networks*, covers microwave network theory which is the theory that describes power flow and can be used with transmission line effects. Topics covered in *Microwave and RF Design: Modules*, focus on designing microwave circuits and systems using modules introducing a large number of different modules. Modules is just another term for a network but the implication is that is is packaged and often available off-the-shelf. Other topics that are important in system design using modules are considered including noise, distortion, and dynamic range. Most microwave and RF designers construct systems using modules developed by other engineers who specialize in developing the modules. Examples are filter and amplifier modules which once designed can be used in many different systems. Much of microwave design is about maximizing dynamic range, minimizing noise, and minimizing DC power consumption. The fifth volume in this series, *Microwave and RF Design: Amplifiers and Oscillators*, considers amplifier and oscillator design and develops the skills required to develop modules.

Volume 1: Microwave and RF Design: Radio Systems

The first book of the series covers RF systems. It describes system concepts and provides comprehensive knowledge of RF and microwave systems. The emphasis is on understanding how systems are crafted from many different technologies and concepts. The reader gains valuable insight into how different technologies can be traded off in meeting system requirements. I do not believe this systems presentation is available anywhere else in such a compact form.

Volume 2: Microwave and RF Design: Transmission Lines

This book begins with a chapter on transmission line theory and introduces the concepts of forward- and backward-traveling waves. Many examples are included of advanced techniques for analyzing and designing transmission line networks. This is followed by a chapter on planar transmission lines with microstrip lines primarily used in design examples. Design examples illustrate some of the less quantifiable design decisions that must be made. The next chapter describes frequency-dependent transmission line effects and describes the design choices that must be taken to avoid multimoding. The final chapter in this volume addresses coupled-lines. It is shown how to design coupled-line networks that exploit this distributed effect to realize novel circuit functionality and how to design networks that minimize negative effects. The modern treatment of transmission lines in this volume emphasizes planar circuit design and the practical aspects of designing

around unwanted effects. Detailed design of a directional coupler is used to illustrate the use of coupled lines. Network equivalents of coupled lines are introduced as fundamental building blocks that are used later in the synthesis of coupled-line filters. The text, examples, and problems introduce the often hidden design requirements of designing to mitigate parasitic effects and unwanted modes of operation.

Volume 3: Microwave and RF Design: Networks

Volume 3 focuses on microwave networks with descriptions based on S parameters and $ABCD$ matrices, and the representation of reflection and transmission information on polar plots called Smith charts. Microwave measurement and calibration technology are examined. A sampling of the wide variety of microwave elements based on transmission lines is presented. It is shown how many of these have lumped-element equivalents and how lumped elements and transmission lines can be combined as a compromise between the high performance of transmission line structures and the compactness of lumped elements. This volume concludes with an in-depth treatment of matching for maximum power transfer. Both lumped-element and distributed-element matching are presented.

Volume 4: Microwave and RF Design: Modules

Volume 4 focuses on the design of systems based on microwave modules. The book considers the wide variety of RF modules including amplifiers, local oscillators, switches, circulators, isolators, phase detectors, frequency multipliers and dividers, phase-locked loops, and direct digital synthesizers. The use of modules has become increasingly important in RF and microwave engineering. A wide variety of passive and active modules are available and high-performance systems can be realized cost effectively and with stellar performance by using off-the-shelf modules interconnected using planar transmission lines. Module vendors are encouraged by the market to develop competitive modules that can be used in a wide variety of applications. The great majority of RF and microwave engineers either develop modules or use modules to realize RF systems. Systems must also be concerned with noise and distortion, including distortion that originates in supposedly linear elements. Something as simple as a termination can produce distortion called passive intermodulation distortion. Design techniques are presented for designing cascaded systems while managing noise and distortion. Filters are also modules and general filter theory is covered and the design of parallel coupled line filters is presented in detail. Filter design is presented as a mixture of art and science. This mix, and the thought processes involved, are emphasized through the design of a filter integrated throughout this chapter.

Volume 5: Microwave and RF Design: Amplifiers and Oscillators

The fifth volume presents the design of amplifiers and oscillators in a way that enables state-of-the-art designs to be developed. Detailed strategies for amplifiers and voltage-controlled oscillators are presented. Design of competitive microwave amplifiers and oscillators are particularly challenging as many trade-offs are required in design, and the design decisions cannot be reduced to a formulaic flow. Very detailed case studies are presented and while some may seem quite complicated, they parallel the level of sophistication required to develop competitive designs.

Case Studies

A key feature of this book series is the use of real world case studies of leading edge designs. Some of the case studies are designs done in my research group to demonstrate design techniques resulting in leading performance. The case studies and the persons responsible for helping to develop them are as follows.

1. Software defined radio transmitter.
2. High dynamic range down converter design. This case study was developed with Alan Victor.
3. Design of a third-order Chebyshev combline filter. This case study was developed with Wael Fathelbab.
4. Design of a bandstop filter. This case study was developed with Wael Fathelbab.
5. Tunable Resonator with a varactor diode stack. This case study was developed with Alan Victor.
6. Analysis of a 15 GHz Receiver. This case study was developed with Alan Victor.
7. Transceiver Architecture. This case study was developed with Alan Victor.
8. Narrowband linear amplifier design. This case study was developed with Dane Collins and National Instruments Corporation.
9. Wideband Amplifier Design. This case study was developed with Dane Collins and National Instruments Corporation.
10. Distributed biasing of differential amplifiers. This case study was developed with Wael Fathelbab.
11. Analysis of a distributed amplifier. This case study was developed with Ratan Bhatia, Jason Gerber, Tony Kwan, and Rowan Gilmore.
12. Design of a WiMAX power amplifier. This case study was developed with Dane Collins and National Instruments Corporation.
13. Reflection oscillator. This case study was developed with Dane Collins and National Instruments Corporation.
14. Design of a C-Band VCO. This case study was developed with Alan Victor.
15. Oscillator phase noise analysis. This case study was developed with Dane Collins and National Instruments Corporation.

Many of these case studies are available as captioned YouTube videos and qualified instructors can request higher resolution videos from the author.

Course Structures

Based on the adoption of the first and second editions at universities, several different university courses have been developed using various parts of what was originally one very large book. The book supports teaching two or three classes with courses varying by the selection of volumes and chapters. A standard microwave class following the format of earlier microwave texts can be taught using the second and third volumes. Such a course will benefit from the strong practical design flavor and modern treatment of measurement technology, Smith charts, and matching networks. Transmission line propagation and design is presented in the context of microstrip technology providing an immediately useful skill. The subtleties of multimoding are also presented in the context of microstrip lines. In such

a class the first volume on microwave systems can be assigned for self-learning.

Another approach is to teach a course that focuses on transmission line effects including parallel coupled-line filters and module design. Such a class would focus on Volumes 2, 3 and 4. A filter design course would focus on using Volume 4 on module design. A course on amplifier and oscillator design would use Volume 5. This course is supported by a large number of case studies that present design concepts that would otherwise be difficult to put into the flow of the textbook.

Another option suited to an undergraduate or introductory graduate class is to teach a class that enables engineers to develop RF and microwave systems. This class uses portions of Volumes 2, 3 and 4. This class then omits detailed filter, amplifier, and oscillator design.

The fundamental philosophy behind the book series is that the broader impact of the material should be presented first. Systems should be discussed up front and not left as an afterthought for the final chapter of a textbook, the last lecture of the semester, or the last course of a curriculum.

The book series is written so that all electrical engineers can gain an appreciation of RF and microwave hardware engineering. The body of the text can be covered without strong reliance on this electromagnetic theory, but it is there for those who desire it for teaching or reader review. The book is rich with detailed information and also serves as a technical reference.

The Systems Engineer

Systems are developed beginning with fuzzy requirements for components and subsystems. Just as system requirements provide impetus to develop new base technologies, the development of new technologies provides new capabilities that drive innovation and new systems. The new capabilities may arise from developments made in support of other systems. Sometimes serendipity leads to the new capabilities. Creating innovative microwave and RF systems that address market needs or provide for new opportunities is the most exciting challenge in RF design. The engineers who can conceptualize and architect new RF systems are in great demand. This book began as an effort to train RF systems engineers and as an RF systems resource for practicing engineers. Many RF systems engineers began their careers when systems were simple. Today, appreciating a system requires higher levels of abstraction than in the past, but it also requires detailed knowledge or the ability to access detailed knowledge and expertise. So what makes a systems engineer? There is not a simple answer, but many partial answers. We know that system engineers have great technical confidence and broad appreciation for technologies. They are both broad in their knowledge of a large swath of technologies and also deep in knowledge of a few areas, sometimes called the “T” model. One book or course will not make a systems engineer. It is clear that there must be a diverse set of experiences. This book series fulfills the role of fostering both high-level abstraction of RF engineering and also detailed design skills to realize effective RF and microwave modules. My hope is that this book will provide the necessary background for the next generation of RF systems engineers by stressing system principles immediately, followed by core RF technologies. Core technologies are thereby covered within the context of the systems in which they are used.

Supplementary Materials

Supplementary materials available to qualified instructors adopting the book include PowerPoint slides and solutions to the end-of-chapter problems. Requests should be directed to the author. Access to downloads of the books, additional material and YouTube videos of many case studies are available at `https://www.lib.ncsu.edu/do/open-education`

Acknowledgments

Writing this book has been a large task and I am indebted to the many people who helped along the way. First I want to thank the more than 1200 electrical engineering graduate students who used drafts and the first two editions at NC State. I thank the many instructors and students who have provided feedback. I particularly thank Dr. Wael Fathelbab, a filter expert, who co-wrote an early version of the filter chapter. Professor Andreas Cangellaris helped in developing the early structure of the book. Many people have reviewed the book and provided suggestions. I thank input on the structure of the manuscript: Professors Mark Wharton and Nuno Carvalho of Universidade de Aveiro, Professors Ed Delp and Saul Gelfand of Purdue University, Professor Lynn Carpenter of Pennsylvania State University, Professor Grant Ellis of the Universiti Teknologi Petronas, Professor Islam Eshrah of Cairo University, Professor Mohammad Essaaidi and Dr. Otman Aghzout of Abdelmalek Essaadi Univeristy, Professor Jianguo Ma of Guangdong University of Technology, Dr. Jayesh Nath of Apple, Mr. Sony Rowland of the U.S. Navy, and Dr. Jonathan Wilkerson of Lawrence Livermore National Laboratories, Dr. Josh Wetherington of Vadum, Dr. Glen Garner of Vadum, and Mr. Justin Lowry who graduated from North Carolina State University.

Many people helped in producing this book. In the first edition I was assisted by Ms. Claire Sideri, Ms. Susan Manning, and Mr. Robert Lawless who assisted in layout and production. The publisher, task master, and chief coordinator, Mr. Dudley Kay, provided focus and tremendous assistance in developing the first and second editions of the book, collecting feedback from many instructors and reviewers. I thank the Institution of Engineering and Technology, who acquired the original publisher, for returning the copyright to me. This open access book was facilitated by John McLeod and Samuel Dalzell of the University of North Carolina Press, and by Micah Vandergrift and William Cross of NC State University Libraries. The open access ebooks are host by NC State University Libraries.

The book was produced using LaTeX and open access fonts, line art was drawn using xfig and inkscape, and images were edited in gimp. So thanks to the many volunteers who developed these packages.

My family, Mary, Cormac, Fiona, and Killian, gracefully put up with my absence for innumerable nights and weekends, many more than I could have ever imagined. I truly thank them. I also thank my academic sponsor, Dr. Ross Lampe, Jr., whose support of the university and its mission enabled me to pursue high risk and high reward endeavors including this book.

Michael Steer
North Carolina State University
Raleigh, North Carolina
mbs@ncsu.edu

List of Trademarks

Contents

CHAPTER 1

Introduction to RF and Microwave Systems

1.1 Introduction

Radio frequency (RF) systems drive the requirements of microwave and RF circuits, and the capabilities of RF and microwave circuits fuel the evolution of RF systems. This interdependence and the trade-offs required necessitate that the successful RF and microwave designer have an appreciation of systems. Today, communications is the main driver of RF system development, leading to RF technology evolution at an unprecedented pace. Similar relationships exist for national security including radar and sensors used in detection and ranging. Other radio systems have less immediate impact on RF technology but are very important for the smaller number of RF engineers working in the fields of navigation, astronomy, defense, and heating. No longer can many years be put aside for methodical trade-offs of circuit complexity, technology development, and architecture choices at the system level. As relationships have become more intertwined, RF communication, radar, and sensor engineers must develop a broad appreciation of technology, communication principles, and circuit design.

This book is the first volume in a series on microwave and RF design. A central aspect of microwave engineering is distributed effects considered in the second volume of his book series [1]. Here the transmission lines are treated as supporting forward and backward traveling voltage and current waves and these are related to electromagnetic effects. The third volume [2] covers microwave network theory which is the theory that describe power flow and can be used to describe transmission line effects. Topics covered in this volume include scattering parameters, Smith charts, and

Table 1-1: Broad electromagnetic spectrum divisions.

Name or band	Frequency	Wavelength
Radio frequency	3 Hz – 300 GHz	100 000 km – 1 mm
Microwave	300 MHz – 300 GHz	1 m – 1 mm
Millimeter (mm) band	110 – 300 GHz	2.7 mm – 1.0 mm
Infrared	300 GHz – 400 THz	1 mm – 750 nm
Far infrared	300 GHz – 20 THz	1 mm – 15 μm
Long-wavelength infrared	20 THz – 37.5 THz	15–8μm
Mid-wavelength infrared	37.5 – 100 THz	8–3 μm
Short-wavelength infrared	100 THz – 214 THz	3–1.4 μm
Near infrared	214 THz – 400 THz	1.4 μm – 750 nm
Visible	400 THz – 750 THz	750 – 400 mm
Ultraviolet	750 THz – 30 PHz	400 – 10 nm
X-Ray	30 PHz – 30 EHz	10 – 0.01 nm
Gamma Ray	> 15 EHz	< 0.02 nm

Gigahertz, GHz = 10^9 Hz; terahertz, THz = 10^{12} Hz; pentahertz, PHz = 10^{15} Hz; exahertz, EHz = 10^{18} Hz.

matching networks that enable maximum power transfer. The fourth volume [3] focuses on designing microwave circuits and systems using modules introducing a large number of different modules. Modules is just another term for a network but the implication is that is is packaged and often available off-the-shelf. Other topics in this chapter that are important in system design using modules are considered including noise, distortion, and dynamic range. Most microwave and RF designers construct systems using modules developed by other engineers who specialize in developing the modules. Examples are filter and amplifier chip modules which once designed can be used in many different systems. Much of microwave design is about maximizing dynamic range, minimizing noise, and minimizing DC power consumption. The fifth volume in this series [4] considers amplifier and oscillator design and develops the skills required to develop modules.

The books in the Microwave and RF Design series are:

- Microwave and RF Design: Radio Systems
- Microwave and RF Design: Transmission Lines
- Microwave and RF Design: Networks
- Microwave and RF Design: Modules
- Microwave and RF Design: Amplifiers and Oscillators

1.2 RF and Microwave Engineering

An RF signal is a signal that is coherently generated, radiated by a transmit antenna, propagated through air or space, collected by a receive antenna, and then amplified and information extracted. An RF circuit operates at the same frequency as the RF signal that is transmitted or received. That is, the frequency at which a circuit operates does not define that it is an RF circuit. The RF spectrum is part of the electromagnetic (EM) spectrum exploited by humans for communications. A broad categorization of the EM spectrum is shown in Table 1-1. Today radios operate from 3 Hz to 300 GHz, although the upper end will increase as technology progresses.

Microwaves refers to the frequencies where the size of a circuit or structure is comparable to or greater than the wavelength of the EM signal. The division is arbitrary but if a circuit structure is greater than $1/20$ of a wavelength, then most engineers would regard the circuit as being a

microwave circuit. For now the microwave frequency range is generally taken as 300 MHz to 300 GHz. At these frequencies distributed effects, sometimes called transmission line effects, must be considered.

One of the key characteristics distinguishing RF signals from infrared and visible light is that an RF signal can be generated with coherent phase, and information can be transmitted in both amplitude and phase variations of the RF signal. Such signals can be easily generated up to 220 GHz. The necessary hardware becomes progressively more expensive as frequency increases. The upper limit of radio frequencies is about 300 GHz today, but the limit is extending slowly above this as technology progresses.

The RF bands are listed in Table 1-2 along with propagation modes and representative applications. The propagation of RF signals in free space follows one or more paths from a transmitter to a receiver at any frequency, with differences being in the size of the antennas needed to transmit and receive signals. The size of the necessary antenna is related to wavelength, with the typical dimensions ranging from a quarter of a wavelength to a few wavelengths if a reflector is used to focus the EM waves. On earth, and dependent on frequency, RF signals propagate through walls, diffract around objects, refract when the dielectric constant of the medium changes, and reflect from buildings and walls. The extent is dependent on frequency. Above ground the propagation at RF is affected by atmospheric loss, by charge layers induced by solar radiation in the upper atmosphere, and by density variations in the air caused by heating as well as the thinning of air with height above the earth. The ionosphere is the uppermost part of the atmosphere, at 60 to 90 km, and is ionized by solar radiation, producing a reflective surface, called the D layer, to radio signals up to 3 GHz. The D layer weakens at night and most radio signals can then pass through this weakened layer. The E layer extends from 90 to 120 km and is ionized by X-rays and extreme ultraviolet radiation, and the ionized regions, which reflect RF signals, form ionized clouds that last only a few hours. The F layer of the ionosphere extends from 200 to 500 km and ionization in this layer is due to extreme ultraviolet radiation. Refraction results from this charged layer rather than reflection, as the charges are widely separated. At night the F layer results in what is called the skywave, which is the refraction of radio waves around the earth. At low frequencies a radio wave penetrates the earth's surface and the wave can become trapped at the interface between two regions of different permittivity, the earth region and the air region. This radio wave is called the **surface wave** or **ground wave**.

Propagating RF signals in air are absorbed by molecules in the atmosphere primarily by molecular resonances such as the bending and stretching of bonds. This bending and stretching converts energy in the EM wave into vibrational energy of the molecules. The transmittance of radio signals versus frequency in dry air at an altitude of 4.2 km is shown in Figure 1-1. The first molecular resonance encountered in dry air as frequency increases is the oxygen resonance, centered at 60 GHz, but below that the absorption in dry air is very small. Attenuation increases with higher water vapor pressure (with a resonance at 22 GHz) and in rain. Within 2 GHz of 60 GHz a signal will not travel far, and this can be used to provide localized communication over a few meters as a local data link. Regions of low attenuation (i.e. high transmittance), are called windows and there are numerous low loss windows.

When a radio signal near 60 GHz passes through air an oxygen molecule of two bound oxygen atoms vibrates and EM energy is transferred to the mechanical energy of vibration and thus heat.

Table 1-2: Radio frequency bands, primary propagation mechanisms, and selected applications.

Band		Frequency wavelength	Propagation mode/applications
TLF	Tremendously low frequency	< 3 Hz $> 100\,000$ km	Penetration of liquids and solids/Submarine communication
ELF	Extremely low frequency	3–30 Hz, 100 000– 10 000 km	Penetration of liquids and solids/Submarine communication
SLF	Super low frequency	30–300 Hz 10 000–1 000 km	Penetration of liquids and solids/Submarine communication
ULF	Ultra low frequency	300–3000 Hz 1 000–100 km	Penetration of liquids and solids/Submarine communication; communication within mines
VLF	Very low frequency	3–30 kHz 100–10 km	Guided wave trapped between the earth and the ionosphere/Navigation, geophysics
LF	Low frequency	30–300 kHz 10–1 km	Guided wave between the earth and the ionosphere's D layer; surface waves, building penetration/Navigation, AM broadcast, amateur radio, time signals, RFID
MF	Medium frequency	300–3000 kHz 1000–100 m	Surface wave, building penetration; day time: guided wave between the earth and the ionosphere's D layer; night time: sky wave/AM broadcast
HF	High frequency	3–30 MHz 100–10 m	Sky wave, building penetration/shortwave broadcast, over-the-horizon radar, RFID, amateur radio, marine and mobile telephony
VHF	Very high frequency	30–300 MHz 10–1 m	Line of sight, building penetration; up to 80 MHz, skywave during periods of high sunspot activity/FM and TV broadcast, weather radio, line-of-sight aircraft communications
UHF	Ultra high frequency	300–3000 MHz 10–1 cm	Line of sight, building penetration; sometimes tropospheric ducting/1G–4G cellular communications, RFID, microwave ovens, radio astronomy, satellite-based navigation
SHF	Super high frequency	3–30 GHz 10–1 cm	Line of sight/5G cellular communications, Radio astronomy, point-to-point communications, wireless local area networks, radar
EHF	Extremely high frequency	30–300 GHz 10–1 mm	Line of sight/5G cellular communications, Astronomy, remote sensing, point-to-point and satellite communications
THF	Terahertz or tremendously high frequency	300–3000 GHz 1000–100 μm	Line of sight /Spectroscopy, imaging

RF signals diffract and so can bend around structures and penetrate into valleys. The ability to diffract reduces with increasing frequency. However, as frequency increases the size of antennas decreases and the capacity to carry information increases. A very good compromise for mobile communications is at UHF, 300 MHz to 4 GHz, where antennas are of convenient size and there is a good ability to diffract around objects and even penetrate walls. This choice can be seen with 1G–4G cellular communication systems operating in several bands from 450 MHz to 3.6 GHz where antennas do not dominate the size of the handset, and the ability to receive calls within buildings and without line of sight to the base station is well known.

RF bands have been further divided for particular applications. The

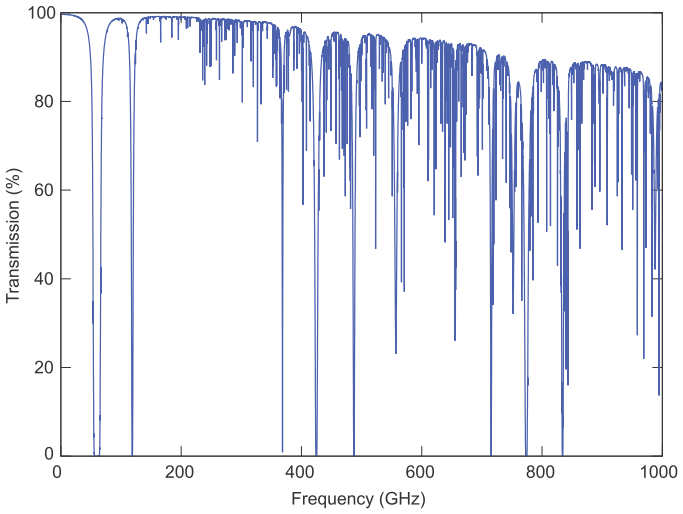


Figure 1-1: Atmospheric transmission at Mauna Kea, with a height of 4.2 km, on the Island of Hawaii where the atmospheric pressure is 60% of that at sea level and the air is dry with a precipitable water vapor level of 0.001 mm. After [5].

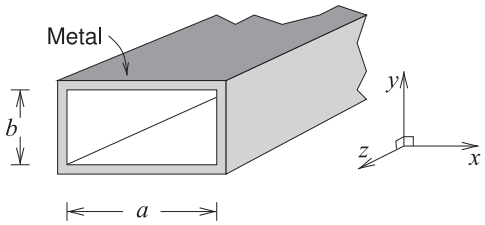


Figure 1-2: Rectangular waveguide with internal dimensions of a and b. Usually $a \approx b$. The EM waves are confined within the four metal walls and propagate in the $\pm z$ direction. Little current flows in the waveguide walls and so resistive losses are small. Compared to coaxial lines rectangular waveguides have very low loss.

frequency bands for radar are shown in Table 1-3. The L, S, and C bands are referred to as having **octave bandwidths**, as the upper frequency of a band is twice the lower frequency. The other bands are half-octave bands, as the upper frequency limit is approximately 50% higher than the lower frequency limit. The same letter band designations are used by other standards. The most important alternative band designation is for the waveguide bands. These bands refer to the useful range of operation of a rectangular waveguide, which is a rectangular tube that confines a propagating signal within four conducting walls (see Figure 1-2). The waveguide bands are shown in Table 1-4 with the conventional letter designation of bands and standardized waveguide dimensions. Compared to coaxial lines rectangular waveguides have very low loss.[1]

1.3 Communication Over Distance

Communicating using EM signals has been an integral part of society since the transmission of the first telegraph signals over wires in the mid 19th century [7]. This development derived from an understanding of magnetic induction based on the experiments of Faraday in 1831 [8] in which he investigated the relationship of magnetic fields and currents. This work of Faraday is now known as Faraday's law, or Faraday's law of induction. It was one of four key laws developed between 1820 and 1835 that described the interaction of static fields and of static fields with currents. These four

[1] A semirigid coaxial line with an outer conductor diameter of 3.5 mm has a loss at 10 GHz of 0.5 db/m while an X-band waveguide has a loss of 0.1 dB/m. At 100 GHz a 1 mm-diameter coaxial line has a loss of 12.5 dB/m compared to 2.5 dB/m loss for a W-band waveguide.

Table 1-3: IEEE radar bands [6]. The mm band designation is also used when the intent is to convey general information above 30 GHz.

Band	Frequency range
L "long"	1–2 GHz
S "short"	2–4 GHz
C "compromise"	4–8 GHz
K_u "kurtz under"	12–18 GHz
K "kurtz" (short in German)	18–27 GHz
K_a "kurtz above"	27–40 GHz
V	40–75 GHz
W	75–110 GHz
F	90–140 GHz
D	110–170 GHz
mm	110–300 GHz

In Table 1-4 the waveguide dimensions are specified in inches (use 25.4 mm/inch to convert to mm). The number in the WR designation is the long internal dimension of the waveguide in hundredths of an inch. The EIA is the U.S.-based Electronics Industry Association. Note that the radar band (see Table 1-3) and waveguide band designations do not necessarily coincide.

Table 1-4: Selected waveguide bands with operating frequencies and internal dimensions (refer to Figure 1-2).

Band	EIA waveguide band	Operating frequency (GHz)	Internal dimensions ($a \times b$, inches)
R	WR-430	1.70–2.60	4.300×2.150
D	WR-340	2.20–3.30	3.400×1.700
S	WR-284	2.60–3.95	2.840×1.340
E	WR-229	3.30–4.90	2.290×1.150
G	WR-187	3.95–5.85	1.872×0.872
F	WR-159	4.90–7.05	1.590×0.795
C	WR-137	5.85–8.20	1.372×0.622
H	WR-112	7.05–10.00	1.122×0.497
X	WR-90	8.2–12.4	0.900×0.400
Ku	WR-62	12.4–18.0	0.622×0.311
K	WR-51	15.0–22.0	0.510×0.255
K	WR-42	18.0–26.5	0.420×0.170
Ka	WR-28	26.5–40.0	0.280×0.140
Q	WR-22	33–50	0.224×0.112
U	WR-19	40–60	0.188×0.094
V	WR-15	50–75	0.148×0.074
E	WR-12	60–90	0.122×0.061
W	WR-10	75–110	0.100×0.050
F	WR-8	90–140	0.080×0.040
D	WR-6	110–170	0.0650×0.0325
G	WR-5	140–220	0.0510×0.0255

laws are the Biot–Savart law (developed around 1820), Ampere's law (1826), Faraday's law (1831), and Gauss's law (1835). These are all static laws and do not describe propagating fields.

1.3.1 Electromagnetic Fields

We now know that there are two components of the EM field, the **electric field**, E, with units of volts per meter (V/m), and the **magnetic field**, H, with units of amperes per meter (A/m). E and H fields together describe the force between charges. There are also two flux quantities that are necessary to understand the interactions between these fields and vacuum or matter. The first is D, the **electric flux** density, with units of coulombs per square meter (C/m^2), and the other is B, the **magnetic flux** density, with units of teslas (T). B and H, and D and E, are related to each other by the properties of the medium, which are embodied in the quantities μ and ε (with the caligraphic letter, e.g. $\mathcal{B}$, denoting a time-domain quantity):

$$\bar{\mathcal{B}} = \mu\bar{\mathcal{H}} \qquad (1.1) \qquad\qquad \bar{\mathcal{D}} = \varepsilon\bar{\mathcal{E}}, \qquad (1.2)$$

where the over bar denotes a vector quantity. The quantity μ is called the **permeability** of the medium and describes the ability to store **magnetic energy** in a region. The permeability in free space (or vacuum) is denoted $\mu_0 = 4\pi \times 10^{-7}$ H/m and the magnetic flux and magnetic field are related as

$$\bar{\mathcal{B}} = \mu_0\bar{\mathcal{H}}. \qquad (1.3)$$

The other material quantity is the **permittivity**, ε, which describes the ability to store energy in a volume and in a vacuum

$$\bar{\mathcal{D}} = \varepsilon_0 \bar{\mathcal{E}}, \tag{1.4}$$

where $\varepsilon_0 = 8.854 \times 10^{-12}$ F/m is the permittivity of a vacuum. The **relative permittivity**, ε_r, is the ratio the permittivity of a material to that of vacuum:

$$\varepsilon_r = \varepsilon/\varepsilon_0. \tag{1.5}$$

Similarly, the **relative permeability**, μ_r, refers to the ratio of permeability of a material to its value in a vacuum:

$$\mu_r = \mu/\mu_0. \tag{1.6}$$

1.3.2 Biot-Savart Law

The Biot–Savart law relates current to magnetic field as, see Figure 1-3,

$$d\bar{H} = \frac{I d\boldsymbol{\ell} \times \hat{a}_R}{4\pi R^2}, \tag{1.7}$$

which has the units of amperes per meter in the SI system. In Equation (1.7) $d\bar{H}$ is the incremental static H field, I is current, $d\boldsymbol{\ell}$ is the vector of the length of a filament of current I, $\hat{a}_R$ is the unit vector in the direction from the current filament to the magnetic field, and R is the distance between the filament and the magnetic field. The $d\bar{H}$ field is directed at right angles to $\hat{a}_R$ and the current filament. So Equation (1.7) says that a filament of current produces a magnetic field at a point. The total magnetic field from a current on a wire or surface can be found by modeling the wire or surface as a number of current filaments, and the total magnetic field at a point is obtained by integrating the contributions from each filament.

1.3.3 Faraday's Law of Induction

Faraday's law relates a time-varying magnetic field to an induced voltage drop, V, around a closed path, which is now understood to be $\oint_\ell \bar{\mathcal{E}} \cdot d\boldsymbol{\ell}$, that is, the closed contour integral of the electric field,

$$V = \oint_\ell \bar{\mathcal{E}} \cdot d\boldsymbol{\ell} = -\oint_s \frac{\partial \bar{\mathcal{B}}}{\partial t} \cdot d\mathbf{s}, \tag{1.8}$$

and this has the units of volts in the SI unit system. The operation described in Equation (1.8) is illustrated in Figure 1-4.

1.3.4 Ampere's Circuital Law

Ampere's circuital law, often called just Ampere's law, relates direct current and the static magnetic field $\bar{\mathcal{H}}$. The relationship is based on Figure 1-5 and Ampere's circuital law is

$$\oint_\ell \bar{H} \cdot d\boldsymbol{\ell} = I_{\text{enclosed}}. \tag{1.9}$$

That is, the integral of the magnetic field around a loop is equal to the current enclosed by the loop. Using symmetry, the magnitude of the magnetic field at a distance r from the center of the wire shown in Figure 1-5 is

$$H = |I|/(2\pi r). \tag{1.10}$$

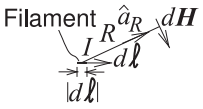


Figure 1-3: Diagram illustrating the Biot-Savart law. The law relates a static filament of current to the incremental H field at a distance.

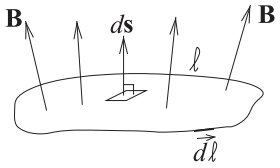


Figure 1-4: Diagram illustrating Faraday's law. The contour ℓ encloses the surface.

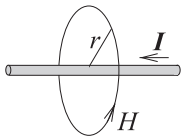


Figure 1-5: Diagram illustrating Ampere's law. Ampere's law relates the current, I, on a wire to the magnetic field around it, H.

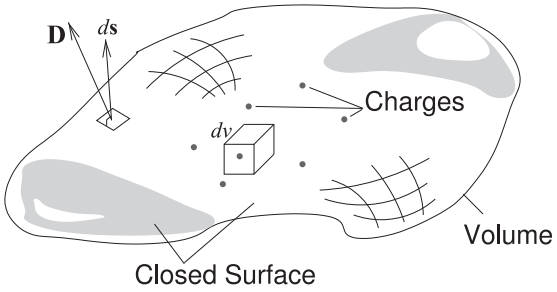


Figure 1-6: Diagram illustrating Gauss's law. Charges are distributed in the volume enclosed by the closed surface. An incremental area is described by the vector $d\mathbf{S}$, which is normal to the surface and whose magnitude is the area of the incremental area.

1.3.5 Gauss's Law

The final static EM law is Gauss's law, which relates the static electric flux density vector, $\bar{D}$, to charge. With reference to Figure 1-6, Gauss's law in integral form is

$$\oint_s \bar{D} \cdot d\mathrm{s} = \int_v \rho_v \cdot dv = Q_{\text{enclosed}}. \tag{1.11}$$

This states that the integral of the electric flux vector, $\bar{D}$, over a closed surface is equal to the total charge enclosed by the surface, Q_{enclosed}.

1.3.6 Gauss's Law of Magnetism

Gauss's law of magnetism parallels Gauss's law which now applies to magnetic fields. In integral form the law is

$$\oint_s \bar{B} \cdot d\mathrm{s} = 0. \tag{1.12}$$

This states that the integral of the magnetic flux vector, $\bar{D}$, over a closed surface is zero reflecting the fact that magnetic charges do not exist.

1.3.7 Telegraph

With the static field laws established, the stage was set to begin the development of the transmission of EM signals over wires. While traveling by ship back to the United States from Europe in 1832, Samuel Morse learned of Faraday's experiments and conceived of an EM telegraph. He sought out partners in Leonard Gale, a professor of science at New York University, and Alfred Vail, "skilled in the mechanical arts," who constructed the telegraph models used in their experiments. In 1835 this collaboration led to an experimental version transmitting a signal over 16 km of wire. Morse was not

Table 1-5: International Morse code.

Symbol	Code	Symbol	Code	Symbol	Code	Symbol	Code	Symbol	Code
1	.----	8	---..	E	.	L	.-..	S	...
2	..---	9	----.	F	..-.	M	--	T	-
3	...--	0	-----	G	--.	N	-.	U	..-
4	....-	A	.-	H	....	O	---	V	...-
5	.....	B	-...	I	..	P	.--.	W	.--
6	-....	C	-.-.	J	.---	Q	--.-	X	-..-
7	--...	D	-..	K	-.-	R	.-.	Y	-.--
								Z	--..

alone in imagining an EM telegraph, and in 1837 Charles Wheatstone opened the first commercial telegraph line between London and Camden Town, England, a distance of 2.4 km. Subsequently, in 1844, Morse designed and developed a line to connect Washington, DC, and Baltimore, Maryland. This culminated in the first public transmission on May 24, 1844, when Morse sent a telegraph message from the Capitol in Washington to Baltimore. This event is recognized as the birth of communication over distance using wires. This rapid pace of transition from basic research into electromagnetism (Faraday's experiment) to a fielded transmission system has been repeated many times in the evolution of wired and wireless communication technology.

The early telegraph systems used EM induction and multicell batteries that were switched in and out of circuit with the long telegraph wire and so created pulses of current. We now know that these current pulses created propagating magnetic fields that were guided by the wires and were accompanied by electric fields. In 1840 Morse applied for a U.S. patent for "Improvement in the Mode of Communicating Information by Signals by the Application of Electro-Magnetism Telegraph," which described "lightning wires" and "Morse code." By 1854, 37,000 km of telegraph wire crossed the United States, and this had a profound effect on the development of the country. Railroads made early extensive use of telegraph and a new industry was created. In the United States the telegraph industry was dominated by Western Union, which became one of the largest companies in the world. Just as with telegraph, the history of wired and wireless communication has been shaped by politics, business interests, market risk, entrepreneurship, patent ownership, and patent litigation as much as by the technology itself.

The first telegraph signals were just short bursts and slightly longer bursts of noise using Morse code in which sequences of dots, dashes, and pauses represent numbers and letters (see Table 1-5).[2] The speed of transmission was determined by an operator's ability to key and recognize the codes. Information transfer using EM signals in the late 19th century was therefore

[2] Morse code uses sequences of dots, dashes, and spaces. The duration of a dash (or "dah") is three times longer than that of a dot (or "dit"). Between letters there is a small gap. For example, the Morse code for PI is ".--. .." . Between words there is a slightly longer pause and between sentences an even longer pause. Table 1-5 lists the international Morse code adopted in 1848. The original Morse code developed in the 1830s is now known as "American Morse code" or "railroad code." The "modern international Morse code" extends the international Morse code with sequences for non-English letters and special symbols.

about 5 **bits per second** (**bits/s**). Morse achieved 10 words per minute.

1.3.8 The Origins of Radio

In the 1850s Morse began to experiment with wireless transmission, but this was still based on the principle of conduction. He used a flowing river, which as is now known is a medium rich with ions, to carry the charge. On one side of the river he set up a series connection of a metal plate, a battery, a Morse key, and a second metal plate. This formed the transmitter circuit. The metal plates were inserted into the water and separated by a distance considerably greater than the width of the river. On the other side of the river, metal plates were placed directly opposite the transmitter plates and this second set of plates was connected by a wire to a galvanometer in series. This formed the receive circuit, and electric pulses established by the transmitter resulted in the charge being transferred across the river by conduction and the pulses subsequently detected by the galvanometer. This was the first wireless transmission using electromagnetism, but it was not radio.

Morse relied entirely on conduction to achieve wireless transmission and it is now known that we need alternating electric and magnetic fields to propagate information over distance without charge carriers. The next steps in the progress to radio were experiments in induction. These culminated in an experiment by Loomis who in 1866 sent the first aerial wireless signals using kites flown by copper wires [9]. The transmitter kite had a Morse key at the ground end and an electric potential would have been developed between the ground and the kite itself. Closing the key resulted in current flow along the wire and this created a magnetic field that spread out and induced a current in the receive kite and this was detected by a galvanometer. However, not much of an electric field is produced and an EM wave is not transmitted. As such, the range of this system is very limited. Practical wireless communication requires an EM wave at a high-enough frequency that it can be efficiently generated by short wires.

1.3.9 Maxwell's Equations

The essential next step in the invention of radio was the development of Maxwell's equations in 1861. Before Maxwell's equations were postulated, several static EM laws were known. These are the Biot–Savart law, Ampere's circuital law, Gauss's law, and Faraday's law. Taken together they cannot describe the propagation of EM signals, but they can be derived from Maxwell's equations. Maxwell's equations cannot be derived from the static electric and magnetic field laws. Maxwell's equations embody additional insight relating spatial derivatives to time derivatives, which leads to a description of propagating fields. Maxwell's equations are

$$\nabla \times \bar{\mathcal{E}} = -\frac{\partial \bar{\mathcal{B}}}{\partial t} - \bar{\mathcal{M}} \qquad (1.13)$$

$$\nabla \cdot \bar{\mathcal{D}} = \rho_V \qquad (1.14)$$

$$\nabla \times \bar{\mathcal{H}} = \frac{\partial \bar{\mathcal{D}}}{\partial t} + \bar{\mathcal{J}} \qquad (1.15)$$

$$\nabla \cdot \bar{\mathcal{B}} = \rho_{mV}\,. \qquad (1.16)$$

Several of the quantities in Maxwell's equation have already been introduced, but now the electric and magnetic fields are in vector form. The other quantities in Equations (1.13)–(1.16) are

- $\bar{\mathcal{J}}$, the **electric current** density, with units of amperes per square meter (A/m^2);
- ρ_V, the **electric charge** density, with units of coulombs per cubic meter (C/m^3);
- ρ_{mV}, the magnetic charge density, with units of webers per cubic meter (Wb/m^3); and
- $\bar{\mathcal{M}}$, the magnetic current density, with units of volts per square meter (V/m^2).

Magnetic charges do not exist, but their introduction through the **magnetic charge** density, ρ_{mV}, and the **magnetic current** density, $\bar{\mathcal{M}}$, introduce an aesthetically appealing symmetry to Maxwell's equations. Maxwell's equations are differential equations, and as with most differential equations, their solution is obtained with particular boundary conditions, which in radio engineering are imposed by conductors. Electric conductors (i.e., electric walls) support electric charges and hence electric current. By analogy, magnetic walls support magnetic charges and magnetic currents. Magnetic walls also provide boundary conditions to be used in the solution of Maxwell's equations. The notion of magnetic walls is important in RF and microwave engineering, as they are approximated by the boundary between two dielectrics of different permittivity. The greater the difference in permittivity, the more closely the boundary approximates a magnetic wall.

Maxwell's equations are fundamental properties and there is no underlying theory, so they must be accepted "as is," but they have been verified in countless experiments. Maxwell's equations have three types of derivatives. First, there is the time derivative, $\partial/\partial t$. Then there are two spatial derivatives, $\nabla\times$, called **curl**, capturing the way a field circulates spatially (or the amount that it curls up on itself), and $\nabla\cdot$, called the **div** operator, describing the spreading-out of a field. In rectangular coordinates, curl, $\nabla\times$, describes how much a field circles around the x, y, and z axes. That is, the curl describes how a field circulates on itself. So Equation (1.13) relates the amount an electric field circulates on itself to changes of the B field in time. So a spatial derivative of electric fields is related to a time derivative of the magnetic field. Also in Equation (1.15) the spatial derivative of the magnetic field is related to the time derivative of the electric field. These are the key elements that result in self-sustaining propagation.

Div, $\nabla\cdot$, describes how a field spreads out from a point. So the presence of net electric charge (say, on a conductor) will result in the electric field spreading out from a point (see Equation (1.14)). In contrast, the magnetic field (Equation (1.16)) can never diverge from a point, which is a result of magnetic charges not existing (except when the magnetic wall approximation is used).

How fast a field varies with time, $\partial\bar{\mathcal{B}}/\partial t$ and $\partial\bar{\mathcal{D}}/\partial t$, depends on frequency. The more interesting property is how fast a field can change spatially, $\nabla\times\bar{\mathcal{E}}$ and $\nabla\times\bar{\mathcal{H}}$—this depends on wavelength relative to geometry. So if the cross-sectional dimensions of a transmission line are less than a wavelength ($\lambda/2$ or $\lambda/4$ in different circumstances), then it will be impossible for the fields to curl up on themselves and so there will be only one solution (with no or minimal spatial variation of the E and H fields) or, in some cases, no solution to Maxwell's equations.

1.3.10 Transmission of Radio Signals

Now the discussion returns to the technological development of radio. About the same time as Loomis's induction experiments in 1864, James Maxwell [10] laid the foundations of modern EM theory in 1861 [11]. Maxwell theorized that electric and magnetic fields are different manifestations of the same phenomenon. The revolutionary conclusion was that if they are time varying, then they would travel through space as a wave. This insight was accepted almost immediately by many people and initiated a large number of endeavors. The period of 1875 to 1900 was a time of tremendous innovation in wireless communication.

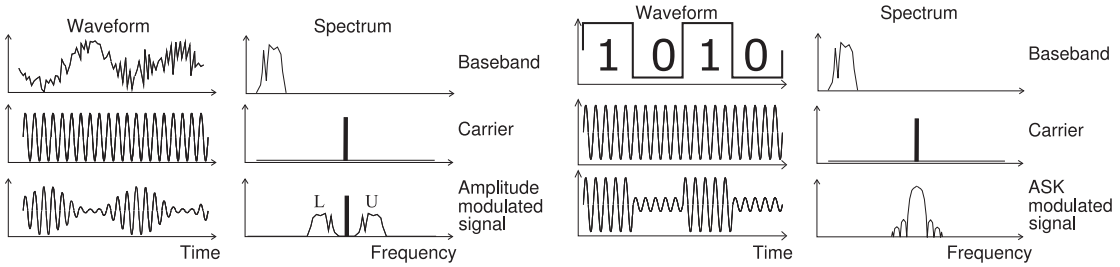
On November 22, 1875, Edison observed EM sparks. Previously sparks were considered to be an induction phenomenon, but Edison thought that he was producing a new kind of force, which he called the etheric force. He believed that this would enable communication without wires. To put this in context, the telegraph was invented in the 1830s and the telephone was invented in 1876.

The next stage leading to radio was orchestrated by D. E. Hughes beginning in 1879. Hughes experimented with a spark gap and reasoned that in the gap there was a rapidly alternating current and not a constant current as others of his time believed. The electric oscillator was born. The spark gap transmitter was augmented with a clockwork mechanism to interrupt the transmitter circuit and produce pulsed radio signals. He used a telephone as a receiver and walked around London and detected the transmitted signals over distance. Hughes noted that he had good reception at 180 feet. Hughes publicly demonstrated his "radio" in 1870 to the Royal Society, but the eminent scientists of the society determined that the effect was simply due to induction. This discouraged Hughes from continuing. However, Hughes has a legitimate claim to having invented radio, mobile digital radio at that, and probably was transmitting pulses on a 100 kHz carrier. In Hugeness's radio the RF carrier was produced by the spark gap oscillator and the information was coded as pulses. It was a small leap to a Morse key-based system.

The invention of practical radio can be attributed to many people, beginning with Heinrich Hertz, who in the period from 1885 to 1889 successfully verified the essential prediction of Maxwell's equations that EM energy could propagate through the atmosphere. Hertz was much more thorough than Hughes and his results were widely accepted. In 1891 Tesla developed what is now called the Tesla coil, which is a transformer with a primary and a secondary coil, one inside the other. When one of the coils was excited by an alternating signal, a large voltage was produced across the terminals of the other coil. Tesla pursued the application of his coils to radio and realized that the coils could be tuned so that the resulting resonance greatly amplified a radio signal.

The next milestone was the establishment of the first practical radio system by Marconi, with experiments beginning in 1894. Oscillations were produced in a spark gap, which were amplified by a Tesla coil. The work culminated in the transmission of telegraph signals across the Atlantic (from Ireland to Canada) by Marconi in 1901. In 1904, crystal radio kits to detect wireless telegraph signals could be readily purchased.

Spark gap transmitters could only send pulses of noise and not voice. One generator that could be amplitude modulated was an alternator. At the end



(a) Amplitude modulation (AM) (b) Amplitude shift keying (ASK)

Figure 1-7: Waveform and spectra of simple modulation schemes. The modulating signal, at the top in (a) and (b), is also called the baseband signal.

of the 19th century, readily available alternators produced a 60 Hz signal. Reginald Resplendent attempted to make a higher-frequency alternator and the best he achieved operated at 1 kHz. Resplendent realized that Maxwell's equations indicated that radiation increased dramatically with frequency and so he needed a much-higher-frequency signal source. Under contract, General Electric developed a 2-kW, 100-kHz alternator designed by Ernst Alexanderson. With this alternator, the first radio communication of voice occurred on December 23, 1900, in a transmission by Fessenden from an island in the Potomac River, near Washington, DC. Then on December 24, 1906, Fessenden transmitted voice from Massachusetts to ships hundreds of miles away in the Atlantic Ocean. This milestone is regarded as the beginning of the radio era.

Marconi subsequently purchased 50 and 200 kW Alexanderson alternators for his trans-Atlantic transmissions. Marconi was a great integrator of ideas, with particular achievements being the design of transmitting and receiving antennas that could be tuned to a particular frequency and the development of a coherer to improve detection of a signal.

1.3.11 Early Radio

Radio works by superimposing relatively slowly varying information, at what is called the **baseband** frequency, on a carrier sinusoid by varying the amplitude and/or phase of the sinusoid. Early radio systems were based on modulating an oscillating carrier either by pulsing the carrier (using for example Morse code)—this modulation scheme is called **amplitude shift keying** (**ASK**)—or by varying the amplitude of the carrier, i.e. **amplitude modulation** (**AM**), in the case of analog, usually voice, transmission. The waveforms and spectra of these modulation schemes are shown in Figure 1-7. The information is contained in the baseband signal, which is also called the modulating signal. The spectrum of the baseband signal extends to DC or perhaps down to where it rolls off at a low frequency. The carrier is a single sinewave and contains no information. The amplitude of the carrier is varied by the baseband signal to produce the modulated signal. In general, there are many cycles of the carrier relative to variations of the baseband signal so that the bandwidth of the modulated signal is relatively small compared to the frequency of the carrier.

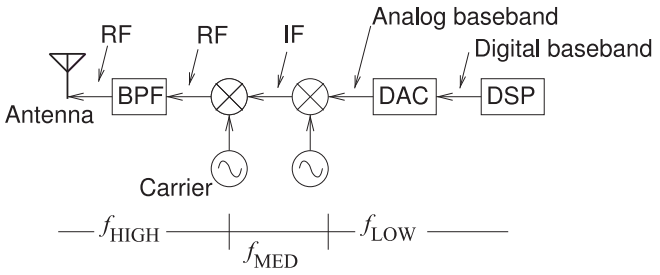


Figure 1-8: A simple transmitter with low, f_{LOW}, medium, f_{MED}, and high frequency, f_{HIGH}, sections. The mixers can be idealized as multipliers, shown as circles with crosses, that boost the frequency of the input baseband or IF signal by the frequency of the carrier.

AM and ASK radios are narrowband communication systems (they use a small portion of the EM spectrum), so to avoid interference with other radios it is necessary to search for an open part of the spectrum to place the carrier signal. In the decade of the 1900s there was little organization and a listener needed to search to find the desired transmission. The technology of the day necessitated this anyway, as the carrier would drift around by 10% or so since it was then not possible to build a stable oscillator. It was not until the *Titanic* sinking in 1912 that regulation was imposed on the wireless industry. Investigations of the *Titanic* sinking concluded that most of the lives lost would have been saved if a nearby ship had been monitoring its radio channels and if the frequency of the emergency channel was fixed. However, a second ship, but not close enough, did respond to *Titanic*'s "SOS" signal. A result of the investigations was the Service Regulations of the 1912 London International Radiotelegraph Convention. These early regulations were fairly liberal and radio stations were allowed to use radio wavelengths of their own choosing, but restricted to four broad bands: a single band at 1500 kHz for amateurs; 187.5 to 500 kHz, appropriated primarily for government use; below 187.5 kHz for commercial use, and 500 kHz to 1500 kHz, also a commercial band. Subsequent years saw more stringent assignment of narrow spectral bands and the assignment of channels. The standards and regulatory environment for radio were set—there would be assigned frequency bands for particular purposes. Very quickly strong government and commercial interests struggled for exclusive use of particular bands and thus the EM spectrum developed considerable value. Entities "owned" portions of the spectrum either through a license or through government allocation.

While most of the spectrum is allocated, there are several open bands where licenses are not required. The **instrumentation, scientific, and medical** (**ISM**) bands at 2.4 and 5.8 GHz are examples. Since these bands are loosely regulated, radios must cope with potentially high levels of interference.

1.4 Radio Architecture

A radio device is comprised of several key reasonably well-defined units. By frequency there are baseband, intermediate frequency (IF), and RF partitions. In a typical device, the information—either transmitted or received, bits or analog waveforms—is fully contained in the baseband unit. In the case of digital radios, the digital information originating in the baseband digital signal processor (DSP) is converted to an analog waveform typically using a digital-to-analog converter (DAC). This architecture is shown for a simple transmitter in Figure 1-8. When the basic information is analog,

say a voice signal in analog broadcast radios, the information is already a baseband analog waveform. This analog baseband signal can have frequency components that range from DC to many megahertz. However, the baseband signal can range from DC to gigahertz in the case of some radars and point-to-point links that operate at tens of gigahertz.

The RF hardware interfaces the external EM environment with the rest of the communication device. The information that is represented at baseband is translated to a higher-frequency signal that can more easily propagate over the air and for which antennas can be more easily built with manageable sizes. Thus the information content is generally contained in a narrow band of frequencies centered at the carrier frequency. The information content generally occupies a relatively small slice of the EM spectrum. The term "generally" is used as it is not strictly necessary that communication be confined to a narrow band: that is, narrow in percentage terms relative to the RF.

The trade-offs in the choice of carrier frequency are that lower-frequency EM signals require larger antennas, typically one-quarter to one-half wavelength long, but propagate over longer distances and tend to follow the curvature of the earth. AM broadcast radio stations operate around 1 MHz (where the wavelength, λ, is 300 m) using transmit antennas that are 100 m high or more, but good reception is possible at hundreds of kilometers from the transmitter. At higher frequencies, antennas can be smaller, a much larger amount of information can be transmitted with a fixed fractional bandwidth, and there is less congestion. An antenna at 2 GHz (where the free-space wavelength λ_0 = 15 cm) is around 4 cm long (and smaller with a dielectric or when folded or coiled), which is a very convenient size for a hand-held communicator.

The concept of an IF is related to the almost universal architecture of transmitters in the 20th century when baseband signals were first translated, or heterodyned, to a band around an IF before a second translation to a higher RF. Initially the IF was just above the audible range and was known as the supersonic frequency. The same progression applies in reverse in a receiver where information carried at RF is first translated to IF before finally being converted to baseband. This architecture resulted in near-optimum noise performance and relatively simple hardware, particularly at RF, where components are much more expensive than at lower frequencies.

The above discussion is a broad description of how radios work. There are many qualifications, as there are many evolving architectures and significant rethinking of the way radios can operate. Architectures and basic properties of radios are trade-offs of the capabilities of technologies, signal processing capability, cost, market dynamics, and politics.

1.5 Conventional Wireless Communications

Up until the mid-1970s most wireless communications were based on centralized high-power transmitters, often operating in a wide-area broadcast mode, and reception (e.g., by a television or radio unit) was expected until the signal level fell below a noise-related threshold. These systems are particularly sensitive to interference, therefore systems transmitting at the same frequency were geographically separated so that a transmitted signal falls below the background noise threshold before there

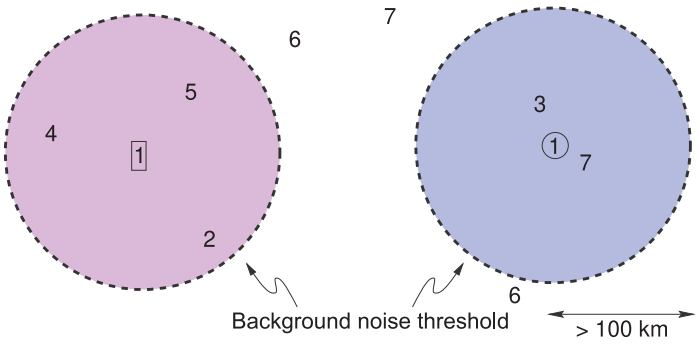


Figure 1-9: Interference in a conventional radio system. The two transmitters, 1, are at the centers of the coverage circles defined by the background noise threshold.

is a chance of it interfering with a neighboring system operating at the same frequency. This situation is illustrated in Figure 1-9. Here there are a number of base stations, each operating at a frequency (or set of frequencies) designated by numerals referring to the frequency of operation, which are correspondingly designated as f_1, f_2, etc. In Figure 1-9 the coverage by two base stations, 1, both operating at the frequency f_1 are shown by the shaded regions. The shading indicates the geographical region over which the signals are above the minimum detectable signal threshold. The frequency reuse factor of these types of systems is low, as there is a large geographical area where there is no reception at a particular frequency. The coverage area will not be circular or constant because terrain is not flat, signals are blocked by and reflected from buildings, and background noise levels vary during the day and signal levels vary from season to season as vegetation coverage changes. Allowances must be made in the allocation of broadcast areas to account for the changing coverage level. At the same time, it is necessary, in conventional radio, for the coverage area to be large so that reception, particularly for mobile devices, is continuous over metropolitan-size areas.

The original mobile radio service in the United States is now called **0G** for zero-generation radio. Very few users could be supported in 0G mobile radio because there were very few channels. The first 0G mobile system, the Mobile Telephone Service introduced in 1946, had six channels. That is, only six calls could be made at any one time. Because of interference this was reduced to three channels. So a metropolitan area such as New York city could only support three calls at the same time. In the three-channel version, the channels were 60 kHz wide and with a little more than 60 kHz guard band between channels. More channels were eventually made available. However the maximum practical frequency at the time was 450 MHz and the spectrum from 1 MHz to 500 MHz was highly sought after. Other uses included AM and FM radio, TV broadcast, military communications, and radar. It was seen by regulatory authorities that it was not in the public interest to support more individual users if that meant that broadcast services that catered to many people had to be compromised. There were no cells, just one large coverage area. In every change in radio generation there have been multiple enhancements to improve capacity. So just providing more bandwidth so there could be more channels was not a viable option. Supporting the transition to 1G was more bandwidth, the concept of cells and handoff, narrow channels, and higher operating frequency (900 MHz to 1 GHz). The continued evolution to fifth generation (5G) radio and the concepts that supported it are described in Chapters 2–5.

1.6 RF Power Calculations

1.6.1 RF Propagation

As an RF signal propagates away from a transmitter the power density reduces conserving the power in the EM wave. In the absence of obstacles and without atmospheric attenuation the total power passing through the surface of a sphere centered on a transmitter is equal to the power transmitted. Since the area of the sphere of radius r is $4\pi r^2$, the power density, e.g. in W/m^2, at a distance r drops off as $1/r^2$. With obstacles the EM wave can diffract, reflect, and follow multiple paths to a receiver where it can combine destructively or constructively. It is the destructive interference that is of concern as this limits the reliable reception of a signal. There is a low probability of perfect cancellation occurring and instead it is found that the power density reduces as $1/r^n$ where n ranges from 2 for free space to 5 for a dense urban environment with many obstacles, no line of sight, and multiple signal paths.

EXAMPLE 1.1 Signal Propagation

A signal is received at a distance r from a transmitter and the received power drops off as $1/r^2$. When $r = 1$ km, 100 nW is received. What is r when the received power is 100 fW?

Solution:

The signal collected by the receiver is proportional to the power density of the EM signal. The received signal power $P_r = k/r^2$ where k is a constant. This leads to

$$\frac{P_r(1\text{ km})}{P_r(r)} = \frac{100\text{ nW}}{100\text{ fW}} = 10^6 = \frac{kr^2}{k(1\text{ km})^2} = \frac{r^2}{(10^3\text{ m})^2}; \quad r = \sqrt{10^{12}\text{ m}^2} = 1000\text{ km} \quad (1.17)$$

EXAMPLE 1.2 Signal Propagation With Obstructions

A transmitter sends a signal to a receiver in a suburban environment that is a distance d away. When $d = 5$ km the signal power received is 100 nW. At what distance from the transmitter is the reliably received signal 1 pW if the received signal power falls off as $1/d^3$.

Solution:

Note that the signal falls off faster than the $1/d^2$ variation of free space. It is not sufficient to know the total power transmitted and instead the power density at a particular distance must be known. The power reliably received, $P_R(5\text{ km})$, at 5 km is 100 nW and this is the power density, $P_D(5\text{ km})$, multiplied by the effective area, A_r, of the receive antenna:

$$P_R(5\text{ km}) = 100\text{ nW} = P_D(5\text{ km})A_r = \frac{k}{d^3} = \frac{k}{(5\text{ km})^3} = \frac{k}{125\text{ km}^3}.$$

Both A_r and k are constants and $k = 12500\text{ nW}\cdot\text{km}^3 = 1.25\cdot10^{-5}\text{ W}\cdot\text{km}^3$. The power received at a distance d is 1 pW when

$$P_R(d) = 1\text{ pW} = 10^{-12}\text{ W} = \frac{k}{d^3} = \frac{1.25\cdot10^{-5}\text{ W}\cdot\text{km}^3}{d^3}$$

$$d^3 = \frac{1.25\cdot10^{-5}\text{ W}\cdot\text{km}^3}{10^{-12}\text{ W}} = 1.25\cdot10^7\text{ km}^3; \quad d = \sqrt[3]{1.25\cdot10^7}\text{ km} = 232.1\text{ km}. \quad (1.18)$$

Table 1-6: Common logarithm formulas. In engineering $\log x \equiv \log_{10} x$ and $\ln x \equiv \log_2 x$.

Description	Formula	Example
Equivalence	$y = \log_b(x) \longleftrightarrow x = b^y$	$\log(1000) = 3$ and $10^3 = 1000$
Product	$\log_b(xy) = \log_b(x) + \log_b(y)$	$\log(0.13 \cdot 978) = \log(0.13) + \log(978)$ $= -0.8861 + 2.990 = 2.104$
Ratio	$\log_b(x/y) = \log_b(x) - \log_b(y)$	$\ln(8/2) = \ln(8) - \ln(2) = 3 - 1 = 2$
Power	$\log_b(x^p) = p \log_b(x)$	$\ln\left(3^2\right) = 2\ln(3) = 2 \cdot 1.0986 = 2.197$
Root	$\log_b\left(\sqrt[p]{x}\right) = \frac{1}{p}\log_b(x)$	$\log\left(\sqrt[3]{20}\right) = \frac{1}{3}\log(20) = 0.4337$
Change of base	$\log_b(x) = \frac{\log_k(x)}{\log_k(b)}$	$\ln(100) = \frac{\log(100)}{\log(2)} = \frac{2}{0.30103} = 6.644$

1.6.2 Logarithm

A cellular phone can reliably receive a signal as small as 100 fW and the signal to be transmitted could be 1 W. So the same circuitry can encounter signals differing in power by a factor of 10^{13}. To handle such a large range of signals a logarithmic scale is used.

Logarithms are used in RF engineering to express the ratio of powers using reasonable numbers. Logarithms are taken with respect to a base b such that if $x = b^y$, then $y = \log_b(x)$. In engineering, $\log(x)$ is the same as $\log_{10}(x)$, and $\ln(x)$ is the same as $\log_e(x)$ and is called the natural logarithm (e = 2.71828 ...). Unfortunately in physics and mathematics (and in programs such as MATLAB), $\log x$ means $\ln x$, so be careful. Common formulas involving logarithms are given in Table 1-6.

1.6.3 Decibels

RF signal levels are usually expressed in terms of the power of a signal. While power can be expressed in absolute terms such as watts (W) or milliwatts (mW), it is much more useful to use a logarithmic scale. The ratio of two power levels P and P_{REF} in bels[3] (B) is

$$P(B) = \log\left(\frac{P}{P_{\text{REF}}}\right), \tag{1.19}$$

where P_{REF} is a reference power. Here $\log x$ is the same as $\log_{10} x$. Human senses have a logarithmic response and the minimum resolution tends to be about 0.1 B, so it is most common to use decibels (dB); 1 B = 10 dB. Common designations are shown in Table 1-7. Also, 1 mW = 0 dBm is a very common power level in RF and microwave power circuits where the m in dBm refers to the 1 mW reference. As well, dBW is used, and this is the power ratio with respect to 1 W with 1 W = 0 dBW = 30 dBm.

Working on the decibel scale enables convenient calculations using power numbers ranging from 10s of dBm to -110 dBm to be used rather than numbers ranging from 100 W to 0.00000000000001 W.

[3] Named to honor Alexander Graham Bell, a prolific inventor and major contributor to RF communications.

Table 1-7: Common power designations: (a) reference power, P_{REF}; (b) power ratio in decibels (dB); and (c) power in dBm and watts.

(a)

P_{REF}	Bell units	Decibel units
1 W	BW	dBW
1 mW = 10^{-3} W	Bm	dBm
1 fW = 10^{-15} W	Bf	dBf

(b)

Power ratio	in dB
10^{-6}	−60
0.001	−30
0.1	−20
1	0
10	10
1000	30
10^{6}	60

(c)

Power	Absolute power
−120 dBm	10^{-12} mW = 10^{-15} W = 1 fW
0 dBm	1 mW
10 dBm	10 mW
20 dBm	100 mW = 0.1 W
30 dBm	1000 mW = 1 W
40 dBm	10^{4} mW = 10 W
50 dBm	10^{5} mW = 100 W
−90 dBm	10^{-9} mW = 10^{-12} W = 1 pW
−60 dBm	10^{-6} mW = 10^{-9} W= 1 nW
−30 dBm	0.001 mW = 1 μW
−20 dBm	0.01 mW = 10 μW
−10 dBm	0.1 mW = 100 μW

EXAMPLE 1.3 Power Gain

An amplifier has a power gain of 1200. What is the power gain in decibels? If the input power is 5 dBm, what is the output power in dBm?

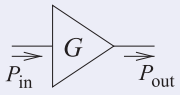
Solution:

Power gain in decibels, $G_{\mathrm{dB}} = 10 \log 1200 = 30.79$ dB.

The output power is $P_{\mathrm{out}}|_{\mathrm{dBm}} = P_{\mathrm{dB}} + P_{\mathrm{in}}|_{\mathrm{dBm}} = 30.79 + 5 = 35.79$ dBm.

EXAMPLE 1.4 Gain Calculations

A signal with a power of 2 mW is applied to the input of an amplifier that increases the power of the signal by a factor of 20.



(a) What is the input power in dBm?

$$P_{\mathrm{in}} = 2\ \mathrm{mW} = 10 \cdot \log\left(\frac{2\ \mathrm{mW}}{1\ \mathrm{mW}}\right) = 10 \cdot \log(2) = 3.010\ \mathrm{dBm} \approx 3.0\ \mathrm{dBm}. \tag{1.20}$$

(a) What is the gain, G, of the amplifier in dB?
The amplifier gain (by default this is power gain) is

$$G = 20 = 10 \cdot \log(20)\ \mathrm{dB} = 10 \cdot 1.301\ \mathrm{dB} = 13.0\ \mathrm{dB}. \tag{1.21}$$

(b) What is the output power of the amplifier?

$$G = \frac{P_{\mathrm{out}}}{P_{\mathrm{in}}}, \quad \text{and in decibels } G|_{\mathrm{dB}} = P_{\mathrm{out}}|_{\mathrm{dBm}} - P_{\mathrm{in}}|_{\mathrm{dBm}} \tag{1.22}$$

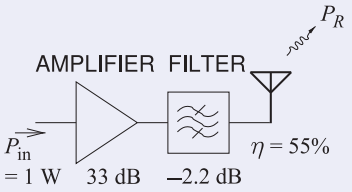
Thus the output power in dBm is

$$P_{\mathrm{out}}|_{\mathrm{dBm}} = G|_{\mathrm{dB}} + P_{\mathrm{in}}|_{\mathrm{dBm}} = 13.0\ \mathrm{dB} + 3.0\ \mathrm{dBm} = 16.0\ \mathrm{dBm}. \tag{1.23}$$

Note that dB and dBm are dimensionless but they do have meaning; dB indicates a power ratio but dBm refers to a power. Quantities in dB and one quantity in dBm can be added or subtracted to yield dBm, and the difference of two quantities in dBm yields a power ratio in dB.

EXAMPLE 1.5 Power Calculations

The output stage of an RF front-end consists of an amplifier followed by a filter and then an antenna. The amplifier has a gain of 33 dB, the filter has a loss of 2.2 dB, and of the power input to the antenna, 45% is lost as heat due to resistive losses. If the power input to the amplifier is 1 W, then:



(a) What is the power input to the amplifier expressed in dBm?

$P_{\text{in}} = 1\text{ W} = 1000\text{ mW}, \quad P_{\text{dBm}} = 10\log(1000/1) = 30\text{ dBm}.$

(b) Express the loss of the antenna in dB.
45% of the power input to the antenna is dissipated as heat.

The antenna has an efficiency, η, of 55% and so $P_2 = 0.55P_1$.
Loss $= P_1/P_2 = 1/0.55 = 1.818 = 2.60$ dB.

(c) What is the total gain of the RF front end (amplifier + filter + antenna)?

$$\begin{aligned}\text{Total gain} &= (\text{amplifier gain})_{\text{dB}} + (\text{filter gain})_{\text{dB}} - (\text{loss of antenna})_{\text{dB}} \\ &= (33 - 2.2 - 2.6)\text{ dB} = 28.2\text{ dB}\end{aligned} \tag{1.24}$$

(d) What is the total power radiated by the antenna in dBm?

$$\begin{aligned}P_R &= P_{\text{in}}|_{\text{dBm}} + (\text{amplifier gain})_{\text{dB}} + (\text{filter gain})_{\text{dB}} - (\text{loss of antenna})_{\text{dB}} \\ &= 30\text{ dBm} + (33 - 2.2 - 2.6)\text{ dB} = 58.2\text{ dBm}.\end{aligned} \tag{1.25}$$

(e) What is the total power radiated by the antenna?

$$P_R = 10^{58.2/10} = (661 \times 10^3)\text{ mW} = 661\text{ W}. \tag{1.26}$$

In Examples 1.3 and 1.4 two digits following the decimal point were used for the output power expressed in dBm. This corresponds to an implied accuracy of about 0.01% or 4 significant digits of the absolute number. This level of precision is typical for the result of an engineering calculation. See Section 2.A.1 of [1] for further discussion of precision and accuracy.

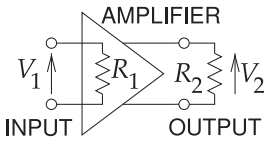
1.6.4 Decibels and Voltage Gain

Figure 1-10(a) is an amplifier with input and output resistances that could be different. If A_v is the voltage gain of the RF amplifier, then

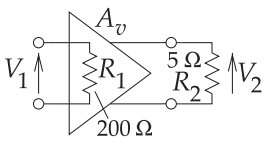
$$V_2 = A_v V_1 \tag{1.27}$$

and the input and output powers will be

$$P_{\text{in}} = \frac{V_1^2}{2R_1} \quad \text{and} \quad P_{\text{out}} = \frac{V_2^2}{2R_2}. \tag{1.28}$$



(a) General amplifier



(b) With $R_1 = 200\ \Omega$ and $R_2 = 5\ \Omega$

Figure 1-10: Amplifiers each with an input resistance R_1 and output resistance R_2.

The '2' in the denominator arises because V_1 and V_2 are peak amplitudes of sinusoids in RF engineering. Thus the power gain is

$$G = \frac{P_{\text{out}}}{P_{\text{in}}} = \frac{V_2^2 2R_1}{V_1^2 2R_2} = \frac{R_1}{R_2}A_v^2. \tag{1.29}$$

The power gain depends on the input and output resistance ratio of the amplifiers and this is commonly used to realize significant power gain even if the voltage gain is quite small. If the input and output resistances of the amplifier are the same, then the power gain is just the voltage gain squared.

In handling this situation some authors have used the unit dBV (decibel as a voltage ratio). This should not be used, decibels should always refer to a power ratio, and it is needlessly confusing to use dBV in RF engineering.

EXAMPLE 1.6 Voltage Gain to Power Gain

Figure 1-10(b) is a differential amplifier with a 200 Ω input resistance and 5 Ω output resistance. If the voltage A_v is 0.6, what is the power gain of the amplifier in dB?

Solution:

The input and output powers are

$$P_{\text{in}} = \tfrac{1}{2}V_1^2/R_1 \quad \text{and} \quad P_{\text{out}} = \tfrac{1}{2}V_2^2/R_2 = \tfrac{1}{2}\frac{(A_v V_1)^2}{R_2}. \tag{1.30}$$

Thus the power gain is

$$G = \frac{P_{\text{out}}}{P_{\text{in}}} = \frac{(A_v V_1)^2}{R_2}\left(\frac{V_1^2}{R_1}\right)^{-1} = \frac{R_1}{R_2}A_v^2 = \frac{200}{5}0.6^2 = 14.4 = 11.58\,\text{dB}. \tag{1.31}$$

The surprising result is that even with a voltage gain of less than 1, a significant power gain can be obtained if the input and output resistances are different. A result used in many RF amplifiers.

1.7 Photons and Electromagnetic Waves

Most of the time it is not necessary to consider the quantum nature of radio waves. Considering electric and magnetic fields, and Maxwell's equations, is sufficient to understand radio systems. Still the underlying physics is that currents in circuits derive from the movement of quantum particles, here electrons, and propagating EM signals are transported by photons, also quantum particles.

In a receiver system, information is translated by a metallic antenna from the propagating photons to moving charges in a conductor. This transfer of information is through the interaction of a photon with charge carriers. So the information transfer is based on the quantized energy of a photon and it is possible that quantum effects could be important. The relative level of the photon energy and the random kinetic energy of an electron is important in determining if quantum effects must be considered.

With all EM radiation, information is conveyed by photons and the energy of a photon is conventionally expressed in terms of electron volts (eV). An electron volt is the energy gained by an electron when it moves across an electric potential of 1 V and 1 eV = 1.602×10^{-19} J (1 J = 6.241509×10^{18} eV).

The energy of a photon is $E = h\nu = hf$, where $h = 6.6260693 \times 10^{-34}\text{J}\cdot\text{s}$ is the Plank constant and ν and f are frequency, with the symbol ν preferred by physicists and f preferred by engineers. So the energy of a photon is proportional to its frequency. At microwave frequencies the energy of a microwave photon ranges from 1.24 μeV at 300 MHz to 1.24 meV at 300 GHz. This is a very small amount of energy.

Physically kT is the amount of energy required to increase the entropy (corresponding to movement) of the electrons by a factor of e [12].

The thermal energy of an electron in joules is $E|_{\text{J}} = kT$ where $k = 1.3806505 \times 10^{-23}$ J/K is the Boltzmann constant and T is the temperature in kelvin. This applies to an electron that is moving in a group of electrons, say in a plasma, in thermodynamic equilibrium. Freely conducting electrons in a conductor are in a plasma. The thermal energy of such an electron is its random kinetic energy with $kT = \frac{1}{2}mv^2$ where m is the mass of the electron and v is its velocity. (The thermal energy of an isolated electron, i.e. in a non-interacting electron gas, is $\frac{3}{2}kT$, but that is not the situation with microwave circuits.) At room temperature ($T = 298$ K) the thermal energy of a conducting electron in a metal is $kT = 25.7$ meV. This is much more than the energy of a microwave photon (1.24 μeV to 1.24 meV). Thus at room temperature discrete quantum effects are not apparent for microwave signals and so microwave radiation (at room temperature) can be treated as a continuum effect.

A photon has a dual nature, as a particle and as an EM wave. The break point as to which nature helps the most in understanding behavior depends on the energy of the photon versus the thermal energy of an electron. When a microwave photon is captured in a metal at room temperature it makes little sense to talk about the photon as increasing the energy state of an individual electron. Instead, it is best to think of the photon as an EM wave with an electric field that accelerates an ensemble of free electrons (in the conduction band). Thus the energy of one photon is transferred to a group of electrons as faster moving electrons but with the energy increase being so small relative to thermal energy that a quantized effect is not apparent. However even a very low power microwave signal has an enormous number of photons (a 1 pW 300 GHz signal has 5×10^9 photons per second) and each photon accelerates the free electrons a little bit with the combined effect being that the electric field of the microwave signal results in appreciable current. This is the view that we use with room temperature circuits and antennas at microwave frequencies; the EM signal (instead of discrete photons) interacts with the free electrons in a conductor and produces current. Quantum effects must be considered when temperature is very low (say below 4 K) or frequency is very high, e.g. the photon energies for red light (400 THz) is 1.7 eV.

1.8 Summary

Today six billion individuals regularly transmit information wirelessly using transmitters and receivers that can be smaller than an infants hand, contain trillions of transistors, and provide connectivity in nearly any location.

The history of radio communications is remarkable and nearly every aspect of electrical engineering is involved. The most important factor of all is that consumers are prepared to part with a large portion of their income to have untethered connectivity. This overwhelming desire for ubiquitous communication surprised even the most optimistic proponents of personal wireless communications. Wireless communication is now a significant part

of every country's economy, and governments are very involved in setting standards and protecting the competitiveness of their own industries.

The history of radio communications has led to the current mode of operation, allocating a narrow slice of the EM spectrum to one or a few users. This choice dictated the need for very stable oscillators and high-rejection filters, for example. The stability of oscillators in consumer products is a few hertz around 1 GHz, a few parts in a billion, a level of precision that is unrivaled for any other physical quantity in a manufactured device.

The RF spectrum is used to support a tremendous range of applications, including voice and data communications, satellite-based navigation, radar, weather radar, mapping, environmental monitoring, air traffic control, police radar, perimeter surveillance, automobile collision avoidance, and many military applications. A big trend is the virtual disappearance of analog radio and now the almost complete use of digital radio. Digital radio is much more tolerant to interference and can use a much smaller slice of the EM spectrum. Another big trend is the tremendous demand for EM spectrum resulting in appreciable use of the spectrum up to 100 GHz and soon beyond that. Currently large parts of the spectrum are allocated for exclusive military use and there is pressure to reduce the spectrum allocated for government use of all kinds.

In RF and microwave engineering there are always considerable approximations made in design, partly because of necessary simplifications that must be made in modeling, but also because many of the material properties required in a detailed design can only be approximately known. Most RF and microwave design deals with frequency-selective circuits often relying on line lengths that have a length that is a particular fraction of a wavelength. Many designs can require frequency tolerances of as little as 0.1%, and filters can require even tighter tolerances. It is therefore impossible to design exactly. Measurements are required to validate and iterate designs. Conceptual understanding is essential; the designer must be able to relate measurements, which themselves have errors, with computer simulations. The ability to design circuits with good tolerance to manufacturing variations and perhaps circuits that can be tuned by automatic equipment are skills developed by experienced designers.

Chapter 2 describes modulation methods and the ideas that led to being able to transmit many bits of data per hertz of bandwidth. High orders of modulation send many bits of information and the higher the order of modulation the more sophisticated the modulation and demodulation schemes must be. Transmitters and receivers that implement the modulation and demodulation methods and up-convert and down-convert, respectively, between the low frequency baseband information and the high frequency radio signals are described in Chapter 3. Antennas reviewed in Chapter 4 are the interface between electronic circuits and freely propagating EM waves. Directional antennas are essential to supporting many users by enabling frequency reuse of the EM spectrum in the same cell. The 1G through to 5G cellular systems are described in Chapter 5. In addition other microwave systems such as radar and WiFi are briefly described and their development has closely followed or just preceded that of cellular radio. The tremendous advances are the result of the synergistic development of system concepts and of digital and microwave hardware.

1.9 References

[1] M. Steer, *Microwave and RF Design, Transmission Lines*, 3rd ed. North Carolina State University, 2019.

[2] ——, *Microwave and RF Design, Networks*, 3rd ed. North Carolina State University, 2019.

[3] ——, *Microwave and RF Design, Modules*, 3rd ed. North Carolina State University, 2019.

[4] ——, *Microwave and RF Design, Amplifiers and Oscillators*, 3rd ed. North Carolina State University, 2019.

[5] "Atmospheric microwave transmittance at mauna kea, wikipedia creative commons."

[6] "IEEE standard 521-2002, *IEEE Standard Letter Designations for Radar-Frequency Bands*, 2002."

[7] T. K. Sarkar, R. Mailloux, A. A. Oliner, M. Salazar-Palma, and D. L. Sengupta, *History of wireless*. John Wiley & Sons, 2006.

[8] "IEEE Virtual Museum," at http://www.ieee-virtual-museum.org Search term: 'Faraday'.

[9] M. Loomis, "Improvement in telegraphing," 1872, US Patent 129,971.

[10] J. Rautio, "Maxwell's legacy," *IEEE Microwave Magazine*, vol. 6, no. 2, pp. 46–53, Jun. 2005.

[11] J. Maxwell, *A Treatise on Electricity and Magnetism*. Clarendon Press (Reprinted, Oxford University Press, 1998), 1873.

[12] P. Atkins and J. DePaula, *Physical Chemistry*, 9th ed. WH Freeman & Company, 2009.

1.10 Exercises

1. Consider a photon at 1 GHz.
   (a) What is the energy of the photon in joules?
   (b) Is this more or less than the random kinetic energy of an electron at room temperature?
2. Consider a photon at various frequencies.
   (a) What is the photon's energy at 1 GHz in terms of electron-volts?
   (b) What is the photon's energy at 10 GHz in terms of electron-volts?
   (c) What is the photon's energy at 100 GHz in terms of electron-volts?
   (d) What is the photon's energy at 1 THz in terms of electron-volts?
3. Consider a photon at 1 THz.
   (a) What is the energy of the photon in terms of electron-volts?
   (b) What is the energy of the photon in joules?
   (c) Is this more or less than the random kinetic energy of an electron at room temperature (300 K)?
   (d) Discuss if it is necessary to consider quantum effects of the 1 THz photon at room temperature.
4. Consider a photon at 10 GHz.
   (a) What is the energy of the photon in terms of electron-volts?
   (b) What is the energy of the photon in joules?
   (c) What is the random kinetic energy of an electron at room temperature (300 K)?
   (d) Calculate the temperature, in kelvins, at which the random kinetic energy of an electron is equal to the energy you calculated in (a).
5. A 10 GHz transmitter transmits a 1 W signal. How many photons are transmitted?
6. A receiver receives a 1 pW signal at 60 GHz. How many photons per second are received?
7. At what frequency is the photon energy equal to the thermal energy of an electron at 300 K?
8. What is the frequency at which the energy of a photon is equal to the thermal energy of an electron at 77 K?
9. What is the wavelength in free space of a signal at 4.5 GHz?
10. Consider a monopole antenna that is a quarter of a wavelength long. How long is the antenna if it operates at 3 kHz?
11. Consider a monopole antenna that is a quarter of a wavelength long. How long is the antenna if it operates at 500 MHz?
12. Consider a monopole antenna that is a quarter of a wavelength long. How long is the antenna if it operates at 2 GHz?
13. A dipole antenna is half of a wavelength long. How long is the antenna at 2 GHz?
14. A dipole antenna is half of a wavelength long. How long is the antenna at 1 THz?
15. Write your family name in Morse code (see Table 1-5).
16. A transmitter transmits an FM signal with a bandwidth of 100 kHz and the signal is received by a receiver at a distance r from the transmitter. When $r = 1$ km the signal power received by the receiver is 100 nW. When the receiver moves

further away from the transmitter the power received drops off as $1/r^2$. What is r in kilometers when the received power is 100 pW. [Parallels Example 1.1]

17. A transmitter transmits an AM signal with a bandwidth of 20 kHz and the signal is received by a receiver at a distance r from the transmitter. When $r = 10$ km the signal power received is 10 nW. When the receiver moves further away from the transmitter the power received drops off as $1/r^2$. What is r in kilometers when the received power is equal to the received noise power of 1 pW? [Parallels Example 1.1]

18. In a legacy, i.e. 0G, broadcast radio system a transmitter broadcasts an AM signal and the signal can be successfully received if the AM signal is 20 dB higher than the 10 fW noise power received. The received signal power when the transmitter and receiver are separated by $r = 1$ km is 100 nW. The received signal power falls off as $1/r^2$ as the receiver moves further away.
    (a) What is the radius of the broadcast circle in which the broadcast signal is successfully received?
    (b) At what distance does the power of the broadcast signal match the noise power?
    (c) If two transmitters both transmit similar AM signals at the same frequency, how far should the transmitters be separated so that the interference received is 10 dB below the noise level?

19. In a legacy radio system a transmitter broadcasts an FM signal and for noise-free reception the FM signal must be 30 dB higher than the received noise power of 10 fW. When the transmitter and receiver are separated by $r = 1$ km the signal power received is 100 nW. The received signal power falls off as $1/r^3$ with greater separation.
    (a) What is the radius of the circle in which the broadcast signal is successfully received?
    (b) At what distance does the power of the broadcast signal match the noise power?
    (c) If two transmitters both transmit similar FM signals at the same frequency and power. One transmitter transmits the desired signal while the second transmits an interfering signal. How far should the transmitters be separated so that the interference received is 10 dB below the noise level?

20. A transmitter broadcasts a signal to a receiver that is a distance d away. The noise power received is 1 pW and when $d = 5$ km the signal power received is 100 nW. What is the radius of the noise threshold circle where the noise and signal powers are equal, when the received signal power falls off as:
    (a) $1/d^2$? [Parallels Example 1.1]
    (b) $1/d^{2.5}$? [Parallels Example 1.2]

21. A signal is transmitted to a receiver that is a distance r away. The noise power received is 100 fW and when r is 1 km the received signal power is 500 nW. What is r when the noise and signal powers are equal when the received signal power falls off as:
    (a) $1/d^2$? [Parallels Example 1.1]
    (b) $1/d^3$? [Parallels Example 1.2]

22. The logarithm to base 2 of a number x is 0.38 (i.e., $\log_2(x) = 0.38$). What is x?

23. The natural logarithm of a number x is 2.5 (i.e., $\ln(x) = 2.5$). What is x?

24. The logarithm to base 2 of a number x is 3 (i.e., $\log_2(x) = 3$). What is $\log_2(\sqrt[2]{x})$?

25. What is $\log_3(10)$?

26. What is $\log_{4.5}(2)$?

27. Without using a calculator evaluate $\log\{[\log_3(3x) - \log_3(x)]\}$.

28. A 50 Ω resistor has a sinusoidal voltage across it with a peak voltage of 0.1 V. The RF voltage is $0.1\cos(\omega t)$, where ω is the radian frequency of the signal and t is time.
    (a) What is the power dissipated in the resistor in watts?
    (b) What is the power dissipated in the resistor in dBm?

29. The power of an RF signal is 10 mW. What is the power of the signal in dBm?

30. The power of an RF signal is 40 dBm. What is the power of the signal in watts?

31. An amplifier has a power gain of 2100.
    (a) What is the power gain in decibels?
    (b) If the input power is -5 dBm, what is the output power in dBm? [Parallels Example 1.3]

32. An amplifier has a power gain of 6. What is the power gain in decibels? [Parallels Example 1.3]

33. A filter has a loss factor of 100. [Parallels Example 1.3]
    (a) What is the loss in decibels?
    (b) What is the gain in decibels?

34. An amplifier has a power gain of 1000. What is the power gain in dB? [Parallels Example 1.3]

35. An amplifier has a gain of 14 dB. The input to the amplifier is a 1 mW signal, what is the output power in dBm?

36. An RF transmitter consists of an amplifier with a gain of 20 dB, a filter with a loss of 3 dB and then that is then followed by a lossless transmit antenna. If the power input to the amplifier is 1 mW, what is the total power radiated by the antenna in dBm? [Parallels Example 1.5]
37. The final stage of an RF transmitter consists of an amplifier with a gain of 30 dB and a filter with a loss of 2 dB that is then followed by a transmit antenna that looses half of the RF power as heat. [Parallels Example 1.5]
    (a) If the power input to the amplifier is 10 mW, what is the total power radiated by the antenna in dBm?
    (b) What is the radiated power in watts?
38. A 5 mW RF signal is applied to an amplifier that increases the power of the RF signal by a factor of 200. The amplifier is followed by a filter that losses half of the power as heat.
    (a) What is the output power of the filter in watts?
    (b) What is the output power of the filter in dBW?
39. The power of an RF signal at the output of a receive amplifier is 1 μW and the noise power at the output is 1 nW. What is the output signal-to-noise ratio in dB?
40. The power of a received signal is 1 pW and the received noise power is 200 fW. In addition the level of the interfering signal is 100 fW. What is the signal-to-noise ratio in dB? Treat interference as if it is an additional noise signal.age gain of 1 has an input impedance of 100 Ω, a zero output impedance, and drives a 5 Ω load. What is the power gain of the amplifier?
41. A transmitter transmits an FM signal with a bandwidth of 100 kHz and the signal power received by a receiver is 100 nW. In the same bandwidth as that of the signal the receiver receives 100 pW of noise power. In decibels, what is the ratio of the signal power to the noise power, i.e. the signal-to-noise ratio (SNR), received?
43. An amplifier with a voltage gain of 20 has an input resistance of 100 Ω and an output resistance of 50 Ω. What is the power gain of the amplifier in decibels? [Parallels Example 1.6]
44. An amplifier with a voltage gain of 1 has an input resistance of 100 Ω and an output resistance of 5 Ω. What is the power gain of the amplifier in decibels? Explain why there is a power gain of more than 1 even though the voltage gain is 1. [Parallels Example 1.6]
45. An amplifier with a volt
46. An amplifier has a power gain of 1900.
    (a) What is the power gain in decibels?
    (b) If the input power is −8 dBm, what is the output power in dBm? [Parallels Example 1.3]
47. An amplifier has a power gain of 20.
    (a) What is the power gain in decibels?
    (b) If the input power is −23 dBm, what is the output power in dBm? [Parallels Example 1.3]
48. An amplifier has a voltage gain of 10 and a current gain of 100.
    (a) What is the power gain as a number?
    (b) What is the power gain in decibels?
    (c) If the input power is −30 dBm, what is the output power in dBm?
    (c) What is the output power in mW?
49. An amplifier with 50 Ω input impedance and 50 Ω load impedance has a voltage gain of 100. What is the (power) gain in decibels?
50. An attenuator reduces the power level of a signal by 75%. What is the (power) gain of the attenuator in decibels?

1.10.1 Exercises By Section

†challenging, ‡very challenging

§1.2 1, 2†, 3†, 4†, 5, 6, 7†, 8†
§1.3 9, 10, 11, 12, 13, 14, 15
§1.5 16, 17, 18‡, 19†, 20†, 21†
§1.6 22, 23, 24, 25, 26, 27, 28, 29, 30 31†, 32, 33, 34, 35, 36†, 37†, 38† 39, 40, 41, 42, 43, 44, 45, 46, 47 48, 49, 50

1.10.2 Answers to Selected Exercises

2(d) 41.36 meV
4(b) 662.6 fJ
11.12 3.25 cm
35 1.301
25 2.096
29 10 dBm
30 10 W
32 7.782 dB
36 50.12 mW
37(b) 3.162 W
42(b) −6 dB

CHAPTER 2

Modulation

2.1 Introduction

Most radio communication systems superimpose slowly varying information on a sinusoidal carrier that is transmitted as a radio frequency (RF) signal. This modulated RF signal is sent through a medium, usually air, by a transmitter to a receiver. In the transmitter information is initially represented at what is called baseband. The process of transferring information from baseband to the much higher frequency carrier wave is called modulation. Most modulation schemes slowly vary the amplitude and/or phase of a sinusoidal carrier waveform. In the receiver the process is reversed using demodulation to extract the baseband information from the varying state, such as the amplitude and/or phase, of the modulated carrier.

Radio has evolved subject to constraints imposed by political, hardware, and compatibility considerations. New schemes generally must be compatible and co-exist with earlier schemes. This chapter discusses the many different modulation schemes that are used in radios. Nearly all modulation schemes are supported in modern radios such as 4G and 5G cellular radios, and many are supported in WiFi. Sometimes this is to provide support for legacy radios while in other situations they are used because simpler modulation formats tolerate higher levels of interference. Indeed the level of so-

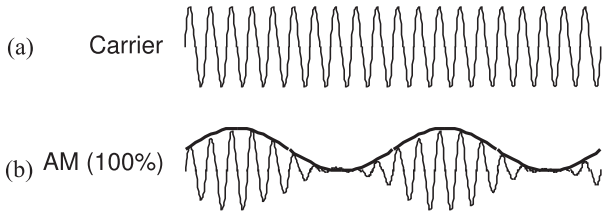


Figure 2-1: AM showing the relationship between the carrier and modulation envelope: (a) carrier; and (b) 100% amplitude modulated carrier.

phistication of modulation methods may need to be frequently changed to accommodate varying interference environments. Legacy analog modulation schemes and the simpler digital modulation schemes where suitable for the relatively unsophisticated hardware of years past. High-order modulation schemes enable many digital bits to be sent in each hertz of bandwidth and are only possible because of the evolution of digital signal processing and because of advances in high-density, low-power digital electronics.

Section 2.2 introduces some of the metrics that are used to compare modulation schemes and Section 2.3 introduces modulation. Section 2.4 describes analog modulation. Then Section 2.5 describes digital modulation followed by sections that deal with the specifics of various digital modulation methods: frequency shift keying (FSK) in Section 2.6; phase shift keying (PSK) in Section 2.8; and quadrature amplitude modulation (QAM) in Section 2.9. Before the discussion of PSK a concept called carrier recovery is discussed in Section 2.7 as the necessity to do this was behind the development of a variety of PSK modulation schemes. This is followed by a discussion of the metrics that can be used to quantify interference and distortion of modulated signals.

Modulation, and the hardware architectures and circuits for modulating and demodulating radio signals, are presented largely in three chapters. There is an overlap of these topics but modulation itself is largely confined to this chapter although some architecture concepts must necessarily be introduced to understand the evolution of modulation schemes. The next chapter, Chapter 3, focuses on architectures and essential circuits for modulators and demodulators.

2.2 Radio Signal Metrics

Radio signals are engineered to trade-off efficient use of the EM spectrum with the complexity and performance of the required RF hardware. Ultimately the goal is to efficiently use spectrum through maximal packing of information, e.g. digital bits, in a given bandwidth while, for mobile radios especially, using as little prime power as possible. The choice of the type of modulation to use is at the core of the communication system design trade-off.

There are two families of modulation methods: analog and digital modulation. In analog modulation the RF signal has a continuous range of values; in digital modulation, the output has a number of discrete states at particular times called clock ticks, say every microsecond. There are just a few modulation schemes, all of which are digital, that achieve the optimum trade-offs of spectral efficiency and ease of use with acceptable hardware complexity. If hardware complexity is not a concern, which modulation scheme is used depends on noise and interference as well as the power required to transmit a signal, and the power required to process a received

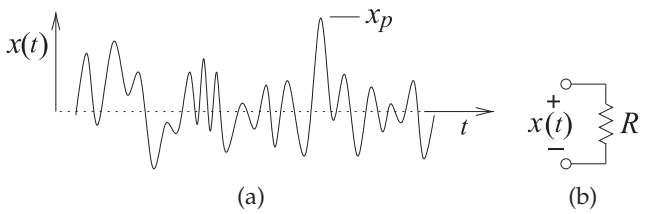


Figure 2-2: Definition of crest factor: (a) arbitrary waveform; and (b) voltage across a resistor.

signal.

This section introduces several metrics that characterize the variability of the amplitude of a modulated signal, and this variability has a direct impact on how analog hardware performs are designed and how efficiently hardware can be used.

2.2.1 Crest Factor and Peak-to-Average Power Ratio

Introduction

In radio engineering crest factor (**CF**) is a metric that describes how the voltage of a modulated carrier signal varies with time, and peak-to-average power ratio (**PAPR**) describes how the instantaneous power of a carrier signal varies with time. Be aware that there is one metric, **peak-to-average ratio** (**PAR**), that is defined differently in the power, communications theory, and microwave communities. In some communities CF is also called the **peak-to-average ratio** (**PAR**). This can leads to problems. Consider, for example, the community that works on smart power metering which combines power measurement, communications theory, and microwave design. The solution to this inevitable confusion is to skip the use of PAR and use unambiguous metrics.

In standards PAR is defined as the ratio of the instantaneous peak value of a signal parameter to its time-averaged value. PAR is used with many signal parameters, e.g. voltage, current, power, and frequency [1].

Crest Factor

CF is the ratio of the maximum signal, such as a voltage, to its root-mean-square (rms) value. Referring to the arbitrary waveform shown in Figure 2-2(a), x_p is the absolute peak value of the waveform $x(t)$, if x_{rms} is its rms value, then the crest factor is [2]

$$\text{CF} = x_p / x_{\text{rms}}. \tag{2.1}$$

More formally,

$$\text{CF} = \frac{\|x\|_\infty}{\|x\|_2}, \tag{2.2}$$

where $\|x\|_\infty$ is the infinity norm, and here is the maximum value of $x(t)$, $\|x\|_\infty = \max[x(t)] = x_p$, and $\|x\|_2$ is just the rms value of $x(t)$:

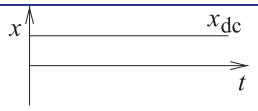
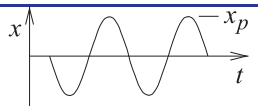
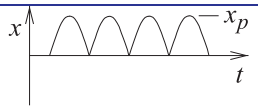
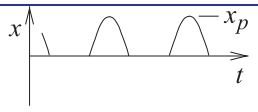
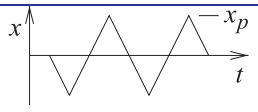
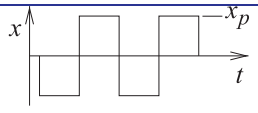
$$x_{\text{rms}} = \|x\|_2 = \lim_{T\to\infty} \sqrt{\frac{1}{T}\int_0^T x(t)\cdot dt}\,. \tag{2.3}$$

Note that CF is a voltage (or current) ratio rather than a power ratio. The CFs of several waveforms are given in Table 2-1.

Peak-to-Average Power Ratio (PAPR)

The peak-to-average power ratio (PAPR) is analogous to CF but for power. If $x(t)$ is the voltage across a resistor, as shown in Figure 2-2(b), then the

Table 2-1: Crest factor (CF) and peak-to-average power ratio (PAPR) of several waveforms. (x_p is the peak value of the waveform.)

Waveform	$x(t)$	Max. value	rms ($x_{\rm rms}$)	CF	PAPR
DC		$x_{\rm dc}$	$x_{\rm dc}$	1	0 dB
Sinewave		x_p	$\frac{x_p}{\sqrt{2}}$	1.414	3.01 dB
Full-wave rectified sinewave		x_p	$\frac{x_p}{\sqrt{2}} = 0.717\ x_p$	1.414	3.01 dB
Half-wave rectified sinewave		x_p	$\frac{x_p}{2}$	2	6.02 dB
Triangle wave		x_p	$\frac{x_p}{\sqrt{3}} = 0.577\ x_p$	1.732	4.77 dB
Square wave		x_p	x_p	1	0 dB

instantaneous peak power in the resistor is

$$P_p = |x_p|^2 / R, \tag{2.4}$$

where again x_p is the peak absolute value of the waveform. P_p is the power of the peak of a waveform treating it as though it was a DC signal. This is appropriate for a slowly varying signal such as a power frequency signal as it is this instantaneous power that determines thermal disruption of a power system. It is not the appropriate power to use with radio signals and a more suitable microwave signal metric is described in Section 2.2.2. The average power dissipated in the resistor is

$$P_{\rm avg} = |x_{\rm rms}|^2 / R. \tag{2.5}$$

Then
$$\text{PAPR} = \frac{P_p}{P_{\rm avg}} = \text{CF}^2 = (x_p/x_{\rm rms})^2 . \tag{2.6}$$

In decibels,
$$\begin{aligned} \text{PAPR}|_{\rm dB} &= 10 \log (\text{PAPR}) \\ &= 20 \log(\text{CF}) = 20 \log (x_p/x_{\rm rms}) . \end{aligned} \tag{2.7}$$

The definition of PAPR above can be used with any waveform and can be used in all branches of electrical engineering. The PAPRs of several waveforms are given in Table 2-1.

EXAMPLE 2.1 Crest Factor and PAPR of an Offset Sinusoid.

What is the crest factor (CF) and peak-to-average power ratio (PAPR) of the signal $x(t) = 0.1 + 0.5\sin(\omega t)$?

Solution:

The signal is a sinusoid offset by a DC term. The peak value of $x(t)$ is $x_p = 0.6$, and the rms value of the signal will be the square root of the rms values squared of the individual DC and sinusoidal components. This applies to any composite signal provided that the components are uncorrelated. So $x_{\rm rms} = \sqrt{0.1^2 + (0.5/\sqrt{2})^2} = 0.3674$. The general solution for a signal $x(t) = a + b\sin(\omega t)$ is, using Equation (2.3),

$$\begin{aligned} x_{\rm rms} &= \sqrt{\lim_{T\to\infty}\frac{1}{T}\int_0^T [x(t)]^2\,.dt} = \sqrt{\lim_{T\to\infty}\frac{1}{T}\int_0^T [a+b\sin(\omega t)]^2\,.dt} \\ &= \sqrt{\lim_{T\to\infty}\frac{1}{T}\int_0^T \left[a^2 + ab\sin(\omega t) + b^2\sin^2(\omega t)\right].dt} \\ &= \sqrt{\lim_{T\to\infty}\frac{1}{T}\left\{\int_0^T a^2.dt + \int_0^T ab\sin(\omega t).dt + \int_0^T b^2\tfrac{1}{2}\left[1+\cos(2\omega t)\right].dt\right\}} \\ &= \sqrt{\lim_{T\to\infty}\frac{1}{T}\left\{a^2T.dt + 0 + \tfrac{1}{2}b^2T\right\}} \end{aligned} \tag{2.8}$$

since the integral of sin and cos over a period is zero. Thus

$$x_{\rm rms} = \sqrt{a^2 + b^2/2} = \sqrt{0.1^2 + \tfrac{1}{2}0.5^2} = 0.3674, \tag{2.9}$$

the crest factor is $$\text{CF} = \frac{x_p}{x_{\rm rms}} = \frac{0.6}{0.3674} = 1.6331, \tag{2.10}$$

and PAPR is $$\text{PAPR} = 20\log(1.6331) = 4.260\ \text{dB}. \tag{2.11}$$

There is a quicker way of calculating PAPR by dealing with the powers directly. The peak power of the waveform is $P_p = x_p^2/R = 0.6^2/R = 0.36/R$, where x is being treated as a voltage across a resistor R. The two parts of $x(t)$, i.e. the DC component and the sinewave, are uncorrelated, so the average power of the combined signal is the sum of the powers of the uncorrelated components, so

$$P_{\rm avg} = \frac{1}{R}\left[0.1^2 + \tfrac{1}{2}0.5^2\right]\frac{1}{R} = \frac{0.1350}{R}. \tag{2.12}$$

Thus, in decibels,

$$\text{PAPR}|_{\rm dB} = 10\log\left(\frac{P_p}{P_{\rm avg}}\right) = \frac{x_p^2}{x_{\rm rms}^2} = 10\log\left(\frac{0.36}{0.135}\right) = 10\log(2.667) = 4.260\ \text{dB}. \tag{2.13}$$

2.2.2 Peak-to-Mean Envelope Power Ratio

Another metric for characterizing signals is the peak-to-mean envelope power ratio (PMEPR) and this is particularly useful for modulated signals. The amount of information sent by a communication signal is proportional to its average power, however, RF hardware must be designed with enough margin to be able to handle peaks in the signal without producing appreciable distortion. The waveform of a narrowband modulated signal appears as a carrier that slowly changes in amplitude and phase. One sinewave of this modulated signal is called a **pseudo-carrier** and the power of one cycle of the pseudo-carrier when the amplitude of the modulated

signal is at its maximum (i.e. at the peak of the envelope) is called the **peak envelope power** (**PEP**) [1] (PEP = P_{PEP}). The ratio of PEP to the average signal power (the power averaged over all time) is called the PMEPR.

Then if the average power of the modulated signal is P_{avg}

$$\mathrm{PMEPR} = \frac{\mathrm{PEP}}{P_{\mathrm{avg}}} = \frac{\mathrm{P_{PEP}}}{P_{\mathrm{avg}}}. \tag{2.14}$$

PMEPR is a good indicator of how sensitive a modulation format is to distortion introduced by the nonlinearity of RF hardware [3].

It is complex to determine the PMEPR for a general modulated signal. Below the mathematics is presented for an AM signal with a sinusoidal modulating signal. Determining the PMEPR otherwise requires numerical integration following the procedure outlined below.

PMEPR of an AM Signal

A good estimate of the PMEPR of an AM signal can be obtained by considering a sinusoidal modulating signal (rather than an actual baseband signal). Let $y(t) = \cos(2\pi f_m t)$ be a cosinusoidal modulating signal with frequency f_m. Then, for AM, the modulated carrier signal is

$$x(t) = A_c\left[1 + m\cos(2\pi f_m t)\right]\cos(2\pi f_c t) \tag{2.15}$$

where m is the modulation index (e.g. 100% AM has $m = 1$). Thus if the power of just one quasi-period of $x(t)$, i.e. one cycle of the pseudo carrier, is considered then $x(t)$ has a power that varies with time.

Consider a voltage $v(t)$ across a resistor of conductance G. The power of the signal is determined by integrating over all time, which is work, and dividing by the time period. This yields the average power:

$$P_{\mathrm{avg}} = \lim_{\tau\to\infty}\int_{-\tau}^{\tau}\frac{1}{2\tau}Gv^2(t)\,dt. \tag{2.16}$$

Now, if $v(t)$ is a sinusoidal, $v(t) = A\cos\omega t$, then

$$\begin{aligned} P_{\mathrm{avg}} &= \lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}A_c^2G\cos^2(\omega t)\,dt \\ &= \lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}A_c^2G\frac{1}{2}\left[1+\cos(2\omega t)\right]\,dt \\ &= \tfrac{1}{2}A_c^2G\left\{\lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}1\,dt + \lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\cos(2\omega t)\,dt\right\} = \tfrac{1}{2}A_c^2G. \end{aligned} \tag{2.17}$$

In the above equation, a useful equivalence has been employed by observing that the infinite integral of a cosinusoid can be simplified to just integrating over one period, $T = 2\pi/\omega$:

$$\lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\cos^n(\omega t)\,dt = \frac{1}{T}\int_{-T/2}^{T/2}\cos^n(\omega t)\,dt, \tag{2.18}$$

where n is a positive integer. In power calculations there are a number of other useful simplifying techniques based on trigonometric identities. Some

of the ones that will be used here are the following:

$$\cos A \cos B = \tfrac{1}{2}\left[\cos(A-B) + \cos(A+B)\right]$$
$$\cos^2 A = \tfrac{1}{2}\left[1 + \cos(2A)\right] \tag{2.19}$$

$$\lim_{\tau\to\infty} \frac{1}{2\tau}\int_{-\tau}^{\tau} \cos\omega t\, dt = \frac{1}{T}\int_{-T/2}^{T/2} \cos(\omega t)\, dt = 0 \tag{2.20}$$

$$\frac{1}{T}\int_{-T/2}^{T/2} \cos^2(\omega t)\, dt = \frac{1}{T}\int_{-T/2}^{T/2} \frac{1}{2}\left[\cos(2\omega t) + \cos(0)\right] \tag{2.21}$$

$$= \frac{1}{2T}\left[\int_{-T/2}^{T/2} \cos(2\omega t)\, dt + \int_{-T/2}^{T/2} 1\, dt\right]$$

$$= \frac{1}{2T}(0+T) = \tfrac{1}{2}\,. \tag{2.22}$$

More trigonometric identities are given in Appendix 1.A.2 of [4]. Also, when cosinusoids $\cos\omega_A t$ and $\cos\omega_B t$, having different frequencies ($\omega_A \neq \omega_B$), are multiplied together, for large τ,

$$\int_{-\tau}^{\tau} \cos\omega_A t \cos\omega_B t\, dt = \int_{-\tau}^{\tau} \tfrac{1}{2}\left[\cos(\omega_A+\omega_B)t + \cos(\omega_A-\omega_B)t\right]\, dt = 0,$$

and if $\omega_A \neq \omega_B \neq 0$, $\displaystyle\int_{-\infty}^{\infty} \cos\omega_A t \cos^n \omega_B t\, dt = 0.$ (2.23)

Now the discussion returns to characterizing an AM signal by considering the long-term average power and the maximum short-term power of the signal. The pseudo-carrier at its peak amplitude is, from Equation (2.15),

$$x_p(t) = A_c\left[1+m\right]\cos(2\pi f_c t)\,. \tag{2.24}$$

Then the power (P_{PEP}) of the peak pseudo carrier is obtained by integrating over one period of the pseudo carrier:

$$P_{\mathrm{PEP}} = \frac{1}{T}\int_{-T/2}^{T/2} Gx^2(t)\, dt = \frac{1}{T}\int_{-T/2}^{T/2} A_c^2 G(1+m)^2\cos^2(\omega_c t)\, dt$$

$$= A_c^2 G\,(1+m)^2\,\frac{1}{T}\int_{-T/2}^{T/2}\cos^2(\omega_c t)\, dt = \frac{1}{2}A_c^2 G\,(1+m)^2\,. \tag{2.25}$$

The **average power** (P_{avg}) of the modulated signal is obtained by integrating over all time, so

$$P_{\mathrm{avg}} = \lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau} Gx^2(t)\, dt$$

$$= A_c^2 G \lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\left\{\left[1+m\cos(\omega_m t)\right]\cos(\omega_c t)\right\}^2\, dt$$

$$= A_c^2 G \lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\left\{\left[1+2m\cos(\omega_m t)+m^2\cos^2(\omega_m t)\right]\cos^2(\omega_c t)\right\}\, dt$$

$$= A_c^2 G\left[\lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\cos^2(\omega_c t)\, dt + \lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau} 2m\cos(\omega_m t)\cos^2(\omega_c t)\, dt\right.$$

$$\left. + \lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau} m^2\cos^2(\omega_m t)\cos^2(\omega_c t)\, dt\right]$$

$$= A_c^2 G\left\{\tfrac{1}{2} + 0 + m^2 \lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\tfrac{1}{4}\left[1+\cos(2\omega_m t)\right]\left[1+\cos(2\omega_c t)\right]\, dt\right\}$$

$$= A_c^2 G \left\{ \tfrac{1}{2} + \frac{m^2}{4} \left[\lim_{\tau\to\infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} 1\, dt + \lim_{\tau\to\infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos(2\omega_m t)\, dt \right.\right.$$
$$\left.\left. + \lim_{\tau\to\infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos(2\omega_c t)\, dt + \lim_{\tau\to\infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos(2\omega_m t)\cos(2\omega_c t)\, dt \right] \right\}$$
$$= A_c^2 G \left[\tfrac{1}{2} + m^2 \left(\tfrac{1}{4} + 0 + 0 + 0 \right) \right] = \tfrac{1}{2} A_c^2 G (1 + m^2/2)\ . \quad (2.26)$$

Thus the rms voltage, $x_{\rm rms}$, can be determined as $P_{\rm avg} = x_{\rm rms}^2 G$. So the PMEPR of an AM signal (i.e., $\rm PMEPR_{AM}$) is

$$\rm PMEPR_{AM} = \frac{P_{PEP}}{P_{avg}} = \frac{\frac{1}{2} A_c^2 G\,(1+m)^2}{\frac{1}{2} A_c^2 G (1 + m^2/2)} = \frac{(1+m)^2}{1 + m^2/2}.$$

For 100% AM described by $m = 1$, the PMEPR is

$$\rm PMEPR_{100\%AM} = \frac{(1+1)^2}{1 + 1^2/2} = \frac{4}{1.5} = 2.667 = 4.26\ dB. \quad (2.27)$$

In expressing the PMEPR in decibels, the formula $\rm PMEPR_{dB} = 10\log(PMEPR)$ is used as PMEPR is a power ratio. As an example, for 50% AM, described by $m = 0.5$, the PMEPR is

$$\rm PMEPR_{50\%AM} = \frac{(1+0.5)^2}{1 + 0.5^2/2} = \frac{2.25}{1.125} = 2 = 3\ dB. \quad (2.28)$$

2.2.3 Two-Tone Signal

In assessing, either through laboratory measurements or simulations, it is common and often necessary to use very simple representations of a baseband signal or even of a modulated signal. This greatly simplifies matters and there is a justified expectation that the performance with the test signal is a good indication of performance with an actual baseband or modulated signal. With simulation at the circuit level it is usually impossible to consider real baseband signals as simulation may not even be possible or simulation may take unacceptable times. Instead it is common to use single-tone, i.e. single sinewave, or two-tone signals. A two-tone signal is a signal that is the sum of two cosinusoids:

$$y(t) = X_A \cos(\omega_A t) + X_B \cos(\omega_B t). \quad (2.29)$$

Generally the frequencies of the two tones are close ($|\omega_A - \omega_B| \ll \omega_A$), with the concept being that both tones fit within the passband of a transmitter's or receiver's bandpass filters. A two-tone signal is not a form of modulation, but is commonly used to characterize the nonlinear performance of RF systems and has an envelope that is similar to that of many modulated signals. The composite signal, $y(t)$, looks like a pseudo-carrier with a slowly varying amplitude, not unlike an AM signal. The tones are uncorrelated so that the average power of the composite signal, $y(t)$, is the sum of the powers of each of the individual tones. The peak power of the composite signal is that of the peak pseudo-carrier, so $y(t)$ has a peak amplitude of $X_A + X_B$. The peak pseudo carrier is the single RF sinusoid where the sinusoid of each sinusoid align as much as possible. Similar concepts apply to three-tone and n-tone signals.

EXAMPLE 2.2 PMEPR of a Two-Tone Signal

What is the PMEPR of a two-tone signal with the tones having equal amplitude?

Solution:

Let the amplitudes of the two tones be X_A and X_B. Now $X_A = X_B = X$, and so the peak pseudo-carrier has amplitude $2X$, and the power of the peak RF carrier is proportional to $\frac{1}{2}(2X)^2 = 2X^2$. The average power is proportional to $\frac{1}{2}(X_A^2 + X_B^2) = \frac{1}{2}(X^2 + X^2) = X^2$, as each tone is independent of the other and so the powers can be added.

$$\text{PMEPR} = \frac{P_{\text{PEP}}}{P_{\text{avg}}} = \frac{2X^2}{X^2} = 2 = 3 \text{ dB}. \tag{2.30}$$

EXAMPLE 2.3 PMEPR of Uncorrelated Signals

Consider the combination of two uncorrelated analog signals, e.g. a two-tone signal. One signal is denoted $x(t)$ and the other $y(t)$, where $x(t) = 0.1 \sin(10^9 t)$ and $y(t) = 0.05 \sin(1.01 \cdot 10^9 t)$. What is the PMEPR of this combined signal?

Solution:

These two signals are uncorrelated and this is key in determining the average power, P_{avg}, as the sum of the powers of each individual signal (k is a proportionality constant):

$$P_{\text{avg}} = \int_{-\infty}^{\infty} x^2(t) \cdot dt + \int_{-\infty}^{\infty} y^2(t) \cdot dt = \frac{k}{2}(0.1)^2 + \frac{k}{2}(0.5)^2 = \frac{k}{2}[0.01 + 0.0025] = 0.00625k.$$

The two carriers are close in frequency so that the sum signal $z(t) = x(t) + y(t)$ looks like a slowly varying signal with a radian frequency near 10^9 rads per second. The peak amplitude of one pseudo-cycle of $z(t)$ is $0.1 + 0.05 = 0.15$. Thus the power of the largest cycle is

$$P_{\text{PEP}} = \tfrac{1}{2}k(0.15)^2 = 0.01125k,$$

and so $$\text{PMEPR} = \frac{P_{\text{PEP}}}{P_{\text{avg}}} = \frac{0.01125}{0.00625} = 1.8 = 2.55 \text{ dB}. \tag{2.31}$$

Summary

The PMEPR is an important attribute of a modulation format and impacts the types of circuit designs that can be used. It is much more challenging to develop power-efficient hardware introducing only low levels of distortion when the PMEPR is high.

It is tempting to consider if the lengthy integrations can be circumvented. Powers can be added if the signal components (the tones making up the signal) are uncorrelated. If they are correlated, then the complete integrations are required. Consider two uncorrelated sinusoids of (average) powers P_1 and P_2, respectively, then the average power of the composite signal is $P_{\text{avg}} = P_1 + P_2$. However, in determining the peak sinusoidal power, the RF cycle where the two largest pseudo-carrier sinusoids align is considered, and here the voltages add to produce a single cycle of a sinewave with a higher amplitude. So peak power applies to just one RF pseudo-cycle. Generally the voltage amplitude of the two sinewaves would be added and then the power calculated. If the uncorrelated carriers are modulated and the modulating signals (the baseband signals) are uncorrelated, then the average power can be determined in the same way, but the peak power calculation is much more complicated. The integrations are the only calculations that can always be relied on and can be used with all modulated signals.

Signals $x(t)$ and $y(t)$ are **uncorrelated** if the integral over all time and time offsets of their product is zero: $C = \int_{-\infty}^{+\infty} x(t)y(t+\tau)\,dt = 0$ for all τ.

The preferred usage of PAR, PAPR, or PMEPR in RF and microwave engineering is currently in a transition phase. The most common usage of PAR and PAPR in electrical engineering refers to the peak of a signal as being the instantaneous peak value, and in the case of PAPR, the instantaneous power of the signal is calculated as if the peak is a DC value. In the past, many RF and microwave publications have taken the peak as the peak power of a sinusoid having an amplitude equal to the peak voltage of the signal and used that to calculate PAR. This usage is inconsistent with the predominant usage in electrical engineering and is a particular problem when using wireless technology in other disciplines. PMEPR is the preferred usage for what RF and microwave engineers intend to refer to when using the term PAR. A reader of RF literature encountering PAR needs to determine how the term is being used. There is no confusion if PMEPR is used.

EXAMPLE 2.4 PAPR and PMEPR of an AM signal

What is the PAPR and PMEPR of a 100% AM signal?

Solution:

The signal is $x(t) = A_c\left[1+\cos 2\pi f_m t\right]\cos 2\pi f_c t$ and the PMEPR of this signal, from Equation (2.27), is 4.26 dB. Now PAPR uses the absolute maximum value of the signal rather than the maximum short-term power of the envelope. The peak value of $x(t)$ is $2A_c$ so the peak power (if the signal is a voltage across a conductance G) is

$$P_{\text{peak,PAPR}} = (2A_c)^2\, G. \tag{2.32}$$

P_{avg} is the same for PAPR and PMEPR for the AM signal, see Equation (2.26), so that

$$\text{PAPR} = \frac{P_{\text{peak,PAPR}}}{P_{\text{avg}}} = \frac{(2A_c)^2\, G}{\frac{1}{2}A_c^2\left(1+\frac{1}{2}\right)} = \frac{4}{3/4} = \frac{16}{3} = 5.333 = 7.27 \text{ dB}. \tag{2.33}$$

So PAPR is 3 dB higher than PMEPR for a 100% modulated AM signal, see Equation (2.27). This is not always the case for other modulation schemes.

2.3 Modulation Overview

There are two families of modulation methods with analog modulation used in early radios including 1G cellular radio, and digital modulation used in modern radios starting with 2G cellular radio. While 1G cellular radio transmitted voice signals using analog modulation, 1G also used a simple type of digital modulation for signaling. With the exception of **ultra-wideband** (**UWB**) **pulse radio** [5], all modern radio modulation schemes slowly vary the amplitude, phase, or frequency of a sinusoidal signal called the carrier. This results in a narrow bandwidth modulated signal perhaps with fractional bandwidth typically in the range of 0.002% to 2%. The early spark-gap wireless telegraph systems were ultra-wideband but they were soon discontinued because they interfered with conventional radios which were soon developed and assigned specific parts, i.e. bands, of the spectrum. The initial pulse radio concept of the 1990s occupied most of the spectrum between 3.1 and 10.6 GHz but was never deployed mainly because capacity was relatively poor. The term ultra-wideband wireless is now widely taken to mean a wireless device such as a radar or radio with a bandwidth which is at least the lesser of 500 MHz or 20% of the carrier frequency [6]. So even the UWB millimeter-wave radios exploiting the high bandwidth available

at millimeter wave frequencies still employ a relatively slowly varying modulation of a carrier.

2.4 Analog Modulation

The waveforms and spectra of the signals with common analog modulation methods are shown in Figure 2-3. The modulating signal is generally referred to as the baseband signal and it contains all of the information to be transmitted and interpreted at the receiver. The waveforms in Figure 2-3 are stylized. They are presented this way so that the effects of modulation can be more easily seen. The baseband signal (Figure 2-3(a)) is shown as having a period that is not much greater than the period of the carrier (Figure 2-3(b)). In reality there would be hundreds or thousands of RF cycles for each cycle of the baseband signal so that the highest frequency component of the baseband signal is a tiny fraction of the carrier frequency. In this situation the spectra shown on the right in Figure 2-3(c–e) would be too narrow to enable any detail to be seen.

2.4.1 Amplitude Modulation

Amplitude Modulation (AM) is the simplest analog modulation method to implement. Here a signal is used to slowly vary the amplitude of the carrier according to the level of the modulating signal. With AM (Figure 2-3(c)) the amplitude of the carrier is modulated, and this results in a broadening of the spectrum of the carrier, as shown in Figure 2-3(c)(ii). This spectrum contains the original carrier component and upper and lower sidebands, designated as U and L, respectively. In AM, the two sidebands contain identical information, so all the information contained in the baseband signal is conveyed if just one sideband is transmitted.

The basic AM signal $x(t)$ has the form

$$x(t) = A_c\,[1 + my(t)]\cos(2\pi f_c t), \tag{2.34}$$

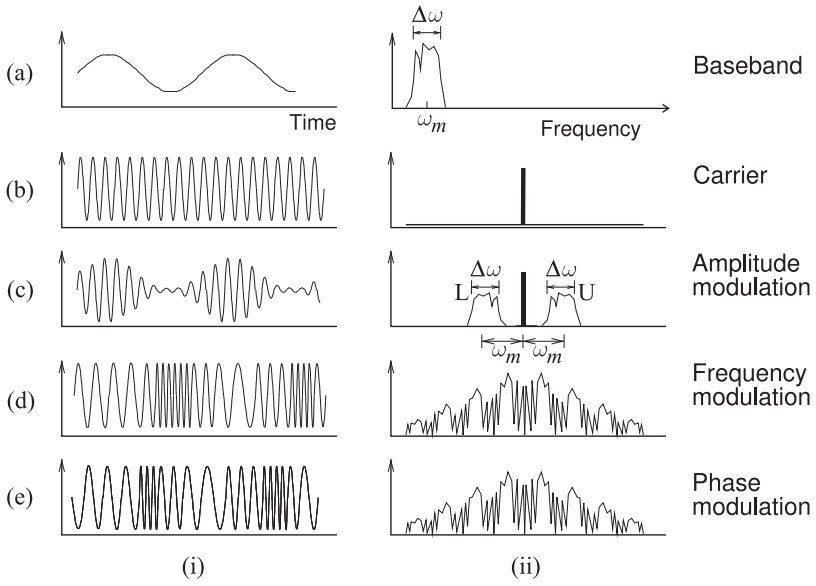


Figure 2-3: Basic analog modulation showing the (i) waveform and (ii) spectrum for (a) baseband signal; (b) carrier; (c) carrier modulated using amplitude modulation; (d) carrier modulated using frequency modulation; and (e) carrier modulated using phase modulation.

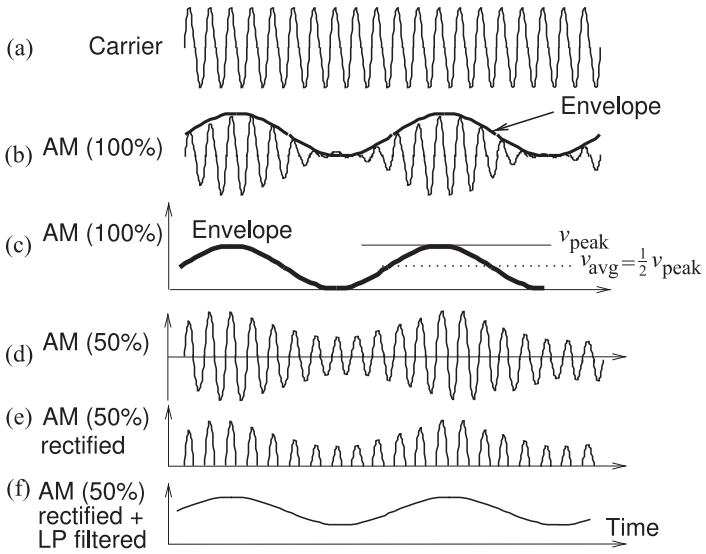


Figure 2-4: AM showing the relationship between the carrier and modulation envelope: (a) carrier; (b) 100% amplitude modulated carrier; (c) modulating or baseband signal; (d) 50% amplitude-modulated carrier; (e) rectified 50% AM modulated signal; and (f) rectified and lowpass (LP) filtered 50% modulated signal. The envelope contains only amplitude information and for AM the envelope is the same as the baseband signal.

where m is the modulation index, $y(t)$ is the baseband information-bearing signal that has frequency components that are much lower than the carrier frequency f_c, and the maximum value of $|y(t)|$ is one. Provided that $y(t)$ varies slowly relative to the carrier, $x(t)$ looks like a carrier whose amplitude varies slowly. To get an idea of how slowly the amplitude varies in an actual system, consider an AM radio that broadcasts at 1 MHz (which is in the middle of the AM broadcast band). The highest frequency component of the modulating signal corresponding to voice is about 4 kHz. Thus the amplitude of the carrier takes 250 carrier cycles to go through a complete amplitude variation. At all times a cycle of the modulated carrier, the pseudo-carrier, appears to be periodic, but in fact it is not quite.

The concept of the envelope of a modulated RF signal is introduced in Figure 2-4. The envelope is an important concept and is directly related to the distortion introduced by analog hardware and to the DC power requirements which determines the battery life for mobile radios. Figure 2-4(a) is the carrier and the amplitude-modulated carrier is shown in Figure 2-4(b). The outline of the modulated carrier is called the envelope, and for AM this is identical to the modulating, i.e. baseband, signal. The envelope is shown again in Figure 2-4(c). At the peak of the envelope, the RF signal has maximum short-term power (considering the power of a single RF cycle). With 100% AM, $m = 1$ in Equation (2.34), there is no short-term RF power when the envelope is at its minimum. The modulated signal with 50% modulation, $m = 0.5$, is shown in Figure 2-4(d) and at all times there is an appreciable RF signal power.

Very simple analog hardware is required to demodulate the basic amplitude modulated signal, that is an AM signal with a carrier and both sidebands. The receiver requires bandpass filtering to select the channel from the incoming radio signal then rectifying the output of the bandpass filter. The waveform after rectification of a 50% AM signal is shown in Figure 2-4(e) and contains frequency components at baseband and sidebands around harmonics of the carrier, and the harmonics of the carrier itself. Lowpass filtering of the rectified waveform extracts the original baseband signal and completes demodulation, see Figure 2-4(f). The only electronics required is a

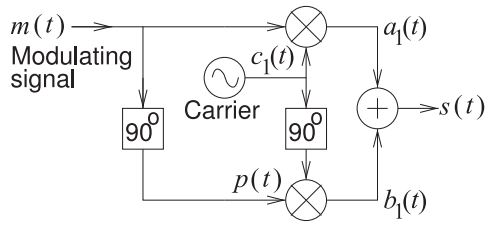


Figure 2-5: Hartley modulator implementing single-sideband suppressed-carrier (SSB-SC) modulation. The "90°" blocks shift the phase of the signal by $+90°$. The mixer indicated by the circle with a cross is an ideal multiplier, e.g. $a_1(t) = m(t) \cdot c_1(t)$.

single diode. The disadvantage is that more spectrum is used than required and the largest signal is the carrier that conveys no information but causes interference in other radios. Without the carrier and both sidebands being transmitted it is necessary to use DSP to demodulate the signal.

It is not possible to represent an actual baseband signal in a simple way and undertake the analytic derivations that illustrate the characteristics of modulation. Instead it is usual to use either a one-tone or two-tone signal, derive results, and then extrapolate the results for a finite bandwidth baseband signal. For a single-tone baseband signal $y(t) = \cos(\omega_m t + \phi)$, then the basic AM modulated signal, from Equation (2.34), is

$$\begin{aligned} x(t) &= [1 + m\cos(\omega_m t + \phi]\cos(\omega_c t) \\ &= \cos(\omega_c t) + \tfrac{1}{2}m[\cos((\omega_c - \omega_m)t - \phi) + \cos((\omega_c + \omega_m)t + \phi)] \end{aligned} \quad (2.35)$$

which has three (radian) frequency components, one at the carrier frequency ω_c, one just below the carrier at $\omega_c - \omega_m$, and one just above at $\omega_c + \omega_M$ (since $\omega_m \ll \omega_c$). The extension to a finite bandwidth baseband signal, see Figure 2-3(a)(ii), is to imagine that ω_m ranges from a lower value $\omega_m - \frac{1}{2}\Delta\omega$ to a higher value $\omega_m + \frac{1}{2}\Delta\omega$. The discrete tones in the modulated signal below and above the carrier then become finite bandwidth sidebands with a lower sideband L centered at $\omega_c - \omega_m$ and an upper sideband U centered at at $(\omega_c + \omega_m)$ each having the same bandwidth, $\Delta\omega$, as the baseband signal, see Figure 2-3(c)(ii).

The AM modulator described so far produces a modulated signal with a carrier and two sidebands. This modulation is called double-sideband (DSB) modulation. There is identical information in each of the sidebands and so only one of the sidebands needs to be transmitted. The carrier contains no information so if only one sideband was transmitted then the received **single-sideband** (**SSB**) **suppressed-carrier SC** (together **SSB-SC**) signal has all of the information needed to recover the original baseband signal. However the simple demodulation process using rectification as described earlier in this section no longer works. The receiver needs to use DSP but the spectrum is used efficiently.

One circuit that implements SSB-SC AM is the **Hartley modulator** shown in Figure 2-5. As will be seen, this basic architecture is significant and used in all modern radios. In modern radios the Hartley modulator, or a variant, takes a modulated signal which is centered at an intermediate frequency and shifts it up in frequency so that it is centered at another frequency a little below or a little above the carrier of the Hartley modulator.

In a Hartley modulator both the modulating signal $m(t)$ and the carrier are multiplied together in a mixer and then also $90°$ phase-shifted versions are mixed before being added together. The signal flow is as follows beginning with $m(t) = \cos(\omega_m t + \phi)$, $p(t) = \cos(\omega_m t + \phi - \pi/2) = \sin(\omega_m t + \phi)$ and

carrier signal $c_1(t) = \cos(\omega_c t)$:

$$\begin{aligned} a_1(t) &= \cos(\omega_m t + \phi)\cos(\omega_c t) = \tfrac{1}{2}[\cos((\omega_c - \omega_m)t - \phi) + \cos((\omega_c + \omega_m)t + \phi)] \\ b_1(t) &= \sin(\omega_m t + \phi)\sin(\omega_c t) = \tfrac{1}{2}[\cos((\omega_c - \omega_m)t - \phi) - \cos((\omega_c + \omega_m)t + \phi)] \\ s(t) &= a_1(t) + b_1(t) = \cos((\omega_c - \omega_m)t - \phi) \end{aligned} \quad (2.36)$$

and so the lower sideband (LSB) is selected. An interesting observation is that the phase, ϕ, of the baseband signal is also translated up in frequency. A feature that is not exploited in AM but is in digital modulation.

2.4.2 Phase Modulation

In phase modulation (**PM**) the phase of the carrier depends on the instantaneous level of the baseband signal. The phase-modulated carrier is shown in Figure 2-3(e)(i) and it looks like the frequency of the modulated carrier is changing. What is actually happening is that when the phase is changing most quickly the apparent frequency of the RF waveform changes. Here, as the baseband signal is decreasing, the phase shift reduces and the effect is to increase the apparent frequency of the RF signal. As the baseband signal increases, the effect is to reduce the apparent frequency of the modulated RF signal. The result is that with PM is that the bandwidth of the time-varying signal is spread out, as seen in Figure 2-6. PM can be implemented using a phase-locked loop (PLL) but further details will be skipped here.

Consider a phase-modulated signal $s(t) = \cos(\omega_c t + \phi(t))$ where $\phi(t)$ is the baseband signal containing the information to be transmitted. The spectrum of $s(t)$ can be determined by simplifying $\phi(t)$ as a sinusoid with frequency $f_m = 2\pi\omega_m$ so that $\phi(t) = \beta\cos(\omega_m t)$ where β is the phase modulation index. (The maximum possible phase change is $\pm\pi$ and then $\beta = \pi$.) The phase-modulated signal becomes

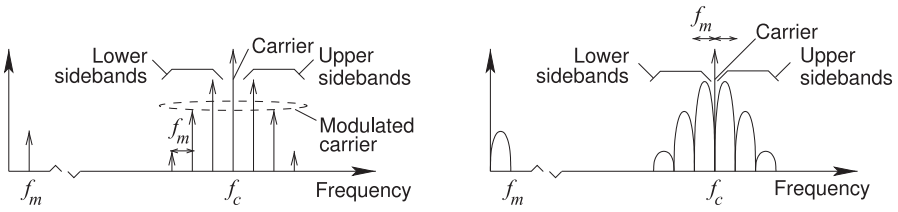
$$\begin{aligned} s(t) &= \cos(\omega_c t + \beta\cos(\omega_m t)]) \\ &= \cos(\omega_c t)\cos(\cos(\beta\omega_m t)) - \sin(\omega_c t)\sin(\cos(\beta\omega_m t)) \end{aligned} \quad (2.37)$$

which has the Bessel function-based expansion

$$\begin{aligned} s(t) &= J_0(\beta)\cos(\omega_c t) \\ &+ J_1(\beta)\cos(\omega_c + \omega_m)t + \pi/2) + J_1(\beta)\cos(\omega_c - \omega_m)t + \pi/2) \\ &+ J_2(\beta)\cos(\omega_c + 2\omega_m)t + \pi) + J_2(\beta)\cos(\omega_c - 2\omega_m)t + \pi] \\ &+ J_3(\beta)\cos(\omega_c + 3\omega_m)t + 3\pi/2) + J_3(\beta)\cos(\omega_c - 3\omega_m)t + 3\pi/2) + \ldots \end{aligned} \quad (2.38)$$

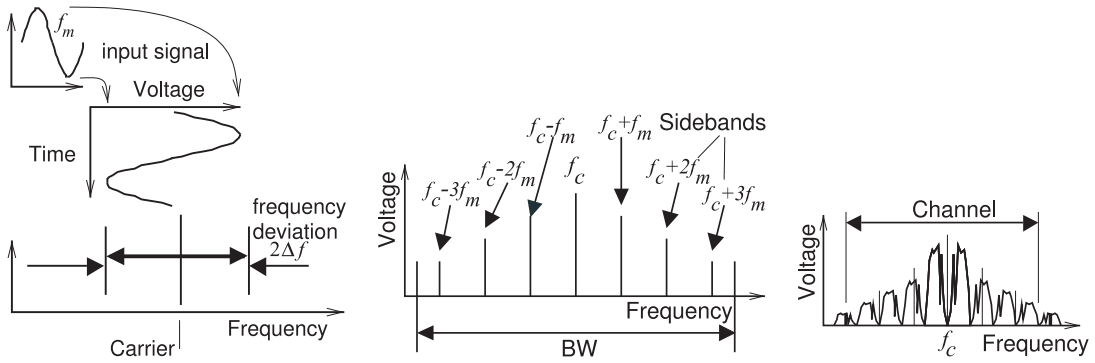
where J_n is the Bessel function of the first kind of order n. The spectrum of this signal is shown in Figure 2-6(a) which consists of discrete tones grouped as lower- and upper-sideband sets centered on the carrier at f_c. The discrete tones in the sidebands are separated from each other and from f_c by f_m. The sidebands have lower amplitude further away from the carrier.

If the modulating signal has a finite bandwidth, approximated by f_m varying from a minimum value, $(f_m - \Delta f)$ up to the maximum frequency $(f_m + \Delta f)$, then the spectrum of the modulated signal becomes that shown in Figure 2-6(b), with the centers of adjacent sidebands separated by f_m and the first sidebands separated from the carrier by f_m as well. This is DSB



(a) With a discrete modulating tone (b) With a finite bandwidth modulating signal

Figure 2-6: Spectrum of a phase-modulated carrier which includes the carrier at f_c and upper and lower sidebands with the spectrum of the discrete modulating signal at f_m.



(a) (b) (c)

Figure 2-7: Frequency modulation: (a) sinusoidal baseband signal shown varying the frequency of the carrier and so FM modulating the carrier; (b) the spectrum of the resulting FM-modulated waveform; and (c) spectrum of the modulated carrier when it is modulated by a broadband baseband signal such as voice.

modulation and there is a carrier (so it is not suppressed). The sidebands do not carry identical information and several, perhaps three below and three above the carrier, are required to enable demodulation of a PM signal. Thus a rather large bandwidth is required to transmit the modulated signal.

2.4.3 Frequency Modulation

The other analog modulation schemes commonly used is frequency modulation (**FM**), see Figure 2-3(d). The signals produced by FM and PM appear to be similar; the difference is in how the signals are generated. In FM, the amplitude of the baseband signal determines the frequency of the modulated carrier. Consider the FM waveform in Figure 2-3(d)(i). When the baseband signal is at its peak value the modulated carrier is at its minimum frequency, and when the signal is at is lowest value the modulated carrier is at its maximum frequency. (Depending on the hardware implementation it could be the other way around.) The result is that the bandwidth of the time-varying signal is spread out, as seen in Figure 2-7.

One way of implementing the FM modulator is to use a voltage-controlled

oscillator (VCO) with the baseband signal controlling the frequency of an oscillator. An FM receiver must compress, in frequency, the transmitted signal to re-create the original narrower bandwidth baseband signal. FM demodulation can be thought of as providing signal enhancement or equivalently noise suppression in a process that can be called analog processing gain. Only the components of the original FM signals are coherently collapsed to a narrower bandwidth baseband signal while noise, being uncorrelated, is still spread out (although rearranged). Thus the ratio of the signal to noise powers increases, as after demodulation only the power of the noise in the smaller bandwidth of the baseband signal is important. Thus compared to AM, FM significantly increases the tolerance to noise that may be added to the signal during transmission. PM has the same property, although the details of modulation and demodulation are different. For both FM and PM signals the peak amplitude of the RF **phasor** is equal to the average amplitude, and so the PMEPR is 1 or 0 dB.

Consider an FM signal $s(t) = \cos([\omega_c + x(t)]t)$ where $x(t)$ is the baseband signal containing the information to be transmitted. The spectrum of $s(t)$ can be determined by simplifying $x(t)$ as a sinusoid with frequency $f_m = 2\pi\omega_m$ so that $x(t) = \beta\cos(\omega_m t)$ where β is the frequency modulation index. The FM signal becomes

$$\begin{aligned} s(t) &= \cos([\omega_c + \beta\cos(\omega_m t)]t) \\ &= \cos(\omega_c t)\cos(\cos(\omega_m t)\beta t) - \sin(\omega_c t)\sin(\cos(\omega_m t)\beta t) \end{aligned} \tag{2.39}$$

which has the Bessel function-based expansion

$$\begin{aligned} s(t) &= J_0(\beta t)\cos(\omega_c t) \\ &- J_1(\beta)\sin(\omega_c + \omega_m)t + \pi/2) - J_1(\beta t)\sin(\omega_c - \omega_m)t + \pi/2) \\ &- J_2(\beta t)\cos(\omega_c - 2\omega_m)t + \pi) + J_2(\beta t)\cos(\omega_c - 2\omega_m)t + \pi] \\ &+ J_3(\beta t)\sin(\omega_c + 3\omega_m)t + 3\pi/2) + J_3(\beta t)\sin(\omega_c - 3\omega_m)t + 3\pi/2) + \ldots \end{aligned} \tag{2.40}$$

where J_n is the Bessel function of the first kind of order n. The spectrum of this signal is shown in Figure 2-7(b) which consists of discrete tones grouped as lower- and upper-sideband sets centered on the carrier at f_c. The discrete tones in the sidebands are separated from each other and from f_c by f_m. The sidebands have lower amplitude further away from the carrier.

If the modulating signal has a finite bandwidth, approximated by $f_m = \omega_m/(2\pi)$ varying from a minimum value $(f_m - \Delta f)$ up to the maximum frequency $(f_m + \Delta f)$, then the spectrum of the modulated signal becomes that shown in Figure 2-7(c) with the centers of adjacent sidebands separated by f_m and the first sidebands from the carrier by f_m as well. This is DSB modulation and there is a carrier (so it is not suppressed but is smaller than with AM). The sidebands do not carry identical information and several, perhaps three on either side of the carrier, are required to enable demodulation of an FM signal. Thus a rather large bandwidth to required to transmit the modulated signal as it is not sufficient to transmit just one sideband to enable demodulation.

Carson's Rule

Frequency- and phase-modulated signals have a very wide spectrum and the bandwidth required to reliably transmit a PM or FM signal is subjective. The best accepted criterion for determining the bandwidth requirement is called Carson's bandwidth rule or just Carson's rule [7, 8].

An FM signal is shown in Figure 2-7. In particular, Figure 2-7(a) shows the FM function. The level (typically voltage) of the baseband signal determines the frequency deviation of the carrier from its unmodulated value. The frequency shift when the modulating signal is a DC value x_m at its maximum amplitude is called the peak frequency deviation, Δf. So, if the modulating signal changes very slowly, the bandwidth of the modulated signal is $2\Delta f$.

A rapidly varying sinusoidal modulating signal produces a modulated signal with many discrete sidebands as seen in Figure 2-7(b). If the modulating baseband signal is broadband, then the sidebands have finite bandwidth as seen in Figure 2-7(c) and many are required to recover the original baseband signal. These sidebands continue indefinitely in frequency but rapidly reduce in power away from the frequency of the unmodulated carrier. Carson's rule provides an estimate of the bandwidth that contains 98% of the energy. If the maximum frequency of the modulating signal is f_m, and the maximum value of the modulating waveform is x_m (which would produce a frequency deviation of Δf if it is DC), then Carson's rule is that the

$$\text{bandwidth required} = 2 \times (f_m + \Delta f)\,. \tag{2.41}$$

Narrowband and Wideband FM

The most common type of FM signal, as used in FM broadcast radio, is called wideband FM, as the maximum frequency deviation is much greater than the highest frequency of the modulating or baseband signal, that is, $\Delta f \gg f_m$. In narrowband FM, Δf is close to f_m. Narrowband FM uses less bandwidth but requires a more sophisticated demodulation technique.

EXAMPLE 2.5 PAPR and PMEPR of FM Signals

Consider FM signals close in frequency but whose spectra do not overlap.

(a) What are PAPR and PMEPR of just one FM signal?
(b) What are PAPR and PMEPR of a signal comprised of two uncorrelated narrowband FM signals each having a small fractional bandwidth and having the same average power.

Solution:

(a) An FM signal has a constant envelope just like a single sinusoid, and so $\text{PAPR} = 1.414 = 3.01$ dB and $\text{PMEPR} = 1 = 0$ dB.
(b) Since the modulation is relatively slow, each of the FM signals will look like single tone signals and the combined signal will look like a two-tone signal. However this is not enough to solve the problem. A thought experiment is required to determine the largest pseudo-carrier when the FM signals combine. If the amplitude of each tone is X, then the amplitude when the FM signal waveforms align is $2X$. (This is the same as the peak of a two-tone signal but arrived at differently.) Then

$$P_{\text{avg}} = \text{sum of the powers of each FM signal} = 2k\tfrac{1}{2}X^2,$$

where k is a proportionality constant. For PAPR,

$$P_p = k(2X)^2 \quad \text{and} \quad \text{PMEPR} = \frac{P_p}{P_{\text{avg}}} = \frac{k(2X)^2}{2k\frac{1}{2}X^2} = 4 = 6.0 \text{ dB}. \tag{2.42}$$

For PMEPR, P_{PEP} = power of the pseudo-carrier $= k\frac{1}{2}(2X)^2$

and $$\text{PMEPR} = \frac{P_p''}{P_{\text{avg}}} = \frac{k\frac{1}{2}(2X)^2}{2k\frac{1}{2}X^2} = 2 = 3.0 \text{ dB}. \tag{2.43}$$

2.4.4 Analog Modulation Summary

Analog modulation was used in the first radios and in 1G cellular radios. Radio transmission using analog modulation, i.e. analog radio, has almost ceased as it does not use spectrum efficiently. Digital modulation along with error correction, can pack much more information in a limited bandwidth. A final comparison of the analog modulation techniques is given in Figure 2-8 emphasizing the PMEPR of AM and FM. The PMEPR of PM is the same as for FM.

One particular event in the development of radio is illustrative of the relationship of technology and business interests. Frequency modulation was invented by Edwin H. Armstrong and patented in 1933 [9, 10]. FM is virtually static free and clearly superior to AM radio. However, it was not immediately adopted largely because AM radio was established in the 1930s, and the adoption of FM would have resulted in the scrapping of a large installed infrastructure (seen as a commercial catastrophe) and so the introduction of FM was delayed by decades. The best technology does not always win immediately! Commercial interests and the large investment in an alternative technology have a great deal to do with the success of a technology [11].

With FM and PM there are two sets of sidebands with one set above the carrier frequency and the other set below. The carrier itself is low-level but is not completely suppressed. Now SSB modulation refers to producing just one of the sideband sets. There is such as thing as SSB FM with just a few sidebands below (or above) the carrier but it is more like a combination of FM and AM [12], and was never deployed.

2.5 Digital Modulation

Digital radio transmits bits by creating discrete states, usually discrete amplitudes and phases of a carrier. The process of creating these discrete states from a digital bitstream is called digital modulation. A state is established at a particular time called a clock tick. What that means is that the

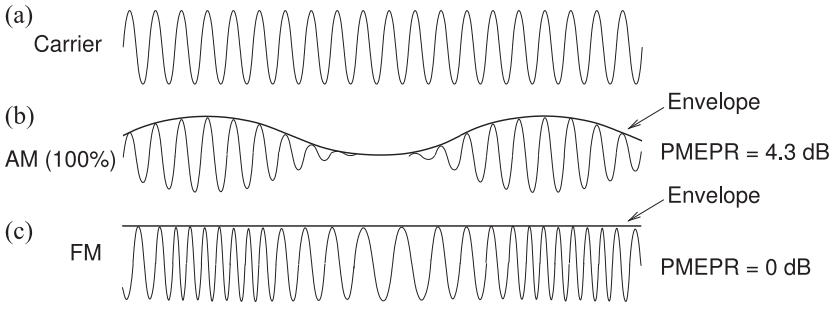


Figure 2-8: Comparison of 100% AM and FM highlighting the envelopes of both: (a) carrier; (b) AM signal; and (c) FM signal with constant envelope.

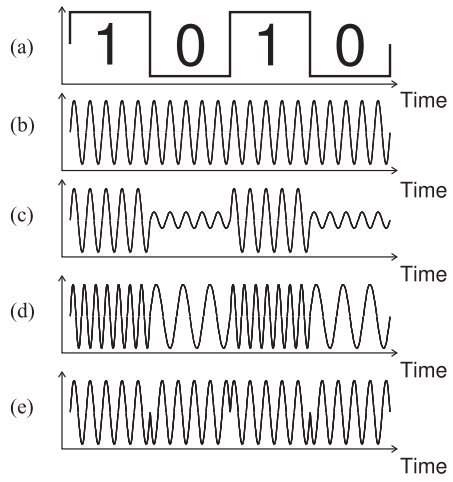


Figure 2-9: Modes of digital modulation: (a) modulating bitstream; (b) **carrier**; (c) carrier modulated using amplitude shift keying (ASK); (d) carrier modulated using frequency shift keying (FSK); and (e) carrier modulated using binary phase shift keying (BPSK).

information in the signal is the state of the waveform, such as the amplitude and the phase of a phasor, at every clock tick such as every microsecond. The time it takes to go from one state to another (a clock tick interval) defines the bandwidth of the modulated signal. For example, if a clock tick is at every microsecond the bandwidth of the modulated signal is about one megahertz as it takes about one microsecond to go from one state to another. The inverse relationship of the interval between clock ticks to bandwidth is only approximate (as will be seen when software-defined radio is considered in a future chapter).

One important digital modulation method does not fit with the description above. This is Frequency Shift Keying (FSK) modulation where the carrier is set to a particular frequency at each clock tick.

The basic digital modulation formats are shown in Figure 2-9. The fundamental characteristic of digital modulation is that there are discrete states, each of which is also known as a symbol, with a symbol defining the value of one or more bits. For example, the states are different frequencies in FSK and different phases in phase shift keying (PSK). With the modulated waveforms shown in Figure 2-9 there are only two states, which is the same as saying that there are two symbols, each symbol having one bit of information (either 0 or 1). With multiple states groups of bits can be represented.

In this section many methods of digital modulation are described. The first few methods are binary modulation methods with just two symbols with one symbol indicating that a single bit is '0' and the other symbol indicating that it is a '1'. Then four-state modulation is introduced with four symbols with each symbol indicating the values of two bits. Higher-order modulation schemes can send more than more bits per symbol and thus more bits per second (bits/s) per hertz of bandwidth. There is a limit to the number of symbols as the "distance" between symbols becomes smaller and the effect of noise, interference, and circuit distortion can cause a symbol to be misinterpreted as another. A modulation method that sends more bits per symbol is said to have higher modulation efficiency. This and other metrics that enable modulation methods to be compared are defined in the next subsection.

2.5.1 Modulation Efficiency

With digital modulation, the information being sent is in the form of bits and it is possible to send more that one bit per second in one hertz of bandwidth. This is because in digital modulation there can be several bits per symbol, however the bandwidth of the modulated signal is determined by the rate of change from one state to another, whereas the number of bits per transition depends on the number of states. It is important for the transition to be no faster than required so as to minimize bandwidth.

The ratio of the **bit rate** in bits per second (bits/s) to the bandwidth (BW) in hertz is called the **modulation efficiency**, η_c, and has the units of bits per second per hertz (bits/s/Hz). The modulation efficiency is also called the channel efficiency, hence the subscript c on η_c. The bits here are the gross bits which includes the information bits and bits added for error correction and others added to aid in identifying the signal, and so η_c is a measure of the performance of the modulation method itself. Thus

$$\text{modulation efficiency} = \eta_c = \frac{\text{gross bit rate}}{\text{bandwidth}}. \tag{2.44}$$

The additional bits added to a bit stream are called **coding bits** and the process of adding the coding bits is called coding. If coding is used, then the information rate is lower than the gross bit rate transmitted. Thus gross bit rate refers to the bits actually transmitted and **information rate** (or **information bit rate**) refers to the bit rate of information transmission. The **link spectrum efficiency** is the information bit rate divided by the bandwidth. Often the term "link" is dropped and just **spectrum efficiency** is used (with units of bits/s/Hz). Thus

$$\text{link spectrum efficiency} = \frac{\text{information bit rate}}{\text{bandwidth}} \leq \eta_c. \tag{2.45}$$

EXAMPLE 2.6 Modulation Efficiency

A radio transmits a bit stream of 2 Mbits/s using a bandwidth of 1 MHz.

(a) What is the modulation efficiency?
(b) If 25% of the bits are used for error correction, what is the modulation efficiency?
(c) With error correction coding, what is the information rate?
(d) With error correction coding, what is the link spectrum efficiency?

Solution:

(a) The gross bit rate is 2 Mbits/s and the bandwidth is 1 MHz. So

$$\eta_c = \text{modulation efficiency} = \frac{\text{gross bit rate}}{\text{bandwidth}} = \frac{2 \text{ Mbits/s}s}{1 \text{ MHz}} = 2 \text{ bits/s/Hz}.$$

(b) The modulation efficiency is unaffected by error correction coding. So the modulation efficiency is unchanged:

$$\eta_c = \text{modulation efficiency} = \frac{\text{gross bit rate}}{\text{bandwidth}} = \frac{2 \text{ Mbits/s}}{1 \text{ MHz}} = 2 \text{ bits/s/Hz}.$$

(c) With 25% of the bits in the gross bit stream being coding bits, the information rate is 75% of 2 Mbits/s or 1.5 Mbits/s.

(d) $$\text{link spectrum efficiency} = \frac{\text{information bit rate}}{\text{bandwidth}} = \frac{1.5 \text{ Mbits/s}}{1 \text{ MHz}} = 1.5 \text{ bits/s/Hz}.$$

2.6 Frequency Shift Keying, FSK

Frequency shift keying (FSK) is one of the simplest forms of digital modulation, with the frequency of the transmitted signal at a clock tick indicating a symbol, usually representing either one or two bits. Binary FSK (BFSK) is illustrated in Figure 2-9(d). It can be implemented by applying a discrete signal to the input of a voltage-controlled oscillator and so was ideally suited to early digital radio as simple high-performance FM modulators were available. Four-state FSK modulation is used in the GSM 2G cellular standard, a legacy standard still widely supported by modern cellular radios and sometimes the only modulation supported by the infrastructure (i.e. basestations) in some regions where it is not economically viable to retrofit old installations.

2.6.1 Essentials of FSK Modulation

The schematic of a binary FSK modulation system is shown in Figure 2-10. Here, a binary bitstream is lowpass filtered and used to drive an FSK modulator, one implementation of which shifts the frequency of an oscillator according to the voltage of the baseband signal. This function can be achieved using a VCO or a PLL circuit, and an FM demodulator can be used to receive the signal. Another characteristic feature of FSK is that the amplitude of the modulated signal is constant, so efficient saturating (and hence nonlinear) amplifiers can be used without the concern of frequency distortion. Not surprisingly, FSK was the first form of digital modulation used in mobile digital radio. A particular implementation of FSK is **Minimum Shift Keying** (**MSK**), which uses a baseband lowpass filter so that the transitions from one state to another are smooth in time and limit the bandwidth of the modulated signal.

The **constellation diagram** is often thought of as being like a phasor diagram and this analogy works most of the time but it does not work for FSK modulation. A phasor diagram describes a phasor that is fixed in frequency. If the phasor is very slowly phase and/or amplitude modulated, then this approximation is good. FSK modulation cannot be represented on a phasor diagram, as the information is in the frequency at the clock ticks and not the than the phase and/or amplitude of a phasor. The symbols of two- and four-state FSK modulation are shown in Figure 2-11 which are called constellation diagrams

As an example consider an FSK-modulated signal with a bandwidth of 200 kHz and a carrier at 1 GHz (this approximately corresponds to the 2G GSM cellular system). This is a 0.02% bandwidth, so the phasor changes very slowly. Going from one FSK state to another takes about 1230 to 3692

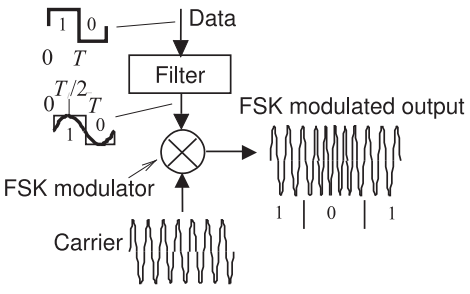
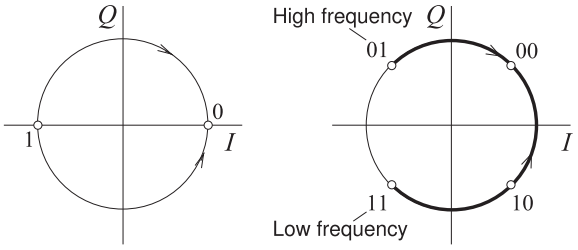


Figure 2-10: The frequency shift keying (FSK) modulation system. In the GSM four-state cellular systemadjacent constellation points differ in frequency by 33.25 kHz.



(a) Two-state FSK (b) Four-state FSK

Figure 2-11: Constellation diagrams of FSK modulation. In two-state FSK a symbol indicates whether a bit is a '0' or a '1'. In four-state FSK there are four symbols and each symbol has a different frequency and indicates the values of two bits.

RF cycles depending on the frequency difference of the transition from one symbol to the next. With a 1 GHz carrier the frequencies of the four symbols are (1 GHz − 33.25 kHz), (1 GHz − 16.62 kHz), (1 GHz + 16.62 kHz), and (1 GHz + 33.25 kHz). This may seem like a very small frequency difference but hardware in the basestation and in the handset can easily achieve a frequency resolution of a few hertz at 1 GHz. In trying to represent FSK modulation on a pseudo-phasor diagram, the frequency is approximated as being fixed and the maximum real frequency shift is arbitrarily taken as being a significant shift of the pseudo-phasor.

In FSK, the states are on a circle in the constellation diagram (see Figure 2-11), with two-state FSK shown in Figure 2-11(a) and four-state FSK shown in Figure 2-11(b). Note that the constellation diagram indicates that the amplitude of the phasor is constant, as FSK modulation is a form of FM modulation. Consider four-state FSK more closely. There are four frequency states ranging from the low-frequency state to the high-frequency state as shown in Figure 2-11(b). In four-state FSK modulation a transition from the low-frequency state to the high-frequency state takes three times longer than a transition between adjacent states. While the '01' and '11' states appear to be adjacent, in reality the frequency transition must traverse through the other frequency states. Filtering of the baseband modulating signal is required to minimize the bandwidth of the modulated four-state FSK signal. This reduces modulation efficiency to less than the theoretical maximum of 2 bits/s/Hz.

In summary, there are slight inconsistencies and arbitrariness in using a phasor diagram for FSK, but FSK does have a defined constellation diagram that is closely related, but not identical, to a phasor diagram. Another difference is that a phasor diagram depends on the amplitude of the RF signal, while a constellation diagram is continuously being re-normalized to the average RF power level to maintain a constant size. With FSK modulation almost the entire modulation and demodulation paths can be implemented using analog circuitry and so was ideally suited to early cellular radios.

2.6.2 Gaussian Minimum Shift Keying

Gaussian minimum shift keying (GMSK) is the modulation scheme used in the **GSM** cellular wireless system and is a variant of **MSK** with waveform shaping coming from a Gaussian lowpass filter. It is a particular implementation of FSK modulation.

The modulation efficiency of GMSK as implemented in the GSM system (it depends slightly on the Gaussian filter parameters) is 1.35 bits/s/Hz. Unfiltered MSK has a constant RF envelope. However filtering is required

to limit the RF bandwidth and this results in amplitude variations of about 30%. This is still very little so one of the fundamental advantages of this modulation scheme is that nonlinear, power-efficient amplification can be used. GMSK is essentially a digital implementation of FM with discrete changes in the frequency of modulation with the input bitstream filtered so that the change in frequency from one state to the next is smooth. It is only at the clock ticks that the modulated signal must have the specified discrete frequency. The phase of the modulating signal is always continuous and there is no information in the phase of the modulated signal.

The ideal transitions in FSK follow a circle from one state to another as shown in Figure 2-11 so that the PMEPR of ideal FSK is 0 dB. With GMSK the transitions do not follow a circle because of the filtering and the transitions also overshoot. As such the amplitude of a GMSK modulated signal varies and the PMEPR of GMSK is 3.01 dB. This is the PMEPR for a single modulated carrier, combining multiple modulated carriers as done in a basestation increases the PMEPR. Statistically the envelopes are less likely to all align if there are multiple carriers. For example, with multi-carrier GMSK, PMEPR = 3.01 dB, 6.02 dB, 9.01 dB, 11.40 dB, 14.26 dB, and 17.39 dB for 1, 2, 4, 8, 16, and 32 carriers respectively. (These values were calculated numerically by simulating a multi-carrier system.)

GMSK and other FSK methods have the advantage that implementation of the baseband and RF hardware is relatively simple. A GMSK transmitter can use conventional frequency modulation. On receive, an FM discriminator, i.e. an FM receiver with sampling, can be used avoiding more complex I and Q demodulation.

2.6.3 Doppler Effect

Frequency is a physical parameter that can be established and measured with great accuracy, down to a few hertz at 1 GHz in a mobile handset for example. Thus if a receiver is stationary the frequency states at the clock ticks of an FSK modulated carrier can be measured with great accuracy. When a receiver and transmitter are moving relative to each other there will be a Doppler shift of the carrier frequency. If the relative velocity of the receiver and transmitter is v_s the Doppler shift will be

$$\Delta f = f v_s / \mathrm{c} \tag{2.46}$$

where f is the frequency of the radio transmission and c is the speed of light. For a receiver moving at 100 km/hr receiving a 1 GHz signal from a fixed transmitter, the Doppler frequency shift is $\Delta f = 92.6$ Hz which is much less than the 33 kHz frequency spacing of adjacent states in the FSK example above. Thus the Doppler shift is not of concern. This effective fixing of the constellation points is one of the advantages of GSM.

2.6.4 Summary

GSM was not the only 2G system. The 2G NADC (for North American Digital Cellular) system modulated the phase of a carrier using phase shift keying. The NADC cellular system had higher modulation efficiency than GSM yet MSK became the dominant 2G system and is still supported as a legacy modulation system in modern cellular radio. The main reason for

this is that GSM was more closely aligned with the business interests of the telephone operators of the day.

2.7 Carrier Recovery

Carrier recovery refers to establishing a local carrier reference signal which accurately reproduces the frequency and, with some modulation methods, the phase of the carrier of the modulated signal. All digital modulation methods require carrier recovery to establish a reference to determine the state of the carrier at the clock ticks. In addition digital modulation methods require that the timing of the clock ticks be established. Since radios using digital modulation all send packets of data, i.e. sequences of symbols, having a known sequence at the beginning of packet transmission enables the timing to be determined.

With FSK modulation the frequency at the clock ticks must be determined. This is relatively simple because the frequency at the clock ticks can be accurately measured as a local clock can be established within a few hertz because of the availability of accurate crystal references. The frequency of the received signal can still be shifted by the Doppler effect of the transmitter or receiver is moving but this is quite small compared to the frequency differences between the received states. With FSK it is not necessary to determine the phase of the carrier.

All digital modulation other than FSK modulates a carrier by shifting the carrier's phase and/or amplitude to a number of discrete states. Recovering the state of this modulated carrier requires that the phase of the carrier be recovered from the receive signal and to do this there must be a constant phase local version of the carrier. The circuits that implement the local version of the carrier are called carrier recovery circuits. These circuits modify a very stable internal oscillator in the receiver that after an initial setting of an approximate frequency, has a frequency and phase that can only change slowly. However, there must be a received signal at all times, because if the received signal falls below the noise level the carrier recovery circuit will try to track the noise. This requirement has led to a number of different modulation schemes that avoid the amplitude of the modulated signal from ever being small during a transition. This is important in 2G and 3G cellular radio but 4G and 5G cellular systems use pilot tones to achieve carrier recovery.

In early digital radios carrier recovery was implemented in analog circuitry and more modern radios implement carrier recovery by splitting the function between an analog oscillator signal that can be assigned to a large number of discrete states (providing course carrier recovery) and DSP of the (coarsely recovered) baseband signal to precisely recover the carrier signal. Thus in modern digital radios the carrier recovery circuit is implemented partially as an analog circuit and partially as a digital circuit.

2.8 Phase Shift Keying Modulation

There are many variations on phase shift keying (**PSK**) modulation with the methods differing by their spectral efficiencies, PMEPR, and suitability for carrier recovery. Compared to FSK more sophisticated digital signal processing is required to demodulate a PSK-modulated signal.

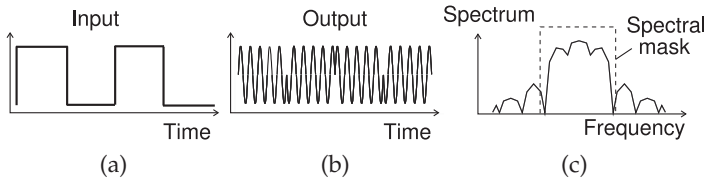


Figure 2-12: Binary PSK modulation: (a) modulating bitstream; (b) the modulated waveform; and (c) its spectrum after smoothing the transitions from one phase state to another.

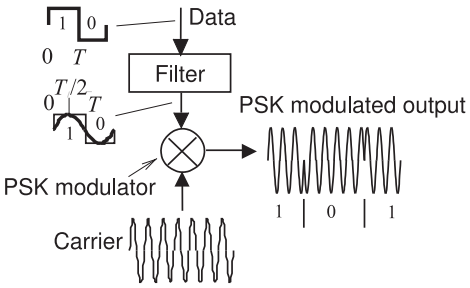


Figure 2-13: A binary phase shift keying (PSK) modulation system.

2.8.1 Essentials of PSK

PSK is an efficient digital modulation scheme and can be simply implemented and demodulated using a phase-locked loop. The simplest scheme is binary PSK (BPSK) with two phase states. The waveform and spectrum of BPSK are shown in Figure 2-12. The incoming baseband bitstream shown in Figure 2-12(a) modulates the phase of a **carrier** producing the modulated signal shown in Figure 2-12(b). The spectrum of the modulated signal is shown in Figure 2-12(c). What is very interesting about this spectrum is that it approximately fills a square. So PSK modulation results in an efficient use of the spectrum. This can be contrasted with the spectrum of an FM signal shown in Figure 2-7(c), which does not fill the channel uniformly. A binary PSK modulation system is shown in Figure 2-13 where the binary input data causes $180°$ phase changes of the carrier. The abrupt changes in phase shown in the output modulated waveform result in more bandwidth than is necessary. However a practical PSK modulator first lowpass filters the binary data before the carrier is modulated. This filtering eliminates the abrupt changes in the phase of the modulated signal and so reduces the required bandwidth. It is the spectrum of this signal that is shown in Figure 2-12(c).

There are many variants of increasing complexity, called orders, of PSK, with the fundamental characteristics being the number of phase states (e.g. with 2^n phase states, n bits of information can be transmitted) and how the phasor of the RF signal transitions from one phase state to another. PSK schemes are designed to shape the spectrum of the modulated signal to fit as much energy as possible within a spectral mask. This results in a modulated carrier whose amplitude varies (and thus has a time-varying envelope). Such schemes require highly linear amplifiers to preserve the amplitude variations of the modulated RF signal.

There are PSK methods that manage the phase transitions to achieve a constant envelope modulated RF signal but these have lower spectral efficiency. Military radios sometimes use this type of modulation scheme as it is much harder to detect and intercept communications if the amplitude of the modulated carrier is constant.

The communication limit of one symbol per hertz of bandwidth,

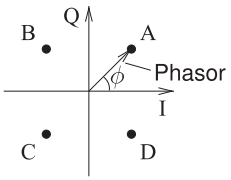


Figure 2-14: Phasor diagram of QPSK modulation. Here there are four discrete phase states of the phasor indictaed by the points A, B, C, and D. The PSK modulator moves the phasor from one phase state to another. The task at the receiver is determining the phase of the phasor.

the **symbol rate**, comes from the **Nyquist signaling theorem**.[1] Nyquist determined that the number of independent pulses that could be put through a telegraph channel per unit of time is limited to twice the bandwidth of the channel. With a modulated RF carrier, this translates to the modulated carrier moving from one state to another in a unit of time equal to one over the bandwidth. The transition identifies a symbol, and hence one symbol can be sent per hertz of bandwidth. More accurately it could be said that the transition is a symbol rather than the end of the transition being a symbol. In PSK modulation the states are the phases of a phasor since the amplitude of the modulated signal is (ideally) constant.

The phase-shifted (i.e. phase-modulated) carrier of a PSK signal can be represented on a phasor diagram. Figure 2-14 is a phasor diagram with four phase states—A, B, C, D—and the phasor moves from one state to another under the control of the modulation circuit. What is shown here is 4-state PSK or quadra-phase shift keying (QPSK) and very often but less accurately called **quadrature phase shift keying**. The states, or symbols, are identified by their angle or equivalently by their rectangular coordinates, called I, for in-phase, and Q, for quadrature phase.

PSK Modulation

In PSK modulation the phase of a carrier signal is set to one of a number of discrete values at the clock ticks. For example, in **QPSK** there are four discrete settings of the phase of the carrier, e.g. $45°$, $135°$, $-135°$, and $-45°$. Converting this to radians the discrete baseband signal is $\phi(t) = \pi/4$, $3\pi/4$, $5\pi/4$, and $7\pi/4$, at the clock ticks. Thus if the bandwidth of the baseband signal is 1 MHz what is shown as $\phi(t)$ are the intended phases of the carrier every microsecond. Wave-shaping or filtering is used to provide a smooth variation of $\phi(t)$ between the clock ticks and so the bandwidth of the modulated signal is constrained. High-order PSK modulation has more discrete states, e.g. 8-PSK has eight discrete phase states.

There are several ways to implement PSK modulation and one uses the quadrature modulator shown in Figure 2-15. The discrete baseband signal $\phi(t)$ could be internal to a DSP which is then interpolated in time and output by the DSP's DAC as two smooth signals $i(t) = \cos(\phi(t))$ and $q(t) = \sin(\phi(t))$. On a phasor diagram $i(t)$ and $q(t)$ at the clock ticks addresses one of QPSK's four states of the carrier's phasor, see Figure 2-14.

For PSK modulation the constellation diagram is very similar to a phasor diagram that is continuously being re-normalized to the average power of

[1] This theorem was discovered independently by several people and is also known as the Nyquist-Shannon sampling theorem, the Nyquist-Shannon-Kotelnikov, the Whittaker-Shannon-Kotelnikov, the Whittaker-Nyquist-Kotelnikov-Shannon (WKS), as well as the cardinal theorem of interpolation theory. The theorem states [13]: "If a function $x(t)$ contains no frequencies higher than B hertz, it is completely determined by giving its ordinates at a series of points spaced $1/(2B)$ seconds apart."

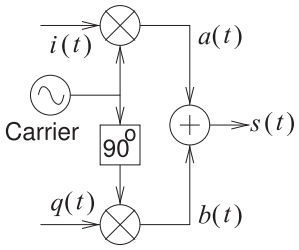


Figure 2-15: Quadrature modulator block diagram. In PSK modulation $i(t)$ and $q(t)$ have the same amplitude and indicate a phase ϕ of the modulated carrier so that $i(t) = \cos[\phi(t)]$ and $q(t) = \sin[\phi(t)]$. The particular example shows two possible values of I_k and Q_k and this indicates QPSK modulation.

the modulated signal. This a subtle but important distinction, for example, a PSK baseband signal has a constellation diagram even though the baseband signal does not have a phasor representation. PSK modulation using the block diagram shown in Figure 2-15 has a carrier that is directly input to the top multiplier and a $90°$ phase-shifted version input to the bottom multiplier. Let the carrier be $\cos(\omega_c t)$ and so the version of the carrier input to the bottom multiplier is $\cos(\omega_c t - \pi/2) = -\sin(\omega_c t)$. So, with $q(t)$ being a $90°$ phase-shifted version of $i(t)$, (using the identities in Section 1.A.2 of [4])

$$a(t) = \cos(\phi(t))\cos(\omega_c t) = \tfrac{1}{2}\left[\cos(\omega_c t - \phi(t)) + \cos(\omega_c t + \phi(t))\right] \quad (2.47)$$

$$b(t) = \sin(\phi(t))[-\sin(\omega_c t)] = -\tfrac{1}{2}\left[\cos(\omega_c t - \phi(t)) - \cos(\omega_c t + \phi(t))\right] \quad (2.48)$$

$$s(t) = a(t) + b(t) = \cos(\omega_c t + \phi(t)). \quad (2.49)$$

Thus $s(t)$ is the single-sideband modulated carrier carrying information in the phase of the modulated carrier. The modulating signal $\phi(t)$ is driven by a digital code that is designed so that $\phi(t)$ changes at a minimum rate (it never has the same value for more than a few clock ticks). Thus there are no low frequency components of $\phi(t)$ and thus there is no modulated signal at or very close to the carrier. Thus the carrier is suppressed but there is a sideband above and below the carrier frequency. This is SSB-SC modulation.

2.8.2 Binary Phase Shift Keying

PSK uses prescribed phase shifts to define symbols, each of which can represent one, two, or more bits. **Binary Phase Shift Keying** (BPSK), illustrated in Figures 2-12 and 2-13, has two phase states and conveys one bit per symbol and is a relatively spectrally inefficient scheme, with a maximum (i.e. ideal) modulation efficiency of 1 bits/s/Hz. The reason why the practical modulation efficiency is less than this number is because the transition from one phase state to the other must be constrained to avoid the modulated signal becoming very small, and also because there are no ideal lowpass filters to filter the input binary data stream. Although it has low modulation efficiency, it is ideally suited to low-power applications. BPSK is commonly used in **Bluetooth**.

The operation of BPSK modulation can be described using the constellation diagram shown in Figure 2-16(a). The BPSK constellation diagram indicates that there are two states. These states can be interpreted as the rms values of $i(t)$ and $q(t)$ at the sampling times corresponding to the bit rate. The distance of a constellation point from the origin corresponds to (normalized) rms power of the pseudo-sinusoid of the modulated carrier at the sampling instant. (Normalization is with respect to the average power.) The curves in Figure 2-16(b) indicate three transitions. The states are at the ends of the transitions. If a 1, in Figure 2-16(b), is assigned to the positive I value and 0

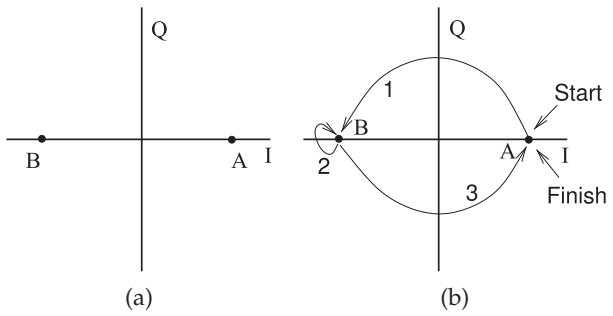


Figure 2-16: BPSK modulation with constellation points A and B: (a) constellation diagram; and (b) constellation diagram with possible transitions from one phase state to the other, or possibly no change in the phase state. In practical systems the transition should not go through the origin, as then the RF signal would drop below the noise level.

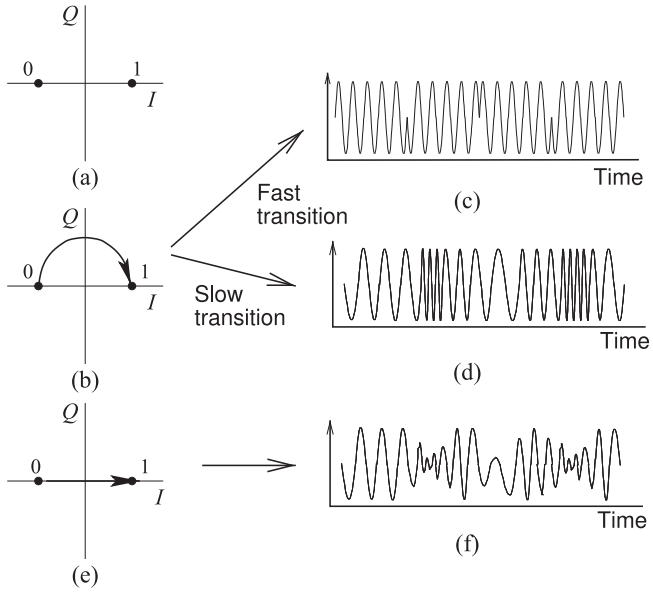
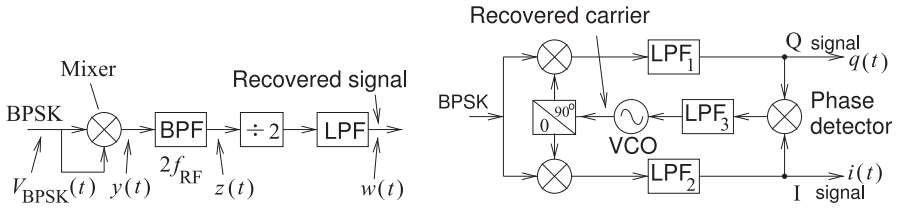


Figure 2-17: BPSK modulation: (a) constellation diagram; (b) constellation diagram with a constant amplitude transition; (c) time-domain waveform if the transition is fast; (d) time-domain waveform if the transition is slow; (e) constellation diagram with transition through the origin; and (f) time-domain waveform if the transition goes through the origin and is slow.

to a negative I value, then the bit sequence represented in Figure 2-16(b) is "1001."

Figure 2-17(a) is the constellation diagram of BPSK, with two symbols denoted as 0 and 1, and the trajectory of the transition from one constellation point to the other depending on the hardware used to implement the BPSK modulator. Figure 2-17(b) shows the transition from the '0' state to the '1' state (and back) while maintaining a constant amplitude. If this transition is very fast, then the waveform produced is as shown in Figure 2-17(c), where there are abrupt phase transitions and these have high spectral content. It is better to slow down the transitions, as then the waveform (shown in Figure 2-17(d)), has smooth transitions and the bandwidth of the modulated carrier is minimal. The preferred smooth transition is obtained by lowpass filtering the baseband signal. That is, the abrupt transitions in the modulated RF signal result in the modulated signal having a broad bandwidth. The graceful transition of BPSK modulation limits the bandwidth of the modulated carrier.

A simple implementation of BPSK modulation would result in direct transition from one state to the others causing the phasor to traverse the origin and the amplitude of the RF signal to become very small and less than the noise level (see Figure 2-17(e)). The resulting modulated RF waveform is



(a) Basic BPSK carrier recovery circuit (b) BPSK Costas loop

Figure 2-18: Block diagram of carrier recovery circuits for BPSK signals.

shown in Figure 2-17(f). This is a problem because the receiver would not be able to track the RF signal.

Carrier Recovery

In a PSK demodulator, a local copy of the carrier must be produced to act as a reference in determining the phase of the modulated signal. The technique that produces the local copy of the unmodulated carrier is called carrier recovery. The circuit that directly implements carrier recovery of a BPSK signal is shown in Figure 2-18(a). At the clock ticks the waveform of a BPSK modulated signal is

$$v_{\text{BPSK}}(t) = A(t)\cos\left(\omega_{\text{RF}}t + n\pi\right), \tag{2.50}$$

where the carrier frequency $f_{\text{RF}} = \omega_{\text{RF}}/(2\pi)$ and n can have a value of 0 or 1. Squaring this produces a signal

$$\begin{aligned} y(t) = v_{\text{BPSK}}^2(t) &= A^2(t)\cos^2\left(\omega_{\text{RF}}t + n\pi\right) = \tfrac{1}{2}A^2(t)\left[1 + \cos\left(2\omega_{\text{RF}}t + n2\pi\right)\right] \\ &= \tfrac{1}{2}A^2(t)\left[1 + \cos\left(2\omega_{\text{RF}}t\right)\right]. \end{aligned} \tag{2.51}$$

This is a signal at twice the carrier frequency with no carrier modulation since $n2\pi$ and 0 radians are indistinguishable. The squaring operation is performed by mixing $v_{\text{BPSK}}(t)$ with itself. Bandpass filtering $y(t)$ produces a signal $z(t)$ at the second harmonic of the carrier. The divide-by-2 block is implemented using a phase-locked loop (PLL). The result is the recovered carrier, $w(t)$, that is used as the timing reference for sampling the demodulated I and Q components at precise times.

A better carrier recovery circuit than that in Figure 2-18(a) and described above is the Costas loop [14] shown in Figure 2-18(b). The BPSK Costas loop implements carrier recovery and I/Q demodulation simultaneously. In Figure 2-18(b) $i(t)$ and $q(t)$ are mixed to produce a signal applied at the input of the third lowpass filter, LPF_3. The main function of this filter is to remove noise and to average the signal coming out of the phase detector. The output of LPF_3 drives a VCO in which the oscillation frequency is controlled by the applied voltage. The quadrature phase shifter then mixes the recovered carrier and a $90°$ shifted version with the BPSK signal.

It is critical that the signal-to-noise ratio (SNR), the ratio of the signal power to the noise power, of the BPSK signal be sufficiently large at all times or else the Costas loop will produce a noisy recovered carrier signal. If the modulated carrier becomes very small, for example when the trajectory on the constellation diagram goes through the origin (where the level of the carrier carrier falls below the noise level), the carrier will not be accurately recovered.

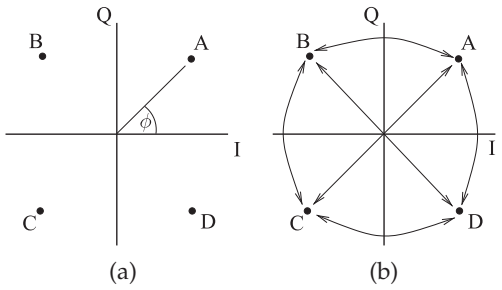


Figure 2-19: QPSK modulation: (a) constellation diagram; and (b) constellation diagram with possible transitions. Each constellation point indicates the phase, ϕ, of the modulated carrier, i.e. $\cos(\omega_c t + \phi)$ where ω_c is the radian frequency of the carrier.

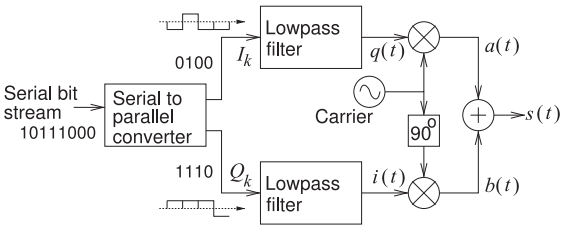


Figure 2-20: QPSK modulator block diagram. I_k and Q_k are similar to a stream of one-bit binary signals but are analog with a either a postitive value or a negative voltage so that after lowpass filtering $i(t)$ and $q(t)$ each have either a positive or a negative voltage at each clock tick.

2.8.3 Quadra-Phase Shift Keying, QPSK

In QPSK wireless systems, modulation efficiency is obtained by sending more than one bit of information per hertz of bandwidth (i.e., more than one bit per symbol). In QPSK information is encoded in four phase states and two bits are required to identify a symbol (i.e., to identify a phase state). The constellation diagram of QPSK is shown in Figure 2-19(a) where the modulated RF carrier has four phase states identified as A, B, C, and D. So a QPSK modulator shifts the phase of the carrier to one of these phase states, and a QPSK demodulator must determine the phase of the received RF signal. The received RF signal is sampled with precise timing as determined by the recovered carrier signal. Thus two bits of information are transmitted per change of phase states. Each change of phase state requires at least 1 Hz of bandwidth with the minimum bandwidth obtained when the transition from one state to another is no faster than that required to reach the new phase state before the sampling instant. QPSK modulation is also referred to as quarternary PSK.

QPSK can be implemented using the modulator shown in Figure 2-20. In Figure 2-20, the input bitstream is first converted into two parallel bitstreams each containing half the number of bits of the original bit stream. Thus a two-bit sequence in the serial bitstream becomes one I_K bit and one Q_K bit. The (I_K, Q_K) pair constitutes the Kth symbol. The bitstreams are converted into waveforms $i(t)$ and $q(t)$ by the wave-shaping circuit.

The constellation diagram of QPSK is the result of plotting I and Q on a rectangular graph as shown in Figure 2-19(a). All possible phase transitions are shown in Figure 2-19(b). In the absence of wave-shaping circuits, $i(t)$ and $q(t)$ have very sharp transitions, and the paths shown in Figure 2-16(b) occur almost instantaneously. This leads to large spectral spreads in the modulated waveform, $s(t)$. So to limit the spectrum of the RF signal $s(t)$, the shape of $i(t)$ and $q(t)$ is controlled; the waveform is shaped, usually by lowpass filtering. So a pulse-shaping circuit changes baseband binary information into a more smoothly varying signal. Each transition or path in Figure 2-16 represents the transfer of a symbol, with the best efficiency that can be obtained in wireless communication being one symbol per hertz

of bandwidth. Each symbol contains two bits so the maximum modulation efficiency of QPSK modulation is 2 bits/s/Hz of bandwidth. What is actually achieved depends on the wave-shaping circuits and on the criteria used to establish the bandwidth of $s(t)$.

Carrier Recovery

Carrier recovery of a QPSK signal is similar to that for a BPSK signal. At the clock ticks an RF QPSK modulated signal

$$v_{\mathrm{QPSK}}(t) = A(t)\cos\left(\omega_{\mathrm{RF}}t + n\pi/2\right); \quad n = 0, 1, 2, 3, \tag{2.52}$$

where the carrier frequency $f_{\mathrm{RF}} = \omega_{\mathrm{RF}}/(2\pi)$. The fourth power of this produces

$$\begin{aligned} v^4_{\mathrm{QPSK}}(t) &= A^4(t)\cos^4\left(\omega_{\mathrm{RF}}t + n\pi/2\right) \\ &= \tfrac{1}{8}A^4(t)\left[3 + 4\cos\left(2\omega_{\mathrm{RF}}t + n\pi\right) + \cos\left(4\omega_{\mathrm{RF}}t + n2\pi\right)\right]. \end{aligned} \tag{2.53}$$

Following bandpass filtering at $4f_{\mathrm{RF}}$ and then dividing the frequency by 4, the carrier is recovered. Circuits implementing this are similar to those for recovering the carrier of BPSK signals. This concept can be extended to carrier recovery for any M-PSK-modulated signal.

EXAMPLE 2.7 QPSK Modulation and Constellation

The bit sequence 110101001100 is to be transmitted using QPSK modulation. Show the transitions on a constellation diagram.

Solution:

The bit sequence 110101001100 must be converted to a two-bit-wide parallel stream of symbols resulting in the sequence of symbols 11 01 01 00 11 00. The symbol 11 transitions to the symbol 01 and then to the symbol 01 and so on. The states (or symbols) and the transitions from one symbol to the next required to send the bitstream 110101001100 are shown in Figure 2-21. QPSK modulation results in the phasor of the carrier transitioning through the origin so that the average power is lower and the PMEPR is high. A more significant problem is that the phasor will fall below the noise floor, making carrier recovery almost impossible.

2.8.4 $\pi/4$ Quadrature Phase Shift Keying

A major objective in digital modulation is to ensure that the RF trajectory from one phase state to another does not go through the origin. The transition is slow, so that if the trajectory goes through the origin, the amplitude of the carrier will be below the noise floor for a considerable time and it will not be possible to recover the carrier reference. This is why the QPSK scheme is not used directly in 2G and 3G cellular radio. (The 4G and 5G cellular radio systems do use QPSK among other modulation schemes and use pilot tones to recover the carrier.) One of the solutions developed to address this problem is the $\pi/4$ quadrature phase shift keying (**π/4-QPSK**) modulation scheme. In this scheme the constellation at each symbol is rotated $\pi/4$ radians from the previous symbol, as shown in Figure 2-22. (In an alternative implementation of $\pi/4$-QPSK modulation the constellation diagram could

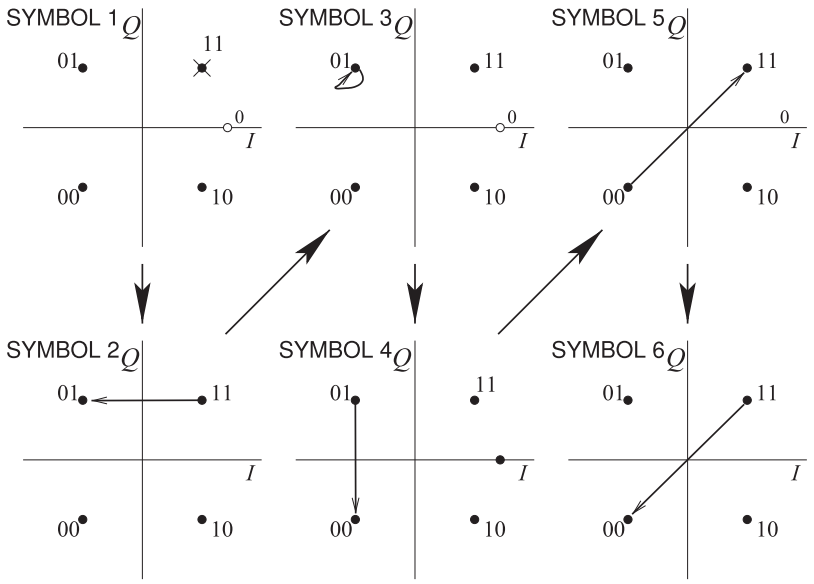


Figure 2-21: Constellation diagram and transitions for the bit sequence 110101001100 sent as the set of symbols 11 01 01 00 11 00 using QPSK. Note that symbols 2 and 3 are identical, so there is no transition. The SYMBOL numbers indicated reference the symbol at the end of the transition (end of the arrow). The assignment of bits to symbols (e.g., assigning the bits '11' to the symbol in the first quadrant) is arbitrary in general but the assignment of symbols is defined in a particular standard.

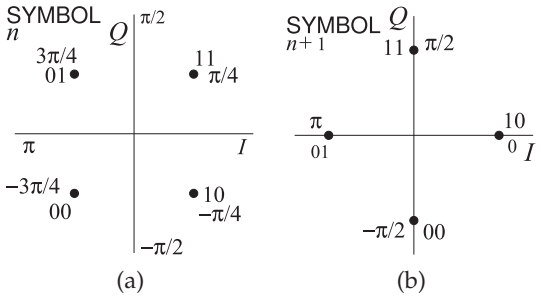


Figure 2-22: Constellation diagram of $\pi/4$-QPSK modulation: (a) initial constellation diagram at one symbol; and (b) the constellation diagram at the time of the next symbol.

rotate by $\pi/4$ continuously rather than switching between conditions as described here.)

One of the unique characteristics of $\pi/4$-QPSK modulation is that there is always a change, even if a symbol is repeated. This helps with recovering the carrier frequency. If the binary bitstream itself (with sharp transitions in time) is the modulation signal, then the transition from one symbol to the next occurs instantaneously and hence the modulated signal has a broad spectrum around the carrier frequency. The transition, however, is slower if the bitstream is filtered, and so the bandwidth of the modulated signal will be less. Ideally the transmission of one symbol per hertz would be obtained. However, in $\pi/4$-QPSK modulation the change from one symbol to the next has a variable distance (and so a transition takes different times) so that the ideal modulation efficiency of one symbol per second per hertz (or 2

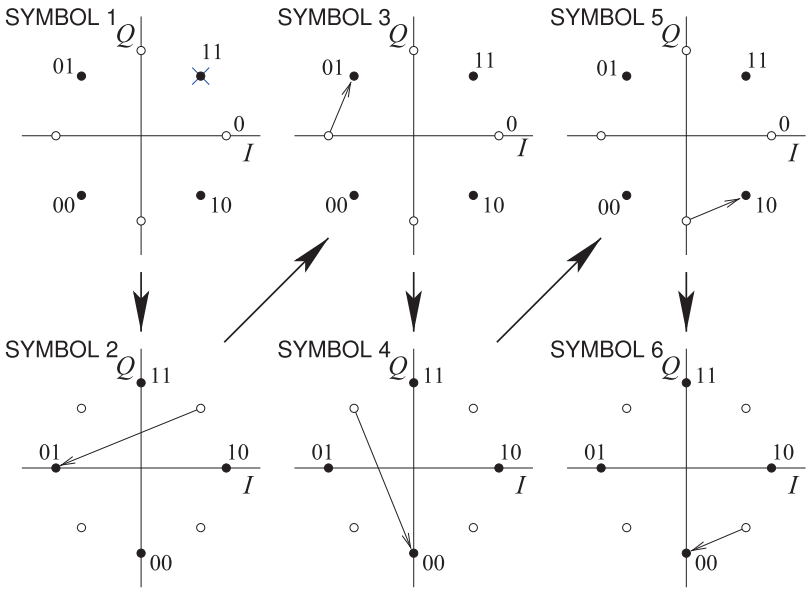


Figure 2-23: Constellation diagram states and transitions for the bit sequence 110101001000 sent as the set of symbols 11 01 01 00 10 00 using $\pi/4$QPSK modulation.

bits/s/Hz) is not obtained. In practice, with realistic filters and allowing for the longer transitions, $\pi/4$-QPSK modulation achieves 1.62 bits/s/Hz.

EXAMPLE 2.8 $\pi/4$-QPSK Modulation and Constellation

The bit sequence 110101001000 is transmitted using $\pi/4$-QPSK modulation. Show the transitions on a constellation diagram.

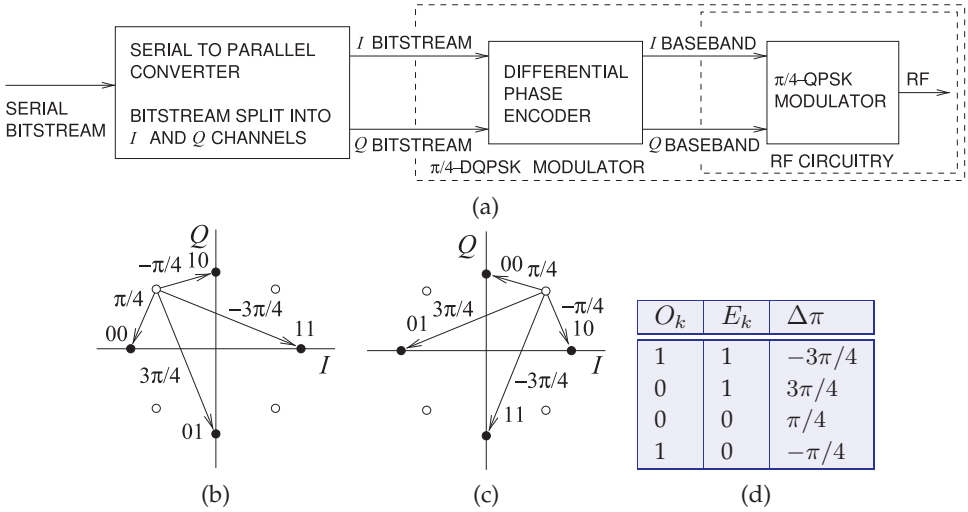
Solution:

The bit sequence 110101001000 must be converted to a two-bit-wide parallel stream of symbols, resulting in the sequence of symbols 11 01 01 00 10 00. The symbol 11 transitions to the symbol 01 and then to the symbol 01 and so on. The constellation diagram of $\pi/4$-QPSK modulation really consists of two QPSK constellation diagrams that are shifted by $\pi/4$ radians, as shown in Figure 2-22. At one symbol (or time) the constellation diagram is that shown in Figure 2-22(a) and at the next symbol it is that shown in Figure 2-22(b). The next symbol uses the constellation diagram of Figure 2-22(a) and the process repeats. The states (or symbols) and the transitions from one symbol to the next that are required to send the bitstream 110101001000 are shown in Figure 2-23.

2.8.5 Differential Quadra Phase Shift Keying, DQPSK

Multiple transmission paths, or **multipaths**, due to reflections result in constructive and destructive interference and can result in rapid additional phase rotations. Thus relying on the phase of a phasor at the symbol sample time, at the clock ticks, to determine the symbol transmitted is prone to error. When an error results at one symbol, this error accumulates when subsequent symbols are extracted. The solution is to use encoding, and one of the simplest encoding schemes is differential phase encoding. In this scheme the information of the modulated signal is contained in changes in phase rather than in the absolute phase. That is, the transition defines the symbol rather than the end point of the transition.

The $\pi/4$-DQPSK modulation scheme is a differentially encoded form of



O_k	E_k	$\Delta\pi$
1	1	$-3\pi/4$
0	1	$3\pi/4$
0	0	$\pi/4$
1	0	$-\pi/4$

(d)

Figure 2-24: A $\pi/4$-DQPSK modulator: (a) differential phase encoder with a $\pi/4$-QPSK modulator; (b) constellation diagram of $\pi/4$-DQPSK; (c) a second constellation diagram; and (d) phase changes in a $\pi/4$-DQPSK modulation scheme. Note that the information is in the phase change rather than the phase state.

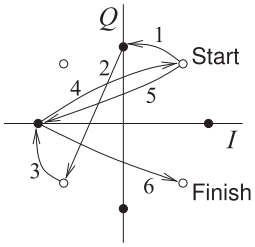


Figure 2-25: Constellation diagram of $\pi/4$-DQPSK modulation showing six symbol intervals coding the bit sequence 000110110101.

$\pi/4$-QPSK. The $\pi/4$-DQPSK scheme incorporates the $\pi/4$-QPSK modulator and an encoding scheme, as shown in Figure 2-24(a). The scheme is defined with respect to its constellation diagram, shown in Figure 2-24(b) and repeated in Figure 2-24(c) for clarity. The D indicates **differential coding**, while the $\pi/4$ denotes the rotation of the constellation by $\pi/4$ radians from one interval to the next. This can be explained by considering Figure 2-24(a). A four-bit stream is divided into two quadrature **nibbles** of two bits each. These nibbles independently control the I and Q encoding, respectively, so that the allowable transitions rotate according to the last transition. The information or data is in the phase transitions rather than the constellation points themselves. The relationship between the symbol value and the transition is given in Figure 2-24(d). For example, the transitions shown in Figure 2-25 for six successive time intervals describes the input bit sequence 000110110101. Its waveform and spectrum are shown in Figure 2-26. More detail of the spectrum is shown in Figure 2-27. In practice with realistic filters and allowing for the longer transitions, $\pi/4$-DQPSK modulation achieves a modulation efficiency of 1.62 bits/s/Hz, the same as $\pi/4$-QPSK, but of course with greater resilience to changes in the transmission path.

In a differential scheme, the data transmitted are determined by

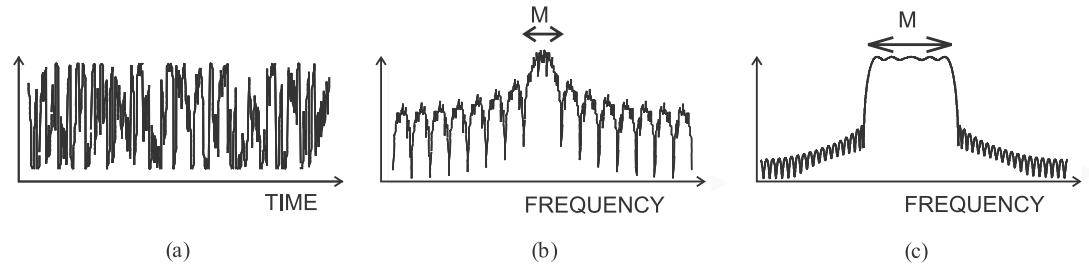


Figure 2-26: Details of digital modulation obtained using differential phase shift keying ($\pi/4$-DQPSK): (a) modulating waveform; (b) spectrum of the modulated carrier, with M denoting the main channel; and (c) details of the spectrum of the modulated carrier focusing on the main channel.

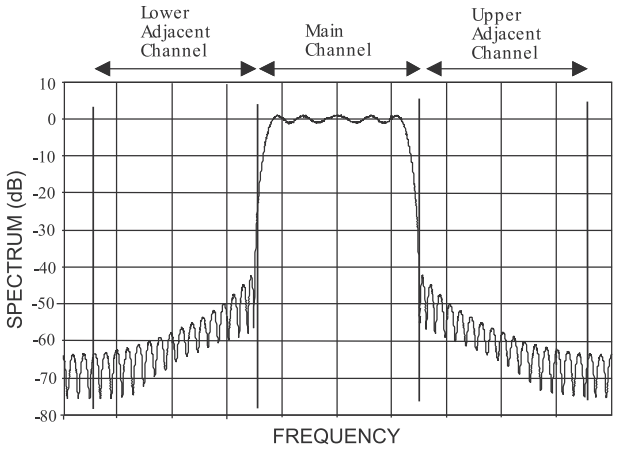


Figure 2-27: Detailed spectrum of a $\pi/4$-DQPSK signal showing the main channel and lower and upper adjacent channels.

comparing a symbol with the previously received symbol, so the data are determined from the change in phase of the carrier rather than the actual phase of the carrier. This process of inferring the data actually sent from the received symbols is called decoding. When $\pi/4$-DQPSK encoding was introduced in the early 1990s the DSP available for a mobile handset had only just reached sufficient complexity. Today, encoding is used with all digital radio systems and is more sophisticated than just the differential scheme of DQPSK. There are new ways to handle carrier phase ambiguity. The sophistication of modern coding schemes is beyond the scope of the hardware-centric theme of this book.

2.8.6 Offset Quadra Phase Shift Keying, OQPSK

The **offset quadra phase shift keying** (**OQPSK**) modulation scheme avoids transitions passing through the origin on the constellation diagram (see Figure 2-29(a)). As in all QPSK schemes, there are two bits per symbol, but now one bit is used to immediately modulate the RF signal, whereas the other bit is delayed by half a symbol period, as shown in Figure 2-28. The maximum phase change for a bit transition is $90°$, and as Q_K is delayed, a total phase change of approximately $180°$ is possible during one symbol. The

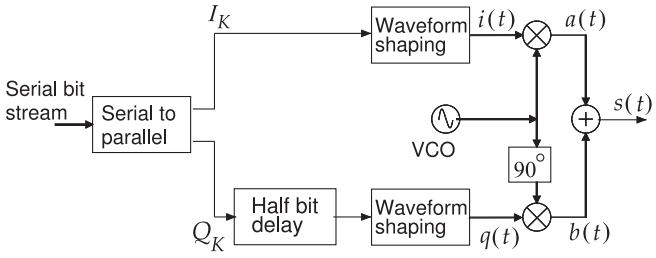


Figure 2-28: Block diagram of an OQPSK modulator.

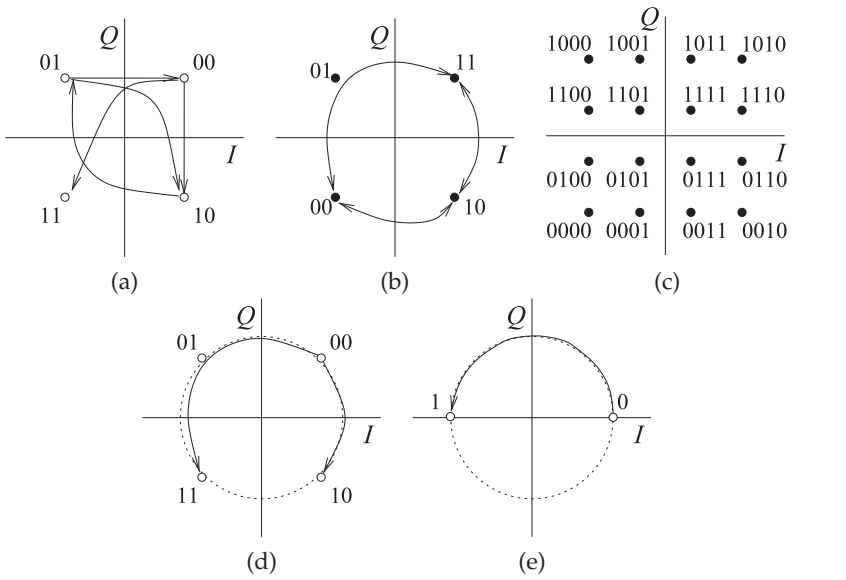


Figure 2-29: Constellation diagrams for various modulation formats: (a) OQPSK; (b) GMSK; (c) 16-QAM; (d) SOQPSK (also FOQPSK); and (e) SBPSK.

constellation diagram is shown in Figure 2-29(a).

The OQPSK modulator can be implemented using relatively simple electronics with a digital delay circuit delaying the Q bit by half a symbol period and lowpass filters shaping the I and Q bits. The OQPSK scheme is also called **staggered quadrature phase shift Keying** (**SQPSK**). Better performance can be obtained by using DSP to shape the I and Q transitions so that they change smoothly and the phasor trajectory nearly follows a circle. Consequently I and Q change together, but in such a manner that the PMEPR is maintained close to 0 dB. Two modulation techniques that implement this are the **shaped offset QPSK** (**SOQPSK**) and the **Feher QPSK** (**FQPSK**) schemes. The constellation diagrams for SOQPSK and FQPSK are shown in Figure 2-29(d). These are constant envelope digital modulation schemes. As with OQPSK, the Q bit is delayed by one-half of a symbol period and the I and Q baseband signals are shaped by a half-sine filter. The advantage is that high-efficiency saturating amplifier designs can be used and battery life extended. There is a similar modulation format called **shaped binary phase shift keying** (**SBPSK**) which, as expected, has two constellation points as shown in Figure 2-29(e). SOQPSK, FQPSK, and SBPSK are **continuous phase modulation** (CPM) schemes, as the phase

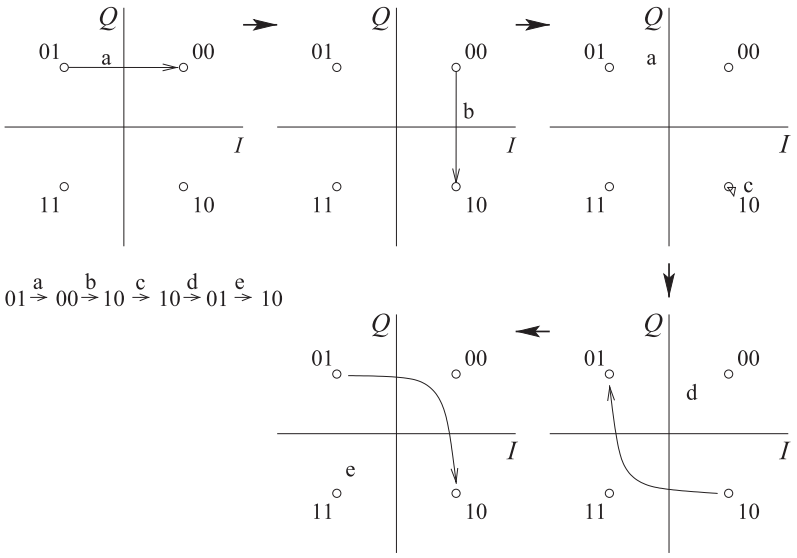


Figure 2-30: Constellation diagram of OQPSK modulation for the bit sequence 010010100110.

never changes abruptly. Instead, the phase changes smoothly, achieving high modulation efficiency and maintaining a constant envelope. Implementation of the receiver, however, is complex. CPM schemes have good immunity to interference and are commonly used in military systems.

EXAMPLE 2.9 OQPSK Modulation

Draw the constellation diagrams for the bit sequence 010010100110 using OQPSK modulation.

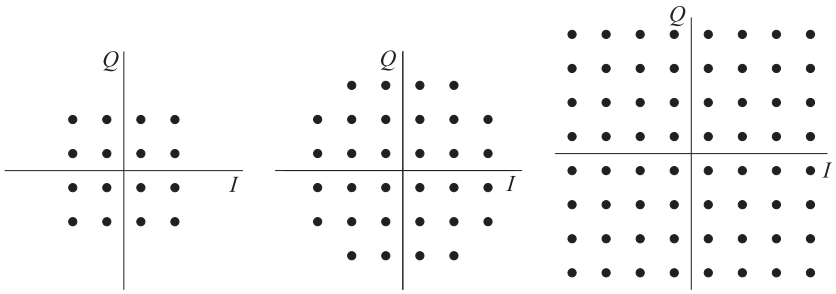
Solution:

The bit sequence is first separated into the parallel stream 01-00-10-10-01-10. The I bit changes first, followed by the Q bit delayed by half of the time of a bit. Five constellation diagrams are shown in Figure 2-30 with the transitions sending the bit sequence.

2.8.7 $3\pi/8$-8PSK, Rotating Eight-State Phase Shift Keying

The $3\pi/8$-8PSK modulation scheme is similar to $\pi/4$-DQPSK in the sense that rotation of the constellation occurs from one time interval to the next. This time, however, the rotation of the constellation from one symbol to the next is $3\pi/8$. This modulation scheme is used in the **enhanced data rates for GSM evolution** (**EDGE**) system, and provides three bits per symbol (ideally) compared to GMSK used in GSM which has two bits per symbol (ideally). With some other changes, GSM/EDGE provides data transmission of up to 128 kbits/s, faster than the 48 kbits/s possible with GSM.

Quadrature modulation schemes with four states, such as QPSK, have two I states and two Q states that can be established by lowpass filtering the I and Q bitstreams. For higher-order modulation schemes such as 8-PSK, this approach will not work. Instead, $i(t)$ and $q(t)$ are established in the DSP unit and then converted using a DAC to generate the analog signals applied



(a) 16 QAM (b) 32 QAM (c) 64 QAM

Figure 2-31: QAM constellation diagrams.

to the hardware modulator. Alternatively the modulated signal is created directly in the DSP and a DAC converts this to an IF and a hardware mixer up-converts this to RF. QAM

2.8.8 Summary

PSK modulation is implemented in many radio standards including all cellular standards after 2G. There was a 2G system that used $\pi/4$DQPSK but that is no longer supported. The modern radio standards support many modulation formats but in high interference situations BPSK, QPSK and 8-PSK have the best performance. While QPSK was dismissed in 2G and 3G because of difficulties with carrier recovery, 4G and 5G have another method for implementing carrier recovery which allows QPSK on its own to be used. GMSK is still supported by modern cellular phones but the infrastructure, i.e. basestations, are starting to be retired.

Most of the modulation schemes described in this section were introduced as optimum trade-offs of modulation efficiency, resistance to interference, and hardware complexity. Some, such as BPSK, draw very little power and are suited to the internet-of-things (IoT) applications which must have a battery lifetime of ten years.

2.9 Quadrature Amplitude Modulation

The digital modulation schemes described so far modulate the phase or frequency of a carrier to convey digital data and the constellation points lie on a circle of constant amplitude. The effect of this is to provide some immunity to amplitude changes to the signal. However, much more information can be transmitted if the amplitude is varied as well as the phase. With considerable signal processing it is possible to reliably use quadrature amplitude modulation (QAM) in which amplitude and phase are both changed.

A 16-state rectangular QAM, 16-QAM, constellation is shown in Figure 2-29(c). Since there are 16 ($= 2^4$) symbols the values of 4 binary bits are uniquely specified by each symbol. In Figure 2-29(c) a gray-scale assignment of 4 bit values is shown. Several QAM schemes are shown in Figure 2-31. These constellations can be produced by separately amplitude modulating an I carrier and a Q carrier. Both carriers have the same frequency but are 90° out of phase. The two carriers are then combined so that the fixed carrier is suppressed. The most common form of QAM is square QAM, or rectangular QAM with an equal number of I and Q states. The most common forms are

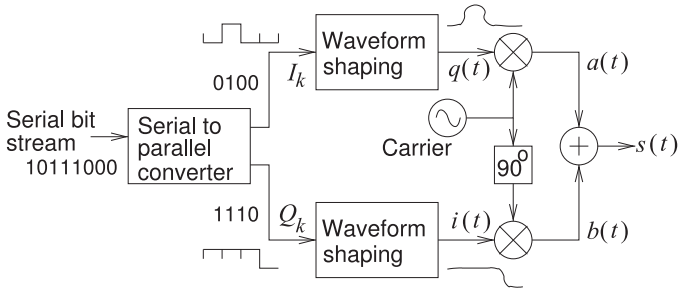


Figure 2-32: QAM modulator block diagram. In QAM modulation $i(t)$ and $q(t)$ address the real and imaginary components of a phasor. The wave-shaping block ensures that the symbol has the correct amplitude and phase at each clock tick.

Modulation	bits/s/Hz
BPSK (ideal)	1
BFSK (actual)	1
QPSK (ideal)	2
GMSK (an actual FSK method)	1.354
$\pi/4$-DQPSK (an actual QPSK method)	1.63
8-PSK (ideal)	3
$3\pi/8$-8PSK (an actual 8PSK method)	2.7
16-QAM (ideal)	4
16-QAM (actual)	2.98
32-QAM (ideal)	4
32-QAM (actual)	3.35
64-QAM (ideal)	6
64-QAM (actual)	4.47
256-QAM (ideal)	8
256-QAM (actual, satellite & cable TV)	6.33
512-QAM (ideal)	9
1024-QAM (ideal)	10
2048-QAM (ideal)	11

Table 2-2: Modulation efficiencies of various modulation formats in bits/s/Hz (bits per second per hertz). The maximum (or ideal) modulation efficiencies obtained by modulation schemes (e.g., BPSK, BFSK, 64-QAM, 256-QAM) result in broad spectra. Actual modulation efficiencies achieved are less in an effort to manage bandwidth. For example, the values for pi/4-DQPSK and 3pi/8-8PSK are actual. This reduction from ideal arises since symbol transitions are of different lengths and length corresponds to time durations. Since the symbol interval is fixed, it is the longest path that determines the bandwidth required.

16-QAM, 64-QAM, and 128-QAM, in 4G, and 256-QAM additionally in 5G. The constellation points are closer together with high-order QAM and so are more susceptible to noise and other interference. Thus high-order QAM can deliver more data, but less reliably than lower-order QAM.

The constellation in QAM can be constructed in many ways, and while rectangular QAM is the most common form, non rectangular schemes exist; for example, having two PSK schemes at two different amplitude levels. While there are sometimes minor advantages to such schemes, square QAM is generally preferred as it requires simpler modulation and demodulation.

One possible architecture of a QAM modulator is shown in Figure 2-32 and this can only be implemented in DSP since it is not sufficient to use analog lowpass filtering to implement the wave-shaping function as the $i(t)$ and $q(t)$ must be precisely the real and imaginary parts of the symbol at each clock tick.

2.10 Digital Modulation Summary

The modulation efficiencies of various digital modulation schemes are summarized in Table 2-2. For example, in 1 kHz of bandwidth the $3\pi/8$-8PSK scheme (supported in 3G cellular radio) transmits 2700 bits.

beta.69: It is critical to control interference in digital radio so that the error

in digital transmission is no more than one bit per symbol. Error correction can then be used to provide error-free digital communications.

The modulation efficiency of an actual modulation method is less than the ideal (see Table 2-2). With digital modulation wave-shaping at baseband is required to constrain the spectrum of the RF-modulated signal. Thus it will take different times for the phasor to make the transition from one symbol to another; to achieve longer transitions in the same time interval requires more bandwidth than that required for shorter transitions. As a result, the modulation efficiency of modulation methods other than binary methods will be less than the ideal. So in a QPSK-like scheme, 2 bits per symbol are achievable, but the longest transition takes the most time, so the bandwidth needs to be increased so that the transition is completed in time (i.e., in a fixed time equal to one over the bandwidth). Various modulation methods have relative merits in terms of modulation efficiency, tolerance to fading (due to destructive interference), carrier recovery, spectral spreading in nonlinear circuitry, and many other issues that are the purview of communication system theorists.

2.11 Interference and Distortion

Demodulation of a received signal is equivalent to reconstructing the original constellation diagram of the modulation signal. Errors are caused by interference, noise, and distortion. Commonly all three effects are lumped together and called interference and treated as though they have noise-like Gaussian randomness. Ideally there is no interference so that a receiver correctly detects the correct symbol upon demodulation. Interference will result in perhaps an incorrect symbol choice and thus error. Errors can be reduced by increasing the signal level at the transmitter thus increasing the signal-to-interference ratio. This comes with a price as increasing the signal level results in higher levels of interference for other radios. The solution arrived at is to ensure that if there is an error, then the incorrectly detected symbol is no more than one symbol away from the actual symbol. This means that there is at most one bit in error while a symbol can carry multiple bits of information. Error correction coding can then be used to eliminate errors.

Émile Baudot used gray codes in telegraphy in 1878 [15]. The name derives from Frank Gray who used them in a pulse code modulation coding scheme [16].

In QAM, symbols are assigned to constellation points using a **Gray code** in which nearest neighbor symbols change by only one bit [17], e.g. see Figure 2-29(c). Thus there is only one bit out of many that will be in error if there is noise and interference. If the error is greater then a lower-order of modulation is used so that if a symbol is incorrectly detected then the incorrect symbol is at most one symbol away from the actual transmitted symbol.

2.11.1 Cochannel Interference

The minimum signal detectable in conventional wireless systems is determined by the **signal-to-interference ratio** (**SIR**) at the input to a receiver, where interference refers to noise as well as interference from other radios. In cellular wireless systems the dominant interference is from other radios and is called cochannel interference. The degree to which cochannel interference can be controlled has a large effect on system capacity. Control of cochannel interference is largely achieved by controlling the power levels

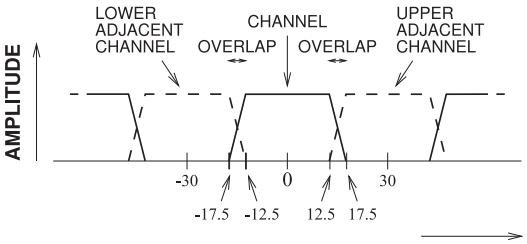


Figure 2-33: Adjacent channels showing overlap in the AMPS and DAMPS cellular systems.

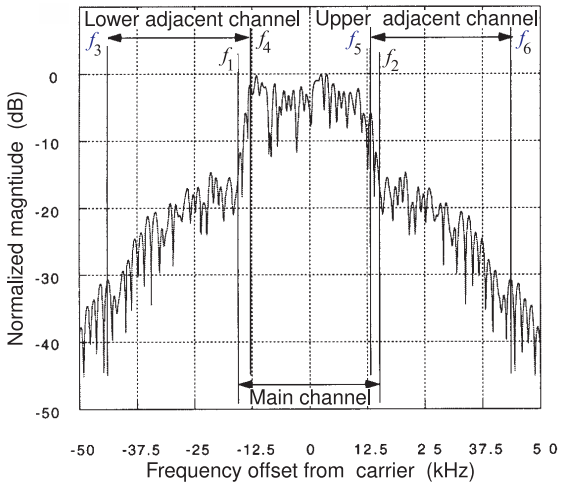


Figure 2-34: Spectrum defining adjacent channel and main channel integration limits using a $\pi/4$-DQPSK modulation scheme.

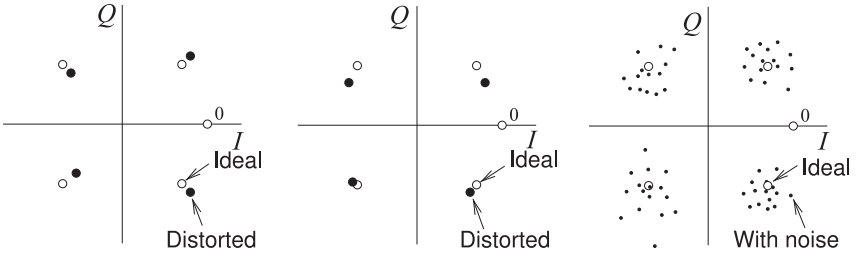
at the base station and at the mobile units.

2.11.2 Adjacent Channel Interference

Adjacent channel interference results from several factors. Since filtering is non ideal, there is inherent overlap of neighboring channels (Figure 2-33). For this reason, adjacent channels are assigned to different cells. The nonlinear behavior of transmitters also contributes to adjacent channel interference. Thus characterization of nonlinear phenomena is important in RF design. The spectrum of a signal modulated using a QPSK scheme is shown in Figure 2-34. The signal between frequencies f_1 and f_2 is due to the digital modulation scheme itself. Most of the signal outside this region is due to nonlinear effects and is called spectral regrowth, a process similar to distortion of a two-tone signal. Using the frequency limits defined in Figure 2-34, the lower channel ACPR is defined as

$$\text{ACPR}_{\text{ADJ,LOWER}} = \frac{\text{Power in lower adjacent channel}}{\text{Power in main channel}} = \frac{\int_{f_3}^{f_4} X(f)df}{\int_{f_1}^{f_2} X(f)df}, \quad (2.54)$$

where $X(f)$ is the spectral power density of the RF signal. Upper channel ACPR, $\text{ACPR}_{\text{ADJ,UPPER}}$, is similarly defined. When ACPR is being referred to without indicating whether it is upper or lower ACPR, the larger (i.e., the worse) of $\text{ACPR}_{\text{ADJ,LOWER}}$ and $\text{ACPR}_{\text{ADJ,UPPER}}$ is used. ACPR is usually expressed in decibels, and while the definition is such that ACPR will be less than one, when expressed in decibels, a positive number is often used (e.g., 20 dB for an ACPR of 0.01 rather than the correct -20 dB, so be careful).



(a) Amplitude distortion. (b) Phase distortion. (c) With noise.

Figure 2-35: Impact of signal impairments on the constellation diagram of QPSK.

2.11.3 Noise, Distortion, and Constellation Diagrams

Noise and nonlinear distortion affect the ability to correctly demodulate signals and determine the transmitted symbols. These distortion effects can be described by their effect on the received constellation diagram, see Figure 2-35, which shows the state of the system at the sampling instant determined by the recovered carrier.

Figure 2-35(a) shows the effect of amplitude distortion errors on the demodulated signal. Sampling of the received signal will be a distorted constellation point that does not correspond to the ideal constellation point. A decision must be made by the DSP unit as to which ideal constellation point corresponds to the distorted constellation point. The effect of phase distortion is shown in Figure 2-35(b). Both amplitude and phase distortion could occur in the transmitter or receiver, or be the result of effects in the signal path. Figure 2-35(c) shows the effect of noise on signal impairment. Again the constellation point extracted from the RF signal is affected by noise and the sampled and ideal constellation points do not coincide. Associating the constellation point extracted from sampling the received RF waveform with the wrong constellation point creates a symbol error and thus a bit error. Errors in recovering the carrier further distort the constellation diagram. All mobile digital radio systems adjust the level of the transmitted RF signal, and additionally in 4G and 5G cellular radio change the order of modulation, so an acceptable BER is obtained. Using more power than necessary reduces battery life and causes additional interference in other radios.

2.11.4 Comparison of GMSK and $\pi/4$DQPSK Modulation

This section presents the results of the type of analysis that is performed to characterize modulation methods. There is a large body of literature documenting the performance of modulation schemes and is usually the result of an assumed error model, that is the type of interference and the statistical description of that interference, and then numerical simulations. This section compares GMSK and $\pi/4$DQPSK modulation methods, the first widely-used cellular digital modulation methods.

The constellations of 4-state GMSK, see Figure 2-11(b), and $\pi/4$DQPSK, see Figure 2-19 are the same except that with $\pi/4$ DQPSK the constellation rotates every clock tick. In GMSK the amplitude of the modulated carrier remains almost constant and the frequency of the carrier varies slowly. With GMSK the symbols correspond to different RF frequencies so it is not possible for symbols with frequencies at opposite ends of the frequency range to transition directly without traversing the other symbol points. This long transition results in reduction of GMSK's modulation efficiency. With

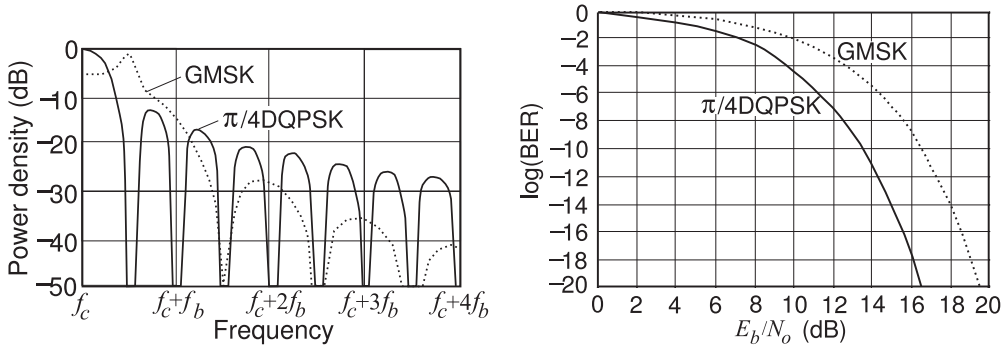


Figure 2-36: Comparison of GMSK and $\pi/4$DQPSK: (a) power spectral density as a function of frequency deviation from the carrier; and (b) BER versus signal-to-noise ratio (SNR) as E_b/N_o (or energy per bit divided by noise per bit).

$\pi/4$DQPSK there are direct transitions and the magnitude of the RF phasor does not stay constant. So while higher modulation efficiency is obtained compared to GMSK, $\pi/4$-DQPSK has a significantly time-varying envelope.

The modulation format used impacts the choice of circuitry, battery life, and the tolerance of the system to noise. Figure 2-36 contrasts the power density versus frequency and **bit error rate** (**BER**) of four-state GMSK and QPSK modulation. In Figure 2-36(a), f_c is the carrier frequency and f_b is the bit frequency, and it is seen that GMSK and QPSK have different spectral shapes. Most of the energy is contained within the bandwidth defined by half the bit frequency (this is the symbol frequency since there are two bits per symbol). At multiples of the bit frequency, the power density with GMSK is much lower than with QPSK, resulting in less interference (**adjacent channel interference** [**ACI**]) with radios in adjacent channels. This is an important metric with radios that is captured by the **adjacent channel power ratio** (**ACPR**), the ratio of the power in the adjacent channel to the power in the main channel. Another important metric is the BER which is increased by noise in the main channel with different modulation formats differing in their susceptibility to interference. The level of noise is captured by the ratio of the power in a bit, E_b, to the noise power, N_o, in the time interval of a bit. This ratio, E_b/N_o (read as E B N O), is directly related to the **signal-to-noise ratio** (**SNR**). In particular, consider the plot of the BER against E_b/N_o shown in Figure 2-36(b), where it can be seen that $\pi/4$DQPSK is less susceptible to noise than GMSK.

2.11.5 Error Vector Magnitude

The error vector magnitude (**EVM**) metric characterizes the accuracy of the waveform at the sampling instances and so is directly related to the BER in digital radio. EVM captures the combined effect of amplifier nonlinearities, amplitude and phase imbalances of separate I and Q signal paths, in-band amplitude ripple (e.g., due to filters), noise, non ideal mixing, non ideal carrier recovery, and DAC inaccuracies.

The EVM is a measure of the departure of a sampled phasor from the ideal

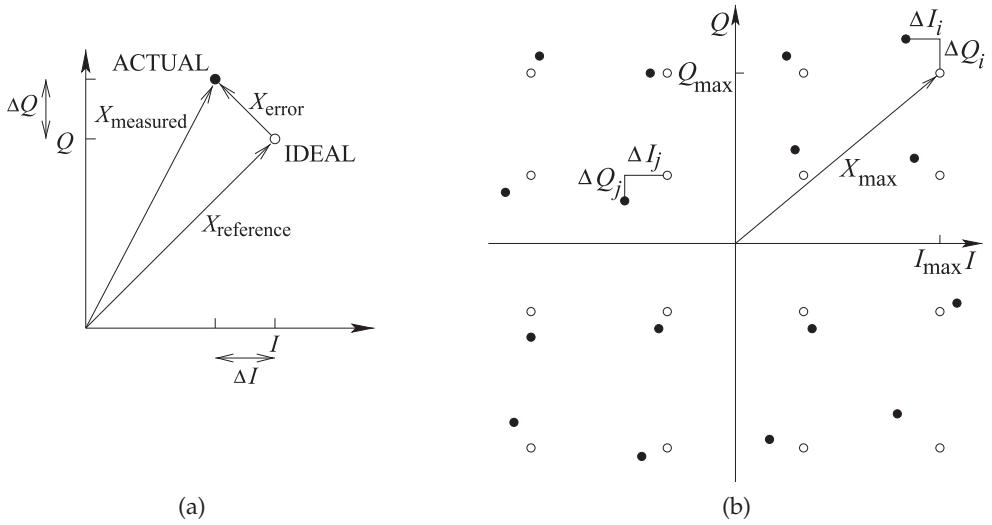


Figure 2-37: Partial constellation diagram showing quantities used in calculating EVM: (a) definition of error and reference signals; and (b) error quantities used when constellation points have different powers.

phasor located at the constellation point (see Figure 2-37(a)). Introducing an error vector, X_{error}, and a reference vector, $X_{\text{reference}}$, that points to the ideal constellation point, the EVM is defined as the ratio of the magnitude of the error vector to the reference vector so that

$$\text{EVM} = \frac{|X_{\text{error}}|}{|X_{\text{reference}}|}. \tag{2.55}$$

Expressing the error and reference in terms of the powers P_{error} and $P_{\text{reference}}$, respectively, enables EVM to be expressed as

$$\text{EVM} = \sqrt{\frac{P_{\text{error}}}{P_{\text{reference}}}}, \tag{2.56}$$

in decibels,
$$\text{EVM}_{\text{dB}} = 10\log\frac{P_{\text{error}}}{P_{\text{reference}}} = 20\log\frac{|X_{\text{error}}|}{|X_{\text{reference}}|}; \tag{2.57}$$

or as a percentage,
$$\text{EVM}(\%) = \frac{X_{\text{error}}}{X_{\text{reference}}} \cdot 100\%. \tag{2.58}$$

If the modulation format results in constellation points having different powers (e.g., with 16-QAM), the constellation point with the highest power is used as the reference and the error at each constellation point is averaged. With reference to Figure 2-37, and with N constellation points,

$$\text{EVM} = \sqrt{\frac{\frac{1}{N}\Sigma_{i=1}^{N}\left(\Delta I_i^2 + \Delta Q_i^2\right)}{X_{\max}^2}}, \tag{2.59}$$

where $|X_{\max}|$ is the magnitude of the reference vector to the most distant constellation point and ΔI_i and ΔQ_i are the I and Q offsets of the actual

constellation point and the ideal constellation point. Note that for QAM the constellation diagram corresponds to a phasor diagram that is being continuously normalized to the average received RF signal level. In the constellation diagram the I and Q coordinates correspond to RMS quantities. Thus $|X_{\max}|$ is an RMS quantity. EVM is traditionally expressed as a percentage.

A similar measure of signal quality is the **modulation error ratio** (**MER**), a measure of the average signal power to the average error power. In decibels it is defined as

$$\text{MER}|_{\text{dB}} = 10\log\frac{\frac{1}{N}\Sigma_{i=1}^{N}\left(I_i^2+Q_i^2\right)}{\frac{1}{N}\Sigma_{i=1}^{N}\left(\Delta I_i^2+\Delta Q_i^2\right)} = 10\log\frac{\Sigma_{i=1}^{N}\left(I_i^2+Q_i^2\right)}{\Sigma_{i=1}^{N}\left(\Delta I_i^2+\Delta Q_i^2\right)}. \quad (2.60)$$

The advantage of the MER is that it relates directly to the SNR.

Another quantity related to both the EVM and MER concepts is the implementation margin, k.. The implementation margin is a measure of the performance of particular hardware and is developed by design groups based on experience with similar designs. The required EVM can be estimated from the hardware implementation margin:

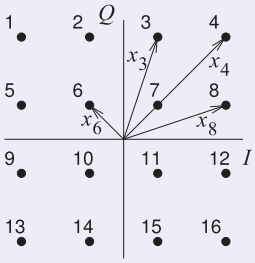
$$\text{EVM}_{\text{required}} = \sqrt{\frac{k}{\text{SNR}\cdot\text{PMEPR}}}. \quad (2.61)$$

In decibels, $$\text{EVM}_{\text{db, required}} = k_{\text{dB}} - \text{SNR}_{\text{dB}} - \text{PMEPR}_{\text{dB}}. \quad (2.62)$$

EXAMPLE 2.10 Modulation Error Ratio

A 16-QAM modulated signal has a maximum RF phasor rms value of 4 V. If the noise on the signal has an rms value of 0.1 V, what is the modulation error ratio of the modulated signal?

Solution:



The distance from the origin to each of the constellation points must be determined, but because of symmetry only the distances $x_3, x_4, x_7(=x_6)$ and x_8 need to be calculated. The maximum RF phasor amplitude is 4 V, so the length $x_4 = 4$, with components

$I_4 = Q_4 = \sqrt{x_4^2/2} = 2.828$

$I_6 = \left(-\frac{1}{3}\right)\cdot I_4 = -0.943 = -Q_6$, so $x_6 = 1.333$

$I_3 = 0.943$; $Q_3 = 2.828$, so $x_3 = 2.981 = x_8$

and $$\text{MER} = \frac{\Sigma_{i=1}^{16}\left(x_i^2\right)}{\Sigma_{i=1}^{16}\left(x_{\text{noise}}^2\right)}.$$

The calculation is simplified by considering just one quadrant and the noise, $x_{\text{noise}} = 0.1$, is the same for each constellation point.

$$\text{MER} = \frac{4\left(x_3^2+x_4^2+x_6^2+x_8^2\right)}{16\cdot x_{\text{noise}}^2}$$

$$= \frac{4\left(2.981^2+4^2+1.333^2+2.981^2\right)}{16\cdot 0.1^2} = \frac{142.2}{0.16} = 888.8.$$

$$\text{MER}|_{\text{dB}} = 10\log(888.8) = 29.5\ \text{dB}$$

Compare this to the SNR calculated at the individual constellation points:
Point 3, SNR = $x_3^2/0.1^2 = 888.6 = 29.5$ dB
Point 4, SNR = $x_4^2/0.1^2 = 1600 = 32.0$ dB
Point 7, SNR = $x_7^2/0.1^2 = 177.7 = 22.5$ dB.

2.12 Summary

All modern modulation methods impress information on a sinusoidal carrier that is at a high enough frequency that it can be easily transmitted. There are many modulation techniques with the choice of which to use based on the technology available to implement the modulation scheme, the tolerance of the modulation scheme to interference, how efficiently the modulation scheme uses the EM spectrum, and the amount of DC power consumed. In military communications it is also important that a modulation scheme produce a noise-like signal that is difficult to detect and intercept.

The first widely adopted modulation schemes produced simple pulses as used in wireless telegraphy. The tolerance to interference was achieved through the relatively slow transmission of bits, and hence redundancy. More information was transmitted when amplitude modulation was introduced to superimpose voice on a carrier. With this scheme, interference was always a problem, and once interference appeared on a signal it could not be removed nor suppressed. The first significant advance in modulation techniques was the invention of frequency modulation. In this scheme a narrowband analog modulating signal (e.g. voice) became a relatively wideband frequency-modulated RF signal. When the modulated signal was received and demodulated, the wide bandwidth of the modulated signal was collapsed to the original relatively narrowband modulating signal. The demodulation process combined correlated components of the modulated signal and the uncorrelated components, noise and interference, were suppressed.

The introduction of digital modulation was a significant advance in the suppression of interference. Digital information was now being transmitted, and errors in the data caused by interference and noise would be completely unacceptable. The solution was to embed error-correcting codes in the data so that if a manageable number of bits were lost in the transmitted signal, the original data could still be fully recovered. As a result, digital radio (using digital modulation) could be used in situations with even more distortion than was acceptable in analog radio (using analog modulation). If interference is low, then today's wireless systems use high-order modulation switching to lower-order (and less spectrally efficient) modulation when necessary to cope with higher interference.

Several important metrics are used to provide a measure of the signal characteristics. The crest factor, peak-to-average ratio, and peak-to-mean envelope power ratio (PMEPR) are all indicators of how much care must be given to nonlinear circuit design, especially to amplifier and mixer design. Amplifiers must operate so that the peak signal is amplified with minimal distortion. It is the peak signal that determines the DC power drawn by an amplifier. However, the average RF output power of an RF front end is determined by the mean of the envelope. So a high PMEPR signal will result in lower amplifier efficiency.

Many of the techniques described in this chapter for modulating and demodulating RF signals were presented as circuit techniques. However many modern phones support multiple standards and hardware implementation would require multiple copies of similar versions of analog circuits. Today it is more cost effective to perform most of the operations in a DSP unit. Most of the time the DSP realization is close to the hardware implementation. An example is carrier recovery. For narrow-band communication signals in wireless communicators, carrier recovery can be performed using a digital implementation of the concepts described for the hardware carrier recovery circuits. While it is more power efficient to implement many of the techniques in hardware, the need to support multiple standards has necessitated the software reconfigurability available with a DSP unit. Which approach is used is the decision of the RF system designer—an experienced engineer with a rich background in wireless technologies. It is therefore important that the aspiring and practicing RF engineer have a broad perspective of RF circuits and of communications theory. Hence the emphasis of this book on a systems approach to RF and microwave design.

Frequency modulation, and the similar PM modulation method were used in the 1G analog cellular radio. With the addition of AM, the three schemes are the bases of all analog radio. Digital cellular radio began with 2G and there were two types of 2G cellular radios with the GSM system using GMSK modulation, a type of FSK modulation, and the NADC system using $\pi/4$-DQPSK modulation. The two 2G systems were incompatible. The 3G cellular radio used two types of QPSK modulation, one for the up-link from handset to basestation, and one from the basestation to a handset. The 1G–3G systems implemented most of the modulation and demodulation functions in analog hardware. With 4G and 5G cellular radio a large number of modulation schemes are supported choosing as high an order of modulation as allowed by the channel conditions. Most of the modulation and demodulation in 4G and 5G are implemented in DSP with just the translation to and from the radio frequency signal implemented in analog hardware.

2.13 References

[1] "American National Standard T1.523-2001, Telecom Glossary 2011," available on-line with revisions at http://glossary.atis.org, 2011, sponsored by Alliance for Telecommunications Industry Solutions.

[2] S. Boyd, "Multitone signals with low crest factor," *IEEE Trans. on Circuits and Systems*, vol. 33, no. 10, pp. 1018–1022, Oct. 1986.

[3] A. Jones, T. Wilkinson, and S. Barton, "Block coding scheme for reduction of peak to mean envelope power ratio of multicarrier transmission schemes," *Electronics Letters*, vol. 30, no. 25, pp. 2098–2099, Dec. 1994.

[4] M. Steer, *Microwave and RF Design, Transmission Lines*, 3rd ed. North Carolina State University, 2019.

[5] D. Porcino and W. Hirt, "Ultra-wideband radio technology: potential and challenges ahead," *IEEE communications magazine*, vol. 41, no. 7, pp. 66–74, 2003.

[6] "FCC (GPO) Title 47, Section 15 of the Code of Federal Regulations SubPart F: Ultra-wideband," http://www.access.gpo.gov/nara/cfr/waisidx_05/47cfr15_05.html.

[7] J. Carson, "Notes on the theory of modulation," *Proc. of the Institute of Radio Engineers*, vol. 10, no. 1, pp. 57–64, Feb. 1922.

[8] L. Couch III, *Digital and Analog Communication Systems*, 6th ed. Prentice-Hall, 2001.

[9] E. Armstrong, "A method of reducing disturbances in radio signaling by a system of frequency modulation," *Proc. of the Institute of Radio Engineers*, vol. 24, no. 5, pp. 689–740, May 1936.

[10] ——, "Radio telephone signaling," US Patent US Patent 1 941 447, 12 26, 1933.

[11] "Armstrong suit over fm settled. r.c.a. and n.b.c. to pay '$1,000,000' ending action begun

by late inventor," New York Times, Dec. 31, 1954.
[12] E. Bedrosian, "The analytic signal representation of modulated waveforms," *Proceedings of the IRE*, vol. 50, no. 10, pp. 2071–2076, 1962.
[13] C. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21, 1949.
[14] J. Costas, "Synchronous communications," *Proc. of the IRE*, vol. 44, no. 12, pp. 1713–1718, Dec. 1956.
[15] F. Hearth, "Origins of the binary code," *Scientific American*, pp. 76–83, Aug. 1972.
[16] F. Gray, "Pulse code modulation," US Patent US Patent 11 111 111, 03 17, 1953.
[17] C. Savage, "A survey of combinatorial gray codes," *SIAM Review*, vol. 39, no. 4, pp. 605–629, 1997.

2.14 Exercises

1. Develop a formula for the average power of a signal $x(t)$. Consider $x(t)$ to be a voltage across a 1 Ω resistor.
2. What is the PAPR of a 5-tone signal when the amplitude of each tone is the same?
3. What is the PMEPR of a 10-tone signal when the amplitude of each tone is the same?
4. Consider two uncorrelated analog signals combined together. One signal is denoted $x(t)$ and the other $y(t)$, where $x(t) = 0.1\sin(10^9 t)$ and $y(t) = 0.05\sin(1.01 \cdot 10^9 t)$. The combined signal is $z(t) = x(t) + y(t)$. [Parallels Example 2.3]
   (a) What is the PAPR of $x(t)$ in decibels?
   (b) What is the PAPR of $z(t)$ in decibels?
   (c) What is the PMEPR of $x(t)$ in decibels?
   (d) Is it possible to calculate the PMEPR of $z(t)$? If so, what is it?
5. Consider two uncorrelated analog signals combined together. One signal is denoted $x(t)$ and the other $y(t)$, where $x(t) = 0.1\sin(10^8 t)$ and $y(t) = 0.05\sin(1.01 \cdot 10^8 t)$. What is the PMEPR of this combined signal? Express PMEPR in decibels. [Parallels Example 2.3]
6. What is PMEPR of a three-tone signal when the amplitude of each tone is the same?
7. What is PMEPR of a four-tone signal when the amplitude of each tone is the same?
8. A tone $x_1(t) = 0.12\cos(\omega_1 t)$ is added to two other tones $x_2(t) = 0.14\cos(\omega_2 t)$ and $x_3(t) = 0.1\cos(\omega_3 t)$ to produce a signal $y(t) = x_1(t) + x_2(t) + x_3(t)$, where $y(t)$, $x_1(t)$, $x_2(t)$ and $x_3(t)$ are voltages across a 100 Ω resistor. Consider that ω_1, ω_2, and ω_3 are 10% apart and that the signals at these frequencies are uncorrelated.
   (a) What is the PMEPR of $x_1(t)$? Express your answer in decibels.
   (b) Sketch $y(t)$.
   (c) The combined signal appears as a pseudo-carrier with a time-varying envelope. What is the power of the largest single cycle of the pseudo-carrier?
   (d) What is the average power of $y(t)$?
   (e) What is the PMEPR of $y(t)$? Express your answer in decibels.
9. Consider two uncorrelated analog signals summed together. One signal is denoted $x(t)$ and the other $y(t)$, where $x(t) = \sin(10^9 t)$ and $y(t) = 2\sin(1.01 \cdot 10^9 t)$ so that the total signal is $z(t) = x(t) + y(t)$. What is the PMEPR of $z(t)$ in decibels? [Parallels Example 2.3]
10. What is the PMEPR of an FM signal at 1 GHz with a maximum modulated frequency deviation of ± 10 kHz?
11. What is the PMEPR of a two-tone signal (consisting of two sinewaves at different frequencies that are, say, 1% apart)? First, use a symbolic expression, then consider the special case when the two amplitudes are equal. Consider that the two tones are close in frequency.
12. What is the PMEPR of a three-tone signal (consisting of three equal-amplitude sinewaves, say, 1% apart in frequency)?
13. A phase modulated tone $x_1(t) = A_1\cos[\omega_1 t + \phi_1(t)$. What is the PMEPR of $x_1(t)$? Express your answer in decibels.
14. What is the PMEPR of an AM signal with 75% amplitude modulation?
15. Two FM voltage signals $x_1(t)$ and $x_2(t)$ are added together and then amplified by an ideal linear amplifier terminated in 50 Ω with a gain of 10 dB and the output voltage of the amplifier is $y(t) = \sqrt{10}\,[x_1(t) + x_2(t)]$.
   (a) What is the PMEPR of $x_1(t)$? Express your answer in decibels?
   (b) What effect does the amplifier have on the PMEPR of the signal?
   (c) If $x_1(t) = A_1\cos[\omega_1(t)t]$ and $x_2(t) = A_2\cos[\omega_2(t)t]$, what is the PMEPR of the output of the amplifier, $y(t)$? Express PMEPR in decibels. Consider that $\omega_1(t)$ and $\omega_2(t)$ are within 0.1% of each other.

16. An FM signal has a maximum frequency deviation of 20 kHz and a modulating signal between 300 Hz and 5 kHz. What is the bandwidth required to transmit the modulated RF signal when the carrier is 200 MHz? Is this considered to be narrowband FM or wideband FM?

17. A high-fidelity stereo audio signal has a frequency content ranging from 50 Hz to 20 kHz. If the signal is to be modulated on an FM carrier at 100 MHz, what is the bandwidth required for the modulated RF signal? The maximum frequency deviation is 5 kHz when the modulating signal is at its peak value.

18. Consider FM signals close in frequency but whose spectra do not overlap. [Parallels Example 2.5]
    (a) What is the PMEPR of just one PM signal? Express your answer in decibels.
    (b) What is the PMEPR of a signal comprised of two uncorrelated narrowband PM signals with the same average power?

19. Consider two nonoverlapping equal amplitude FM signals having center frequencies within 1%.
    (a) What is the PMEPR in dB of just one FM modulated signal?
    (b) What is the PMEPR in dB of a signal comprising two FM signals of the same power?

20. Consider a signal $x(t)$ that is the sum of two uncorrelated signals, a narrowband AM signal with 50% modulation, $y(t)$, and a narrow-band FM signal, $z(t)$. The center frequencies of $y(t)$ and $z(t)$ are within 1%. The carriers have equal amplitude. Express answers in dB.
    (a) What is the PAPR of the AM signal $x(t)$?
    (b) What is the PAPR of the FM signal $z(t)$?
    (c) What is the PAPR of $x(t)$?
    (d) What is the PMEPR of the AM signal $x(t)$?
    (e) What is the PMEPR of the FM signal $z(t)$?
    (f) What is the PMEPR of $x(t)$?

21. Two phase modulated tones $x_1(t) = A_1 \cos[\omega_1 t + \phi_1(t)$ and $x_2(t) = A_2 \cos[\omega_2 t + \phi_2(t)$ are added together as $y(t) = x_1(t) + x_2(t)$. What is the PMEPR of $y(t)$ in decibels. Consider that ω_1 and ω_2 are within 0.1% of each other.

22. A radio uses a channel with a bandwidth of 25 kHz and a modulation scheme with a gross bit rate of 100 kbits/s that is made of an information bit rate of 60 kbits/s and a code bit rate of 40 kbits/s.
    (a) What is the modulation efficiency in bits/s/Hz?
    (b) What is the spectral efficiency in bits/s/Hz?

23. A cellular communication system uses $\pi/4$-DQPSK modulation with a modulation efficiency of 1.63 bits/s/Hz to transmit data at the rate of 30 kbits/s. This would be the spectral efficiency in the absence of coding. However, 25% of the transmitted bits are used to implement a forward error correction code.
    (a) What is the gross bit rate?
    (b) What is the information bit rate?
    (c) What is the bandwidth required to transmit the information and code bits?
    (d) What is the spectral efficiency in bits/s/Hz?

24. A radio uses a channel with a 5 MHz bandwidth and uses 256-QAM modulation with a modulation efficiency of 6.33 bits/s/Hz. The coding rate is 3/4 (i.e. of every 4 bits sent 3 are data bits and the other is an error correction bit).
    (a) What is gross bit rate in Mbits/s?
    (b) What is information rate in Mbits/s?
    (c) What is the spectral efficiency in bits/s/Hz?

25. The following sequence of bits 0100110111 is to be transmitted using QPSK modulation. Take these data in pairs, that is, as 01 00 11 01 11. These pairs, one bit at a time, drive the I and Q channels. Show the transitions on a constellation diagram. [Parallels Example 2.7]

26. The following sequence of bits 0100110111 is to be transmitted using $\pi/4$-DQPSK modulation. Take these data in pairs, that is, as 01 00 11 01 11. These pairs, one bit at a time, drive the I and Q channels. Use five constellation diagrams, with each diagram showing one transition or symbol. [Parallels Example 2.8]

27. The following sequence of bits 0100110111 is transmitted using OQPSK modulation. Take these data in pairs, that is, as 01 00 11 01 11. These pairs, one bit at a time, drive the I and Q channels. Show the transitions on a constellation diagram.

28. Draw the constellation diagram of OQPSK.

29. Draw the constellation diagrams of $3\pi/8$-8DPSK and explain the operation of this system and describe its advantages.

30. How many bits per symbol can be sent using $3\pi/8$-8PSK?

31. How many bits per symbol can be sent using 8-PSK?

32. How many bits per symbol can be sent using 16-QAM?

33. Draw the constellation diagram of OQPSK modulation showing all possible transitions. You may want to use two diagrams.

34. What is the PMEPR of a 5-tone signal when the amplitude of each tone is the same?
35. Draw the constellation diagram of 64QAM.
36. How many bits per symbol can be sent using 32QAM?
37. How many bits per symbol can be sent using 16QAM?
38 How many bits per symbol can be sent using 2048QAM?
39. Consider a two-tone signal and describe intermodulation distortion in a short paragraph and include a diagram.
40. A 16-QAM modulated signal has a maximum RF phasor amplitude of 5 V. If the noise on the signal has an rms value of 0.2 V, what is the EVM of the modulated signal? [Parallels Example 2.10]
41. Consider a digitally modulated signal and describe the impact of a nonlinear amplifier on the signal. Include several negative effects.
42. A carrier with an amplitude of 3 V is modulated using 8-PSK modulation. If the noise on the modulated signal has an rms value of 0.1 V, what is the EVM of the modulated signal? [Parallels Example 2.10]
43. Consider a 32-QAM modulated signal which has a maximum I component, and a maximum Q component, of the RF phasor of 5 V. If the noise on the signal has an RMS value of 0.1 V, what is the modulation error ratio of the modulated signal in decibels? Refer to Figure 2-31(b). [Parallels Example 2.10]

2.14.1 Exercises By Section

§12.2	1, 2, 3, 4, 5, 6, 7, 8, 9		20, 21	§12.9	32, 33, 34, 35, 36, 37, 38
	10, 11, 12	§12.5	22, 23, 24	§12.11	39, 40, 41, 42, 43
§12.4	13, 14, 15, 16, 17, 18, 19	§12.8	25, 26, 27, 28, 29, 30, 31		

2.14.2 Answers to Selected Exercises

5	2.55 dB	15	no effect	22(a)	4 bit/s/Hz
7(e)	3.78 dB	20(a)	6 dB	43	36.02 dB
8	0.00022 W	20(e)	0 dB		

CHAPTER 3

Transmitters and Receivers

3.1 Introduction

An essential function of a radio transmitter is modulation which is implemented by a modulator which converts the information at baseband to a finite bandwidth modulated radio signal centered at or near the carrier frequency. At the block level a modulator is described by its architecture and at a lower level by its circuit implementation. In a receiver a demodulator demodulates the radio signal extracting the baseband information.

The architectures of transmitters and receivers have changed significantly since the first radio signals were transmitted. Largely this is the result of increased integration of hardware but also due to advances in modulation concepts. In modern radios, as used with 4G and 5G cellular radio, many of the modulation functions once only implemented with analog hardware are now implemented digitally in a digital signal processing (**DSP**) unit. Only the final translation to the frequency of the radio signal is implemented in analog hardware. This concept is known as **software-defined radio** (SDR) which also has the additional implication that the analog hardware can be reconfigured under software control. SDR has two significant impacts.

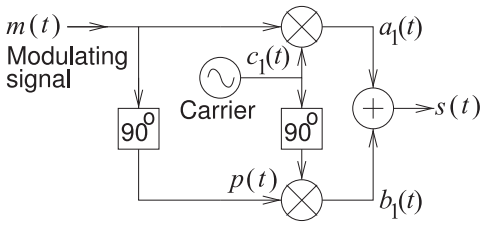


Figure 3-1: Hartley modulator implementing single-sideband suppressed-carrier (SSB-SC) modulation. The "90°" blocks shift the phase of the signal by $+90°$. The mixer indicated by the circle with a cross is an ideal multiplier, e.g. $a_1(t) = m(t) \cdot c_1(t)$.

One is that it is feasible to support a large number of modulation schemes including legacy modulation schemes, without any changes to the analog hardware. The second is that by simplifying the functionality of the analog hardware it is possible to optimize analog hardware for maximum efficiency. Demodulation has also been significantly impacted by SDR with a similar transfer of functions from analog to digital hardware.

In some designs a receiver and a transmitter share components and then the RF front end is called a transceiver. This chapter is concerned with the architectures of receivers, transmitters, and transceivers. Design choices are made to maximize the tolerance to interference, manage hardware complexity, optimize power efficiency, and enable various radios to simultaneously operate together.

The presentation in this chapter closely follows the historical development of transmitters and receivers. As well, modulators and demodulators at different stages of evolution are considered together.

Section 3.2 introduces single-sideband and double-sideband modulation. Then early modulator and demodulator technology is considered in Section 3.3. These were circuits that could be implemented with just a few vacuum tubes or transistors and, in one case, a demodulator could be implemented with a single diode. With increasing circuit sophistication an architectural approach to modulator and demodulator design became possible. These are considered in Section 3.4. Demodulation requires local generation of a radio signal's carrier in a process called carrier recovery. This is discussed in Section 3.5. Following this is a discussion of more modern transmitters and receivers as used in 2G and 3G cellular radio in Sections 3.6 and 3.7 respectively. Current radios, e.g. 4G and 5G, are software defined radios (SDRs) where most of the modulation and demodulation is implemented in DSP and there is considerable software-based adaptation of the analog hardware that now undertakes the task of translating between and low intermediate frequency and the frequency of the radio signal. Sections 3.8–3.13 present the many aspects of SDR transmitters and receivers.

3.2 Single-Sideband and Double-Sideband Modulation

The simplest implementation of analog modulation results in a modulated carrier signal whose spectrum consists of the carrier and upper and lower sidebands. It is possible to eliminate one of the sidebands in AM modulation, or one of the sidebands sets in PM and FM modulation, producing single sideband modulation SSB. This however needs to be implemented in DSP. At the same time the carrier can be suppressed resulting in suppressed-carrier modulation or together SSB-SC modulation.

The simplest system that implements SSC-SC modulation is the **Hartley modulator** [1, 2], shown in Figure 3-1. This circuit results in **single-**

sideband (SSB) modulation or more precisely single-sideband modulation **suppressed-carrier (SSB-SC) modulation**. This circuit is used in all modern radios taking a modulated signal which is centered at an intermediate frequency and shifting it up in frequency so that its is centered at another frequency a little below or a little above the carrier of the Hartley modulator.

Both the modulating signal $m(t)$and the carrier are multiplied together in a mixer and then also $90°$ phase-shifted versions are also mixed before being added together. The signal flow is as follows beginning with $m(t) = \cos(\omega_{m1}t)$, $p(t) = \cos(\omega_{m1}t - \pi/2) = \sin(\omega_{m1}t)$ and carrier signal $c_1(t) = \cos(\omega_c t)$:

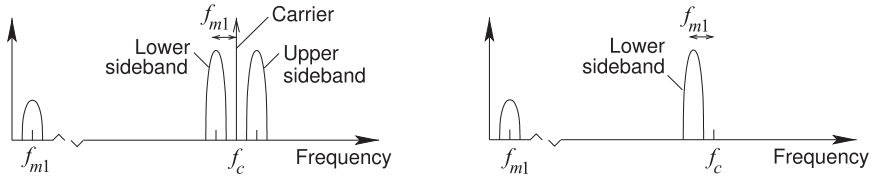
$$\begin{aligned} a_1(t) &= \cos(\omega_{m1}t)\cos(\omega_c t) = \tfrac{1}{2}[\cos((\omega_c - \omega_{m1})t) + \cos((\omega_c + \omega_{m1})t)] \\ b_1(t) &= \sin(\omega_{m1}t)\sin(\omega_c t) = \tfrac{1}{2}[\cos((\omega_c - \omega_{m1})t) - \cos((\omega_c + \omega_{m1})t)] \\ s_1(t) &= a_1(t) + b_1(t) = \cos((\omega_c - \omega_{m1})t) \end{aligned} \tag{3.1}$$

and so the lower sideband (USB) is selected.

That is, if a finite bandwidth modulating signal $m(t)$ was mixed only once with the carrier $c_1(t)$, the spectrum of the output $a_1(t)$ would include upper and lower sidebands as well as the carrier as shown in Figure 3-2(a). With the Hartley modulator the spectrum of Figure 3-2(b) is obtained.

In digital modulation both the upper and lower sideband are retained but the carrier is suppressed. Both sidebands are required to recover the signal but the spectrum is used efficiently as the modulating signal is complex with two components, think of real/imaginary parts, or amplitude/phase information. Since a DSP unit is required the resulting modulated signal must be at a relatively low frequency and then a second frequency conversion stage is required to shift the modulated signal to the desired operating frequency, see Figure 3-3. This second stage is a SSB-SC modulator, however since the input to the second stage is a DSB-SC signal, the final RF signal is a DSB-SC signal.

All the concepts introduced in this section are still used in modern radios, just that now they are mostly implemented in a DSP unit rather than in analog hardware.



(a) Double sideband (DSB) (b) Single-sideband suppressed carrier (SSB-SC)

Figure 3-2: Spectrum of a modulated carrier with a modulating signal of finite bandwidth f_m with f_{m1} being the center frequency of the baseband signal.

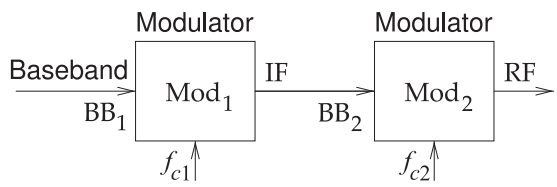


Figure 3-3: Two-stage modulator with the baseband signal, BB_1, input to the first modulator, Mod_1, producing the intermediate frequency signal IF_1. This becomes the baseband signal, BB_2, for the second modulator, Mod_2, producing the radio frequency signal, RF.

3.3 Early Modulation and Demodulation Technology

Early modulators and demodulators are considered here in part because the terms associated with the historical transmitters and receivers are still used today, but also because the early trade-offs influenced the architectures used today. Today transmitters and receivers use DSP technology, very stable LOs, and sophisticated clock recovery schemes. This was not always so. One of the early problems was demodulating a signal when the frequency of transmitter oscillators, i.e. carriers, drifted by up to 10%. Radio at first used AM and the carrier was sent with the information-carrying sidebands. With this signal, a simple single-diode rectifier circuit connected to a bandpass filter could be used, but the reception was poor. To improve performance it was necessary to lock an oscillator in the receiver to the carrier and then amplify the received signal. Here some of the early schemes that addressed these problems are discussed. There were many more variants, but the discussion covers the essential ideas.

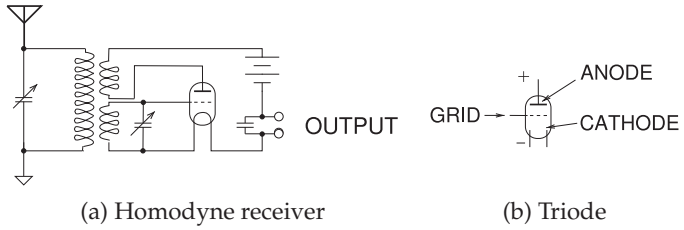
3.3.1 Heterodyne Receiver

The heterodyning principle mixes a single-tone signal, the LO, with a finite bandwidth signal to produce a lower-frequency version of the information-bearing signal. With the LO frequency set appropriately, the low-frequency signal would be in the audio range. If the information-bearing signal is an AM signal, then the low-frequency version of the signal is the original audio signal, which is the envelope of the AM signal. This type of receiver is called a **tuned radio frequency** (**TRF**) **receiver**, and performance is critically dependent on the stability of the LO and the selectivity of the receive filter. The TRF receiver required the user to adjust a tunable capacitor so that, with a fixed inductor, a tunable bandpass filter was created. Such a filter has a limited Q and a bandwidth that is wider than the bandwidth of the radio channel.[1] Even worse, a user had to adjust both the frequency of the bandpass filter and the frequency of the LO. The initial radios based on this principle were called **audions**, used a triode vacuum tube as an amplifier, and were used beginning in 1906. They were an improvement on the **crystal detectors** (which used a single diode with filters), but there was a need for something better.

3.3.2 Homodyne Receiver

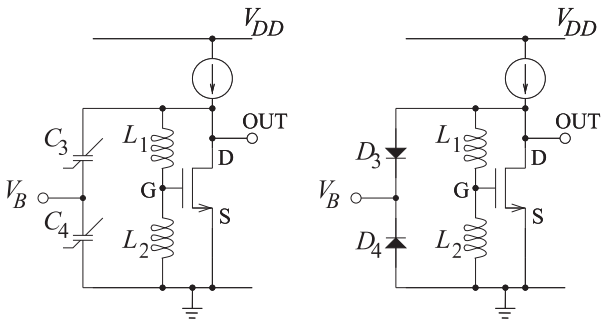
The homodyne [3], syncrodyne (for synchronous heterodyne) [4], and autodyne (for automatic heterodyne) circuits were needed improvements on the audion and are based on the regenerative circuit invented by Edwin Armstrong in 1912 while he was an electrical engineering student at New York City's Columbia University [5]. Armstrong's circuit fed the input signal into an amplifying circuit and a portion of this signal was coupled back into the input circuit so that the signal was amplified over and over again. This is a positive feedback amplifier. A small input RF signal was amplified to such a large extent that it resulted in the amplifying circuit becoming nonlinear and consequently it rectified the amplitude modulated RF signal.

[1] Q is the **quality factor** and is the ratio of the energy stored to the energy resistively lost in each cycle. Good frequency selectively in a filter requires high-Q components.



(a) Homodyne receiver (b) Triode

Figure 3-4: Colebrook's original homodyne receiver: (a) circuit with an antenna, tunable band-pass filter, and triode amplifier; and (b) triode vacuum tube.



(a) (b)

Figure 3-5: A common source Hartley voltage-controlled oscillator (VCO): (a) with nonlinear capacitors; and (b) with diodes which have a variable capacitance when reverse biased.

Colebrook used this principle and developed the original homodyne receiver shown in Figure 3-4(a). This serves to illustrate the operation of the family of regenerative receivers. The antenna shown on the left-hand side is part of a resonant circuit that is in the feedback path of a triode oscillator. The triode vacuum tube is shown in Figure 3-4(b). Here the grid coils (which control the flow of carriers between the bottom cathode[2] and top anode) are weakly coupled to the anode circuit. When an AC signal appears at the top anode, the part within the passband of the tuned circuit is fed back to the grid and the signal is reinforced. The radio signals of the day were AM and had a relatively large carrier, so the oscillator locked on to the carrier. The AM sidebands were then successfully heterodyned down to the desired audio frequencies.

The **autodyne** worked on a slightly different principle in that the oscillation frequency was tuned to a slightly different frequency from the carrier. Still, the autodyne combined the functions of an oscillator and detector in the same circuit.

3.3.3 FM Modulator

FM modulation can be implemented using a voltage-controlled oscillator (VCO) with the baseband signal controlling the frequency of an oscillator. A VCO can be very simple circuit and so was easily implemented in early radio. The circuits in Figure 3-5 are known as common source Hartley VCOs, where in Figure 3-5(a) the controllable elements are the nonlinear capacitors. These are generally implemented as reverse-biased diodes, as shown in Figure 3-5(b) where the bias, V_B, changes the capacitance of the reverse-biased diodes. Changing the capacitance changes the resonant frequency of the feedback loop formed by the inductors and the diode capacitances (called varactors).

[2] The **cathode** is heated (the heater circuit is not shown) and electrons are spontaneously emitted in a process called **thermionic emission**.

3.3.4 FM Demodulator

An FM demodulator is often implemented using a phase-locked loop with an error signal used to control the frequency of a voltage-controlled oscillator (VCO) with the loop arranged so that the VCO tracks the received signal. The desired baseband signal, i.e. the demodulated FM signal, being the loop's error signal.

3.3.5 Superheterodyne Receiver

The superheterodyne receiver was invented by Edwin Armstrong in 1918 [6]. The key concept was to **heterodyne** down in two stages, use fixed filters, and use a tunable LO. The receiving antenna was connected to a bandpass filter that allowed several channels to pass. This relaxed the demands on the receive filter, but also filters with higher selectivity could be constructed if they did not need to be tuned. The filtered received signal is then mixed with an offset LO to produce what is called a **supersonic** signal—a signal above the audio range—and hence the name of this architecture. The performance of the superheterodyne (or superhet) receive architecture has only recently been achieved at cellular frequencies using direct conversion architectures.

3.3.6 Summary

Early radio used AM and FM modulation and both modulation and demodulation could be performed with very simple circuits. However they used a lot of spectrum for the information that was transmitted.

3.4 Receiver and Transmitter Architectures

The essential function of a radio transmitter architecture is taking low-frequency information, the baseband signal, and transferring that information to much higher frequencies by superimposing the baseband signal on a high-frequency carrier, i.e a sinewave. This could be done by slowly varying the amplitude, frequency, and/or phase of the carrier in what is called modulating the carrier to produce a modulated carrier (signal). A caveat here is that there are other less common ways of transferring the baseband information such as rapidly changing the frequency of the carrier at a faster rate than the maximum frequency of the baseband signal in a scheme called frequency hoping. Frequency hoping is particularly useful when operating in hostile situations, such as military communications, when the interference environment cannot be controlled. In reality there is a very large number of ways of producing a radio signal that carries the baseband information. The central theme of this book is to focusing on modulation schemes that slowly modulate characteristics of a carrier signal and other schemes are introduced as an exception. This section describes early architectural developments that led to the development of the regulations that still guide modern architectures; concepts that efficiently pack as much information as possible into a fixed RF bandwidth.

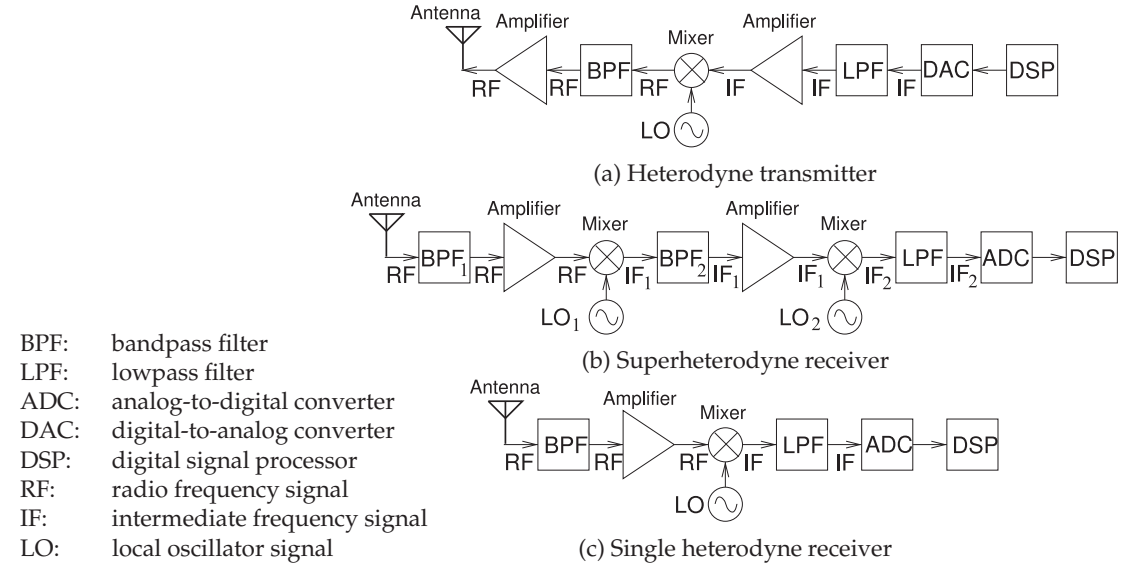
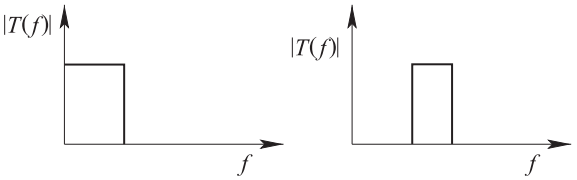


Figure 3-6: RF front ends: (a) a one-stage transmitter; (b) a receiver with two mixing (or heterodyning) stages; and (c) a receiver with one heterodyne stage.



(a) Lowpass, LPF (b) Bandpass, BPF

Figure 3-7: Ideal filter responses where $T(f)$ is the transmission response as a function of frequency f.

3.4.1 Radio as a Cascade of Two-Ports

The front end of an RF communication receiver or transmitter combines a number of subsystems in cascade. The design of the RF front end requires trade-offs of noise generated by the circuit, of frequency selectivity, and of power efficiency, which translates into battery life for a communication handset. There are only a few receiver and transmitter architectures that achieve the optimum trade-offs. The essentials of these architectures are shown in Figure 3-6. These architectures achieve frequency selectivity using bandpass filters (BPFs) and lowpass filters (LPFs) which ideally have the responses shown in Figure 3-7. The corner frequencies of these filters and, in the case of the BPFs, their center frequencies, are adjusted to minimize interfering signals and noise that passes through the system. The three architectures shown in Figure 3-6 have antennas that interface between circuits and the outside world. Antennas generally have a broad bandwidth, much greater than the bandwidth of an individual communication channel.

3.4.2 Heterodyne Transmitter and Receiver

First consider the transmitter architecture shown in Figure 3-6(a). In a transmitter, a low-frequency information-bearing signal is translated to a

frequency that can be more easily radiated. The information is contained in the baseband signal, which in many modern systems is generated as a digital signal within the digital signal processor (DSP). Then the **digital-to-analog converter**, (**DAC**), converts the digital baseband to an analog signal called the analog baseband, identified in Figure 3-6(a) as the intermediate frequency (IF) signal at the output of the DAC. Older systems generate the IF signal using analog hardware as this reduces the power consumed by the digital electronics. The IF then passes through a lowpass filter (LPF) to remove the harmonics resulting from the DAC process, and the signal is then amplified before being applied to the mixer. There are several types of mixers, but the central concept is multiplying the IF signal by a much larger local oscillator (LO) signal at frequency f_{LO}. The LO will be a single cosinusoid and if its amplitude is A_{LO}, then the LO signal is $x_{\mathrm{LO}} = A_{\mathrm{LO}} \cos(2\pi f_{\mathrm{LO}})$. While the IF signal will have a finite bandwidth, the operation of the mixer can be illustrated by considering that the IF is a single cosinusoid with amplitude A_{IF} and frequency f_{IF}, that is the RF signal is $x_{\mathrm{IF}} = A_{\mathrm{IF}} \cos(2\pi f_{\mathrm{IF}})$. Then the output of the mixer is

$$\begin{aligned} x_{\mathrm{RF}} &= x_{\mathrm{IF}} \times x_{\mathrm{LO}} = A_{\mathrm{IF}} A_{\mathrm{LO}} \cos(2\pi f_{\mathrm{IF}}) \cos(2\pi f_{\mathrm{LO}}) \\ &= \tfrac{1}{2} A_{\mathrm{RF}} A_{\mathrm{LO}} \left\{ \cos[2\pi(f_{\mathrm{LO}} - f_{\mathrm{IF}})] + \cos[2\pi(f_{\mathrm{LO}} + f_{\mathrm{IF}})] \right\}. \end{aligned} \tag{3.2}$$

In Equation (3.2) a trigonometric expansion has been used. The LO is chosen so that its frequency is close to that of the desired RF so that multiplication by the mixer results in an output that has one component at $f_{\mathrm{RF}\Delta} = f_{\mathrm{LO}} - f_{\mathrm{IF}}$ and one at $f_{\mathrm{RF}\Sigma} = f_{\mathrm{LO}} + f_{\mathrm{IF}}$. One of these is selected by the BPF, is further amplified, and then delivered to the antenna and radiated.

3.4.3 Superheterodyne Receiver Architecture

The first receiver architecture to be considered is the superheterodyne (or superhet) receiver architecture shown in Figure 3-6(b). Heterodyning refers to the use of a mixer and a superheterodyne circuit has two mixers. The antenna collects information from the environment at RF, and immediately this is bandpass filtered (by the BPF_1 block) to eliminate most of the interfering signals and noise. Thus the first BPF reduces the range of voltages presented to the first amplifier and so reduces the chance that the amplifier will distort the desired signal.

The RF at the output of the leftmost bandpass filter, BPF_1, still has a spectrum that is much broader than that of the desired communication signal. For example, in 3G radio the communication channel is 5 MHz wide but the first BPF could be 50 MHz wide and the RF could be 1 GHz or 2 GHz. So it is still necessary to use additional frequency selectivity to isolate the single required channel. The optimum choice at this stage is to follow the first BPF with an amplifier to boost the level of the signal. This also boosts the level of the noise that was captured by the antenna along with the signal, but it means that the noise added by the circuitry after the first amplifier has much less importance. The next block in the receiver is the mixer that shifts the information down in frequency to the first intermediate frequency, IF_1. The local oscillator, LO_1, is chosen so that its frequency is close to that of the RF so that multiplication by the mixer results in a low frequency at the output (i.e., at $f_{\mathrm{IF1}} = f_{\mathrm{LO}} - f_{\mathrm{RF}}$), and one at a frequency nearly twice that of the LO (i.e., at the sum frequency, $f_{\Sigma} = f_{\mathrm{LO}} + f_{\mathrm{RF}}$). The center frequency and

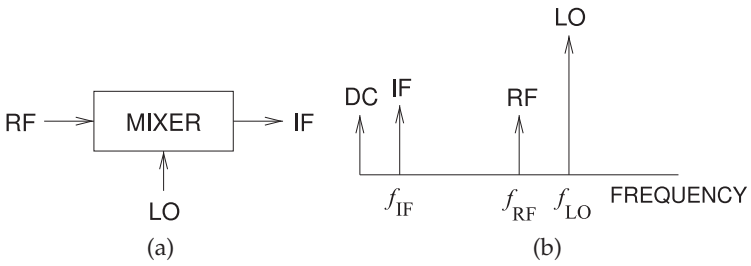


Figure 3-8: Simple mixer circuit: (a) block diagram; and (b) spectrum.

bandwidth of the second bandpass filter, BPF_2, is chosen so that only the signals around f_{IF1} pass through. Thus the main function of the first mixer stage and BPF_2 in the superheterodyne receiver is to convert the information at the RF down to a lower frequency, here at IF_1. The operation of the mixer is shown in Figure 3-8.

In the superheterodyne receiver architecture the output of the first mixer, f_{IF1}, is still at too high a frequency for the signal to be directly converted into digital form where it can be processed. So it is natural to ask why the frequency translation was not all the way down to baseband. The main reason for this is that there is substantial noise on an LO at frequencies very close to the oscillation frequency, f_{LO1}. This noise drops off quickly away from the oscillation frequency and its level at frequency f is proportional to $(|f - f_{\mathrm{LO1}}|)^n, n = 1, 2, \ldots$. This noise will appear at the output of the mixer and will be substantial if the RF and LO are very close in frequency. So the optimum trade-off is to shift the frequency of the information-bearing signal in two stages. Following further amplification, the second stage of mixing converts the information-bearing component of the signal centered at the first intermediate frequency, IF_1, to the second intermediate frequency, IF_2, which is usually at the baseband frequency or slightly above it. The baseband signal is now analog and this is converted to digital form by the ADC, and then the signal, now the digital baseband signal, can be digitally processed by the DSP.

Since the second mixer operates at much lower frequencies than the first mixer, the phase noise on LO_2 does not overlap the signal at IF_2. The reason why this architecture is called a superheterodyne receiver architecture is because when this architecture was first used, IF_2 was an audio signal and IF_1 was above the audible range and so was a supersonic signal. Thus the super in superheterodyne initially referred to the supersonic IF.

3.4.4 Single Heterodyne Receiver

The second receiver architecture shown in Figure 3-6(c) has a single heterodyne or mixing stage. If the LO has very low noise the IF will be at a lower frequency than in the superheterodyne architecture shown in Figure 3-6(b). A high-performance ADC is then required to convert the signal to digital form and deliver it to the DSP unit. Substantial digital processing power is required to translate the signal to a digital baseband signal. Alternatively a high-performance subsampling ADC could be used that in effect performs mixing during conversion. The advantage of this architecture is a simplified RF section and is of particular advantage when multiple RF communication standards are supported as the DSP and ADC can be common while the analog RF hardware for each band is considerably simplified.

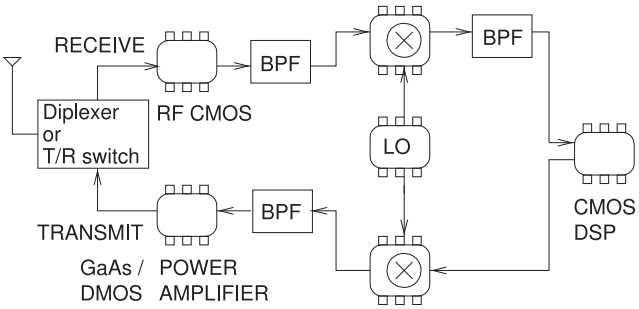


Figure 3-9: RF front end organized as multiple chips. This corresponds to a combination of the receive architecture shown in Figure 3-6(c) and the transmitter architecture shown in Figure 3-6(a).

3.4.5 Transceiver

The major transistor- or diode-based **active elements** in the RF **front end** of both the transmitter and receiver are the **amplifiers**, **mixers**, and **oscillators**. These subsystems have much in common, using nonlinear devices to convert power at DC to power at RF. In the case of mixers, power at the LO is also converted to power at RF. The front end of a typical cellphone is shown in Figure 3-9. The components here are generally implemented in a module and use different technologies for the various elements, optimizing cost and performance.

Return now to the mixer-based **transceiver** (for transmitter and receiver) architecture shown in Figure 3-9. Here, a single antenna is used, and either a **diplexer**[3] (a combined lowpass and highpass filter) or a switch is used to separate the (frequency-spaced) transmit and receive paths. If the system protocol requires transmit and receive operations at the same time, a diplexer is required to separate the transmit and receive paths. This filter tends to be large, lossy, or costly (depending on the technology used). Consequently a transistor **switch** is preferred if the transmit and receive signals operate in different time slots. In the receive path the low-level received signal is amplified and the initial amplifier needs to have very good noise performance. Once the signal is larger the noise performance of the amplifiers is less critical. The initial amplifier is called a **low-noise amplifier** (**LNA**). The amplified receive signal is then bandpass filtered and frequency down-converted by a mixer to IF that can be sampled by an ADC to produce a digital signal that is further processed by DSP. Variants of this architecture include one that has two mixing stages (as in the superheterodyne receiver shown in Figure 3-6(b)), and another with no mixing that relies instead on direct conversion of the receive signal using, as one possibility, a subsampling ADC. In the transmit path, the architecture is reversed, with a DAC driven by the DSP chip that produces an information-bearing signal at the IF which is then frequency up-converted by a mixer, bandpass filtered, and amplified by what is called a power amplifier to generate the tens to hundreds of milliwatts required.

[3] A duplexer separates transmitted and received signals and is often implemented as a filter called a diplexer. A **diplexer** separates two signals on different frequency bands so that they can use a common element such as an antenna. If the transmitted and received signals are in different time slots, the duplexer can be a switch.

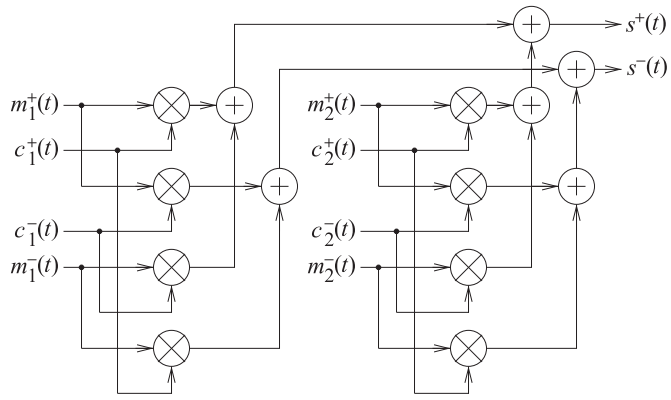


Figure 3-10: Differential implementation of a Hartley SSB-SC modulator: $s^+(t)$, $s^-(t)$ are the positive and negative components of the differential modulated signal $s(t)$; $(c_1^+(t),\ c_1^-(t))$ is the differential carrier; $(m_1^+(t),\ m_1^-(t))$ is the differential quadrature-modulated IF carrier; $(c_2^+(t),\ c_2^-(t))$ is the differential carrier shifted $90°$; and $m_2^+(t)$ and $m_2^-(t)$ are the $90°$ phase-shifted versions of $m_1^+(t)$ and $m_1^-(t)$, respectively.

3.4.6 Hartley Modulator

The Hartley modulator [1, 2], shown in Figure 3-1, results in SSB**single-sideband (SSB) modulation** or more precisely SSB **suppressed-carrier (SSB-SC) modulation**. This is one of the great inventions and variants of this circuit are used in all modern radios. In the Hartley modulator the modulating signal $m(t)$ and the carrier are multiplied together in a mixer and then also $90°$ phase-shifted versions are also mixed before being added together. The signal flow is as follows beginning with $m(t) = \cos(\omega_{m1}t)$, $p(t) = \cos(\omega_{m1}t - \pi/2) = \sin(\omega_{m1}t)$ and carrier signal $c_1(t) = \cos(\omega_c t)$:

$$\begin{aligned} a_1(t) &= \cos(\omega_{m1}t)\cos(\omega_c t) = \tfrac{1}{2}[\cos((\omega_c - \omega_{m1})t) + \cos((\omega_c + \omega_{m1})t)] \\ b_1(t) &= \sin(\omega_{m1}t)\sin(\omega_c t) = \tfrac{1}{2}[\cos((\omega_c - \omega_{m1})t) - \cos((\omega_c + \omega_{m1})t)] \\ s_1(t) &= a_1(t) + b_1(t) = \cos((\omega_c - \omega_{m1})t) \end{aligned} \tag{3.3}$$

and so the lower sideband (LSB) is selected. That is, if a finite bandwidth modulating signal $m(t)$ was mixed only once with the carrier $c_1(t)$, the spectrum of the output $a_1(t)$ would include upper and lower sidebands as well as the carrier as shown in Figure 3-2(a). With the Hartley modulator the spectrum of Figure 3-2(b) is obtained.

Every frequency component of $m(t)$ needs to be shifted by $90°$ and this can be done using a polyphase circuit or digitally using a Hilbert transform. To select the upper sideband the summer in Figure 3-1 is replaced by a subtraction block and then the output becomes $s'(t) = a_1(t) - b_1(t) = \cos((\omega_c + \omega_{m1})t)$.

Another circuit that implements SSB-SC modulation is the **Weaver modulator** [7]. It uses only lowpass filters and mixers and is often used to implement SSB-SC in a digital-signal processor.

3.4.7 The Hartley Modulator in Modern Radios

In RF engineering, especially when RFICs are used, it is common to implement the phase shift of the modulating signal digitally using the Hilbert transform and to use differential signals. Then the Hartley SSB-SC modulator shown in Figure 3-1 is implemented as shown in Figure 3-10 where $(m_1^+(t), m_1^-(t))$ is the differential form of the analog signal corresponding to the modulated signal $s(t)$ at the output of the generic quadrature modulator in Figure 3-11. With the large number of modulation formats supported in modern cellular communication standards it is indeed

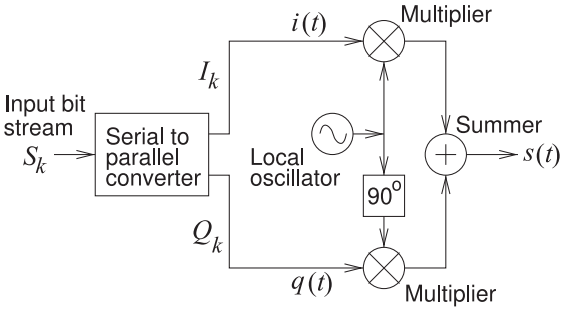


Figure 3-11: Quadrature modulator block diagram. An input bitstream, S_k, is divided into two bitstreams, I_k and Q_k, which are applied to the multipliers as (possibly filtered) waveforms $i(t)$ and $q(t)$. The output of the multipliers (appropriately filtered) are summed to yield a modulated carrier signal, $s(t)$. The $90°$ block shifts the carrier by $90°$.

fortunate that the quadrature modulators can be implemented digitally followed possibly by wave-shaping. The differential signal $(m_2^+(t), m_2^-(t))$ is the phase-shifted form of $(m_1^+(t), m_1^-(t))$ in which every frequency component is shifted by $90°$ usually implemented as the Hilbert transform of the digital forms of $(m_1^+(t), m_1^-(t))$ (before wave-shaping).

3.5 Carrier Recovery

Demodulation requires the generation of a local version of the carrier of a received radio signal. This is simple with a DSB RF signal as the carrier is sent with the radio signal. DSB FM is a little different as strictly the carrier is not sent in the radio signal but the average frequency of the received signal is the carrier frequency. With all SSB schemes it is essential to create the carrier in a process called carrier recovery. This section considered carrier recovery for digitally modulated signals.

In modern radios carrier recovery is done in two stages. First a crude estimate of the carrier is generated and used to crudely demodulate the received signal. The crudely demodulated signal is then sampled to yield a digital demodulated signal. Then, for 2G and 3G radio, a fine carrier recovery procedure using a digital implementation of what has been described is used. With 4G and 5G radios the crude carrier recovery procedure is used then pilot tones transmitted along with the signal are used to develop a low-frequency version of the carrier

3.6 Modern Transmitter Architectures

Modern transmitters maximize both spectral efficiency and electrical efficiency by using quadrature modulation and suppressing the carrier in the radio signal. Electrical efficiency must be achieved together with tight specifications on allowable distortion, and designs must achieve this with minimum manual adjustments. The discussion in this section focuses on narrowband communications when the modulated RF carrier can be considered as a slowly varying RF phasor. Modern radios must implement various order modulation schemes and also must implement legacy modulation methods.

3.6.1 Quadrature Modulator

Most digital modulation schemes set the amplitude and phase of a carrier. The one exception is frequency shift keying (FSK) modulation, which changes the frequency of the carrier and has much in common with FM.

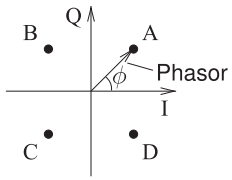


Figure 3-12: Phasor diagram with four discrete phase states that can be set by combining I and Q signals.

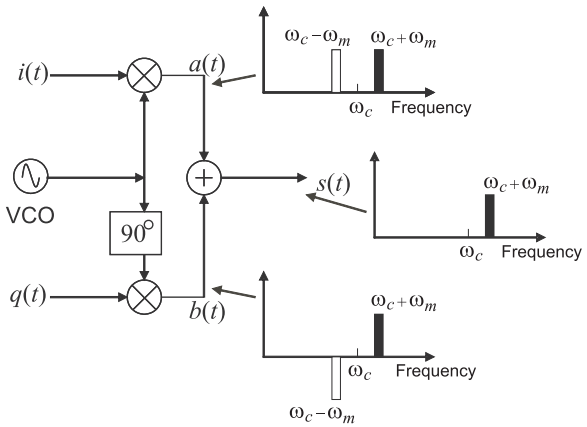


Figure 3-13: Quadrature modulator showing intermediate spectra.

So, in most digital modulation schemes, setting the amplitude and phase of a carrier addresses a point on a phasor diagram. A phasor diagram with four discrete states is shown in Figure 3-12. The circuit that implements this digital modulation is shown in Figure 3-11.

3.6.2 Quadrature Modulation

Quadrature modulation describes the frequency conversion process in which the real and imaginary parts of the RF phasor are varied separately. A subsystem that implements quadrature modulation is shown in Figure 3-13. This is quite an ingenious circuit. The operation of this subsystem is described by what is known as the generalized quadrature modulation equation:

$$s(t) = i(t)\cos\left[\omega_c t + \varphi_i(t)\right] + q(t)\sin\left[\omega_c t + \varphi_q(t)\right], \tag{3.4}$$

where, $i(t)$ and $q(t)$ embody the particular modulation rule for amplitude, $\varphi_i(t)$ and $\varphi_q(t)$ embody the particular modulation rule for phase, and ω_c is the carrier radian frequency. In terms of the signals identified in Figure 3-13, the quadrature modulation equation can be written as

$$s(t) = a(t) + b(t) \tag{3.5}$$
$$a(t) = i(t)\cos\left[\omega_c t + \varphi_i(t)\right] \tag{3.6}$$
$$b(t) = q(t)\sin\left[\omega_c t + \varphi_q(t)\right]. \tag{3.7}$$

Figure 3-13 shows that both $a(t)$ and $b(t)$ have two bands, one above and one below the frequency of the carrier, ω_c. However, there is a difference. The LO (here designated as the VCO) is shifted 90° (perhaps using an RC delay line) so that the frequency components of $b(t)$ have a different phase relationship to the carrier than those of $a(t)$. When $a(t)$ and $b(t)$ are combined, the carrier content is canceled, as is one of the sidebands provided that $q(t)$ is a 90° phase-shifted version of $i(t)$. This is exactly what is desired:

the carrier should not be transmitted, as it contains no information. Also, it is desirable to transmit only one sideband, as it contains all of the information in the modulating signal. This type of modulation is SSB-SC modulation. In the next section, frequency modulation is used to demonstrate SSB-SC operation.

3.6.3 Frequency Modulation

Frequency modulation is considered here to demonstrate SSB-SC operation. Let $i(t)$ and $q(t)$ be finite bandwidth signals centered at radian frequency ω_m with their phases $\phi_i(t)$ and $\phi_q(t)$ chosen so that $(\varphi_q(t) - \varphi_i(t))$ is 90° on average. This is shown in Figure 3-13, where ω_m represents the frequency components of $i(t)$ and $q(t)$. With reference to Figure 3-13,

$$i(t) = \cos(\omega_m t) \quad \text{and} \quad q(t) = -\sin(\omega_m t), \tag{3.8}$$

and the general quadrature modulation equation, Equation (3.4), becomes

$$s(t) = i(t)\cos(\omega_c t) + q(t)\sin(\omega_c t) = a(t) + b(t), \tag{3.9}$$

where
$$a(t) = \frac{1}{2}\{\cos[(\omega_c - \omega_m)t] + \cos[(\omega_c + \omega_m)t]\} \tag{3.10}$$

and
$$b(t) = \frac{1}{2}\{\cos[(\omega_c + \omega_m)t] - \cos[(\omega_c - \omega_m)t]\}. \tag{3.11}$$

Thus the combined frequency modulated signal at the output is

$$s(t) = a(t) + b(t) = \cos[(\omega_c + \omega_m)t], \tag{3.12}$$

and the carrier and lower sideband are both suppressed. The lower sideband, $\cos[(\omega_c - \omega_m)t]$, is also referred to as the image which may not be exactly zero because of circuit imperfections. In modulators it is important to suppress this image, and in demodulators it is important that undesired signals at the image frequency not be converted along with the desired signals.

3.6.4 Polar Modulation

In polar modulation, the $i(t)$ and $q(t)$ quadrature signals are converted to polar form as amplitude $A(t)$ and phase $\phi(t)$ components. This is either done in the DSP unit or, if a modulated RF carrier is all that is provided, using an envelope detector to extract $A(t)$ and a limiter to extract the phase information corresponding to $\phi(t)$. Two polar modulator architectures are shown in Figure 3-14. In the first architecture, Figure 3-14(a), $A(t)$ and $\phi(t)$ are available and $A(t)$ is used to amplitude modulate the RF carrier, which is then amplified by a power amplifier (PA). The phase signal, $\phi(t)$, is the input to a phase modulator implemented as a PLL. The output of the PLL is fed to an efficient amplifier operating near saturation (also called a saturating amplifier). The outputs of the two amplifiers are combined to obtain the large modulated RF signal to be transmitted.

In the second polar modulation architecture, Figure 3-14(b), a low-power modulated RF signal is decomposed into its amplitude- and phase-modulated components. The phase component, $\phi(t)$, is extracted using a limiter that produces a pulse-like waveform with the same zero crossings as the modulated RF signal. Thus the phase of the RF signal is captured. This

is then fed to a saturating amplifier whose gain is controlled by the carrier envelope, or $A(t)$. Specifically, $A(t)$ is extracted using an envelope detector, with a simple implementation being a rectifier followed by a lowpass filter with a corner frequency equal to the bandwidth of the modulation. $A(t)$ then drives a switching (and hence efficient) power supply that drives the **saturating power amplifier**.

3.7 Modern Receiver Architectures

This section discusses transmitter and receiver architectures in the generations before software defined radio (as used in 4G and 5G). These are architectures that could be implemented in analog hardware.

3.7.1 Receiver Architectures

It is more challenging to achieve a high performance for a receiver than it is for a transmitter.

Communication receivers most commonly use mixing of the RF signal with a fixed signal called an LO to produce a lower-frequency replica of the modulated RF signal. Some receiver architectures use one stage of mixing, while others use two stages. In cellular systems, the receiver must be sensitive enough to detect signals of 100 fW or less.

Some of the architectures used in modern receivers are shown in Figure 3-15. Figure 3-15(a) is the superheterodyne architecture in much the same form that it has been used for a century. Key features of this architecture are that there are two stages of mixing, and filtering is required to suppress spurious mixing products. Each mixing stage has its own VCO. The receiver progressively reduces the frequency of the information-bearing signal. The image rejection mixer in the dashed box achieves rejection of the image frequency to produce an IF (or baseband frequency) that can be directly sampled. However, it is difficult to achieve the required amplitude and phase balance. Instead, the architecture shown in Figure 3-15(b) is sometimes used. The filter between the two mixers can be quite large. For example, if the

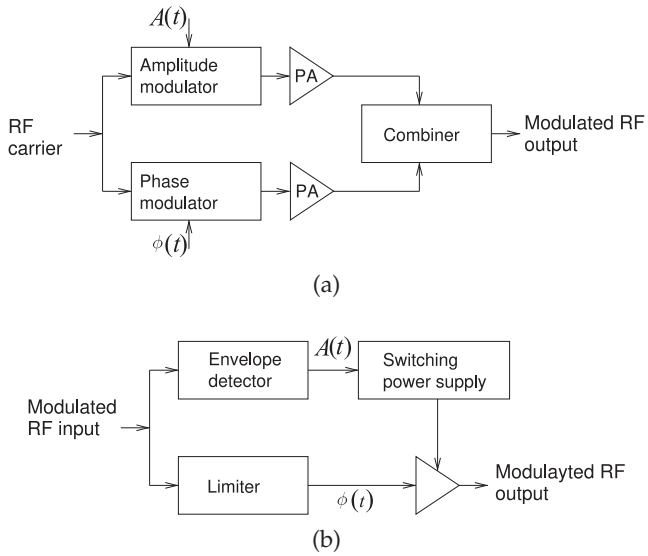
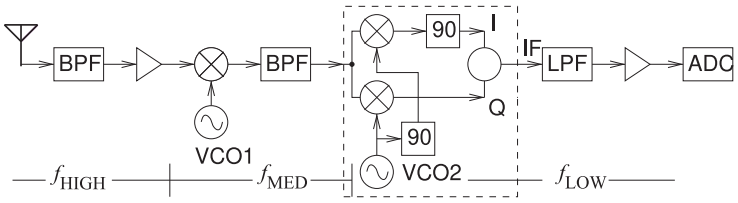
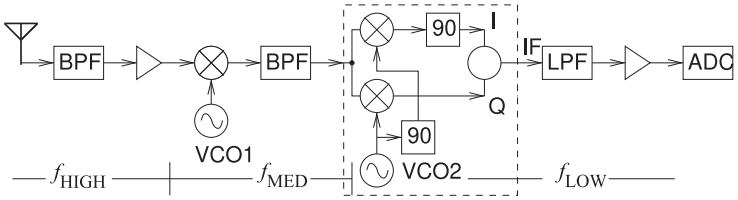


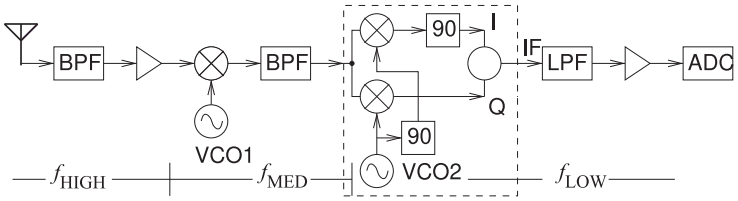
Figure 3-14: Polar modulator architectures: (a) amplitude- and phase-modulated components amplified separately and combined; and (b) the amplitude used to modulate a power supply driving a saturating amplifier with phase modulated input.



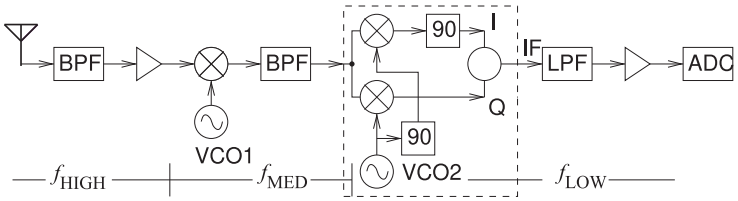
(a) Superheterodyne, Hartley image rejection receiver



(b) Superheterodyne receiver



(c) Dual conversion receiver



(d) Low-IF or zero-IF receiver

Figure 3-15: Architectures of modern receivers: (a) superheterodyne receiver using the **Hartley** architecture for image rejection; (b) superheterodyne receiver; (c) dual-conversion receiver; and (d) low-IF or zero-IF receiver. BPF, bandpass filter; LPF, lowpass filter; ADC, analog-to-digital converter; VCO, voltage-controlled oscillator; 90, 90° phase shifter; f_{HIGH}, f_{MED}, and f_{LOW} indicate relatively high-, medium-, and low-frequency sections.

incoming signal is 1 GHz, the frequency of the signal after the first mixer could be 100 MHz.

Filters are smaller and have higher performance at higher frequencies. This is exploited in the dual-conversion receiver shown in Figure 3-15(c). This is similar to the traditional superheterodyne architecture except that the IF between the two mixers is high. For example, if the incoming signal is 1 GHz, the output of the first mixer could be 3 GHz. This architecture also enables broad radio operation with the band selected by choosing the frequencies of the two local oscillators.

The low-IF or zero-IF receiver shown in Figure 3-15(d) uses less hardware and is common in less demanding communication applications. In high-performance systems, such as the cellular phone system, this architecture requires more design time as well as calibration circuitry to trim the I and Q paths so that they are closely matched.

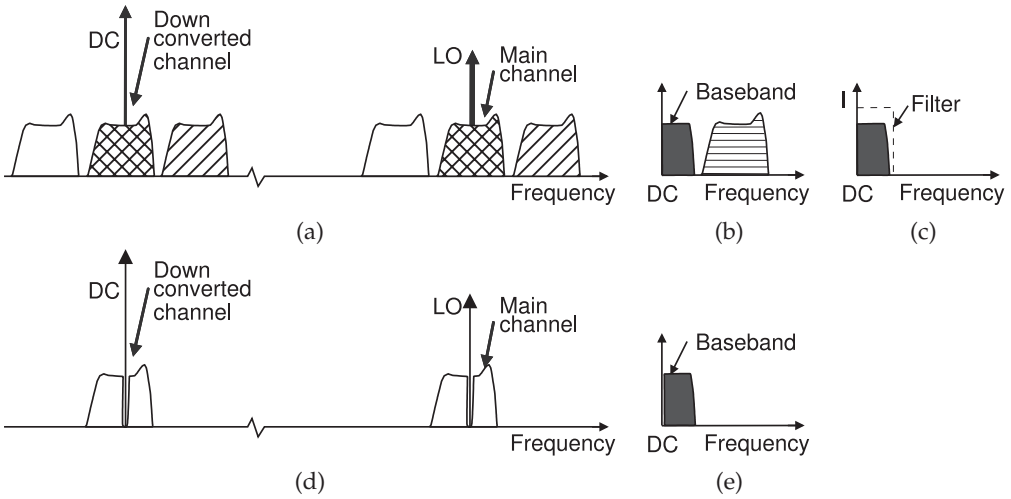
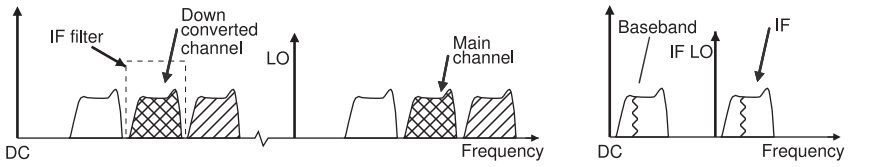


Figure 3-16: Frequency conversion using homodyne mixing: (a) the spectrum with a large LO and the low frequency products after down conversion; (b) the baseband spectrum showing only positive frequencies; (c) the baseband spectrum after mixing; (d) down conversion spectra when the radio signal has no spectral content at the carrier frequency; and (e) the lowpass filtered down-converted signal in (d). .

3.7.2 Homodyne Frequency Conversion

Homodyne mixing and detection is one of the earliest wireless receiver technologies and is used in AM radio. In homodyne mixing, the carrier of a modulated signal is regenerated and synchronized in phase with the incoming carrier frequency. Mixing the carrier with the RF signal results in an IF signal centered around zero frequency.

The signal spectra in homodyne mixing is shown in Figure 3-16. In Figure 3-16(a), the RF signals are shown on the right-hand side and the baseband signals are shown on the left-hand side. It is usual to show both positive and negative frequencies at the lower frequencies so that the conversion process is more easily illustrated. Of course negative frequencies do not exist. The characteristic of homodyne mixing is that the LO corresponds to the carrier and is in the middle of the desired RF channel. RF signal components mix with the LO, and it appears that the entire RF spectrum is down-shifted around DC. Of course, the actual baseband spectrum is only defined for positive frequencies, so the negative-frequency baseband signals and the positive-frequency baseband signals combine to yield the detected baseband spectrum shown in Figure 3-16(b). With other modulation schemes, this possible loss of information is avoided using quadrature demodulation. An amplitude modulated signal has identical modulation sidebands, so the collapsing of positive and what is shown as negative frequencies at baseband results in no loss of information. Then a simple amplitude detection circuitry, such as a rectifier, is used and the rectified signal was (typically) passed directly to a speaker.



(a) Conversion from RF to IF (b) Conversion to baseband

Figure 3-17: Frequency conversion using superheterodyne mixing.

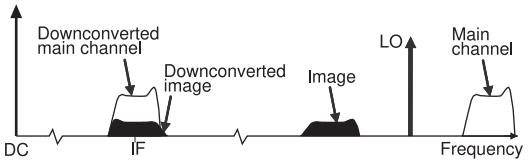


Figure 3-18: Frequency conversion using heterodyne mixing showing the effect of image distortion with the down-converted image overlapping the down-converted main channel.

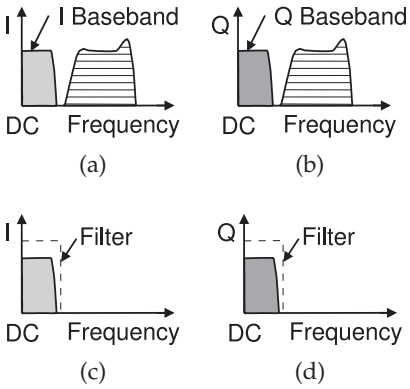


Figure 3-19: Frequency conversion using direct conversion quadrature mixing: (a) the baseband spectrum at the I output of the receiver; (b) the baseband spectrum at the Q output of the receiver; and (c and d) the spectrum of the I and Q channels following.

3.7.3 Heterodyne Frequency Conversion

In heterodyne mixing, the locally generated LO and the main RF channel are separated in frequency, as shown in Figure 3-17(a). In this figure the RF signals (shown as three discrete channels on the right-hand side of the spectrum) mix with the LO to produce signals at a lower frequency. This lower frequency is usually not the final baseband frequency, and so is called the intermediate frequency (IF). The IF of the main channel is at the difference frequency of the RF signal and the LO. There are several important refinements to this. The first of these is concerned with limiting the number of signals that can mix with the LO. This is done using an RF preselect filter. To see the difficulties introduced by the image channel, consider the frequency conversion to an intermediate frequency described in Figure 3-18. Filtering reduces the level of the image channel as shown in Figure 3-18(a). Note that the main channel and its image are equidistant from the LO, see Figure 3-18. Both down-convert to the same IF frequency. In the worst-case scenario, the IF image could be larger than that of the desired channel.

3.7.4 Direct Conversion Receiver

Zero-IF direct conversion receivers are similar to quadrature homodyne receivers in that the LO is placed near the center of the RF channel The important characteristic of direct conversion receivers is that there is only one level of mixing. The conversion process is described in Figure 3-19. A particular advantage of direct conversion is that the relatively large IF filters are eliminated. They are invariably implemented as quadrature

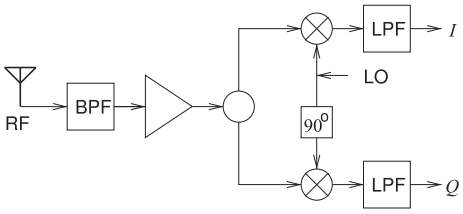


Figure 3-20: Direct conversion quadrature demodulator.

demodulators, see Figure 3-20. used in cellular phones as it uses uses little DC power, and hence extends battery life, and is compatible with monolithic ICs. Direct conversion is now the preferred method of down conversion

The main nonideality of this design is the DC offset in the down-converted spectrum. DC offset results mostly from self-mixing, or rectification, of the LO. This DC offset can be much larger than the down-converted signal itself. One way of coping with the DC offset is to highpass filter the down-converted signal, but highpass filtering requires a large passive component (e.g., a series capacitor), at least to avoid the dynamic range problems of active filters. Highpass filtering the down-converted signal necessarily throws away information in the signal spectrum, and it is only satisfactory to do this if there is very little information around DC to begin with.

The primary design effort with zero-IF converters is overcoming the DC offset problem, and to a lesser extent coping with the jitter of the LO. However there is a scheme used in 4G and 5G that overcomes these limitations. This will be discussed in following chapters.

3.7.5 Low-IF Receiver

In a low-IF receiver, single-stage heterodyne mixing is used to down-convert the modulated RF carrier to a frequency just above DC, perhaps a few hundred kilohertz or a few megahertz, depending on the bandwidth of the RF channel. In doing this, the DC offset problem of a direct conversion receiver is avoided. This frequency offset can be just a few hundred hertz to be effective. Low IF conversion is now the preferred method of down-conversion in cellular phones as it requires very little batter usage, does not require large IF filters, and is compatible with monolithic ICs. Sometimes low-IF conversion is referred to as direct down conversion but there is a subtle difference.

3.7.6 Subsampling Analog-to-Digital Conversion

Subsampling receivers overcome the DC offset problem typical of other direct conversion receivers. The idea is to sample the modulated RF signal at a subharmonic of the carrier of the RF signal to be converted. The sampling rate must be at least twice the bandwidth of the baseband signal and the track-mode bandwidth must be greater than the carrier frequency. Thus the sampling aperture is the critical parameter and must be several times smaller than the period of the carrier. Fortunately the aperture times of CMOS tracking circuits are adequate. It is critical that an RF preselect filter be used to eliminate unwanted interferers and noise outside the communication band. Aliasing of signals outside the **Nyquist** bandwidth onto the baseband signal is a consequence of subsampling. Adjacent channel signals are converted without aliasing, but these will lie outside the bandwidth of the

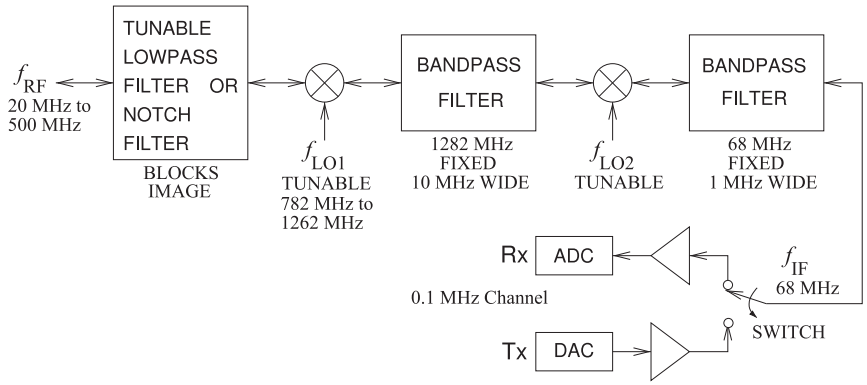


Figure 3-21: Bilateral double-conversion transceiver for wideband operation of an emergency or military radio.

baseband signal. Flicker noise on the sampling clock is multiplied by the subsampling ratio and appears as additional noise at baseband. This was at one time a very attractive option but has been out-performed by the low-IF receiver.

3.7.7 First IF-to-Baseband Conversion

In a superheterodyne conversion architecture there are two heterodyne stages, with the IF of the first stage in modern systems in the range of 20 to 200 MHz. The assignment of frequencies is known as frequency planning, and this is treated as proprietary by the major radio vendors. This IF is then converted to a much lower IF, typically around 100 kHz to a few megahertz above the center frequency of the baseband signal. This frequency is generally called baseband, but strictly it is not because the signal is still offset in frequency from DC. Some direct conversion architectures leave the first heterodyne mixing stage in place and use direct conversion of the first IF to baseband (true baseband—around DC).

3.7.8 Bilateral Double-Conversion Receiver

The receivers considered so far are suitable for narrowband communications typical of point-to-point and consumer mobile radio. There are many situations where the range of received or transmitted RF signals covers a very wide bandwidth, such as with emergency radios, television, and military communications. Typically, however, the instantaneous bandwidth is small. If narrowband RF front-end architectures are used, a switchable filter bank would be required and this would result in an impractically large radio. One solution to covering very wide RF bandwidths is the double-conversion transceiver architecture shown in Figure 3-21. The frequency plan of a typical radio using 0.1 MHz channels between 20 MHz and 500 MHz is shown. The key feature of this radio is that bidirectional mixers are used. Following the RF chain from left to right, the RF is first mixed up in frequency, bandpass filtered using a high-Q distributed filter, and then down-converted to a lower frequency that can be sampled directly by an ADC. A much higher performance passive (and hence bidirectional) filter can be realized

at gigahertz frequencies than at a few tens of megahertz. On transmit, the function is similar, with the mixers and LO reused. As a receiver, the notch filter or lowpass filter is used to block the image frequency of the first mixer so that only the upper sideband IF is presented to the first bandpass filter. The lowpass or notch filter may be fixed, although, with the plan shown, there must be at least two states of the filters. On transmit, the lowpass or notch filters prevent the image frequency from being radiated.

3.8 Introduction to Software Defined Radio

The 2G and 3G cellular radio schemes used very few modulation methods. With the introduction of 4G and now 5G many modulation methods must be supported with the highest-order modulation scheme used determined by the level of noise and interference on a channel. This support for multiple modulation methods is only possible if most of the demodulation process is performed in DSP where software controls demodulation. Such a radio is called a software defined radio (SDR). In an SDR many of the functions that traditionally would be performed using analog hardware are instead implemented digitally and only the RF functions close to the transmit/receive antenna are implemented in analog form. An extreme concept for the transmitter side of the radio is to implement all of the functions digitally with a final digital-to-analog converter (DAC) connected to an antenna. Then the performance is limited by the maximum frequency and output power of the DAC. This ultimate SDR provides maximum flexibility, for example, it is easy to change modulation schemes, but would also contribute to very high battery drain. So instead there is a trade-off with many aspects such as modulation and fine tuning of the RF carrier frequency done digitally while, with a transmitter, the final up-conversion done using analog hardware which only needs to be able to support the bandwidth of the modulated signal but otherwise is agnostic to the type of modulation used.

All radios today use quadrature modulation, a type of digital modulation, in which the real (in-phase) and imaginary (quadrature-phase) components of the phasor of the carrier are varied independently. While it is possible to vary the amplitude and phase of a carrier tone separately this is not common.

Quadrature demodulation recovers the original signals that varied the in-phase and quadrature-phase components of the carrier. Demodulation can proceed in stages with an initial analog separation of RF in-phase and quadrature signals which are sampled by an analog-to-digital converter (ADC) and demodulation completed in a digital signal processor.

The following sections discuss various aspects of an SDR radio. Section 3.9 begins with a description of quadrature modulation in a way that helps in the understanding of an SDR transmitter. In Section 3.10 a specific example of an SDR transmitter is presented in which the time-domain and frequency-domain signals are followed through first a digital signal processor (DSP) and then an analog up-converter to produce an RF signal. This discussion is followed by a description of an SDR receiver with the general SDR quadrature demodulator in Section 3.11 and then a specific example in Section 3.12.

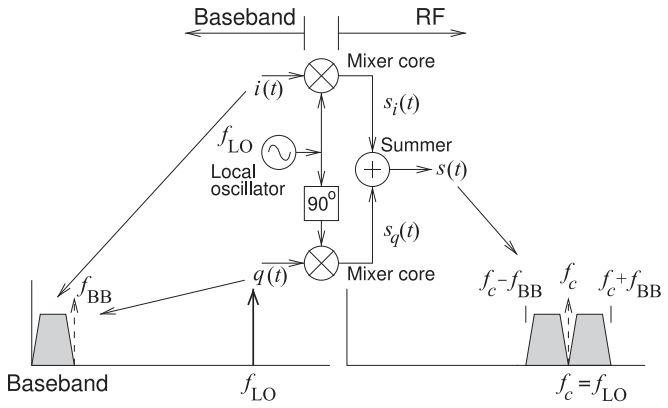


Figure 3-22: Double-sideband suppressed-carrier (DSB-SC) modulation. Quadrature modulator with independent modulating baseband signals $i(t)$, the in-phase component input to the in-phase mixer core, and $q(t)$, the quadrature component driving the quadrature-phase mixer core. This results in a modulated signal with two sidebands around the local oscillator or carrier frequency. Each RF sideband has a mix of the information in $i(t)$ and $q(t)$.

3.9 SDR Quadrature Modulator

The SDR transmitter uses two-stage modulation with DSB-SC modulation implemented in DSP to produce an IF signal which is output using DACs to produce analog I and Q channel IF signals. These IF signals are then input to an analog quadrature modulator implementing SSB-SC modulation with the resulting radio signal being a DSB-SC radio signal. As of the time of this writing the digital portion was implemented in what is called a baseband chip and the analog portion implemented in an RF modem chip. One can expect that eventually these would be combined into a single chip. As far as the RF modem chip is concerned, the IF signals input to the up-converter are baseband signals and this is how they are often referred to when the focus is on the RF modem chip.

Quadrature modulation, see Figure 3-22, comprises two mixer cores which are driven by a modulating in-phase component $i(t)$ and a modulating quadrature-phase component $q(t)$ where in-phase and quadrature phase refer to the phase of the local oscillator input to the mixer cores. Here $i(t)$ and $q(t)$ are baseband signals with a spectrum extending from (near) DC to f_{BB} and in today's radios they are produced internally in a DSP unit. The finite bandwidth $i(t)$ and $q(t)$ signals contain I-channel and Q-channel information respectively. The top mixer core is driven directly by the local oscillator and the other, the quadrature mixer core, is also driven by the local oscillator but now it is phase shifted by $90°$, i.e. it is in quadrature. This scheme produces double-sideband suppressed carrier DSB-SC modulation and $s(t)$ is the modulated output signal with each sideband having bandwidth f_{BB}. The block-level schematic illustrates the basic architecture of a quadrature modulator which is expanded if the signals are differential signals with additional variations according to whether the mixers are implemented as analog multipliers or as switches controlled by the LO. The whole structure shown is referred to as a mixer and each mixer core on its own is also often referred to as a mixer. This operation can be implemented without error in DSP.

3.9.1 Analog Quadrature Modulator

The second stage of an SDR transmitter implements DSB-SC modulation using analog circuitry producing an RF signal.

An analog quadrature modulator using multipliers is shown in Figure

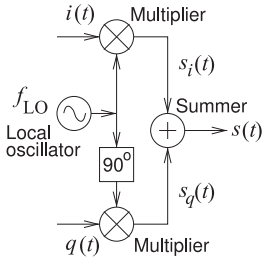
3-23(a) and consists of two multipliers each of which has two inputs and one output with the outputs summed yielding a modulated output signal $s(t)$. One particular characteristic of a quadrature modulator is that the LO at frequency $f_{\rm LO}$ is directly input to one of the multipliers but the second multiplier is driven by a version of the LO with a 90° phase lag, i.e. the LO input to the first mixer is in-phase and the LO input to the second multiplier has quadrature-phase (the phase is shifted by 90°). This second LO is also called the quadrature LO. The 90° phase difference of the two LOs is where the quadrature in quadrature modulator comes from. So if the LO is $\sin(2\pi f_{\rm LO})$, $i(t)$ is multiplied by $\sin(2\pi f_{\rm LO})$. Then $q(t)$ is multiplied by $\sin(2\pi f_{\rm LO} - \pi/2) = -\cos(2\pi f_{\rm LO})$. The second inputs of the multipliers in Figure 3-23(a) are the signals $i(t)$ and $q(t)$ with i indicating that the signal is driving the in-phase multiplier and q indicating that the signal is driving the quadrature-phase multiplier. The signals $i(t)$ and $q(t)$ may be independent, or the frequency components of $q(t)$ may phase lag $i(t)$ by 90° but otherwise be the same as $i(t)$. These two options yield modulated output signals with different bandwidths.

Transistor-Based Multiplier

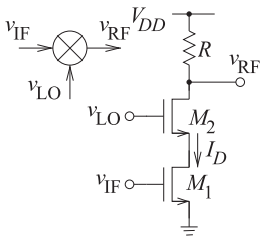
There are several ways to implement the mixer core in Figure 3-23(a) with the most common being as a multiplier or as a switch. Both can be conveniently implemented using transistors. The analog multiplier shown in Figure 3-23(b) is based on a cascode amplifier with one input applied to the gate of transistor $\rm M_1$. Instead of the gate of $\rm M_2$ being held at a DC voltage as with a cascode amplifier, the gate of $\rm M_2$ is also an input. Approximately, the drain current, I_D, of $\rm M_1$ is proportional to the gate voltage $v_{\rm IF}$ and the voltage gain of $\rm M_2$, i.e. $v_{\rm RF}/v_{\rm LO}$ is proportional to I_D. Thus the RF output voltage $v_{\rm RF}$ is proportional to the product of $v_{\rm IF}$ (which in the modulator is either $i(t)$ or $q(t)$) and $v_{\rm LO}$. So when $v_{\rm IF}$ and $v_{\rm LO}$ are sinewaves the output $v_{\rm LO}$ will be the trigonometric expansion of the product of two sinewaves and this product will also comprise two sinewaves at the sum and difference frequencies. Then circuit symmetry is used to select just one of these.

Polyphase Filter

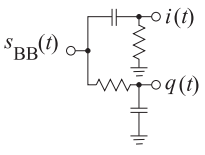
A polyphase filter, such as the one-stage polyphase filter in in Figure 3-23(c), takes an input analog input signal and outputs two signals that are the same except that the frequency components are shifted by 90°. This circuit can be used to produce the quadrature LO signal or to shift the frequency components of a baseband signal. More commonly a polyphase filter is



(a) Modulator core

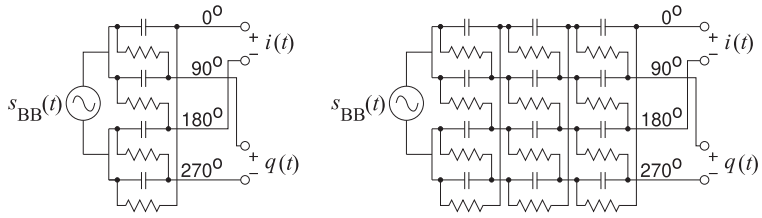


(b) Simple multiplier



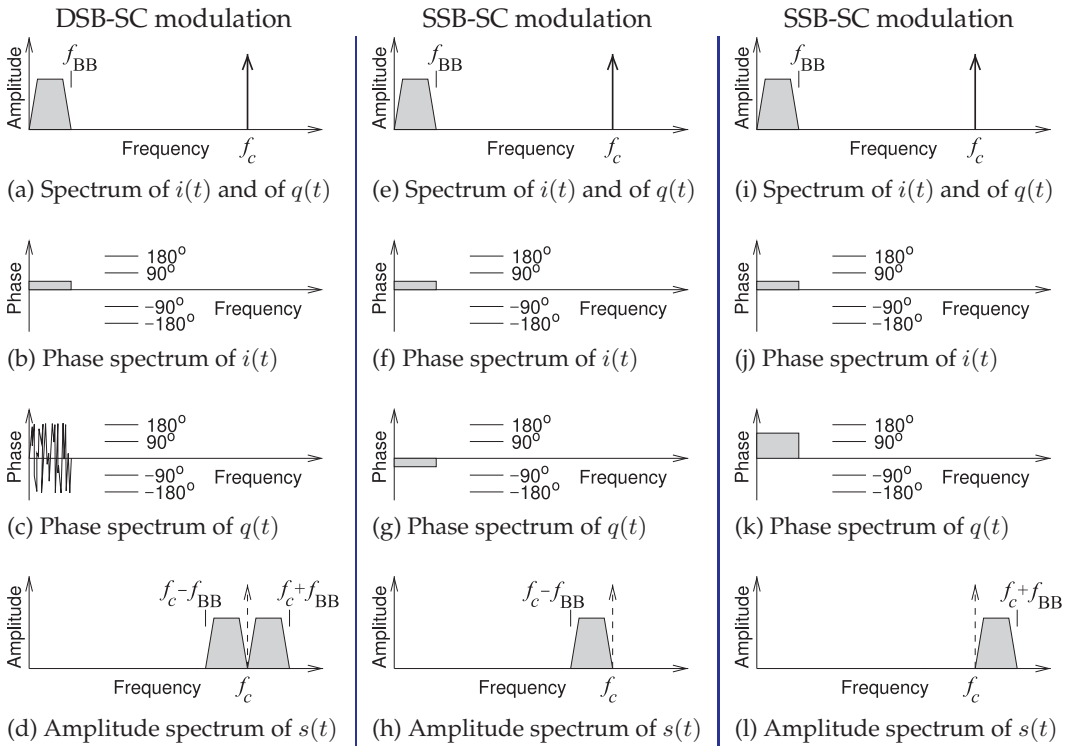
(c) Polyphase filter

Figure 3-23: Quadrature modulator. The simple modulator in (b) is based on a FET cascode amplifier and the polyphase filter in (c) has a 90° phase difference between $i(t)$ and $q(t)$.



(a) One-stage (b) Three-stage

Figure 3-24: Differential polyphase filters.



(a) Spectrum of $i(t)$ and of $q(t)$ (e) Spectrum of $i(t)$ and of $q(t)$ (i) Spectrum of $i(t)$ and of $q(t)$

(b) Phase spectrum of $i(t)$ (f) Phase spectrum of $i(t)$ (j) Phase spectrum of $i(t)$

(c) Phase spectrum of $q(t)$ (g) Phase spectrum of $q(t)$ (k) Phase spectrum of $q(t)$

(d) Amplitude spectrum of $s(t)$ (h) Amplitude spectrum of $s(t)$ (l) Amplitude spectrum of $s(t)$

Figure 3-25: Spectra of the baseband $i(t)$ and $q(t)$ signals and the modulated $s(t)$ signal, see Figure 3-23(a), for double sideband (DSB) and single sideband (SSB) suppressed-carrier (SC) modulation. Here the carrier frequency $f_c = f_{\rm LO}$, the LO frequency in Figure 3-23(a).

realized in differential form as shown in Figure 3-24(a). The polyphase filters in Figures 3-23(c) and 3-24(a) are narrowband but the bandwidth can be increased using more stages, see Figure 3-24(b).

Double Sideband Modulation

When $i(t)$ and $q(t)$ are independent, effectively pseudo-random, signals the result is double sideband (DSB) modulation, see Figures 3-25(a–d) with each sideband having the bandwidth of the baseband signals. The amplitude spectra of $i(t)$ and $q(t)$ will be the same as shown in Figure 3-25(a) and each

has a bandwidth f_{BB}.[4] As seen in Figure 3-25(b) the frequency components of the $i(t)$ spectrum are arbitrarily assigned a $45°$ phase. Since $i(t)$ and $q(t)$ are independent the phase of $q(t)$ relative to the phase of $i(t)$ is random, see Figure 3-25(c). The modulated signal $s(t)$ then has a sideband below the carrier frequency f_c and a sideband above f_c, see Figure 3-25(d), for a total bandwidth $2f_{\mathrm{BB}}$.

The multipliers in a quadrature modulator are implemented as mixer circuits and one type of mixer in particular is the multiplicative mixer shown in Figure 3-23(a). Ideally a multiplicative mixer multiplies two sinewaves together to produce the trigonometric expansion of the product of two sinewaves. For example, $\sin(A) \cdot \sin(B) = \frac{1}{2}[\cos(A-B) - \cos(A+B)]$. Following the signal paths in Figure 3-23(a) and considering the frequency component $A_i(f_i)\sin(\omega_i t)$ of $i(t)$ at radian frequency $\omega_i = 2\pi f_i$ and a frequency component $A_q(f_q)\sin(\omega_q t)$ of $q(t)$ at radian $\omega_q = 2\pi f_q$ the modulated signal with the LO frequency replaced by f_c (with radian carrier frequency $\omega_c = 2\pi f_c$) is

$$\begin{aligned}
s(t) &= s_i(t) + s_q(t) \\
&= [A_i(f_i)\sin(\omega_i t)\sin(\omega_c t)] + [A_q(f_q)\sin(\omega_q t)\sin(\omega_c t - \pi/2)] \\
&= [A_i(f_i)\sin(\omega_i t)\sin(\omega_c t)] - [A_q(f_q)\sin(\omega_q t)\cos(\omega_c t)] \\
&= \tfrac{1}{2}\left[A_i(f_i)\cos\omega_c - \omega_i)t - A_i(f_i)\cos(\omega_c + \omega_i)t\right. \\
&\quad \left. -A_q(f_q)\sin(\omega_c + \omega_q)t + A_q(f_q)\sin(\omega_c - \omega_q)t\right] \\
&= \tfrac{1}{2}\left[A_i(f_i)\cos(\omega_c - \omega_i)t + A_q(f_q)\sin(\omega_c - \omega_q)t\right] \\
&\quad - \tfrac{1}{2}\left[A_i(f_i)\cos(\omega_c + \omega_i)t + A_q(f_q)\sin(\omega_c + \omega_q)t\right]. \qquad (3.13)
\end{aligned}$$

The DSB-SC modulated signal is the signal in Equation (3.13) summed for all f_i and f_q components from DC to f_{BB}.

The expansion in Equation (3.13) can be repeated for all of the frequency components of $i(t)$ and $q(t)$. So while the expansion is only performed for discrete frequencies, all that is necessary is that the multiplier be practically ideal, something that is typically achieved by an analog multiplier to better than 1%. If the DSB-SC signal was produced digitally then multiplication can be precisely implemented and the DSB-SC modulation is ideal although the maximum frequency is limited by the capabilities of the digital circuitry. A DSP-based DSB-SC modulation has a low carrier frequency as keeping the (digital) carrier frequency low reduces DC power requirements. If $i(t)$ and $q(t)$ are independent, Equation (3.13) indicates a lower modulated sideband at the range of frequencies $(f_{\mathrm{LO}} - f_i)$ and $(f_{\mathrm{LO}} - f_q)$ and an upper modulated sideband at the range of frequencies $(f_{\mathrm{LO}} + f_i)$ and $(f_{\mathrm{LO}} + f_q)$ for all f_i and f_q from 0 (DC) to f_{BB}. That is, this is DSB-SC modulation, as seen in Figure 3-25(d). In demodulation both sidebands are needed to recover $i(t)$ and $q(t)$.

Single Sideband Modulation

When $i(t)$ and $q(t)$ are the same signal except that every frequency component of $q(t)$ is shifted by $90°$ the result is single sideband (SSB) modulation and the modulated output signal has a bandwidth f_{BB}. The

[4] The short-term spectra will be different because $i(t)$ and $q(t)$ are different signals but over a long time interval the envelope of the amplitude spectra will become similar.

carrier itself does not exist in the output with a quadrature modulator using multipliers so this modulator implements SSB suppressed-carrier (SSB-SC) modulation. The modulated output signal is obtained with $q(t) = A_i(f_i)\sin(\omega_i - \pi/2) = -A_i(f_i)\cos(\omega_i)$. Then Equation (3.13) becomes (but now $f_{\rm LO}$ is used to distinguish it from the carrier frequency which is defined by the characteristics of the modulating signal)

$$\begin{aligned} s(t) &= s_i(t) + s_q(t) \\ &= [A_i(f_i)\sin(\omega_i t)\sin(\omega_{\rm LO}t)] - [A_i(f_i)\cos(\omega_i t)\sin(\omega_{\rm LO}t - \pi/2)] \\ &= A_i(f_i)\left[\sin(\omega_i t)\sin(\omega_{\rm LO}t) + \cos(\omega_i t)\cos(\omega_{\rm LO}t)\right] \\ &= \tfrac{1}{2}A_i(f_i)\left\{\cos[(\omega_{\rm LO} - \omega_i)t] - \cos[(\omega_{\rm LO} + \omega_i)t]\right. \\ &\quad \left. + \cos[(\omega_{\rm LO} + \omega_i)t] + \cos[(\omega_{\rm LO} - \omega_i)t]\right\} \\ &= A_i(f_i)\cos[(\omega_{\rm LO} - \omega_i)]t. \end{aligned} \tag{3.14}$$

Equation (3.14) indicates that just the lower sideband is present and this is SSB-SC modulation as seen in Figure 3-25(h) and the bandwidth of the modulated output signal is $f_{\rm BB}$. The original $i(t)$ signal can be recovered from this one sideband but that is because $q(t)$ contains exactly the same information as $i(t)$ (although phase shifted).

If instead the phase of each frequency component of $q(t)$ led the same frequency component of $i(t)$ by $+90°$ then $s(t)$ would comprise the upper sideband and this is still this would be SSB-SC modulation, see Figures 3-25(i–l). For SSB-SC modulation each frequency component of $q(t)$ must have a phase that differs from the corresponding component of $i(t)$ by $90°$. A lumped-element circuit that realizes this is the polyphase filter, see Figure 3-23(c), but the phase shift can also be realized in DSP.

Earlier, just before Equation (3.14), it was stated that the frequency of the carrier was defined by the characteristics of the modulated signal which in turn depends on the characteristics of the modulating signal. This modulating signal, the $i(t)$ input to the SSB-SC modulator, could also be modulated as is usually the case in SDR where DSB-SC modulation in done in a DSP and this is followed by SSB-SC modulation done at RF using analog hardware. Identifying the correct RF carrier is required for demodulation. In identifying the carrier there are two situations to consider. If the input, $i(t)$, of the SSB-SC modulator is not modulated, e.g. it is just a baseband signal, then the carrier frequency is just the frequency of the LO of the SSB-SC modulator as shown in Figure 3-26(a), i.e. $f_c = f_{\rm LO}$. If the input signal to the SSB-SC modulator is itself a DSB-SC signal (produced by a DSB-SC modulator) so that it has its own intermediate carrier frequency $f_{c,\rm IF}$, then the carrier frequency $f_c = f_{\rm LO} - f_{c,\rm IF}$. This situation is shown in Figure 3-26(b). (Note that the carrier frequency would be above $f_{\rm LO}$ if the frequency components of $q(t)$ were advanced in phase by $90°$ relative to the phase of the frequency components of $i(t)$.)

3.9.2 Summary

This section discussed quadrature modulation and showed how the same circuit can be used for DSB and for SSB modulation. The difference is in whether or not $i(t)$ and $q(t)$ are related. In modern radios DSB is implemented in DSP to produce an IF modulated signal with the spectra shown in Figure 3-25(d) and f_c is very low, perhaps even $f_c = f_{BB}$. Then

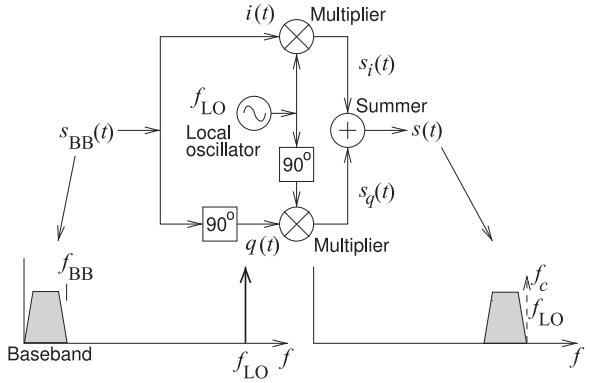
this DSB-SC signal becomes the baseband of an analog SSB modulator that produces the RF modulated signal. This RF modulated signal is a DSB signal with a (suppressed) carrier in the middle of the spectrum.

3.10 Case Study: SDR Transmitter

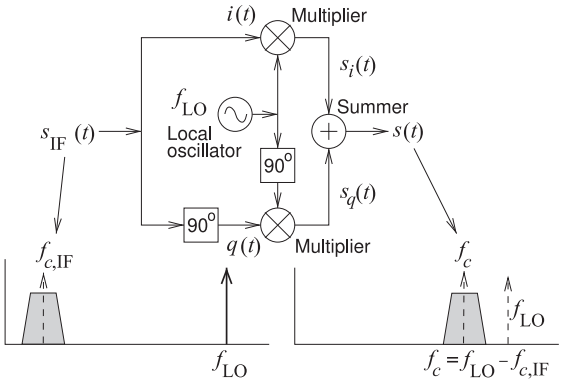
An SDR transmitter combines DSB-SC modulation having a relatively low intermediate frequency carrier with a broadband analog SSB-SC modulation to produce a DSB-SC RF signal. There are many possible implementations. This case study presents system simulations of an SDR transmitter with specific parameters. First a DSB-SC quadrature modulator is studied with sinusoidal in-phase and quadrature baseband signals and then a SSB-SC modulator is examined. This is followed by the study of a direct analog modulation of a digital signal. The final study corresponds to a typical SDR transmitter with DSP-based intermediate frequency DSB-SC modulation followed by SSB-SC analog modulation producing a DSB-SC modulated RF signal.

3.10.1 Analog Quadrature Modulator

The quadrature modulator of Figure 3-23(a) is simulated here with a 10 MHz sinewave for $i(t)$, a 15 MHz sinewave for $q(t)$, and a 1 GHz sinewave LO. The local oscillator is a 1 GHz sinewave. The resulting RF waveform at the



(a) SSB-SC modulator with baseband input signal $s_{\mathrm{BB}}(t)$



(b) SSB-SC modulator with IF input signal $s_{\mathrm{IF}}(t)$

Figure 3-26: Quadrature modulator with the $i(t)$ component derived directly from the input signal and $q(t)$ derived from a $90°$ negatively phase-shifted input signal. The IF signal in (b) is typically a DSB-SC signal with intermediate carrier frequency $f_{c,\mathrm{IF}}$ set in DSP.

output, $s(t)$, is shown in Figure 3-27. The waveform is plotted on a 1 μs scale so that there are 10 cycles of $i(t)$ and 15 cycles of $q(t)$ which are modulated on 1,000 cycles of the 1 GHz RF carrier.

The spectra of the signals at the outputs of the two multipliers, i.e. of $s_i(t)$ and $s_q(t)$, are shown in Figure 3-28(a). Each spectrum has two peaks offset from the 1 GHz LO frequency by 10 MHz for the I signal, $s_i(t)$, and by 15 MHz for the Q signal, $s_q(t)$. The amplitudes of each spectra are symmetrical around the LO frequency but there is a difference in the phase of the Fourier transforms of $s_i(t)$ and $s_q(t)$. The phases of the Fourier transforms of the signals are shown in Figure 3-28(b). The upper and lower sideband phases of the 10 MHz I channel signal are equal but the phases of the upper and lower sidebands of the 15 MHz Q channel signal differ by $180°$, i.e. the upper sideband of $s_q(t)$ is the negative of the lower sideband of $s_q(t)$. In contrast the phases of the upper and lower realized sidebands of $s_i(t)$ are equal.

The amplitude and phase spectra of the combined signal, $s(t) = s_i(t) + s_q(t)$, are shown in Figures 3-28(c and d) and are as expected from combining the spectra of the components. The spectrum of $s(t)$, consists of two pairs of peaks with the peaks of one pair being above and below the LO frequency by 10 MHz, and the peaks of the other pair being offset by 15 MHz. The components are in a lower sideband and an upper sideband and hence this modulation is DSB modulation. Also there is not a component of the output signal at the local oscillator frequency, f_{LO}. The middle of the output signal bandwidth is generally the carrier frequency f_c which here is the same as f_{LO}. Thus the carrier does not exist in the output and so this is called suppressed carrier (SC) modulation. Together this is double sideband suppressed carrier (DSB-SC) modulation. Suppression of the carrier is a property of the multiplicative mixer but some types of modulators, e.g. amplitude modulators, have the carrier in the RF output signal and hence the suppressed-carrier distinction being used with quadrature modulation.

Note that each of the sidebands of the DSB-SC modulated signal have both I and Q information. With finite and equal bandwidth $i(t)$ and $q(t)$ signals, the I and Q information in the RF modulated signal will have overlapping lower sideband and upper sideband. It is only possible to recover and separate the original $i(t)$ and $q(t)$ signals if both sidebands are used in demodulation.

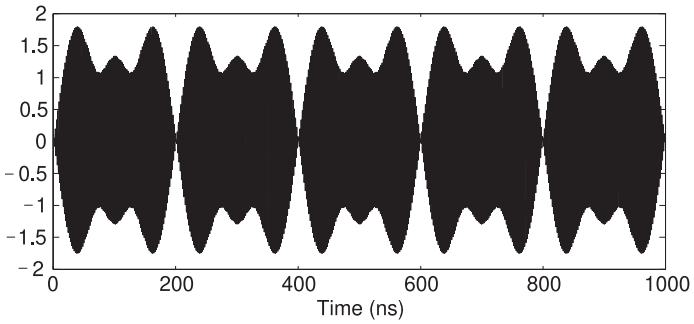


Figure 3-27: A 1 GHz carrier modulated by a 10 MHz sinusoidal $i(t)$ and a 15 MHz sinusoidal $q(t)$.

(a) Spectra of the intermediate modulated carrier signals $s_i(t)$ and $s_q(t)$

(b) Phase of the Fourier transform of intermediate modulated carrier signals $s_i(t)$ and $s_q(t)$

(c) Spectrum of the modulated carrier signal $s(t)$

(d) Phase of Fourier transform of the modulated carrier signal $s(t)$

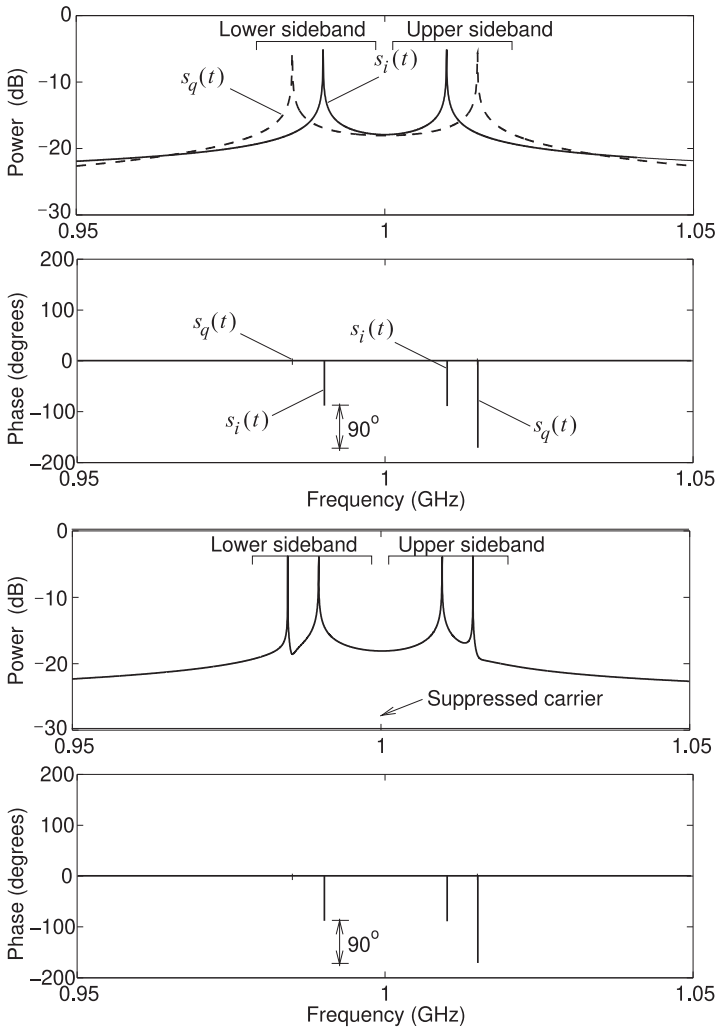
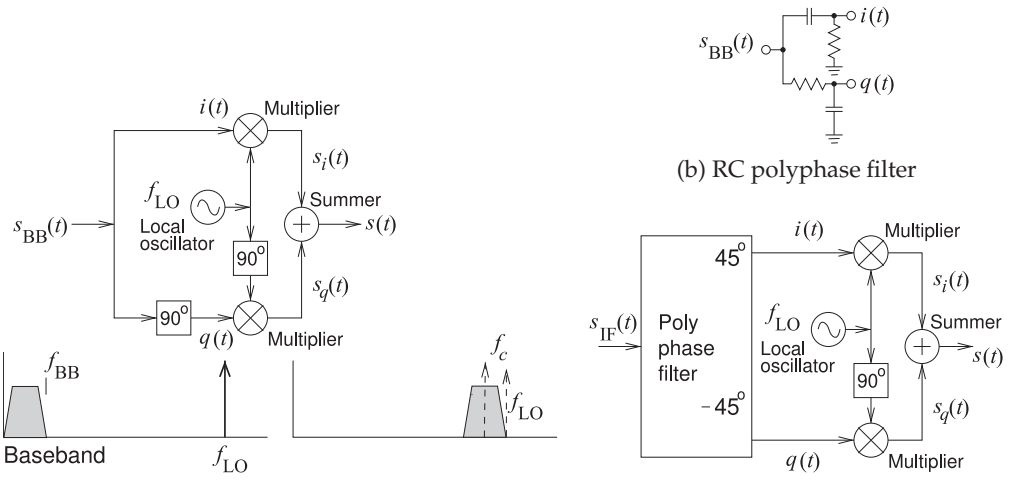


Figure 3-28: Spectrum of signals in the quadrature modulator of Figure 3-23(a) with 10 MHz in-phase, $i(t)$, and 15 MHz quadrature-phase, $q(t)$, modulating signals. (Simulated time = 65.536 μs, using a 2^{23} point FFT.)

3.10.2 Single-Sideband Suppressed-Carrier (SSB-SC) Modulation

Examination of the phase differences of the lower and upper sideband components with the DSB-SC modulator considered in the previous section leads to the design of a SSB-SC modulator. This is obtained when $i(t)$ and $q(t)$ are the same signal except that the frequency components of $q(t)$ lag those of $i(t)$ by $90°$. Thus the phase components of $s_q(t)$ shown in Figure 3-28(b) are shifted by $-90°$ so that the lower sideband components of $s_i(t)$ and $s_q(t)$ (now having the same frequency) combine constructively but the upper sideband components cancel. The variation of the quadrature modulator that implements this is shown in Figure 3-29(a) with the baseband signal

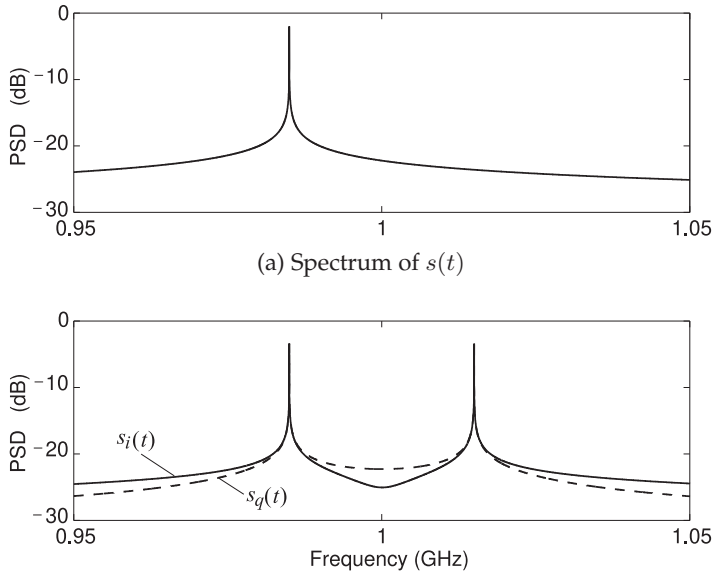


(a) SSB-SC modulator with $f_c = f_{LO} - \frac{1}{2}f_{BB}$

(b) RC polyphase filter

(c) Modulator with polyphase filter

Figure 3-29: Quadrature modulator as a single-sideband, suppressed-carrier (SSB-SC) modulator. At the design frequency the polyphase filter in (b) the phase of $i(t)$ is advanced by $45°$ and the phase of $q(t)$ is retarded by $45°$.



(a) Spectrum of $s(t)$

(b) Spectra of $s_i(t)$ and $s_q(t)$

Figure 3-30: Spectra of the modulated signal with $f_{LO} = 1$ GHz. (Simulated time = 65.536 μs, 2^{23} point FFT.)

$s_{BB}(t) = i(t)$, and $q(t)$ is the same signal except that $q(t)$ lags $i(t)$ by $90°$. The signals in this modulator will now be examined.

With $s_{BB} = i(t)$ being a 15 MHz sinusoidal signal, the $q(t)$ is a 15 MHz sinewave that lags $i(t)$ by $90°$. The spectrum at the output of the quadrature modulator of Figure 3-29(a and c) is as shown in Figure 3-30(a). Now there is only one sideband. Insight into the operation of SSB modulation is obtained by examining the spectra of the signals at the output of the multipliers, see Figure 3-30(b). It is seen that the amplitude spectra of $s_i(t)$ and of $s_q(t)$ are

(a) Phases of $s_i(t)$ and $s_q(t)$

(b) Phase difference of $s_i(t)$ and $s_q(t)$

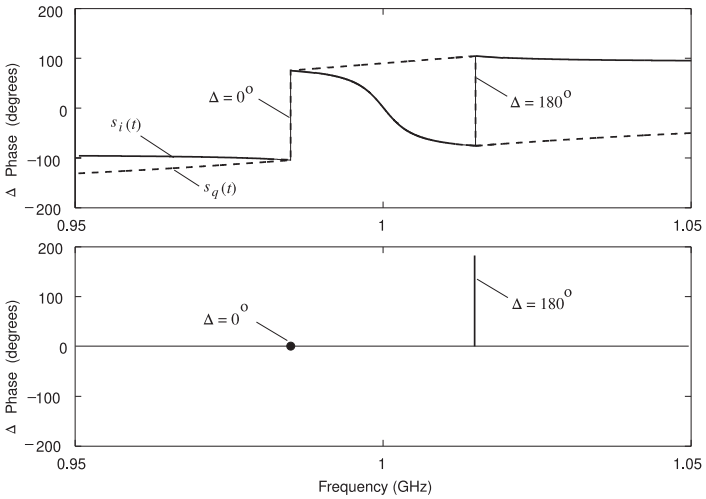


Figure 3-31: Phase of the modulated signal. (Simulated time = 65.536 μs, 2^{23} point FFT.)

essentially the same and both have upper and lower sidebands. However there is a difference in the phases of their Fourier transforms, see Figure 3-31. It is seen that at 15 MHz offset from the 1 GHz LO there is no differences in the phases of $s_i(t)$ and $s_q(t)$ components in the lower sideband but there is a $180°$ difference in the upper sideband, see Figure 3-31(b). Thus when $s_i(t)$ and $s_q(t)$ are summed their lower sideband components add but their upper sideband components cancel thus suppressing the upper sideband in the combined signal $s(t)$. This illustrates the most important metric of a SSB modulator as having I/Q paths that are matched as any imbalance in implementation that results in an amplitude or phase difference of the I and Q paths will result in a spurious (upper) sideband. Provided that the balance is good a filter is not required to suppress an upper sideband.

The SSB-SC modulator was modeled here with a sinusoidal input signal, $s_{\rm BB}(t)$, but upper sideband suppression will also be obtained with a finite bandwidth baseband signal. This is achieved if every frequency component of $s_{\rm BB}(t)$ is $90°$ phase-shifted to become $q(t)$ so that $q(t)$ is the quadrature-phase version of the in-phase $i(t)$. A circuit that implements this is the polyphase filter and one type is the RC circuit shown in Figure 3-29(b). This has a finite bandwidth but this can be broadened by using multiple stages, e.g. see Figure 3-24(b). Incorporating the polyphase filter in the quadrature modulator leads to the implementation in Figure 3-29(c).

3.10.3 Digital Quadrature Modulation

A digital quadrature modulator is shown in Figure 3-32. The input bitstream S_k is converted into two independent binary bitstreams I_k and Q_k which are then lowpass filtered, or in general wave-shaped, to provide two independent analog signals $i(t)$ and $q(t)$. That is, each pair of bits in the S_k bitstream becomes one I_k bit and one Q_k bit. The binary waveforms of I_k and Q_k are then filtered to obtain the analog signals $i(t)$ and $q(t)$ each of which has the same baseband bandwidth indicated in the spectrum on the lower

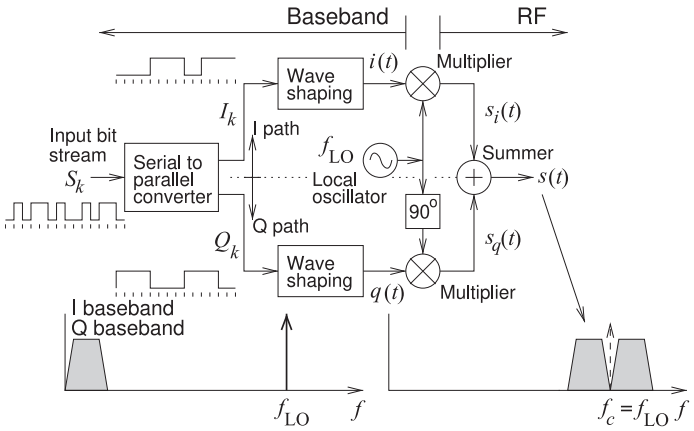


Figure 3-32: Digital quadrature modulators with independent binary I and Q channels yielding double sideband, suppressed carrier (DSB-SC) modulation.

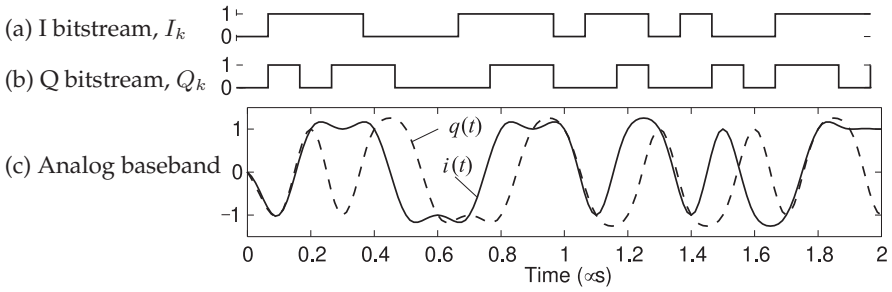


Figure 3-33: Baseband signals for a four state binary quadrature modulator. The I_k and Q_k bitstreams are derived from the random bitstream S_k = 00 11 10 11 01 00 10 11 11 00 10 11 00 10 01 00 11 11 10 and taken as (I_k,Q_k) pairs. The initial setting is $i(0) = 0 = q(0)$.

left in Figure 3-32. The in-phase baseband signal, $i(t)$, is multiplied by the local oscillator signal having frequency $f_{\rm LO}$. Also, the quadrature baseband signal, $q(t)$, is multiplied by the local oscillator signal having frequency $f_{\rm LO}$ but now the LO is delayed by $90°$. Summing $s_i(t)$ and $s_q(t)$, the outputs of the multipliers, yields the double sideband suppressed carrier RF signal $s(t)$ which has twice the bandwidth of each of the analog baseband signals as indicated in the lower right in Figure 3-32.

The modulator of Figure 3-32 results in four states of the modulated carrier signal. That is, if the modulated carrier signal $s(t)$ is sampled at a time indicated by the common clock of I_k and Q_k, the sample of the carrier will have a particular amplitude and phase. Since this is QPSK modulation the amplitudes at these multiple sampling instances will be the same but the phases will have four values addressed by the coordinate (I_k, Q_k).

Now consider a particular quadrature multiplier where S_k is a 20 Mbit/s random bitstream so that I_k and Q_k are the 10 Mbit/s bit streams shown in Figure 3-33(a and b). The I_k and Q_k bitstreams are then level shifted so that the transitions are between -1 and 1 instead of between 0 and 1. The level-shifted bitstreams are then lowpass filtered to obtain the analog baseband signals $i(t)$ and $q(t)$ shown in Figure 3-33(c). The lowpass filter used here is a raised cosine filter which is commonly used in digital modulation rather than a lumped-element lowpass filter and here has an effective corner frequency

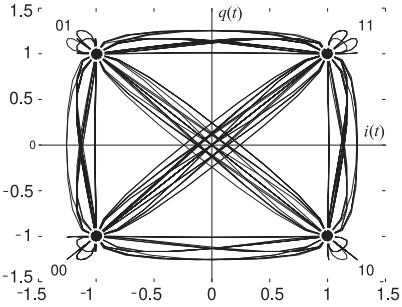


Figure 3-34: Constellation diagram for the binary quadrature modulator. This is also the phasor diagram (with appropriate scaling) of the modulated carrier signal with constellation points sampled at 0.1 μm intervals shown as large dots. There are four constellation points corresponding to four symbols and each symbol represents two bits of information.

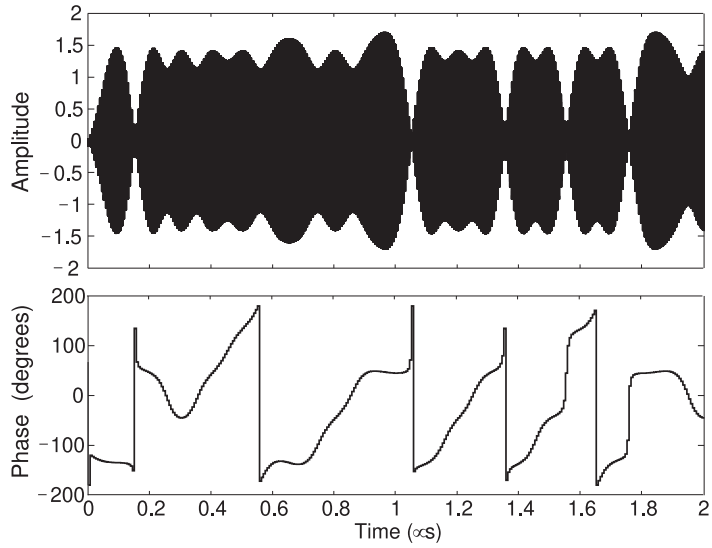
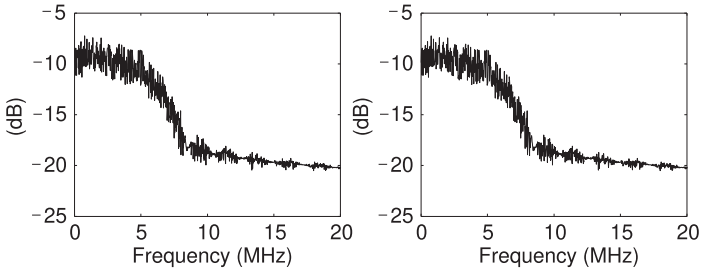


Figure 3-35: Amplitude and phase of the modulated carrier signal $s(t)$ as the output of a binary modulator with a 20 Mbit/s digital baseband signal. Note that the rapid phase transitions occur when the amplitude of the modulated carrier goes to zero. (The rapid switch between $180°$ and $-180°$ when the amplitude is not zero is a continuous smooth phase change.)

of 7 MHz. In practice the raised cosine filter is implemented in DSP so that the signals $i(t)$ and $q(t)$ follow DACs.

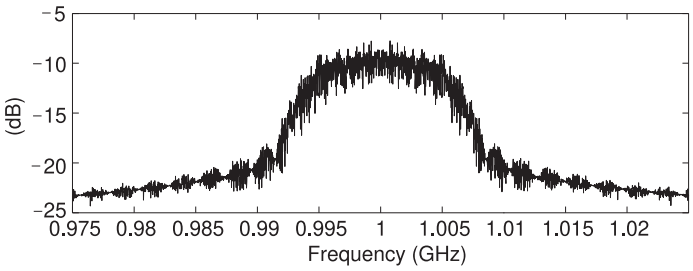
One feature of the raised cosine filter is that the filtered response at the clock-derived sampling times is exact. That is, if the filtered analog baseband signals are sampled every 0.1 μs then the sampled values of $i(t)$ and $q(t)$ will be exactly either $+1.00$ or -1.00 and (after level-shifting by adding one and multiplying by half) the original bitstream is exactly recovered as 0 or 1. If an analog lowpass filter was used the sampled values would not be exactly right. The raised cosine filter introduces no sampling distortion and also the transitions are minimal, i.e. they have the minimum bandwidth compared to what would be obtained if an analog filter was used for wave-shaping. The bandwidth of the filtered signal obtained using a lumped-element filter would be 10 MHz or more and even then the carrier samples would never correspond exactly with the constellation points. Plotting $q(t)$ against $i(t)$ yields the transitions shown in Figure 3-34. Samples of the baseband signal every 0.1 μs coincides with one of the four states and enables two bits of information to be recovered.

The next stage of the DSB-SC modulator multiplies the $i(t)$ signal by a 1 GHz LO and the $q(t)$ signal is multiplied by a $90°$ phase-shifted LO. Then the outputs of the multipliers are summed to produce the modulated RF signal $s(t)$ shown in Figure 3-35. The LO here is in the middle of the RF bandwidth and so here the carrier frequency is the same as the LO frequency.



(a) Spectrum of $i(t)$

(b) Spectrum of $q(t)$



(c) Spectrum of modulated carrier $s(t)$

Figure 3-36: Spectra of signals in the digital modulator with a 20 Mbit/s input stream, S_k, and a 1 GHz LO. (1024 symbols, a time duration of 102.4 μs, and using a 524,288 point FFT)

Sampling the RF signal $s(t)$ every 0.1 μs provides the amplitudes and phases of the carrier with each sample coinciding with one of the constellation points. This modulation scheme is generally called QPSK modulation for quadra-phase shift keying but it is sometimes, but less accurately, called quadrature phase shift keying. In QPSK modulation each constellation point corresponds to one of four phases of the RF signal: $45°$, $135°$, $-45°$, or $-135°$.

QPSK modulation was originally implemented with simpler hardware so the QPSK term is used rather than 4-QAM for four-state quadrature amplitude modulation which more accurately reflects the modulation process described in this section, it does not have to be done this way if only the phase of the carrier is to be adjusted.

Plotting the phasor of the RF carrier signal will result in a phasor diagram identical to the transitions shown in the constellation diagram of Figure 3-34 although it may be necessary to scale the amplitude of the phasor signal to match the values in the constellation diagram. Note that the constellation diagram does not change when the average signal level changes. With the sampling clock aligned to the original clock for the I_k and Q_k bitstreams, samples of the RF phasor will precisely coincide with one of the constellation points in Figure 3-34 (and hence $s(t)$ is on the way to being demodulated).

Digital radio sends data in finite length packets and here the packet is 2048 bits long. With 2 bits per symbol in QPSK modulation there are 1024 symbols and with a symbol interval of 0.1 μs, the duration of the packet is 102.4 μs.

The spectra of the baseband and the output RF signals in the modulator are shown in Figure 3-36. The spectra of $i(t)$ and $q(t)$, Figures 3-36(a and b) respectively, are not identical because each bitstreams is independent. For each bitstream the spectra extends down to DC. It would be possible to use a coding scheme for the bitstreams that ensures that there is not a DC component and this aids automatic carrier recovery in pre-4G cellular systems. Instead 4G and 5G systems use separate mechanisms enabling carrier recovery. The bandwidth of the in-phase and quadrature-phase

(a) Spectrum of the modulated carrier signal, $s(t)$

(b) Spectrum of the in-phase (I) component of the modulated carrier signal, $s_i(t)$

(c) Spectrum of the quadrature-phase (Q) component of the modulated carrier signal, $s_q(t)$

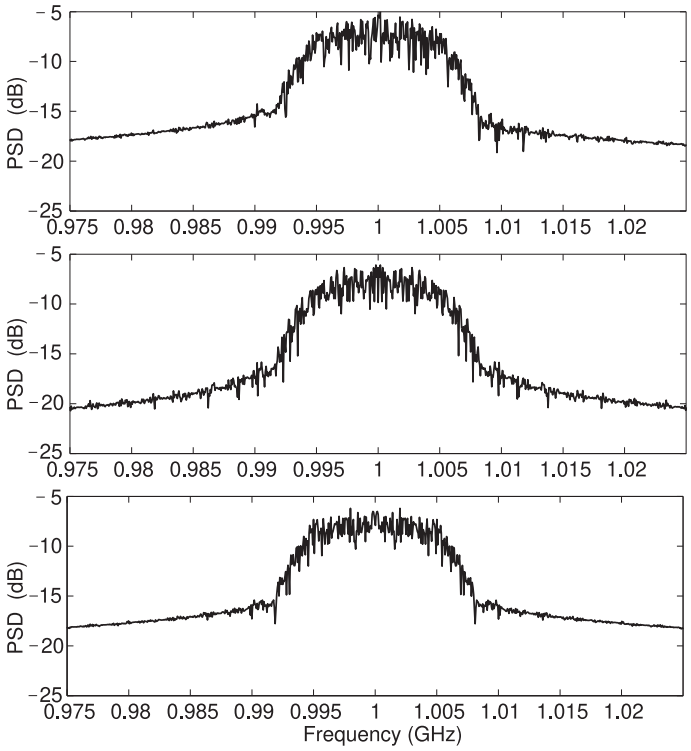


Figure 3-37: Spectra of the digitally modulated RF signals for 1024 symbols (4096 bits), a time duration of 102.4 μs, and using a 524,288 point FFT.

analog baseband signals are slightly less than 10 MHz. The spectrum of the RF output signal is shown in Figure 3-36(c) and the bandwidth of this double-sideband modulated signal is slightly less than 20 MHz. The RF signal has two sidebands, one below 1 GHz and one above even though there is no clear demarcation such as a dip at 1 GHz. The continuous spectrum through 1 GHz is a consequence of the baseband spectra extending to DC.

As was noted previously, and in the absence of noise, the constellation points are faithfully replicated if the RF signal is sampled at 0.1 μs intervals (provided that the phase and the frequency of the carrier has been replicated accurately). The replication of the constellation points is a property of the raised cosine filter which also results in reduced bandwidth baseband analog signals than would be obtained if analog lowpass filtering was used. Also note that (with the DSB-SC quadrature modulation described here) the carrier frequency, the center of the RF signal spectrum, is 1 GHz, the same as the LO frequency.

The spectra of all of the RF signals are shown in Figure 3-37. The spectrum of the RF output, Figure 3-37(a), is centered at 1 GHz and this is the carrier frequency and is also the LO frequency for this DSB-SC transmitter. A curiosity is examining the phases of the RF signals, however little insight is obtained from the phase plots in Figure 3-38.