

Spectrum and Network Measurements

2nd Edition

Robert A. Witte

Spectrum and Network Measurements

Spectrum and Network Measurements

2nd Edition

Robert A. Witte



Edison, NJ
theiet.org



Published by SciTech Publishing, an imprint of the IET.
www.scitechpub.com
www.theiet.org

© 1993, 2001, 2014 by Robert A. Witte

Second edition by SciTech Publishing, 2014
First edition published by SciTech Publishing, 2006 (1-884932-16-9)
First edition published by Noble Publishing, 2001
First edition published by Prentice Hall, 1993

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the web at copyright.com. Requests to the Publisher for permission should be addressed to The Institution of Engineering and Technology, Michael Faraday House, Six Hills Way, Stevenage, Herts, SG1 2AY, United Kingdom.

While the author and publisher believe that the information and guidance given in this work are correct, all parties must rely upon their own skill and judgement when making use of them. Neither the author nor publisher assumes any liability to anyone for any loss or damage caused by any error or omission in the work, whether such an error or omission is the result of negligence or any other cause. Any and all such liability is disclaimed.

Cover Photo: Keysight Technologies

10 9 8 7 6 5 4 3 2 1

ISBN 978-1-61353-014-6 (hardback)
ISBN 978-1-61353-036-8 (PDF)

Typeset in India by MPS Limited
Printed in the USA by Integrated Books International
Printed in the UK by TJ International Ltd, Cornwall

*To the four women who have shaped my life:
Eileen, Joyce, Sara, and Rachel*

Contents

Preface	xvii
Abbreviations	xxi
1 Introduction to Spectrum and Network Measurements	1
1.1 Signals and Systems	1
1.2 Time Domain and Frequency Domain Relationships	2
1.3 System Transfer Function	3
1.4 Advantages of using Frequency Domain Measurements	4
1.5 Spectrum Measurements	5
1.6 Network Measurements	7
1.7 Combined Spectrum/Network Analyzers	9
1.8 Modular Instruments	10
Bibliography	10
2 Decibels	11
2.1 Definition of the Decibel	11
2.2 Cardinal Values	13
2.3 Absolute Decibel Values	14
2.4 Gain and Loss Calculations	17
2.5 Decibels and Percent	19
2.6 Error Expressed in Decibels	20
Bibliography	21
3 Fourier Theory	23
3.1 Periodicity	23
3.2 Fourier Series	24
	vii

3.3	Fourier Series of a Square Wave	25
3.4	Fourier Series of Other Waveforms	30
3.5	Fourier Transform	31
3.6	Fourier Transform of a Pulse	32
3.7	Inverse Fourier Transform	33
3.8	Fourier Transform Relationships	33
3.9	Discrete Fourier Transform	36
3.10	Limitations of the DFT	38
3.11	Fast Fourier Transform	39
3.12	Relating Theory to Measurements	39
3.13	Finite Measurement Time	40
	Bibliography	41
4	Fast Fourier Transform Analyzers	43
4.1	The Bank-of-Filters Analyzer	43
4.2	Frequency Resolution	44
4.3	The FFT Analyzer	45
4.4	Sampled Waveform	46
4.5	Sampling Theorem	47
4.6	FFT Properties	51
4.7	Controlling the Frequency Span	52
4.8	Band Selectable Analysis	53
4.9	Leakage	55
4.10	Hanning Window	55
4.11	Flattop Window	58
4.12	Uniform Window	59
4.13	Exponential Window	60
4.14	Selecting a Window Function	62
4.15	Oscillator Characterization	62
4.16	Spectral Maps	63
4.17	Time Domain Display	65
4.18	Network Measurements	65
4.19	Phase	66

4.20	Electronic Filter Characterization	67
4.21	Cross-Power Spectrum	68
4.22	Coherence	70
4.23	Correlation	72
4.24	Autocorrelation	73
4.25	Cross-Correlation	75
4.26	Histogram	76
4.27	Real-Time Bandwidth	78
4.28	Real-Time Bandwidth and RMS Averaging	79
4.29	Real-Time Bandwidth and Transients	80
4.30	Overlap Processing	81
4.31	Swept Sine	83
4.32	Octave Measurements	84
	Bibliography	85
5	Swept Spectrum Analyzers	87
5.1	The Wave Analyzer	87
5.2	Heterodyne Block Diagram	88
5.3	The Swept Spectrum Analyzer	89
5.4	Practical Considerations	91
5.5	Input Section	91
5.6	Resolution Bandwidth	92
5.7	Sweep Limitations	92
5.8	Specialized Sweep Modes	95
5.9	Local Oscillator Feedthrough	95
5.10	Digital IF Section	96
5.11	Types of Detectors	97
5.12	The Tracking Generator	98
5.13	FFT versus Swept Measurements	98
5.14	Modern Spectrum Analyzer Block Diagrams	99
5.15	Real-Time Spectrum Analyzer	101
5.16	Types of Spectrum Analyzers	103
	Bibliography	104

6	Modulation Measurements	107
6.1	The Carrier	107
6.2	Amplitude Modulation	108
6.3	AM Measurements	113
6.4	Zero-Span Operation	114
6.5	Other Forms of Amplitude Modulation	115
6.6	Angle Modulation	115
6.7	Narrowband Angle Modulation	118
6.8	Wideband Angle Modulation	119
6.9	FM Measurements	122
6.10	Combined AM and FM	123
6.11	Digital Modulation	125
6.12	Quadrature Modulation	126
6.13	Common Digital Modulation Formats	127
6.14	Error Vector Magnitude	131
6.15	Channel Measurements	133
	Bibliography	135
7	Distortion Measurements	137
7.1	The Distortion Model	137
7.2	Single-Tone Input	138
7.3	Two-Tone Input	139
7.4	Higher-Order Models	142
7.5	The Intercept Concept	142
7.6	Harmonic Distortion Measurements	145
7.7	Use of Low-Pass Filter on Source	146
7.8	Intermodulation Distortion Measurements	147
7.9	Distortion Internal to the Analyzer	148
	Bibliography	149
8	Noise and Noise Measurements	151
8.1	Statistical Nature of Random Noise	151
8.2	Mean, Variance, and Standard Deviation	152

8.3	Power Spectral Density	153
8.4	Frequency Distribution of Noise	153
8.5	Equivalent Noise Bandwidth	155
8.6	Noise Units and Decibel Relationships	157
8.7	Noise Measurement	158
8.8	Automatic Noise Level Measurement	159
8.9	Noise Floor	159
8.10	Correction for Noise Floor	160
8.11	Phase Noise	161
	Bibliography	165
9	Pulse Measurements	167
9.1	Spectrum of a Pulsed Waveform	167
9.2	Effective Pulse Width	169
9.3	Line Spectrum	170
9.4	Pulse Spectrum	171
9.5	Pulsed RF	174
9.6	Pulse Desensitization	175
	Bibliography	176
10	Averaging and Filtering	177
10.1	Predetection Filtering	177
10.2	Predetection Filters	179
10.3	Postdetection Filtering	180
10.4	Postdetection Filters	181
10.5	Averaging	183
10.6	Variance Ratio	183
10.7	General Averaging	184
10.8	Linear Weighting	185
10.9	Exponential Weighting	185
10.10	Averaging in Spectrum and Network Analyzers	187
10.11	RMS Average	188

10.12	Vector Averaging	188
10.13	Smoothing	191
10.14	Averaging versus Filtering	191
	Bibliography	193
11	Transmission Lines	195
11.1	The Need for Transmission Lines	195
11.2	Distributed Model	196
11.3	Characteristic Impedance	196
11.4	Propagation Velocity	197
11.5	Generator, Line, and Load	197
11.6	Impedance Changes	201
11.7	Sinusoidal Voltages	202
11.8	Complex Reflection Coefficient	203
11.9	Return Loss	203
11.10	Standing Waves	204
11.11	Input Impedance of a Transmission Line	207
11.12	Measurement Error Due to Impedance Mismatch	209
11.13	Insertion Gain and Loss	212
11.14	Line Losses	216
11.15	Coaxial Lines	216
	Bibliography	217
12	Measurement Connections	219
12.1	The Loading Effect	219
12.2	Maximum Voltage and Power Transfer	220
12.3	High-Impedance Inputs	220
12.4	Active High-Impedance Probes	223
12.5	Z_0 Impedance Inputs	223
12.6	Input Connectors	224
12.7	Z_0 Terminations	225
12.8	Power Dividers and Splitters	225
12.9	Attenuators	228

12.10 Return Loss Improvement	230
12.11 The Classical Attenuator Problem	232
12.12 Impedance Matching Devices	234
12.13 Measurement Filters	236
Bibliography	238
13 Two-Port Networks	241
13.1 Sinusoidal Signals	241
13.2 The Transfer Function	243
13.3 Improved Two-Port Model	244
13.4 Impedance Parameters	245
13.5 Admittance Parameters	246
13.6 Hybrid Parameters	246
13.7 Transmission Parameters	247
13.8 Scattering Parameters	247
13.9 Transfer Function and S_{21}	250
13.10 Why S-Parameters?	250
Bibliography	251
14 Network Analyzers	253
14.1 Basic Network Measurements	253
14.2 Oscilloscope and Sweep Generator	253
14.3 Network Measurements Using a Spectrum Analyzer	254
14.4 Vector Network Analyzer	255
14.5 Directional Bridges and Couplers	257
14.6 S-Parameter Test Set	257
14.7 Modern Vector Network Analyzer Configurations	259
14.8 Sweep Limitations	260
14.9 Power Sweep	262
14.10 Flexible Source Frequency	262
14.11 VNA Time Domain Measurements	264
14.12 Nonlinear VNA Measurements	265
Bibliography	265

15	Vector Network Measurements	267
15.1	Distortionless Transmission	267
15.2	Nonlinearity	269
15.3	Linear Distortion	269
15.4	Importance of Linear Phase	270
15.5	Group Delay	273
15.6	Normalization	275
15.7	Measurement Plane	278
15.8	Reflection Measurements	279
15.9	Directional Bridges and Couplers	284
15.10	Reflection Configuration	287
15.11	Reflection Normalization	288
15.12	Error in Reflection Measurements	289
15.13	Vector Error Correction	290
15.14	Normalization Revisited	291
15.15	Two-Term Error Correction	291
15.16	Three-Term Error Correction	291
15.17	Two-Port Error Correction	293
	Bibliography	295
16	EMC Measurements	297
16.1	Electromagnetic Compatibility	297
16.2	Radiated Emissions	298
16.3	Antennas	300
16.4	Near Field and Far Field	303
16.5	EMI Receiver Requirements	304
16.6	Peak, Quasi-Peak, and Average Detectors	305
16.7	Conducted Emissions	306
16.8	Line Impedance Stabilization Network	306
16.9	EMC Troubleshooting	309
16.10	Near-Field Probes	310
16.11	Current Probe	312
16.12	Preamplifiers	313
	Bibliography	314

17 Analyzer Performance and Specifications	315
17.1 Source Specifications	315
17.2 Receiver Characteristics	317
17.3 Spectrum Analyzer Dynamic Range	319
17.4 Network Analyzer Specifications	321
Bibliography	324
Appendix A	325
Index	329

Preface

*When you can measure what you are speaking about,
and express it in numbers, you know something about it.*

—Lord Kelvin, May 3, 1883

This book is about the theory and practice of spectrum and network measurements in electronic systems. It is intended for readers who have a background in electrical engineering and use spectrum analyzers and network analyzers to characterize electronic signals or systems.

This is the book that I wish someone had handed me when I started my career as an electrical engineer. This work was formed from thousands of interactions with my fellow engineers at HP, Agilent Technologies, and now Keysight Technologies about this or that measurement challenge. My target reader is the recent electrical engineering graduate or an engineer recently tossed into the challenge of performing spectrum or network measurements. For inspiration, I often think about the electrical engineering students I have taught and then write with the goal of helping them apply that big pile of electrical theory in their heads.

Since the first edition was written, the body of knowledge in this area has grown dramatically. Rather than triple the size of the book, I chose to keep it focused on the core measurement principles and list key references for further study. This second edition does reflect the dramatic impact of digital technology, driving significant change in the systems being measured, and the technology used inside the measuring instruments. Every chapter of the book has been impacted by this important shift.

The concept of wireless communication has been around for decades, evolving from spark gap transmitters to handheld digital mobile phones. Spark gap transmissions relied on Morse code (the original digital format), occupied wide spectrum bandwidth, and were relatively inefficient. Over time, communication systems adopted AM and FM analog modulation techniques to implement amplitude modulation and frequency modulation broadcast radio, two-way radio, and early cellular telephones. More recently, digital formats have emerged as the most efficient and versatile modulation schemes. It has been fascinating to witness the explosion in wireless communications devices, and it is not over yet.

A merging of wireless and digital technology is producing an unprecedented level of electronic connectivity in our society. The increasing usage of wireless devices has caused a

corresponding high demand for engineers and technicians who understand radio frequency and microwave circuits and systems. Despite these recent changes in technology, the fundamentals of signals propagating through circuits and through the air have not changed. The basic theory of *signals and systems* and the measurements that accompany it still apply. Concepts such as Fourier analysis, transmission lines, intermodulation distortion, signal-to-noise ratio, and scattering parameters (S-parameters) represent a critical foundation for this new era of wireless development. The purpose of this book is to enable the reader to understand that basic theory, to relate it to measured results, and to apply it in creating new RF and microwave designs.

Although some of the internal functions of spectrum analyzers and network analyzers are discussed, the real emphasis of the book is on the theory and practice of frequency domain measurements. Enough theory is provided so that the reader can understand how a particular measurement is made, what the possible sources of error are, and the significance of the results. Many numerical examples are given to aid the reader in understanding the material and to help relate theory and practice.

The book can certainly be read cover to cover, but it is also organized into independent chapters and subchapters. This allows the reader to read selectively and enhances the usefulness of the book as a reference.

Chapter 1 is an introduction to spectrum and network measurements. Decibels are an often used and misused concept, so Chapter 2 is devoted to that topic. Fourier theory, the theoretical basis for spectrum analysis, is summarized in Chapter 3. The two main types of spectrum analyzers (fast Fourier transform analyzers and swept analyzers) are discussed in Chapters 4 and 5. Chapters 6 through 9 cover several important measurement applications: modulated signals, signal distortion, noise, and pulsed waveforms. Averaging and filtering are covered together in Chapter 10.

Chapters 11 and 12 cover transmission lines and measurement connection techniques. Chapter 13 introduces two-port network theory, which is the basis for network analysis. Chapters 14 and 15 cover network analyzers, focusing on using vector network analyzers for transmission and reflection measurements. Electromagnetic compatibility (EMC) measurements are covered in Chapter 16, which is new to this edition. Chapter 17 ends the book with a discussion of instrument performance and specifications.

Additional support material for this book is available online at <http://www.electronic-measurement.com>.

Acknowledgments

Many people contribute to the undertaking of writing a book such as this one, either directly or indirectly. My appreciation goes to my friend and colleague Ken Wyatt for helping me with this book, especially the chapter on EMC measurements (Chapter 16). Ken is a world-class expert on EMC, so check out his work at <http://www.emc-seminars.com>. My gratitude also goes to Joe Gorin, James Gitre, and Bob Cutler, who provided valuable feedback on all or portions of the book. Thank you!

I continue to be impressed with the wealth of measurement knowledge embedded in the application notes from Agilent Technologies. (Note that the electronic measurement

business of Agilent has been launched as a new company, Keysight Technologies). These app notes were of tremendous value to me, and I have referenced many of them at the end of each chapter. Most of the authors are unnamed, but I am grateful for their contributions.

My appreciation goes to the companies that have supplied photographs and other graphics for this book: Keysight Technologies, ETS-Lindgren and Beehive Electronics.

Portions of Chapter 4 (Sections 4.21 through 4.32) were contributed by the R&D department of the Lake Stevens Instrument Division of Hewlett-Packard (now Keysight Technologies).

Thank you to Dudley Kay and the entire crew at SciTech Publishing for producing the second edition and hanging in there with me to get it done.

My thanks to Gary Breed, Dennis Ford, and Martina Voigt of Noble Publishing for bringing the first edition of this book back into print and giving it a second life. These colleagues helped me in a variety of ways to complete the first edition: Jerry Daniels, Glenn Engel, Bryan Hoog, Roy Mason, Harry Plate, Bill Spaulding, Joe Tarantino, and Ken Wyatt.

Bob Witte

bob.witte@electronic-measurements.com

Abbreviations

AC – alternating current
ACPR – adjacent channel power ratio
ACLR – adjacent channel leakage ratio
AM – amplitude modulation
BPSK – binary phase-shift keying
BNC – Bayonet Neill Concelman (connector)
CE – conducted emissions
CISPR – International Special Committee on Radio Interference
CW – continuous wave
DANL – displayed average noise level
dBm – decibels relative to one milliwatt
dBV – decibels relative to one volt
DC – direct current (or 0 Hz)
DSB – double sideband
DUT – device under test
EDF – forward directivity error
ESF – forward source match error
ERF – forward reflection tracking error
ELF – forward load match error
ETF – forward transmission tracking error
EXF – forward crosstalk error
EDR – reverse directivity error
ESR – reverse source match error
ERR – reverse reflection tracking error
ELR – reverse load match error
ETR – reverse transmission tracking error

xxii Abbreviations

EXR – reverse crosstalk error
EVM – error vector magnitude
EM – electromagnetic
EMC – electromagnetic compatibility
EMI – electromagnetic interference
EVM – error vector magnitude
FCC – Federal Communications Commission
FDM – frequency division multiplexing
FFT – fast Fourier transform
FM – frequency modulation
FMT – frequency mask trigger
FPGA – field programmable gate array
IEC – International Electrotechnical Commission
IF – intermediate frequency
IFFT – inverse fast Fourier transform
IBW – effective impulse bandwidth
IMD – intermodulation distortion
LISN – line impedance stabilization network
LO – local oscillator
LSB – lower sideband
LTI – linear time invariant (system)
MER – modulation error ratio
NBW – equivalent noise bandwidth
OBW – occupied bandwidth
PDF – probability density function
PM – phase modulation
POI – probability of intercept
PRN – pseudo-random noise
PRF – pulse repetition frequency
PSD – power spectral density
QAM – quadrature amplitude modulation
16QAM – 16-state quadrature amplitude modulation
QPSK – quadrature phase-shift keying
RBW – resolution bandwidth
RCE – relative constellation error
RE – radiated emissions

RF – radio frequency
RL – return loss
RTBW – real-time bandwidth
RTSA – real-time spectrum analyzer
RMS – root mean square
SSB – single sideband
SNA – scalar network analyzer
SNA – spectrum/network analyzer
SHI – second harmonic intercept
SOI – second-order intercept
SOLT – short-open-load-through (calibration)
SWR – standing wave ratio
TDR – time domain reflectometry
THD – total harmonic distortion
TOI – third-order intercept
TRL – through-reflect-line (calibration)
USB – upper sideband
VBW – video bandwidth
VCO – voltage-controlled oscillator
VNA – vector network analyzer
VR – variance ratio
VSWR – voltage standing wave ratio

Introduction to Spectrum and Network Measurements

In this chapter we'll review some of the basic system concepts that relate to spectrum and network measurements. The basic function of a spectrum analyzer and network analyzer will be introduced, along with a few example measurements.

1.1 Signals and Systems

An electrical system normally has one or more input ports and one or more output ports. Electrical devices such as filters, attenuators, and amplifiers fall into this category. Figure 1-1 shows a system with a single input, $x(t)$, and a single output, $y(t)$.

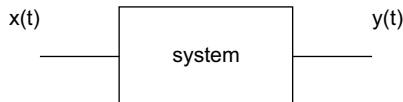


Figure 1-1 A simple system having one input, $x(t)$, and one output, $y(t)$.

A more complex system, a phase lock loop, is shown in Figure 1-2. Although there is still only one input and one output, there are several blocks or subsections of the system, each having its own input and output. Each block of the system may be considered as another system. When designing or testing such a system, an engineer thinks in terms of the individual blocks and the signals flowing between the blocks. Measurement instrumentation is used in the design phase when the engineer verifies the performance of the individual blocks and signals. Later, the signals and system blocks may be measured during manufacturing to verify functionality and performance. Also, the system may be measured as part of maintaining it in the field.

Network measurements characterize the circuit blocks of the system, whereas spectrum measurements characterize the signals present. For example, in Figure 1-2, the frequency content of the output, $y(t)$, might be a critical parameter in the performance of the system and could be measured using a spectrum analyzer. Similarly, the transfer characteristics of the low-pass (loop) filter might be of interest, which could be measured with a network analyzer.

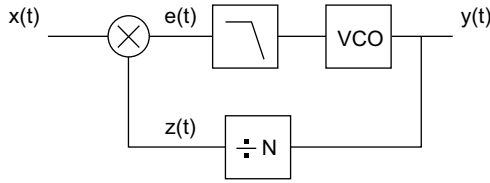


Figure 1-2 A phase lock loop is a complex system with multiple blocks and multiple signals.

1.2 Time Domain and Frequency Domain Relationships

A common way to describe an electrical signal is its *time domain representation*, the voltage or current as a function of time, as shown in Figure 1-3a. The blocks in a system can be characterized in the time domain by measuring the step response, pulse response, or the response at the output due to some other input signal. An oscilloscope displays the time domain representation of a signal.

Another way to describe a signal is using its *frequency domain representation*, the amplitude of the signal as a function of frequency, as shown in Figure 1-3b. The frequency domain representation must include both magnitude and phase information to fully represent a signal. Fourier theory relates the time domain and frequency domain representations.

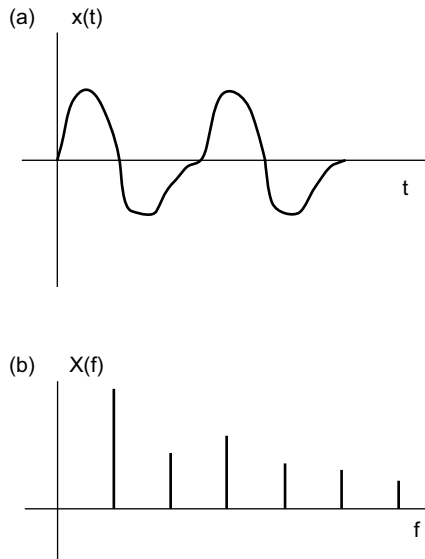


Figure 1-3 (a) A signal shown as a function of time. (b) A signal shown as a function of frequency.

Appropriate use of the Fourier series, Fourier transform, and the discrete Fourier transform (DFT) allow the transformation of a time domain function, $x(t)$, into a frequency domain function, $X(f)$. Figure 1-4 shows a commonly used method of relating the time and frequency domains in one three-dimensional plot.

The spectrum analyzer is a common electronic instrument for measuring the frequency content of a signal and displaying it in the frequency domain. Thus, the spectrum analyzer is to the frequency domain as the oscilloscope is to the time domain.

Network measurements also make use of a frequency domain representation to characterize a system. However, network measurements are performed by applying a stimulus at the input to the system and measuring the resulting output signal. This stimulus must cover a wide range of frequencies for the output signal to adequately represent the frequency domain performance of the system. Most often, the stimulus is a sine wave source swept through the frequency range of interest, but other signals can be used, such as a broadband noise source.

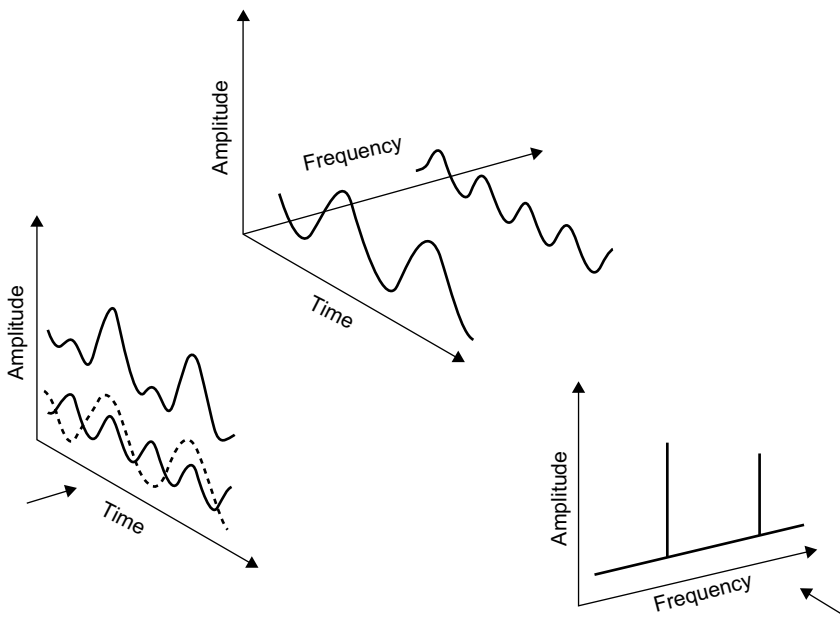


Figure 1-4 A three-dimensional approach that shows the relationship between the time domain and frequency domain.

1.3 System Transfer Function

The stimulus signal, $X(f)$, is applied to the input of a system, and the output, $Y(f)$, is measured (Figure 1-5). The transfer function is the ratio of the output over the input, both as a function of frequency.

$$\text{transfer function: } H(f) = \frac{Y(f)}{X(f)}$$

This implies a simple model of the system. That is, the input signal and the transfer function completely determine the output signal, with no loading effects present. Two-port parameters (discussed in Chapter 13) provide for a more complete model of a system.

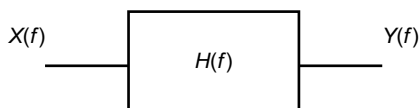


Figure 1-5 The transfer function of a system describes its behavior in the frequency domain.

Often a system is assumed to be *linear time invariant* (LTI). For a system to be linear, it must allow two input signals to be summed to create the corresponding summed output signal.

That is, if

input $x_1(t)$ produces output $y_1(t)$

and

input $x_2(t)$ produces output $y_2(t)$

then the scaled and summed input

$a_1x_1(t) + a_2x_2(t)$ produces output $a_1y_1(t) + a_2y_2(t)$

where a_1 and a_2 are real scalar values.

For a system to be *time invariant*, the output that is produced in response to an input does not change with time. That is, if the input $x(t)$ produces the output $y(t)$, then a time-delayed version of $x(t)$, which is $x(t - t_0)$, will produce the delayed output $y(t - t_0)$.

If a system is both linear and time invariant, it meets the criteria of LTI. Of course, many systems do not strictly meet this criteria. For example, practical electronic circuits often introduce distortion products due to the nonlinear behavior of the circuit. It is often these imperfections in circuit performance that limit the overall system performance, which means they are an important parameter to understand and measure.

1.4 Advantages of using Frequency Domain Measurements

Why use frequency domain measurement techniques? The answer varies with the application, but frequency domain measurements have several distinct advantages.

Narrowband frequency domain measurements provide greater sensitivity than time domain measurements. Since the measurement bandwidth can be narrowed almost arbitrarily, frequency domain analyzers can greatly reduce the amount of noise present in the measurement. Similarly, narrowband measurements can remove large interfering signals at undesired frequencies. Consider the measurement of harmonic distortion of a near-perfect sine wave. A spectrum analyzer can ignore the large fundamental frequency when measuring the harmonic level. A time domain measurement with an oscilloscope must simultaneously measure the fundamental and the much smaller harmonics in the signal. Harmonic distortion

measurements with an oscilloscope are limited to a few percent, while spectrum analyzers routinely allow 0.01% distortion measurements.

Some systems are inherently frequency domain oriented. For instance, the frequency division multiplexing (FDM) systems used in telecommunications systems operate by sandwiching together multiple signals in the frequency domain. Cellular telephones and other wireless mobile devices operate within a given frequency band, with the radio spectrum divided up for use by the various cell sites. Frequency domain measurements are a natural way to characterize these signals and systems.

Even systems that are not usually thought of as being inherently frequency domain in nature may still require frequency domain measurements. For instance, stray capacitance and resistive losses in a high-speed digital circuit may limit the bandwidth of the circuit and the speed of a digital pulse. A network analyzer can determine the bandwidth of the circuit by measuring its transfer function in the frequency domain.

Multiple signals are usually easier to separate in the frequency domain than in the time domain. For instance, suppose the output of a switching power supply contains significant levels of the 60 Hz line frequency (and its harmonics) and the switching frequency of the power supply. Whichever of these is the largest will be discernible by a time domain measurement. Usually, if there are multiple frequencies present, it will be difficult to view them with an oscilloscope. A spectrum analyzer, on the other hand, can separate these frequency components and measure them accurately.

1.5 Spectrum Measurements

A signal is characterized using a spectrum analyzer, as shown in Figure 1-6. The measurement is usually as simple as connecting the analyzer to the source of the signal. However, loading effects and other sources of measurement error may need to be considered. The frequency spectrum of the signal will appear on the analyzer's display. An example is shown in Figure 1-7.

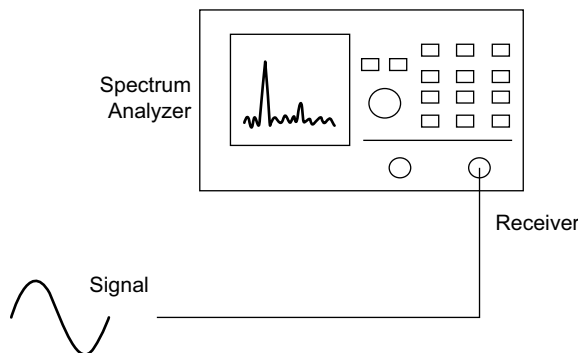


Figure 1-6 A spectrum measurement is performed by applying the signal to be analyzed to the input of a spectrum analyzer.

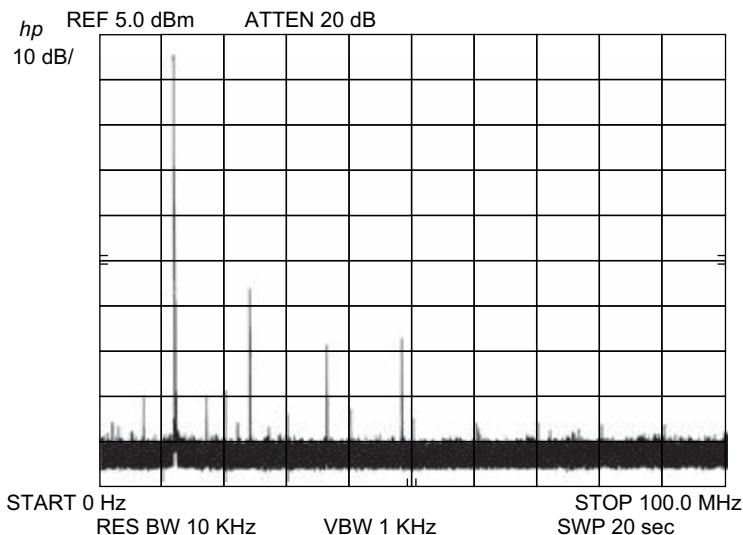


Figure 1-7 A typical spectrum analyzer measurement showing the harmonic content of a signal generator output.

The complexity of the measurement varies according to the application. In a simple case, the spectrum analyzer may be used to measure the amplitude and frequency of a signal spectral line. More often, the spectral content of the signal includes multiple responses such as harmonics, modulation sidebands, and spurious responses. Noise levels can also be measured (if the measured noise is greater than the analyzer's noise), and the noise level can be displayed as a function of frequency.

The standard vertical scale on a spectrum analyzer is logarithmic and marked in decibels. This allows a large dynamic range to be displayed on a reasonable-sized screen. Many analyzers also provide a linear vertical scale for users that prefer to work in terms of volts. The horizontal scale is, of course, frequency. It is most often a linear frequency scale, but a logarithmic frequency scale is used in some applications.

Spectrum analyzers are available in a wide range of configurations with a corresponding wide range of performance. Frequency range is the most fundamental parameter to use to categorize spectrum analyzers. Different measurement technologies are most effective in different frequency bands. *Fast Fourier transform (FFT) spectrum analyzers* are primarily intended for audio and mechanical measurements, using analog-to-digital conversion and the FFT to cover from near 0 Hz to a few hundred kHz. *Swept spectrum analyzers* use traditional radio frequency receiver circuits to sweep the frequency range of interest. These analyzers are typically offered in frequency ranges that start at a few Hz on the low end and extend to 50 GHz or higher. Figure 1-8 shows a high-performance 26.5 GHz spectrum analyzer that combines the FFT measurement technique with the traditional swept approach. Besides frequency range, other factors such as cost, dynamic range, sensitivity, accuracy, and feature set vary from analyzer to analyzer.



Figure 1-8 A mid-performance spectrum analyzer with a frequency range of 20 Hz to 26.5 GHz. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

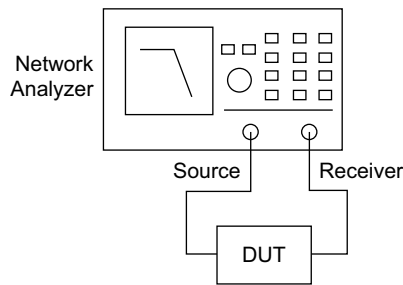


Figure 1-9 A network analyzer provides a signal source to the device under test (DUT) and measures the response at the device's output.

1.6 Network Measurements

A network is characterized in the frequency domain by connecting the source of a network analyzer to the input of the network and the analyzer's receiver to the output of the network (Figure 1-9). Thus, the network analyzer provides its own stimulus for the device under test (DUT).

The transfer function is the most common network measurement (Figure 1-10). The gain or loss of an attenuator, a filter, an amplifier, or other circuit as a function of frequency is an important design parameter. The transfer function is normally displayed with a logarithmic vertical scale (in decibels). The horizontal axis is frequency and may be logarithmic (resulting in a Bode plot) or linear. Other functions such as the phase, group delay, real part, or imaginary part of the transfer function may also be displayed.

Reflection measurements characterize the input or output behavior of DUT. This includes such parameters as return loss, reflection coefficient, impedance, and standing wave ratio, all as a function of frequency. Reflection measurements usually require the use of specialized accessories such as a directional bridge, directional coupler, or *S*-parameter test set.

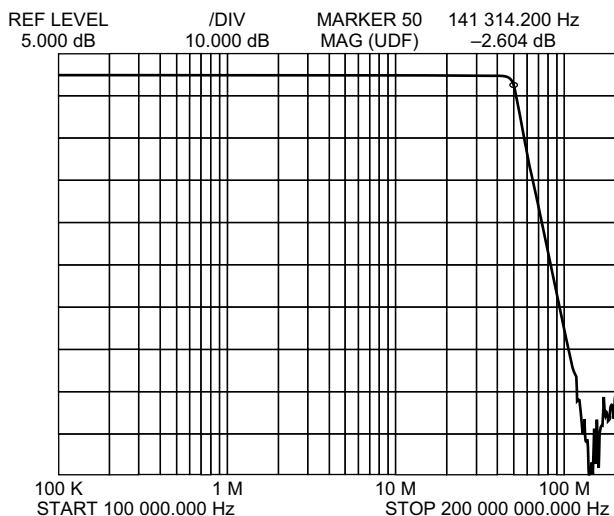


Figure 1-10 A typical network measurement showing the transmission characteristics of a low-pass filter.

Network analyzers are available in two main varieties: *scalar* and *vector*. Scalar network analyzers (SNAs) provide only magnitude information (no phase information) and have tended to be less expensive to implement. As vector network analyzers (VNAs) have decreased in cost, the VNA has largely displaced the SNA in the marketplace. A typical VNA is shown in Figure 1-11.

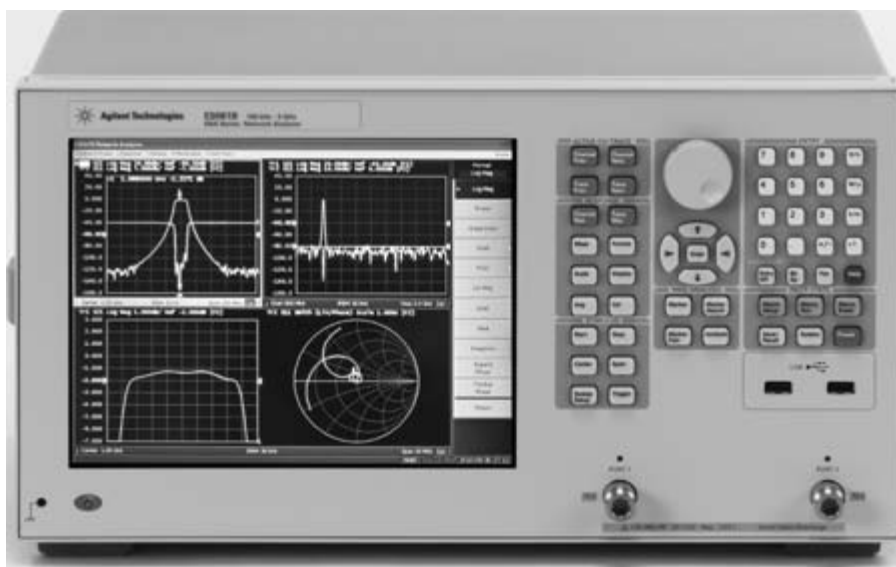


Figure 1-11 A 2-port network analyzer with a frequency range of 100 kHz to 3 GHz. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

1.7 Combined Spectrum/Network Analyzers

Some instruments combine the spectrum analyzer and network analyzer in one instrument (Figure 1-12). This hybrid approach is a natural one, since the same instrument user often needs to perform both spectrum and network measurements. Furthermore, the block diagrams and technologies used in the two types of instruments are similar enough that combining the two instruments can be done at a reasonable cost.

The instrument shown in Figure 1-12 has a compact, handheld form factor, which is especially useful for field service applications.

Why doesn't a network analyzer inherently have the ability to also make spectrum measurements? Usually, a network analyzer design will take advantage of the fact that the frequency of the stimulus signal is known since it is supplied by the network analyzer. This allows the use of a simpler receiver block diagram rather than one that must reject images and other off-carrier frequency components. This is unlike the spectrum case, where a signal is often unknown and complex with multiple frequencies present. Thus, a network analyzer can have a simpler and less expensive block diagram.



Figure 1-12 This RF analyzer offers both spectrum and network measurements in one compact instrument. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

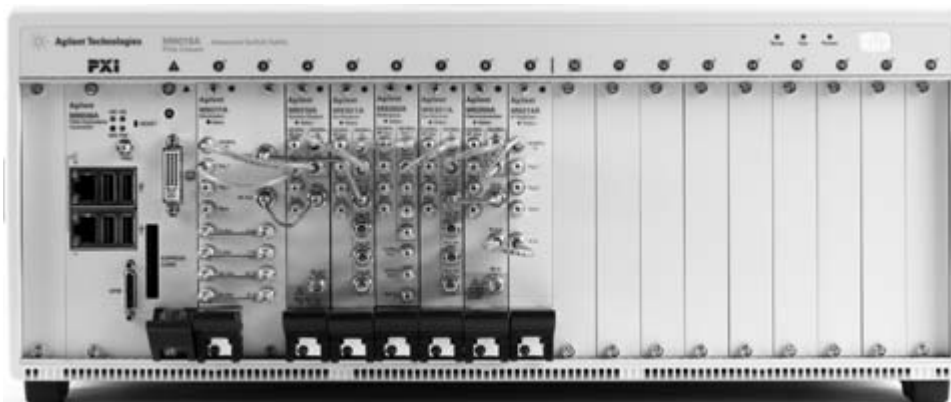


Figure 1-13 This PXI frame is shown with an embedded controller and set of modules that implement a flexible vector signal analyzer. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

1.8 Modular Instruments

Spectrum analyzers and network analyzers are available in modular form factors, with the PXI standard being the most common (Figure 1-13). Industry standard modular systems have several advantages over traditional bench instruments: flexible configuration (especially number of channels), smaller physical size, measurement speed, and tighter integration between instruments. The main disadvantage is that they are much less of a turnkey system, requiring the end user to configure and optimize the test system. Modular instruments often have less measurement precision compared to bench instruments, due to the smaller physical size.

Bibliography

- Oliver, Bernard M., and John M. Cage. *Electronic Measurements and Instrumentation*. New York: McGraw-Hill Book Company, 1971.
- Oppenheim, Alan V., and Alan S. Willsky. *Signals and System*, 2d ed. Englewood Cliffs, NJ: Prentice Hall, Inc., 1996.
- PXI Systems Alliance, *PXI Hardware Specification*, Revision 2.2, <http://pxisa.org>, September 2004.
- Schwartz, Mischa. *Information, Transmission, Modulation, and Noise*, 3rd ed. New York: McGraw-Hill Book Company, 1980.
- Ziemer, Rodger E., William H. Tranter, and D. Ronald Fannin. *Signals and Systems: Continuous and Discrete*, 4th ed. Upper Saddle River, NJ: Prentice Hall, Inc., 1998.

Decibels

Decibels are used to specify ratios of powers and voltages in a logarithmic fashion. Absolute levels can also be specified by supplying suitable reference values. Decibels are commonly used for gain and loss calculations in electronic systems.

Generally, spectrum and network analyzers display measurement results with their displays calibrated in decibels. The popularity of the decibel in such applications is due to its ability to compress logarithmically widely varying signal levels. For example, a 1 V signal and a 0.1 mV signal can both be represented on a display with 100 dB of range. To show these two signals simultaneously with reasonable clarity on a linear scale is impractical.

Decibels also are useful for gain and loss calculations, where multiplication operations are transformed into (easier) additions.

2.1 Definition of the Decibel

The fundamental definition of the decibel (dB) is in terms of a power ratio. Two powers, P_1 and P_2 , can be related in dB by

$$A_{\text{(dB)}} = 10 \log(P_2/P_1) \quad (2-1)$$

where \log indicates the base 10 logarithm.

The subscript (dB) is used to indicate that the numerical result is in decibels. As shown, P_2 is expressed relative to P_1 . Reversing P_1 and P_2 changes the sign of the result in decibels.

If the powers P_1 and P_2 resulted from a pair of voltages across a pair of resistors, then

$$A_{\text{(dB)}} = 10 \log \frac{(V_2^2/R_2)}{(V_1^2/R_1)} \quad (2-2)$$

$$A_{\text{(dB)}} = 10 \log(V_2/V_1)^2 + 10 \log(R_1/R_2) \quad (2-3)$$

$$A_{\text{(dB)}} = 20 \log(V_2/V_1) + 10 \log(R_1/R_2) \quad (2-4)$$

The first term is the voltage form of the decibel equation, and the second term accounts for differences in the two resistances. If the two resistances are equal, this equation can be further simplified.

$$A_{(\text{dB})} = 20 \log(V_2/V_1) \quad (2-5)$$

The last equation has taken on a life of its own and is often used as the defining equation for the decibel. Strictly speaking, the decibel is defined in terms of power only. If the resistances associated with each of the powers (voltages) are equal, then the power equation and the voltage equation are consistent. If the voltage equation is used when the resistances are not equal, incorrect results will occur.

Despite this potential problem, the voltage equation is widely used in situations where the two resistances are not equal. For instance, the voltage equation is often used to specify the voltage gain of operational amplifier circuits. In these circuits, the input impedance is usually very high and the output impedance is usually low. The voltage form of the decibel equation can be used successfully in such a case as long as power gain is not inferred from it.

Example 2.1

Calculate the ratio of P_2 and P_1 and express in decibels. $P_1 = 2 \text{ W}$, $P_2 = 12 \text{ W}$. Exchange P_1 and P_2 and recalculate.

The linear ratio is $A = P_2/P_1 = 12/2 = 6$. Expressed in decibels

$$A_{(\text{dB})} = 10 \log(P_2/P_1) = 10 \log(12/2) = 7.78 \text{ dB}$$

With P_1 and P_2 reversed,

$$A_{(\text{dB})} = 10 \log(P_1/P_2) = 10 \log(2/12) = -7.78 \text{ dB}$$

Solving the decibel equations for the power or voltage ratio results in

$$A = \frac{P_2}{P_1} = 10^{(A_{(\text{dB})})/10} \quad (2-6)$$

$$A = \frac{V_2}{V_1} = 10^{(A_{(\text{dB})})/20} \quad (2-7)$$

Example 2.2

The voltage gain of a circuit (the ratio of the output voltage to the input voltage) is 25 dB. If the output voltage is 5 V, what is the input voltage?

$$V_2/V_1 = 10^{(25/20)} = 17.78$$

$$V_1 = V_2/17.78 = 5/17.78 = 0.281 \text{ V}$$

2.2 Cardinal Values

It is worth summarizing some of the common cardinal values for decibels (Table 2-1). Although precise calculations would be best accomplished using a computer, these ratios can provide a more intuitive, working knowledge of decibels.

Table 2-1 Common Cardinal Values for Decibels

Voltage Ratio	Power Ratio	Decibels
1	1	0 dB
1.414	2	3 dB
2	4	6 dB
3.16	10	10 dB
10	100	20 dB

Some mathematical identities can be used to develop some rules of thumb for working with decibels.

Rule 1. Changing the sign of the decibel value corresponds to taking the reciprocal of the linear ratio.

If

$$A_{(\text{dB})} = 10 \log(A)$$

then

$$-A_{(\text{dB})} = 10 \log(1/A)$$

Rule 2. Adding two decibel values is equivalent to multiplying their corresponding linear ratios.

$$10 \log(A_1) + 10 \log(A_2) = 10 \log(A_1 \times A_2)$$

Example 2.3

Use Table 2-1 and the foregoing rules to express the following linear ratios in terms of decibels: (a) $V_2/V_1 = 20$; (b) $P_2/P_1 = 0.5$; (c) $V_2/V_1 = 40$.

- (a) $V_2/V_1 = 10 \times 2$. From Table 2-1, the ratios of 10 and 2 can be converted to decibel values of 20 dB and 6 dB. Using rule 2, $A_{(\text{dB})} = 20 \text{ dB} + 6 \text{ dB} = 26 \text{ dB}$.
 - (b) The reciprocal of P_2/P_1 is 2, which in decibels is 3 dB. Thus, by rule 1, the decibel value is -3 dB .
 - (c) $V_2/V_1 = 10 \times 2 \times 2$. From Table 2-1, these can be converted to decibel values of 20 dB, 6 dB, and 6 dB. Summing these provides the complete result, $A_{(\text{dB})} = 32 \text{ dB}$.
-

2.3 Absolute Decibel Values

The original definition of the decibel allows only for representing ratios of two values in decibel form. By providing a reference value (either a voltage or a power), decibels can be used to refer to absolute voltage or power values.

$$P_{(\text{dB})} = 10 \log(P/P_{\text{REF}}) \quad (2-8)$$

$$V_{(\text{dB})} = 20 \log(V/V_{\text{REF}}) \quad (2-9)$$

dBm

The most common power reference for spectrum and network measurements is 1 mW, resulting in dBm.

$$P_{(\text{dBm})} = 10 \log(P/0.001) \quad (2-10)$$

Note that this definition does not depend on the impedance that dissipates the power.

It is convenient to develop the voltage form of the equation, since many power measurements are actually calibrated voltage measurements. To do so, the impedance level must be specified since it relates the voltage and power levels. It follows that these equations are valid only for the specified impedances.

The voltage reference produces 1 mW of power in a resistor of the appropriate impedance. For $R = 50 \Omega$

$$V_{\text{REF}} = \sqrt{PR} = \sqrt{0.001 \times 50} = 0.2236 \text{ V} \quad (2-11)$$

$$P_{(\text{dBm})} = 20 \log(V_{\text{RMS}}/0.2236) \quad \text{for } 50 \Omega \quad (2-12)$$

For $R = 75 \Omega$,

$$P_{(\text{dBm})} = 20 \log(V_{\text{RMS}}/0.2739) \quad \text{for } 75 \Omega \quad (2-13)$$

Note that the reference voltage and the voltage to be converted to dBm are both in root mean square (RMS) volts. The symbol $P_{(\text{dBm})}$ was used even though the decibel equation uses voltage to emphasize that dBm is defined in terms of power. The voltage form of the equation is valid only for one particular impedance, whereas the power form of the equation is independent of impedance. A particular dBm value will always indicate the same power level but will correspond to different voltages for different impedances.

Example 2.4

Express the following voltages and powers in terms of dBm: (a) $P = 25 \mu\text{W}$; (b) $V_{\text{RMS}} = 1 \text{ V}$, 50Ω impedance; (c) $V_{\text{RMS}} = 1 \text{ V}$, 75Ω impedance.

(a) $P_{(\text{dBm})} = 10 \log(25 \times 10^{-6}/0.001) = -16.0 \text{ dBm}$

(b) $P_{(\text{dBm})} = 20 \log(1/0.2236) = 13.0 \text{ dBm}$

(c) $P_{(\text{dBm})} = 20 \log(1/0.2739) = 11.24 \text{ dBm}$

dBW

For higher power applications, a power reference of 1 W may be used, resulting in dBW.

$$P_{(\text{dBW})} = 10 \log(P/1.0) = 10 \log(P) \quad (2-14)$$

dBV

The most common voltage reference is 1 V (RMS), resulting in dBV.

$$V_{(\text{dBV})} = 20 \log(V_{\text{RMS}}/1) = 20 \log(V_{\text{RMS}}) \quad (2-15)$$

Measurements in dBV are based on voltage only. A particular dBV value will always have a corresponding voltage value, independent of the impedance present. This means that a constant dBV value will supply differing amounts of power to different impedances. This runs counter to the previous assertion that the decibel is strictly defined in terms of power.

dBmV

Another voltage reference used for decibel measurements is 1 mV RMS, resulting in dBmV.

$$V_{(\text{dBmV})} = 20 \log(V_{\text{RMS}}/0.001) \quad (2-16)$$

dB μ V

Another voltage reference used for decibel measurements is 1 μ V RMS, resulting in dB μ V.

$$V_{(\text{dB}\mu\text{V})} = 20 \log(V_{\text{RMS}}/0.000001) \quad (2-17)$$

dBm/dBV Conversions

To convert between dBm and dBV, the impedance must be specified. Both dBm and dBV can be computed using the voltage form of the decibel equation but with different voltage references. Due to the logarithmic nature of decibels, for any particular impedance dBm and dBV differ by a constant.

$$P_{(\text{dBm})} = V_{(\text{dBV})} + 10 \log[1/(0.001 \times R)] \quad (2-18)$$

For $R = 50 \Omega$ and 75Ω

$$P_{(\text{dBm})} = V_{(\text{dBV})} + 13.01 \quad \text{for } 50 \Omega \quad (2-19)$$

$$P_{(\text{dBm})} = V_{(\text{dBV})} + 11.25 \quad \text{for } 75 \Omega \quad (2-20)$$

The equations showing a power on the left side and a voltage on the right side may seem inconsistent at first. However, they are the logarithmic equivalent of $P = V^2/R$ and serve to emphasize the differing nature of dBm (which is based on power) and dBV (which is based on voltage).

Figure 2-1 allows convenient conversion from volts to dBV and dBm.

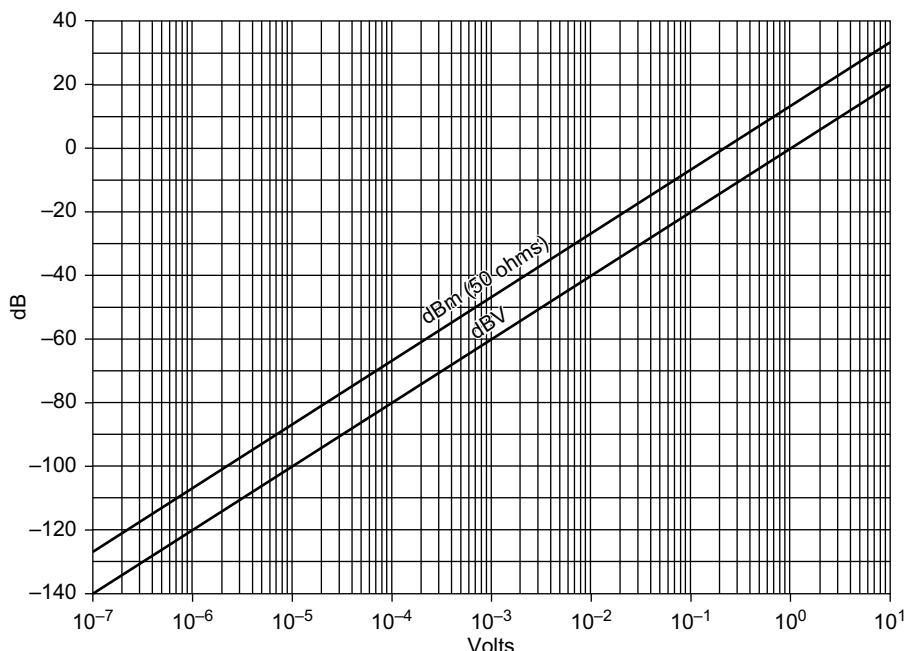


Figure 2-1 Plot of dBm and dBV versus RMS voltage.

Example 2.5

Convert the following measured values to dBV and dBm, as possible, within the limitations of the information provided: (a) 0.1 V RMS across 50 Ω ; (b) 0.5 V RMS, unknown impedance; (c) 5 mW into 75 Ω ; (d) 30 μ W, unknown impedance.

- (a) $V_{(\text{dBV})} = 20 \log(0.1) = -20 \text{ dBV}$; $P_{(\text{dBm})} = 20 \log(0.1/0.2236) = -6.99 \text{ dBm}$.
 (b) $V_{(\text{dBV})} = 20 \log(0.5) = -6.02 \text{ dBV}$; dBm cannot be determined without knowing either the power or the impedance.
 (c) $P_{(\text{dBm})} = 10 \log(0.005/0.001) = 6.99 \text{ dBm}$; $V_{(\text{dBV})} = P_{(\text{dBm})} - 11.25 = -4.26 \text{ dBV}$.
 (d) $P_{(\text{dBm})} = 10 \log(30 \times 10^{-6}/0.001) = -15.23 \text{ dBm}$; dBV cannot be determined without knowing either the voltage or the impedance.
-

High-Impedance Measurements

Although dBm and dBV are relatively straightforward concepts, confusion can occur in high-impedance measurements. For example, many spectrum analyzers compute and display the dBm value corresponding to the measured *voltage* assuming a 50 Ω (or other) impedance, even though the high-impedance input is being used. The measurement is misleading

since the voltage form of the dBm equation is used even though the impedance is not 50 Ω . This is done as an aid to the instrument user, assuming either that an appropriate termination (load) has been installed at the input to the analyzer or that the user knows how to interpret the potentially confusing data. When the answer is displayed as dBm, the user should make sure that something in the measurement system or device under test provides the appropriate load impedance. Some analyzers with high-impedance inputs allow the user to specify the impedance to be used in computing dBm.

2.4 Gain and Loss Calculations

The power gain of a system is the ratio of the output power to the input power.¹

$$G_P = P_2/P_1 \quad (2-21)$$

where

P_2 = output power

P_1 = input power

Power gain is often specified in terms of decibels.

$$G_{P(\text{dB})} = 10 \log(P_2/P_1) \quad (2-22)$$

If P_2 is greater than P_1 , the system exhibits actual power gain. The ratio P_2/P_1 is greater than unity and is positive when expressed in decibel form. If P_2 is less than P_1 , the system has a power gain of less than unity and actually exhibits a loss. When expressed in decibels, the gain is negative. If P_1 and P_2 are equal, the gain is 1, or in dB, 0 dB.

In ratio form, the power loss is

$$L_P = P_1/P_2 = 1/G_P \quad (2-23)$$

Using decibels,

$$L_{P(\text{dB})} = 10 \log(P_1/P_2) = 10 \log(1/G_P) \quad (2-24)$$

A loss is the negative of the corresponding gain, when both are expressed in decibels. For example, a loss of 10 dB is the same as a gain of -10 dB.

Voltage Gain

Gain can also be expressed using voltage, resulting in voltage gain. Again, the warnings apply about using voltage ratios expressed in decibels when the two impedances involved are not equal.

$$G_V = V_2/V_1 \quad (2-25)$$

¹ The definition of power gain is sometimes further refined into several different definitions: operating power gain, transducer power gain, available power gain, and insertion power gain, see Carson (1975).

where

V_2 = output voltage

V_1 = input voltage

In decibels, the voltage gain is

$$G_{V(\text{dB})} = 20 \log(V_2/V_1) \quad (2-26)$$

Example 2.6

Compute the gain and loss (both ratio and dB) for a circuit having an input power of 0.40 mW and an output power of 0.25 mW.

The power gain

$$G_P = 0.25/0.40 = 0.625$$

The power loss

$$L_P = 1/G_P = 1.6$$

In decibels

$$G_{P(\text{dB})} = 10 \log(0.625) = -2.04 \text{ dB}$$

$$L_{P(\text{dB})} = 2.04 \text{ dB}$$

Multiple Blocks

When multiple circuits are cascaded together, decibels are often used to simplify the gain calculations. The electronic system shown in Figure 2-2 has three individual blocks, each with its own gain. The total gain of this system can be computed using

$$G_T = P_{\text{OUT}}/P_{\text{IN}} = G_{P1} \cdot G_{P2} \cdot G_{P3} \quad (2-27)$$

In terms of decibels,

$$G_{T(\text{dB})} = 10 \log(G_{P1} \cdot G_{P2} \cdot G_{P3}) \quad (2-28)$$

$$G_{T(\text{dB})} = 10 \log(G_{P1}) + 10 \log(G_{P2}) + 10 \log(G_{P3}) \quad (2-29)$$

$$G_{T(\text{dB})} = G_{1(\text{dB})} + G_{2(\text{dB})} + G_{3(\text{dB})} \quad (2-30)$$

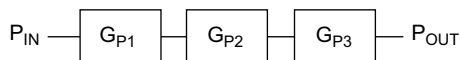


Figure 2-2 Decibels can be used to simplify gain calculations of multiple blocks.

Thus, when decibels are used for gain calculations, multiplication operations are transformed into additions. The output power may be expressed in absolute decibels (such as dBm).

$$G_T = P_{\text{OUT}}/P_{\text{IN}} \quad (2-31)$$

$$G_{T(\text{dB})} = P_{\text{OUT}(\text{dB})} - P_{\text{IN}(\text{dB})} \quad (2-32)$$

$$P_{\text{OUT}(\text{dB})} = G_{T(\text{dB})} + P_{\text{IN}(\text{dB})} \quad (2-33)$$

Expanding the total gain into its individual components,

$$P_{\text{OUT}(\text{dB})} = G_{1(\text{dB})} + G_{2(\text{dB})} + G_{3(\text{dB})} + P_{\text{IN}(\text{dB})} \quad (2-34)$$

Example 2.7

Compute the total system gain in dB for the system shown in Figure 2-3. If the input power is 150 μW , what is the output power in dBm?

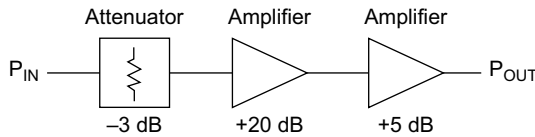


Figure 2-3 A simple system with gain and loss (see Example 2.7).

The total system gain in dB is $G_{(\text{dB})} = -3 \text{ dB} + 20 \text{ dB} + 5 \text{ dB} = 22 \text{ dB}$. The input power is 150 μW , which is

$$10 \log(150 \times 10^{-6}/0.001) = -8.24 \text{ dBm}$$

$$P_{\text{OUT}(\text{dB})} = -8.24 \text{ dBm} + 22 \text{ dB} = 13.76 \text{ dBm}$$

2.5 Decibels and Percent

Often, decibels are used to compare the relative sizes of two signals on a spectrum analyzer. The smaller of the two signals can be described as being a certain number of dB down from the larger signal, which acts as a reference. Modulation measurements and harmonic distortion measurements are often stated this way.

It may also be desirable to express the size of the smaller signal as a percent of the larger one. Since

$$A_{(\text{dB})} = 10 \log(P_2/P_1) \quad (2-35)$$

$$\frac{P_2}{P_1} = 10^{(A_{(\text{dB})}/10)} \quad (2-36)$$

(P_2/P_1) is just the ratio of the two signal powers but is not expressed in percent. It is understood that the ratio must be multiplied by 100 to get percent.

Similarly, for voltage

$$\frac{V_2}{V_1} = 10^{(A_{\text{dB}}/20)} \quad (2-37)$$

Example 2.8

A signal is 42 dB smaller than another signal. What percent of the second signal is the first signal (in terms of voltage)?

$$A_{\text{dB}} = -42\text{dB}, V_2/V_1 = 10^{(-42/20)} = 0.00794, \text{ which is } 0.794\%.$$

2.6 Error Expressed in Decibels

The smaller of the two signals may actually be a source of error in a measurement. For example, the smaller signal may be a spurious response occurring at the same frequency as the (desired) larger signal. Depending on how the two signals add together, an error is produced in the measurement. In most analyzer measurements, the smaller signal may add constructively or destructively (or somewhere in between), depending on the relative phases of the signals. Adding the signals together gives a maximum bound on the error and subtracting them gives a minimum bound.

The error of the combination of the two signals relative to the desired signal is

$$\frac{V_1 \pm V_2}{V_1} = 1 \pm \frac{V_2}{V_1} \quad (2-38)$$

In decibel form,

$$\text{error}_{\text{dB}} = 20 \log(1 \pm V_2/V_1) \quad (2-39)$$

This is the error induced in V_1 (expressed in dB), due to the presence of V_2 . If V_2 is zero, then error_{dB} is 0 dB, indicating that the decibel value of V_1 has no error in it.

Example 2.9

If the smaller signal in Example 2.8 introduces an error into the large signal, express this error in dB. If the large signal is -20 dBV, what is the measured signal (including the error)?

$$\text{error}_{\text{dB}} = 20 \log(1 \pm 0.00794) = \pm 0.069 \text{ dB}$$

If the error is positive,

$$V_{\text{dBV}} = -20 \text{ dBV} + 0.069 \text{ dB} = -19.931 \text{ dBV}$$

If the error is negative,

$$V_{\text{dBV}} = -20 \text{ dBV} - 0.069 \text{ dB} = -20.069 \text{ dBV}$$

Table 2-2 Error Due to Interfering Signal

Interfering Signal Level db	Relative to Desired Signal		Error Introduced Into Desired Signal	
	Power %	Voltage %	Error(dB) +	Error(db) –
0	100.00%	100.00%	6.0206	–∞
–1	79.43%	89.13%	5.5350	–19.2715
–2	63.10%	79.43%	5.0780	–13.7365
–3	50.12%	70.79%	4.6495	–10.6907
–4	39.81%	63.10%	4.2489	–8.6585
–5	31.62%	56.23%	3.8755	–7.1773
–6	25.12%	50.12%	3.5287	–6.0412
–7	19.95%	44.67%	3.2075	–5.1405
–8	15.85%	39.81%	2.9108	–4.4096
–9	12.59%	35.48%	2.6376	–3.8063
–10	10.00%	31.62%	2.3866	–3.3018
–20	1.00%	10.00%	0.8279	–0.9151
–30	0.10%	3.16%	0.2704	–0.2791
–40	0.010%	1.00%	0.0864	–0.0873
–50	0.0010%	0.32%	0.0274	–0.0275
–60	0.00010%	0.10%	0.0087	–0.0087
–70	0.000010%	0.032%	0.0027	–0.0027
–80	0.0000010%	0.010%	0.0009	–0.0009
–90	0.00000010%	0.0032%	0.0003	–0.0003
–100	0.000000010%	0.0010%	0.0001	–0.0001

Table 2-2 is a list of useful decibel relationships. The first column lists the difference between two signal levels, expressed in decibels. The corresponding percent (either power or voltage) is shown in the next two columns, and the worst case errors introduced into the larger signal by the smaller signal (according to equation (2-39)) are listed in the last two columns.

Example 2.10

Using Table 2-2, predict the error introduced into a –10 dBm signal by an interfering signal that is 20 dB lower in power level.

The interfering signal is –20 dB relative to the desired signal. From Table 2-2, an error of +0.8279 dB or –0.9151 dB will be introduced depending on the phase of the interfering signal. So the measured signal level is within the range of –9.1721 dBm to –10.9151 dBm.

Bibliography

Carson, Ralph S. *High Frequency Amplifiers*. New York: John Wiley & Sons, Inc., 1975.

Fourier Theory

The most common way of representing signals is in the time domain. Another representation of a signal is via the frequency domain, which is inherent in spectrum measurements. In the frequency domain, the signal is described in terms of its frequency content, plotting the amount of power present at each frequency. A complete frequency domain representation includes both the magnitude and phase of the signal. The frequency domain is related to the time domain by a body of knowledge generally known as Fourier theory, named for Jean Baptiste Joseph Fourier (1768–1830). This includes the series representation known as the *Fourier series* and the transform techniques known as the *Fourier transform*. Discrete (digitized) signals can be transformed into the frequency domain using the *discrete Fourier transform* (DFT).

3.1 Periodicity

A signal or function is *periodic* if it meets the following criterion:

$$x(t) = x(t + T) \quad \text{for all } t \quad (3-1)$$

where

T = period of the function

In other words, a periodic function can be shifted in time by exactly one period and the resulting new function will look the same as the original one. A periodic function of time repeats itself every T seconds (Figure 3-1).

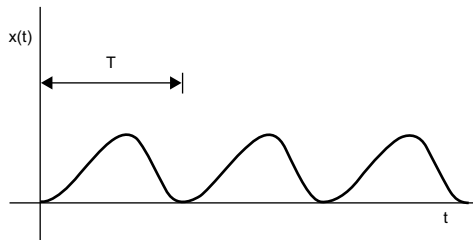


Figure 3-1 A periodic signal repeats every T seconds.

3.2 Fourier Series

Most periodic signals can be represented by a series expansion of sines and cosines. There are some mathematical limitations on the represented signal, but physically realizable signals meet these constraints.¹

The Fourier series representation of a periodic function has the form²

$$x(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos 2\pi n f_0 t + b_n \sin 2\pi n f_0 t) \quad (3-2)$$

where

$$a_n = \frac{2}{T} \int_{-T/2}^{T/2} x(t) \cos 2\pi n f_0 t \, dt \quad (3-3)$$

$$b_n = \frac{2}{T} \int_{-T/2}^{T/2} x(t) \sin 2\pi n f_0 t \, dt \quad (3-4)$$

where

f_0 = fundamental frequency in hertz

T = period of the signal

T and f_0 are related by

$$f_0 = \frac{1}{T} \quad (3-5)$$

The frequency in rad/sec (ω_0) is

$$\omega_0 = 2\pi f_0 \quad (3-6)$$

Using the Fourier series, a periodic signal can be expanded into a summation of sines and cosines. The weighting of these sines and cosines are given by the a_n and b_n coefficients. These coefficients are found by integrating (over one period) the function multiplied by the sine or cosine associated with that coefficient. The sine and cosine terms are all harmonically related to the fundamental frequency, ω_0 . The $a_0/2$ term is simply the average (DC) value of the waveform and can often be found by inspection.

It may be inconvenient to work with separate sine and cosine terms, so the two terms can be combined into one sinusoid with an appropriate magnitude and phase angle.

$$x(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \sqrt{a_n^2 + b_n^2} \cos(2\pi n f_0 t + \theta_n) \quad (3-7)$$

¹ We will take a less than rigorous mathematical approach.

² The reader should be aware that there are several different ways of defining the Fourier series, with subtle differences in formulation.

where

$$\theta_n = \tan^{-1}(-b_n/a_n)$$

Alternatively, a_n and b_n can be combined into a complex coefficient that gives the complex form of the Fourier series. Instead of sines and cosines, a complex exponential is used:

$$x(t) = \sum_{n=-\infty}^{\infty} c_n e^{j2\pi n f_0 t} \quad (3-8)$$

where

$$c_n = \frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-j2\pi n f_0 t} dt \quad (3-9)$$

The two Fourier series representations are related by

$$c_n = (a_n - j b_n)/2 \quad (3-10)$$

The complex coefficient can also be expressed in magnitude/phase format:

$$c_n = |c_n| \angle \theta_n \quad (3-11)$$

Note that the complex form of the Fourier series is usually shown with n ranging from negative infinity to positive infinity whereas the original form restricts n to positive values. The complex form is chosen in anticipation of the Fourier transform, which includes negative frequencies. The factor of 2 that appears in equation (3-10) accounts for the presence of twice as many terms (both positive and negative) in the complex form. Frequency domain representations that include only positive frequencies are called *single sided*; those that include both positive and negative frequencies are called *double sided*.

3.3 Fourier Series of a Square Wave

As an example of the significance and utility of the Fourier series, the coefficients of a square wave will be determined. In addition, the square wave is a common signal in electrical systems (Figure 3-2a).

$$\begin{aligned} a_n &= \frac{2}{T} \int_{-T/2}^{T/2} x(t) \cos(2\pi n f_0 t) dt \\ &= \frac{2}{T} \int_{-T/2}^0 (-1) \cos(2\pi n f_0 t) dt + \frac{2}{T} \int_0^{T/2} (1) \cos(2\pi n f_0 t) dt \\ &= \frac{2}{T} \left[-\frac{1}{2\pi n f_0} \sin(2\pi n f_0 t) \Big|_{-T/2}^0 + \frac{1}{2\pi n f_0} \sin(2\pi n f_0 t) \Big|_0^{T/2} \right] \\ &= 0 \end{aligned}$$

$$\begin{aligned}
 b_n &= \frac{2}{T} \int_{-T/2}^{T/2} x(t) \sin(2\pi n f_0 t) dt \\
 &= \frac{2}{T} \int_{-T/2}^0 (-1) \sin(2\pi n f_0 t) dt + \frac{2}{T} \int_0^{T/2} (1) \sin(2\pi n f_0 t) dt \\
 &= \frac{2}{T} \left[\frac{1}{2\pi n f_0} \cos(2\pi n f_0 t) \Big|_{-T/2}^0 - \frac{1}{2\pi n f_0} \cos(2\pi n f_0 t) \Big|_0^{T/2} \right] \\
 &= \frac{1}{n\pi} (2 - 2 \cos n\pi) \\
 &= \frac{4}{n\pi} \quad \text{for } n \text{ odd} \\
 &= 0 \quad \text{for } n \text{ even}
 \end{aligned}$$

The Fourier series for the square wave is

$$x(t) = \frac{4}{\pi} \sin(2\pi f_0 t) + \frac{4}{3\pi} \sin(6\pi f_0 t) + \frac{4}{5\pi} \sin(10\pi f_0 t) + \dots \quad (3-12)$$

Therefore, the ideal square wave has only odd harmonics. With the particular phase chosen for the square wave, the a_n (cosine) terms are all zero, while the odd b_n (sine) terms remain nonzero. If the phase of the square wave were changed relative to $t = 0$, a_n could be nonzero but only for the odd harmonics. Similarly, at just the right phase b_n could become zero.

The square wave and its harmonics can be examined graphically, which helps show their relationship. Figure 3-2b shows the first three harmonics of the square wave. Figures 3-2c–f show a square wave constructed from a finite number of its harmonics. Note how the harmonics tend to fill in the square wave as each additional harmonic is added to the plot. It takes an infinite number of harmonics to produce a perfect square wave, but in practice the

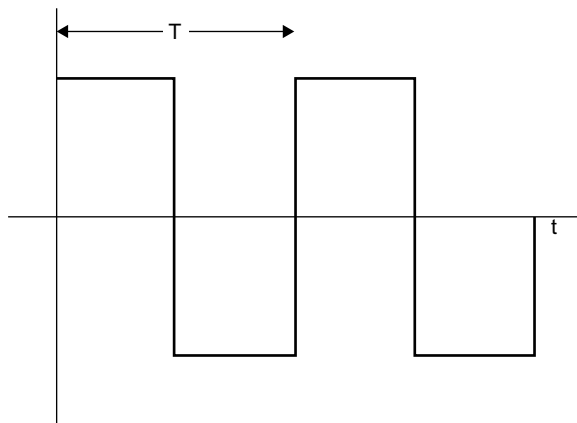


Figure 3-2a The square wave.

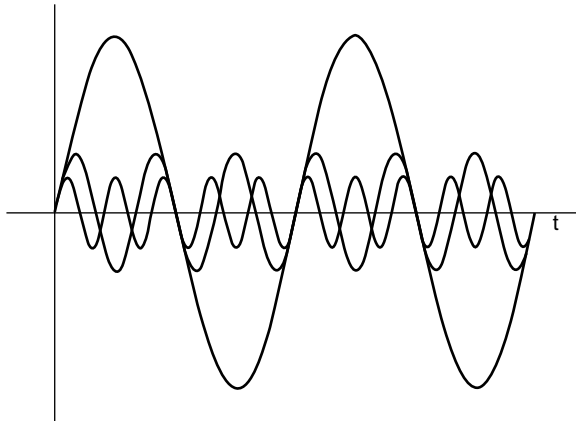


Figure 3-2b The fundamental, third harmonic, and fifth harmonic of the square wave.

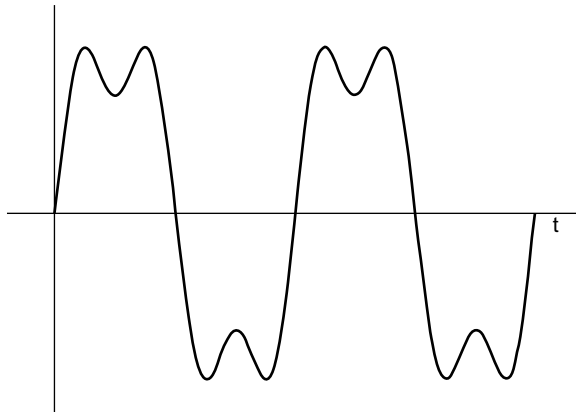


Figure 3-2c The square wave with only the fundamental and third harmonic included.

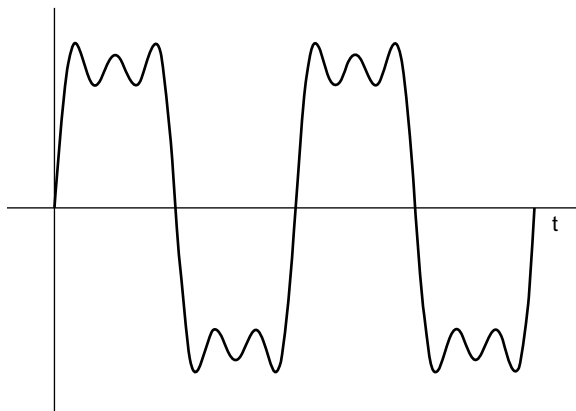


Figure 3-2d The square wave with up to the fifth harmonic included.

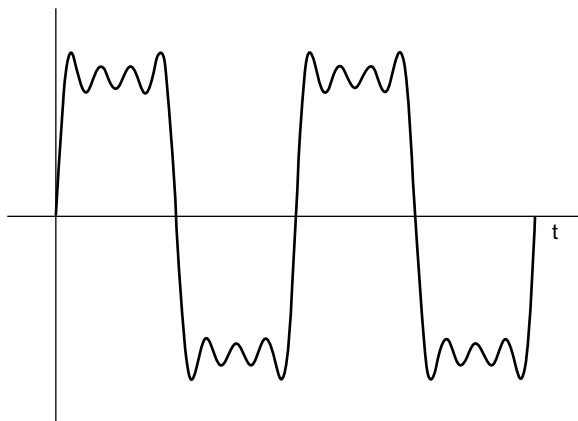


Figure 3-2e The square wave with up to the seventh harmonic included.

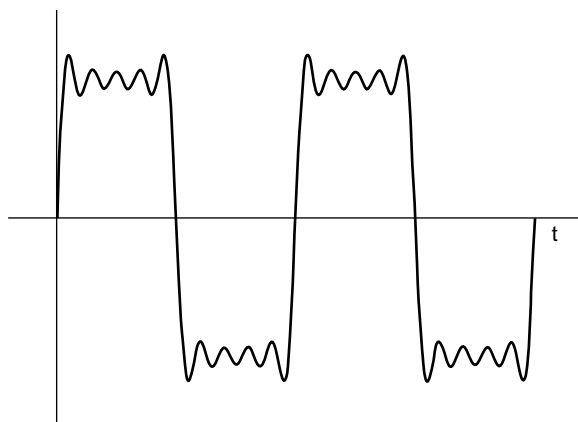


Figure 3-2f The square wave with up to the ninth harmonic included.

higher harmonics are often ignored since their amplitudes are small compared with the fundamental.

Note how the harmonics (which are sine terms in this case) are just the right phase to fill in the square wave. Had the square wave been shifted to the left by 90° , the sine terms would have been useless in filling in the square wave shape and cosine terms would have been prescribed by the previous mathematics. If the complex form of the Fourier series was used, the magnitude of c_n would remain the same with changes in the waveform's phase but the phase of c_n would change.

Although the Fourier series is a mathematical technique, an intuitive feel can be acquired by looking at the waveform graphically.

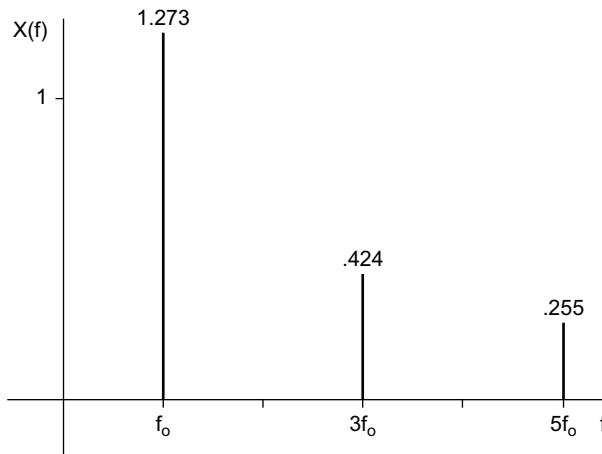


Figure 3-3 The frequency domain representation of a square wave showing the fundamental, third harmonic, and fifth harmonic.

Another way to show the square wave in the frequency domain is to plot the amplitude of each harmonic as voltage versus frequency (Figure 3-3). Since each harmonic appears as a single vertical line, they are called *spectral lines* and such a frequency domain plot is called a *line spectra*. The Fourier series will always result in a frequency domain representation with only line spectra, since the series form includes only the fundamental and harmonic frequencies. This contrasts with frequency domain spectrums, which are continuous and will be encountered later.

Practical Considerations

The amplitude of each odd harmonic of the square wave is $1/n$ times the fundamental, where n is the harmonic number. When examined on a spectrum analyzer, we can expect to see each harmonic reduced by this factor. The amplitude of the harmonic corresponds to its voltage, so using decibels, the n -th harmonic will be $20 \log(1/n)$ decibels relative to the fundamental.

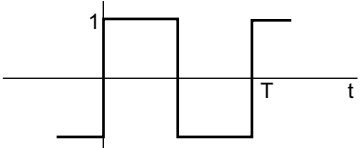
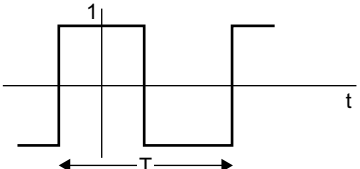
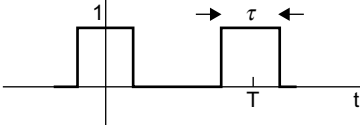
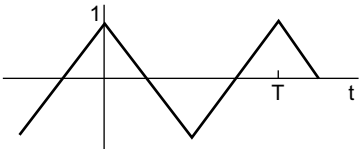
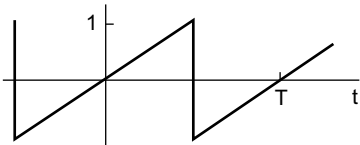
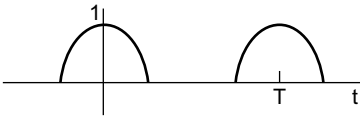
Ideally, the even harmonics are nonexistent. This is true if the square wave being measured is perfect. If the waveform is not perfectly symmetrical or has other forms of distortion, the even harmonics will be nonzero, which is true of most practical measurements. A spectrum analyzer can easily detect such imperfections in a square wave even though the square wave may look undistorted when measured with an oscilloscope.

By the nature of the mathematical formula used, a_n and b_n in the Fourier series represent the zero-to-peak value of the particular harmonic. Spectrum analyzers, however, are normally calibrated to measure the root mean square (RMS) value of a spectral line, usually expressed in dBm or RMS volts. Thus, to correlate the Fourier series representation and the typical measured result, it is necessary to multiply the Fourier series coefficient by $1/\sqrt{2}$ and, if desired, convert to dBm.

3.4 Fourier Series of Other Waveforms

The Fourier series representation of other periodic waveforms can be determined using the techniques given. For convenience, the Fourier series representations of some common waveforms are tabulated in Table 3-1.

Table 3-1 Fourier Series of Waveforms

 <p>Square Wave (Odd)</p>	$\frac{4}{\pi} \sum_{\substack{n=1 \\ \text{odd}}}^{\infty} \frac{1}{n} \sin\left(\frac{2\pi nt}{T}\right)$
 <p>Square Wave (Even)</p>	$\frac{4}{\pi} \sum_{\substack{n=1 \\ \text{odd}}}^{\infty} \frac{(-1)^{(n-1)/2}}{n} \cos\left(\frac{2\pi nt}{T}\right)$
 <p>Pulse Train</p>	$\frac{\tau}{T} + \frac{2\tau}{T} \sum_{n=1}^{\infty} \frac{\sin\left(\frac{\pi n \tau}{T}\right)}{\frac{\pi n \tau}{T}} \cos\left(\frac{2\pi nt}{T}\right)$
 <p>Triangle Wave</p>	$\frac{8}{\pi^2} \sum_{\substack{n=1 \\ \text{odd}}}^{\infty} \frac{1}{n^2} \cos\left(\frac{2\pi nt}{T}\right)$
 <p>Sawtooth Wave</p>	$\frac{2}{\pi} \sum_{\substack{n=1 \\ \text{odd}}}^{\infty} \frac{(-1)^{n+1}}{n} \sin\left(\frac{2\pi nt}{T}\right)$
 <p>Half-Wave Cosine</p>	$\frac{1}{\pi} + \frac{1}{2} \cos\left(\frac{2\pi t}{T}\right) - \frac{2}{\pi} \sum_{\substack{n=2 \\ \text{odd}}}^{\infty} \frac{(-1)^{n/2}}{n^2 - 1} \cos\left(\frac{2\pi nt}{T}\right)$

Example 3.1

Determine the amplitude and frequency of the fundamental of the waveform shown in Figure 3-4. If the signal is a voltage present across 50Ω , what is the power level in dBm of the fundamental? Determine the amplitude of the second harmonic and express it in decibels relative to the fundamental.

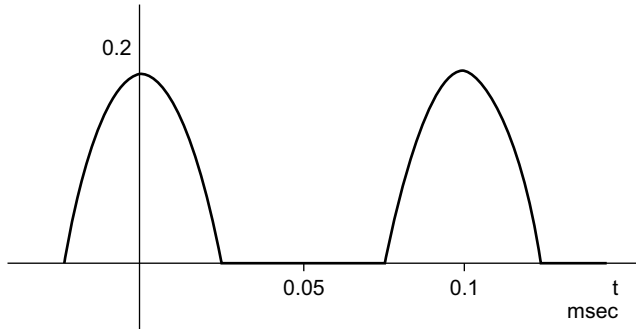


Figure 3-4 The half-wave rectified sine wave is a periodic signal.

From Table 3-1, the first few terms of the Fourier series of the half cosine wave are

$$x(t) = \frac{1}{\pi} + \frac{1}{2} \cos(2\pi t/T) + \frac{2}{3\pi} \cos(4\pi t/T)$$

The waveform shown in Figure 3-4 has a peak voltage of 0.2 V, so the Fourier series is multiplied by 0.2.

$$x(t) = 0.2 \left[\frac{1}{\pi} + \frac{1}{2} \cos(2\pi t/T) + \frac{2}{3\pi} \cos(4\pi t/T) \right]$$

The frequency of the fundamental is $1/T = 1/(0.1 \text{ msec}) = 10 \text{ kHz}$.

The amplitude of the fundamental is $0.2(1/2) = 0.1 \text{ V}$ zero-to-peak. Converting this value to RMS gives $0.707 \times 0.1 = 0.0707 \text{ V}$. Using equation (2-12), the amplitude in dBm (50Ω) is $20 \log(0.0707/0.223) = -9.98 \text{ dBm}$.

The amplitude of the second harmonic is $0.2(2/3\pi) = 0.0424 \text{ V}$ zero-to-peak, or 0.030 V RMS. Expressed as decibels relative to the fundamental, the second harmonic is $20 \log(0.030/0.0707) = -7.45 \text{ dB}$.

3.5 Fourier Transform

Although the Fourier series representation of a signal is very powerful, it is limited to periodic signals. Signals that are not periodic may be represented in the frequency domain by the Fourier transform. The Fourier transform of a time domain signal $x(t)$ is

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt \quad (3-13)$$

where

$$\begin{aligned} X(f) &= \text{frequency domain representation of the signal} \\ x(t) &= \text{time domain representation of the signal} \\ f &= \text{frequency} \end{aligned}$$

The Fourier transform transforms a time domain signal into a continuous frequency domain representation. Recall that the Fourier series representation, by definition, contains only the fundamental frequency and its harmonics. Not only are these discrete frequencies, but they are also harmonically related. The Fourier transform can represent discrete frequencies but is more often used to represent continuous functions in the frequency domain. Thus, a one-time event (e.g., a pulse) in the time domain can also be represented in the frequency domain.

Mathematically, the frequency domain representation is a complex function, containing both magnitude and phase information. Although many spectrum measurements are performed just using the magnitude of the signal, phase is required for a full representation of the signal.

3.6 Fourier Transform of a Pulse

As an example and because it is a common electrical signal, we will determine the Fourier transform of a single pulse (Figure 3-5a).

$$\begin{aligned} X(f) &= \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt \\ &= \int_{-T/2}^{T/2} e^{-j2\pi ft} dt = \left. \frac{e^{-j2\pi ft}}{-j2\pi f} \right|_{-T/2}^{T/2} \\ &= \frac{e^{j\pi f T} - e^{-j\pi f T}}{j2\pi f} = T \frac{\sin(\pi T f)}{\pi T f} \end{aligned} \quad (3-14)$$

The frequency domain representation for a pulse is of the form $(\sin x)/x$ (Figure 3-5b). Notice that the function is continuous and extends over the entire frequency axis, both positive and negative. Thus, a perfect pulse occupies an infinite bandwidth. However, the amplitude of the frequency content tends to decrease with increasing frequency, and, in practice, a finite bandwidth can be assumed.

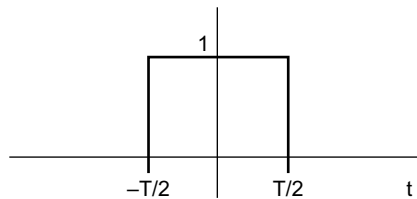


Figure 3-5a A single time domain pulse.

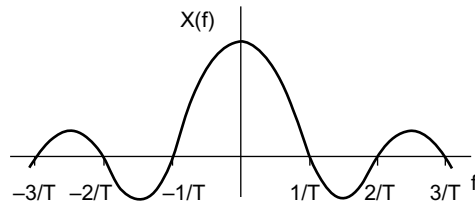


Figure 3-5b The spectrum of a single pulse.

The zero crossings of $X(f)$ are often used as a means of estimating the bandwidth of the pulse. Most of the pulse's energy is in the main lobe, which exists at frequencies below $f = 1/T$. As the width of the time domain pulse is decreased, T becomes smaller. In the frequency domain, as T becomes smaller, the first zero crossing moves out to a higher frequency. Therefore, the narrower the pulse, the wider the bandwidth in the frequency domain. This should make sense intuitively, since a narrower pulse requires higher frequency content to recreate the waveform in the time domain. This is true of signals in general—the faster the voltage changes in the time domain, the wider the bandwidth in the frequency domain.

3.7 Inverse Fourier Transform

The *inverse Fourier transform* converts the frequency domain representation (obtained by the Fourier transform) back into the time domain representation. The inverse transform is given by

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{2\pi ft} df \quad (3-15)$$

Thus, Fourier theory provides a means of transforming a time domain signal into the frequency and (just as important) provides a means of getting the frequency domain representation back into the time domain.

The time domain and frequency domain representations of a signal are known as *transform pairs*. They are unique in that each time domain representation has only one frequency domain representation and vice versa. A table of common Fourier transform pairs is given in Table 3-2.

3.8 Fourier Transform Relationships

Many mathematical operations in the time domain have a corresponding operation in the frequency domain. These relationships are often used to reduce the difficulty of finding a transform of a particular function. These relationships also lend insight into how the time and frequency domain relate. Table 3-3 is a compilation of commonly used Fourier transform relationships.

Table 3-2 Fourier Transform Pairs

	$x(t)$	$\delta(t)$	$x(f)$	1
Unit Impulse				
Constant				$\delta(f)$
Unit Step				$\frac{1}{2}\delta(f) + \frac{1}{j2\pi f}$
Pulse				$\frac{T \sin(\pi T f)}{\pi T f}$

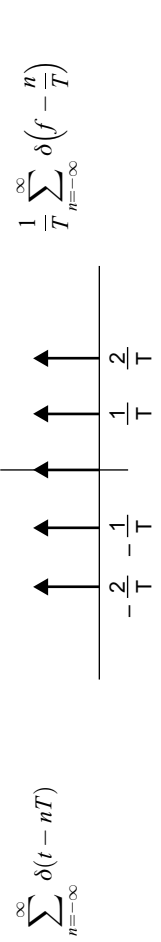
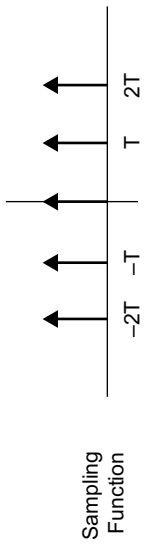
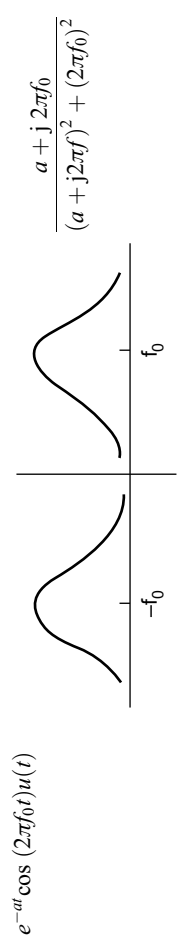
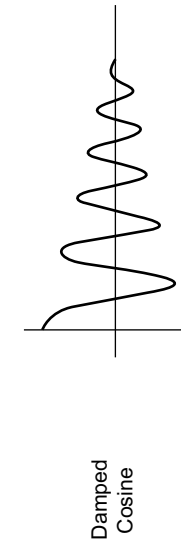
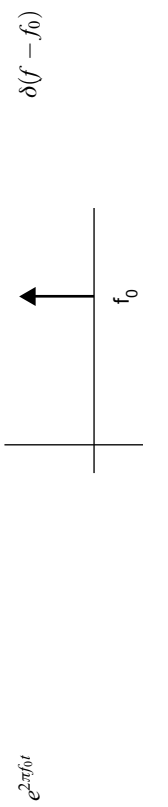
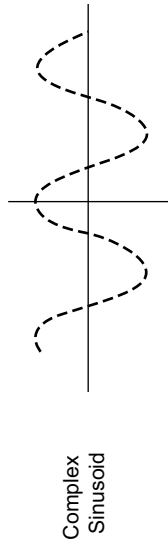
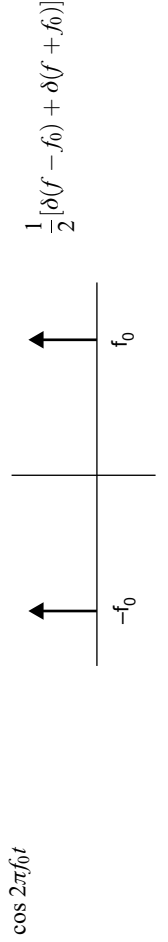
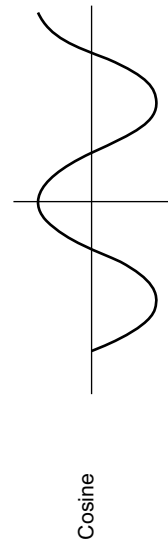
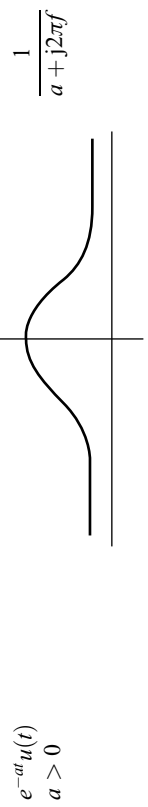
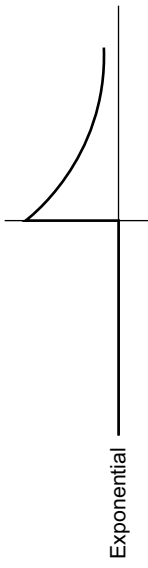


Table 3-3 Properties of the Fourier Transform

	$x(t)$	$X(f)$
Magnitude scaling	$Ax(t)$	$AX(f)$
Time scaling	$x(at)$	$\frac{1}{ a }X\left(\frac{f}{a}\right)$
Linearity	$x_1(t) + x_2(t)$	$X_1(f) + X_2(f)$
Time delay	$x(t - t_0)$	$e^{-j2\pi ft_0}X(f)$
Time derivative	$\frac{d^n}{dt^n}x(t)$	$(j2\pi f)^n X(f)$
Modulation	$x(t) \cos(2\pi f_0 t)$	$\frac{1}{2}[X(f - f_0) + X(f + f_0)]$
Complex modulation	$e^{j2\pi f_0 t} x(t)$	$X(f - f_0)$
Multiplication	$x_1(t) x_2(t)$	$\int_{-\infty}^{\infty} X_1(\lambda) X_2(f - \lambda) d\lambda$
Convolution	$\int_{-\infty}^{\infty} x_1(\lambda)x_2(t - \lambda) d\lambda$	$X_1(f) X_2(f)$
Symmetry	$X(t)$	$x(-f)$

3.9 Discrete Fourier Transform

The Fourier transform is mostly an analysis tool, a powerful means of understanding how signals behave in a system. It is not directly used in a measurement system to produce the frequency domain representation of a signal. The DFT is a discrete version of the Fourier transform. It allows a sampled time domain signal to be transformed into a sampled frequency domain form. Digitizing a real-world signal in the time domain and performing a DFT produces the frequency domain representation of the signal. Thus, the DFT goes beyond being just an analysis tool to being a way to implement the measurement.

We had previously introduced the complex form of the Fourier series. It is rewritten here with a slight change of variable (the period T has become t_p and harmonic number n is replaced by k).

$$c_k = \frac{1}{t_p} \int_{-t_p/2}^{t_p/2} x(t) e^{-j2\pi k f_0 t} dt \quad (3-16)$$

Consider the periodic waveform shown in Figure 3-6a. Suppose that a sampled version of one period of this waveform is available (Figure 3-6b). The Fourier series can be applied to this sampled waveform, with the minor change that the time domain waveform is not continuous. This means that $x(t)$ will be replaced by $x(nT)$, where T is the time between

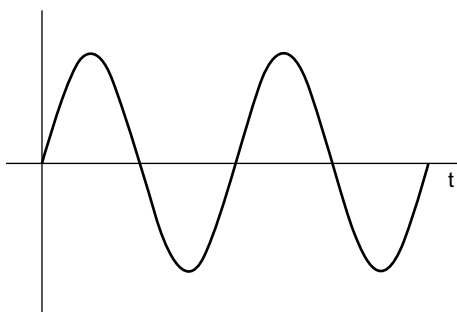


Figure 3-6a A periodic signal to be sampled.

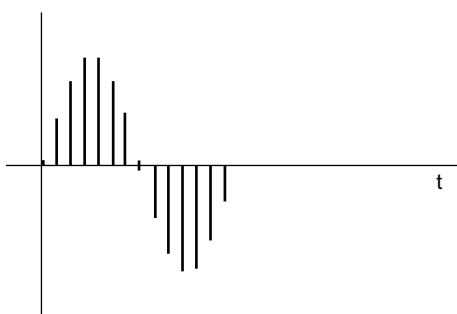


Figure 3-6b The sampled version of one period of the signal.

samples. Also, instead of an integration, a discrete summation of the sampled waveform will be performed with the result multiplied by the time between samples, T .

$$c_k = \frac{T}{t_p} \sum_{n=0}^{N-1} x(nT) e^{-j2\pi k f_0 n T} \quad (3-17)$$

Note that the range of n was chosen to be from 0 to $N - 1$, producing N samples. This particular range is not mandatory but is customary for defining the DFT. The fundamental frequency, f_0 , is also the spacing between the discrete frequency points. We will rename f_0 as F and attempt to provide consistent notation. Finally, the DFT is usually defined to be N times the complex Fourier series coefficient.³

$$X(kF) = N c_k \quad (3-18)$$

$$X(kF) = \frac{NT}{t_p} \sum_{n=0}^{N-1} x(nT) e^{-j2\pi k F n T} \quad (3-19)$$

³ This is only a scale factor and does not affect the frequency content of the DFT. In instrumentation use, the DFT must have appropriate scale factors added to properly calibrate the instrument.

Since the number of samples, N , times the sample time, T , equals the period, t_p , the equation simplifies to give the common form of the DFT.

$$X(kF) = \sum_{n=0}^{N-1} x(nT)e^{-j2\pi kFnT} \quad (3-20)$$

where

- N = number of samples
- F = spacing of the frequency domain samples
- T = sample period in the time domain

In instrumentation use, the input to the DFT is a record of time domain data obtained by sampling the signal being analyzed. The sample rate, f_s , is equal to $1/T$. After N time domain samples are collected, the DFT algorithm uses the time domain samples to produce N frequency domain samples, spaced F Hz apart. These N frequency domain samples are not totally independent. The set of samples numbered less than $N/2$ are redundant with the samples numbered above $N/2$. For an N point DFT, only the samples up to and including $N/2$ frequency domain points are normally retained. In general, these points are complex numbers, providing vector information.

Remember that we started the derivation with the Fourier series and not the Fourier transform. As the number of time domain samples, N , increases (and therefore the number of frequency domain samples increases), we can stop considering the DFT as a small set of spectral lines and start thinking of it as a good approximation to the continuous Fourier transform.

The inverse of the DFT, the *inverse discrete Fourier transform* (IDFT), is given by

$$x(nT) = \frac{1}{N} \sum_{k=0}^{N-1} X(kF) e^{j2\pi F T k n} \quad (3-21)$$

The IDFT provides a means for converting the discrete frequency domain information back into a discrete time domain waveform. As one might imagine, the DFT and IDFT have properties that are very similar to their continuous counterparts.

3.10 Limitations of the DFT

The DFT is only an approximation to the Fourier transform. It differs from the continuous Fourier transform in several important ways.

Obviously, due to the quantized nature of the DFT, it is valid at only certain frequencies. The frequency resolution of the DFT can be increased by using a larger number of samples.

The theory behind the DFT implicitly assumes that the waveform was periodic. Whether this is the case or not, the mathematics of the DFT will treat the sampled waveform as if it repeats. This causes a phenomenon known as leakage, which is an important limitation of the DFT, but one that can be minimized by proper use of time domain windowing. Leakage is discussed further in Chapter 4.

Since the DFT is performed with digital arithmetic, it is subject to the limitations imposed by the particular algorithm chosen. In particular, finite arithmetic effects due to the number of bits used can limit the dynamic range and noise performance of the DFT.

3.11 Fast Fourier Transform

The *fast Fourier transform* (FFT) is a very quick and efficient algorithm for implementing a DFT. The original basis for the FFT was developed by J. W. Cooley and J. W. Tukey in 1965. Although it is often implied that there is just one FFT, in reality an entire class of algorithms are commonly referred to as the FFT. An FFT algorithm gains a significant speed advantage over the DFT by carefully selecting and organizing intermediate results. Ignoring finite arithmetic effects, the results are the same whether an FFT or a DFT is used.

The number of computations required for a DFT is on the order of N^2 , where N is the number of samples, or record length. The FFT, on the other hand, requires $N \log_2 N$ computations (\log_2 indicates the base 2 logarithm). The most common FFT algorithms require N to be a power of 2. A typical record length in a spectrum analyzer might be 2^{10} , or 1024. This means a DFT would require over 1 million computations, whereas an FFT would require only 10,240 computations. Assuming all computations take the same amount of time, the FFT could be computed in less than 1% of the DFT computation time. Clearly, this is a substantial time savings and explains why the FFT dominates in instrumentation use.

Examining the details of how and why an FFT is implemented is beyond the scope of this book. For our purposes, we will consider the FFT to be simply an efficient implementation of a DFT. For more information, see Oppenheim and Schaffer (1975).

3.12 Relating Theory to Measurements

When the instrument user attempts to relate Fourier theory to an actual measurement, some notable differences will appear. The major differences are summarized here:

1. The spectrum analyzer normally shows a one-sided spectrum, whereas the Fourier transform and perhaps the Fourier series (depending on which form is used) show a two-sided spectrum.
2. The frequency resolution (resolution bandwidth) of the spectrum analyzer determines the width and shape of discrete spectral lines. Ideally, the lines are infinitely thin, but they appear with a finite width due to the resolution bandwidth of the analyzer.
3. Other distortion and noise effects generated internal to the spectrum analyzer will affect the measurement. For example, the noise floor of the analyzer may obscure low-level frequency components or distortion products may appear as additional spectral lines.

In particular, it can be a problem relating the amplitude predicted by Fourier theory to the measured amplitude. In an attempt to reconcile theory and measurement, let us consider a simple, but instructive case: the cosine. We apply both the Fourier series and the Fourier transform to this signal and then compare the results with a practical spectrum measurement.

Consider the time domain waveform

$$v(t) = V_0 \cos 2\pi f_0 t \quad (3-22)$$

The RMS value, as measured by an RMS-reading voltmeter, would be 0.707 times the zero-to-peak value. This value should agree with the spectrum analyzer measurement:

$$V_{\text{RMS}} = 0.707 V_0 \quad (3-23)$$

The Fourier series for this voltage waveform can easily be found by inspection.

$$v(t) = a_1 \cos 2\pi f_0 t \quad (3-24)$$

where

$$a_1 = V_0$$

This implies a single spectral line at f_0 , with a zero-to-peak amplitude of V_0 . From Table 3-2, the Fourier transform of the waveform is

$$V(f) = \frac{V_0}{2} [\delta(f - f_0) + \delta(f + f_0)] \quad (3-25)$$

Since the Fourier transform is a two-sided representation, with both positive and negative frequencies, the frequency domain representation indicates two impulse functions: one at $+f_0$ and the other at $-f_0$. The amplitudes of each of these impulse functions is $V_0/2$. This amplitude is doubled to convert the double-sided amplitude to the equivalent single-sided amplitude. Thus, the zero-to-peak amplitude equal to V_0 agrees with the Fourier series analysis and, if multiplied by 0.707 to obtain the RMS value, agrees with the voltmeter reading and a spectrum analyzer reading.

3.13 Finite Measurement Time

The discussion of the Fourier series and the Fourier transform both involved integrals that cover all time, that is, from $-\infty$ to $+\infty$. Therefore, to ascertain correctly the frequency domain representation of a signal, the time domain function must be known for all time. For theoretical analysis, this does not present a problem, but real-world measurements occur in a finite time. Normally, the spectrum analyzer user simply performs the measurement over some convenient time interval and assumes that the time interval chosen adequately represents the signal. Mathematically speaking, the signal is assumed to be *stationary*.⁴

The characteristics of many signals are constant over time in which case such an assumption is justified. By definition, a periodic signal repeats over and over again for all time, producing a constant spectrum. Some other signals change quite rapidly and should not be assumed to have constant spectrums. As an example consider a radio transmitter. If the modulating signal is a person's voice, the spectrum of the signal will change quickly and unpredictably as the radio operator speaks different words. A measurement taken at any particular time will not represent the signal over all time. However, if a constant audio tone modulates the radio signal, the spectrum is constant.

When measuring a signal's spectrum, we should consider the possibility that the signal's spectral content may be varying. If this variation is slow compared with the duration of the measurement, it is not of concern. However, if the signal varies fast enough, the spectrum analyzer may not produce the desired result. In particular, traditional swept spectrum

⁴ A signal is stationary if its statistical nature does not change with time, which implies that its spectrum is constant. For a more rigorous discussion, see Oppenheim and Willsky (1996).

analyzers can have very slow sweep rates, depending on the measurement setup. These analyzers may not be suitable for measuring fast changing signals. FFT-based analyzers and real-time spectrum analyzers acquire the measured signal much faster and are much more effective in capturing changes in spectral content. This will be explored further in Chapters 4 and 5.

Bibliography

Agilent Technologies. "Fundamentals of Signal Analysis," Application Note 243, Publication Number 5952-8898E, June 2000.

Brigham, E. Oran. *The Fast Fourier Transform and Its Applications*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1988.

Cooley, J. W., and Tukey, J. W., "An Algorithm for the Machine Calculation of Complex Fourier Series," *Math. Computation*, Vol. 19 (1965).

Hayt, William H. Jr., Jack E. Kemmerly, and Steven M. Durbin. *Engineering Circuit Analysis*, 7th ed. New York: McGraw-Hill Book Company, 2007.

McGillem, Clare D., and George R. Cooper. *Continuous and Discrete Signal and System Analysis*. New York: Holt, Rinehart and Winston, Inc., 1974.

Oppenheim, Alan V., and Alan S. Willsky. *Signals and Systems*, 2d ed. Upper Saddle River, NJ: Prentice Hall, Inc., 1996.

Oppenheim, Alan V., and Ronald W. Schaffer. *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1975.

Ramirez, Robert W. *The FFT, Fundamentals and Concepts*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1985.

Schwartz, Mischa. *Information, Transmission, Modulation, and Noise*, 3rd ed. New York: McGraw-Hill Book Company, 1980.

Ziemer, Rodger E., William H. Tranter, and D. Ronald Fannin. *Signals and Systems: Continuous and Discrete*, 4th ed. Upper Saddle River, NJ: Prentice Hall, Inc., 1998.

Fast Fourier Transform Analyzers

The *fast Fourier transform* (FFT) can be used to implement a spectrum analyzer by digitizing the input waveform and performing an FFT on the time domain signal to obtain the frequency domain representation. What seems to be a simple measurement technique often turns out to be much more complicated in practice. Given reasonable computational power, usually in the form of a digital signal processor (DSP), field programmable gate array (FPGA) or custom integrated circuit, the FFT analyzer can provide significant speed improvement over the more traditional swept analyzer. The classic FFT analyzer (also called a *dynamic signal analyzer*) covers the frequency range from DC up to a few hundred kilohertz. These analyzers are typically applied to audio and mechanical measurements.¹

4.1 The Bank-of-Filters Analyzer

The bank-of-filters technique is not common in general electronic instrumentation but has been used in some applications such as low-frequency audio meters (1/3-octave spectrum analyzers). This technique is included here to provide a theoretical base for discussing more practical spectrum analyzer block diagrams.

One simple approach to implementing a spectrum analyzer is to connect a bank of electronic filters together, each with its own output device (Figure 4-1). For a small number of filters, this technique has the advantage of simplicity. Also, this measurement technique is quite fast and can result in a real-time measurement system.

Each of the electronic filters is a band-pass filter tuned to a different center frequency. The bandwidths and center frequencies of the filters are aligned as shown in Figure 4-2 to provide complete coverage of the frequency range of interest with minimal overlap of filter shapes. Ideally, infinitely steep “brick wall” filters would be used to provide zero overlap between filter passbands. The outputs of the filters are connected to detectors that convert the AC (sine wave) signal into a DC level, which is displayed by a meter. Alternatively, the detector outputs could be multiplexed together and plotted on a graphical display.

¹ Mechanical measurements, including vibration and structural analysis, represent an important use of FFT-based spectrum analyzers and are covered in more depth by Agilent Technologies (1997, 2000).

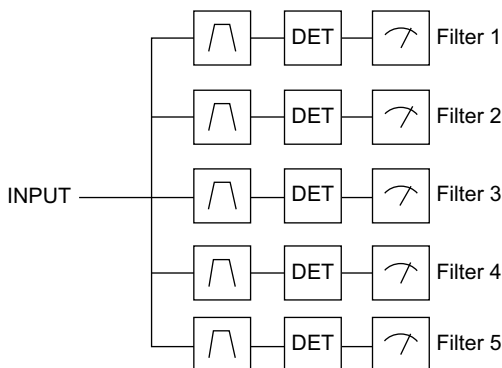


Figure 4-1 The bank-of-filters spectrum analyzer uses a set of filters to determine the frequency content of a signal.

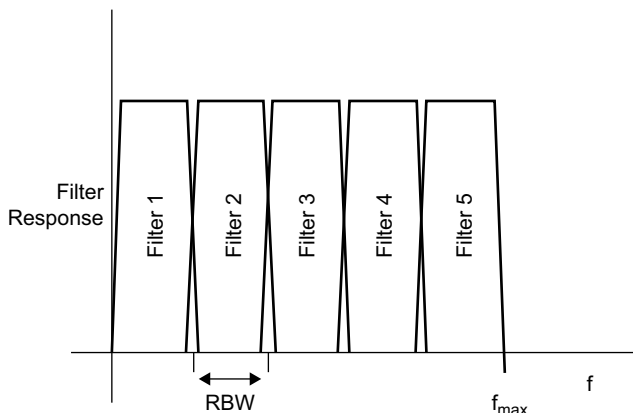


Figure 4-2 The filters are adjacent in the frequency domain, aligned for minimum overlap.

4.2 Frequency Resolution

Each filter is designed to pass only one small range of frequencies onto the detector. Thus, each filter/detector/meter combination displays the energy present over that particular range of frequencies. If two frequencies are present within the same filter, both of them will affect the meter reading. (The exact meter reading will depend on the type of detector used.) The analyzer cannot resolve two signals in the same filter. Thus, the resolution bandwidth (RBW) determines the frequency resolution of the analyzer.

For example, consider Figure 4-3. Frequency components f_1 and f_2 appear in the passband of the same filter. Therefore, they cannot be resolved in frequency. Frequency components f_3 and f_4 , on the other hand, do not appear within the same filter, and each will be measured individually. The frequency of f_3 and f_4 are known to the extent that they are within the passband of their respective filters. Thus, their frequencies are known to within the RBW.

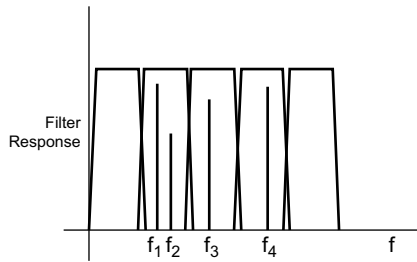


Figure 4-3 The bank-of-filters analyzer response shown with some representative spectral lines.

Assuming extremely sharp brickwall filters precisely positioned such that the edges of their passbands just touch, the resolution bandwidth of a bank-of-filters analyzer is given by

$$RBW = f_{\max}/m \quad (4-1)$$

where

$$\begin{aligned} f_{\max} &= \text{maximum frequency of the analyzer} \\ m &= \text{number of filters} \end{aligned}$$

This equation can be used to explain the major limitation of the bank-of-filters analyzer. Suppose that a spectrum analyzer with a frequency resolution (resolution bandwidth) of 100 Hz must cover the 0 to 100 kHz frequency range. The number of filters required is

$$m = f_{\max}/B_{\text{res}} = 100 \text{ kHz}/100 = 1000 \quad (4-2)$$

Not only would this be a large number of filters to implement, but also building a steep-walled 100 Hz wide filter with a center frequency near 100 kHz would be very difficult. For this reason, the bank-of-filters analyzer is used mainly where a much wider resolution bandwidth is acceptable.

4.3 The FFT Analyzer

As mentioned previously, the fast Fourier transform can be used to determine the frequency domain representation of a time domain signal. The signal must be digitized in the time domain; then the FFT algorithm is executed to find the spectrum. Figure 4-4 shows a simplified block diagram of an FFT analyzer. The input signal is first passed through a variable attenuator to provide various measurement ranges. Then the signal is low-pass filtered to remove undesirable high-frequency content that is beyond the frequency range of the instrument. The waveform is sampled and converted to digital form by the combination of the sampler circuit and the analog-to-digital converter (ADC). The microprocessor (or other digital circuitry) receives the sampled waveform, computes the spectrum of the waveform using the FFT, and writes the results on the display.

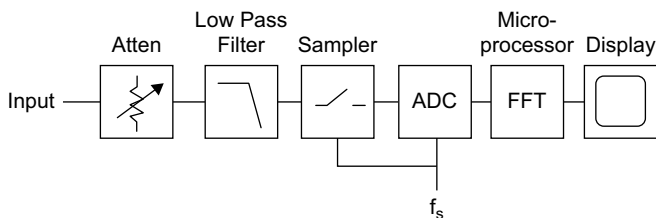


Figure 4-4 The simplified block diagram of the fast Fourier transform spectrum analyzer.

The FFT analyzer accomplishes the same thing that the bank-of-filters analyzer does, but without the need for many band-pass filters. Instead, the FFT analyzer uses digital signal processing to implement the equivalent of many individual filters. When considering the operation of the FFT analyzer, it is appropriate to think in terms of a bank of parallel filters, each filtering a portion of the frequency spectrum. A typical FFT spectrum analyzer is shown in Figure 4-5.

Conceptually, the FFT approach is simple and straightforward: digitize the signal and compute the spectrum. In practice, some effects must be accounted for to make the measurement meaningful.

4.4 Sampled Waveform

In a sampled system, the time domain waveform (Figure 4-6a) is effectively multiplied by the sample function (Figure 4-6b) to produce the sampled waveform (Figure 4-6c).



Figure 4-5 A typical four-channel FFT-based spectrum analyzer. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

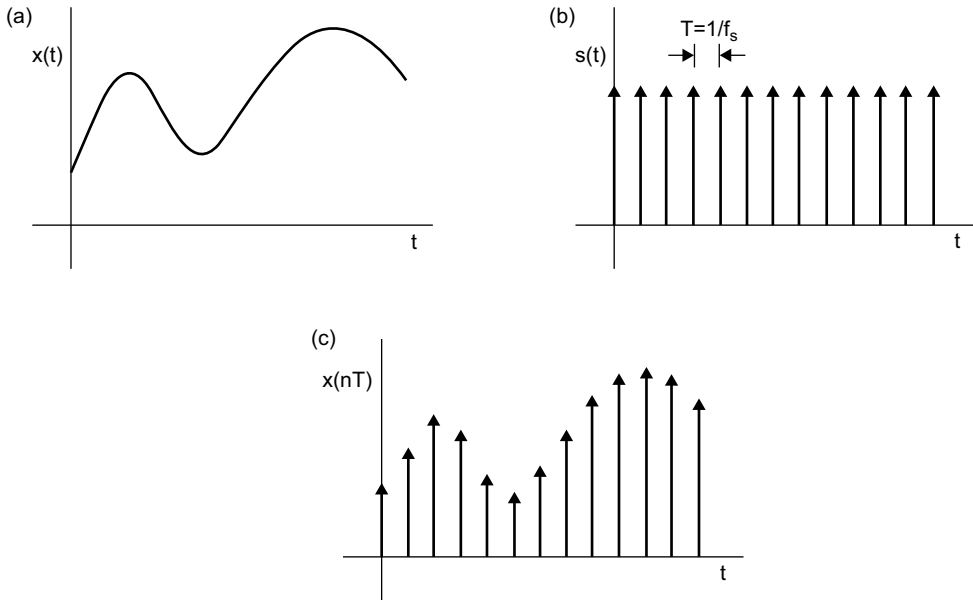


Figure 4-6 (a) A time domain waveform. (b) The sampling function. (c) The sampled waveform.

The sampling function is shown as a series of impulse functions, spaced at $T = 1/f_s$, where f_s is the sample rate of the system.

$$s(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT) \tag{4-3}$$

When these impulse functions are multiplied with the original waveform, they produce a new series of impulse functions with each one weighted according to the original waveform.

$$x(nT) = \sum_{n=-\infty}^{\infty} x(t) \delta(t - nT) \tag{4-4}$$

The sampled analog waveform is converted into a sequence of digital numbers using an ADC. The output of the ADC is an array or record of numbers representing the sampled waveform. The sampled and digitized version of the waveform still retains the shape and information content of the unsampled waveform, if the sample rate is sufficiently high.

4.5 Sampling Theorem

The waveform must be sampled often enough to produce a digitized time record that faithfully represents the original waveform. The *sampling theorem* states that a baseband signal must be sampled at a rate greater than twice the highest frequency present in the signal. The minimum acceptable sample rate is called the *Nyquist rate*. Thus,

$$f_s > 2f_{\max} \tag{4-5}$$

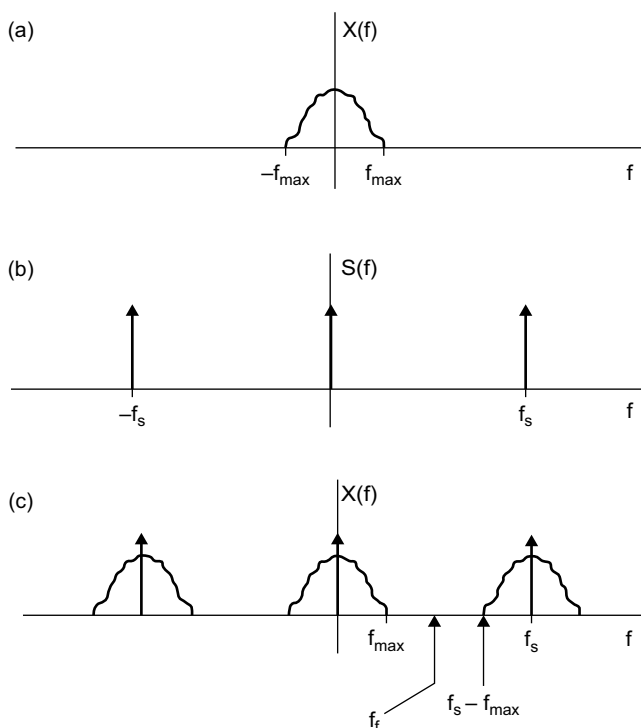


Figure 4-7 (a) The spectrum of the unsampled waveform. (b) The spectrum of the sampling function. (c) The spectrum of the sampled waveform.

where

$$f_s = \text{sample rate}$$

$$f_{\max} = \text{highest frequency of interest}$$

Figure 4-7a shows the frequency spectrum, $X(f)$, of a signal, $x(t)$, with a maximum frequency of f_{\max} . The frequency spectrum of the sampling function, as given by Table 3-1, is an infinite number of impulse functions spaced every f_s in frequency (Figure 4-7b). The spectrum of the sampled waveform can be derived by convolving² $X(f)$ with $S(f)$, which results in the original spectrum $X(f)$ appearing centered around each impulse function of $S(f)$ (Figure 4-7c).

This type of spectrum is always found in sampled systems—the baseband signal is repeated at integer multiples of the sample frequency. Notice that the spectrum between 0 and f_s is symmetrical about $f_s/2$, which is also called the *folding frequency*, f_f . The original signal can be recovered by applying a low-pass filter with a cutoff frequency of f_f , as long as the frequency content centered around f_s does not encroach on the baseband signal. Stated mathematically, the following condition must be met:

$$f_s - f_{\max} > f_f \tag{4-6}$$

² For a discussion of the fine points of convolution, see McGillem and Cooper (1974).

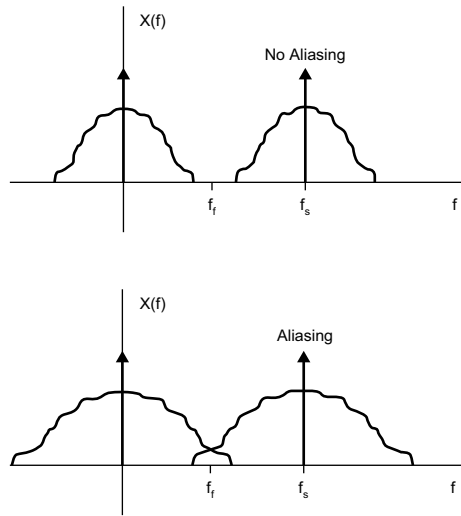


Figure 4-8 Aliasing occurs when the sample rate is not high enough.

which is just a restatement of the sampling theorem since

$$f_s - f_{\max} > f_s/2 \quad (4-7)$$

$$f_s/2 > f_{\max} \quad (4-8)$$

$$f_s > 2f_{\max} \quad (4-9)$$

Figure 4-8 shows the spectra of two sampled signals: one where the sampling theorem is met and another that violates the sampling theorem. Notice that when the sampling theorem is violated unwanted frequency components show up below f_f . This phenomenon is known as *aliasing*, since these undesirable frequency components appear under the alias of another (baseband) frequency.

To prevent aliasing in an FFT analyzer, two conditions must be met:

1. The input signal must be band limited. In other words, there must exist an f_{\max} above which no frequency components are present.³ This is usually accomplished by inserting a low pass filter, commonly known as an anti-alias filter, in the signal path. (This is the low-pass filter shown in Figure 4-4.)
2. The input signal must be sampled at a rate that satisfies the sampling theorem.

The sampling frequency required by the sampling theorem is the minimum theoretical value that can reconstruct the signal properly. In practice, it is necessary to use a sampling frequency somewhat higher than this value. Figure 4-9 shows the frequency response of a

³ In practice, frequency components above f_{\max} are allowed to exist but must be sufficiently attenuated so that they do not affect the measurement.

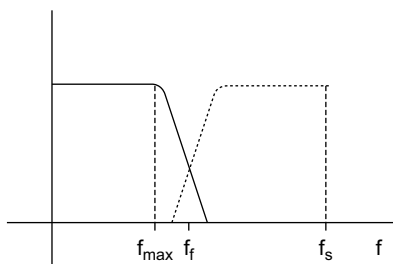


Figure 4-9 The response of the anti-alias filter requires that the sample rate be somewhat higher than the sampling theorem states.

practical low-pass filter. The filter will have a finite slope above its cutoff frequency, f_{max} . The mirrored response of the filter above the folding frequency is also shown. The overlap between the filter response and its mirrored response represent the region where aliasing can occur. The system is designed so that the folding frequency (and the sampling frequency) are large enough that the anti-alias response has room to roll off. Thus, f_{max} , the highest frequency that the analyzer will measure, must be significantly less than f_f . For practical filter implementations, f_s is typically 2.5 times f_{max} .

As shown, aliasing can be explained in the frequency domain, but it is also helpful to consider it briefly in the time domain. Figure 4-10 shows a set of sample points that fit two different waveforms. One of the waveforms has a frequency that violates the sampling theorem; the other does not. (The higher-frequency waveform violates the sampling theorem, of course.) Unless an anti-alias filter removes the unwanted alias frequency, the two sampled sine waves will be indistinguishable when processed digitally.

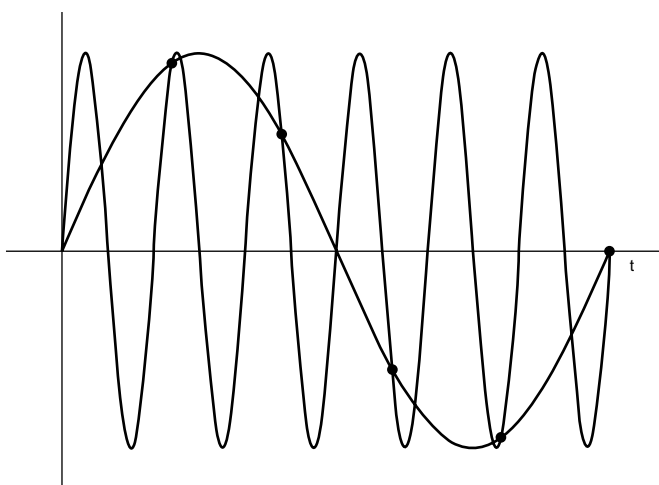


Figure 4-10 Aliasing in the time domain.

4.6 FFT Properties

The FFT is a record-oriented algorithm. A time record, N samples long, is the input, and the frequency spectrum, N samples long, is the output. Recall from Chapter 3 that N is often restricted to being a power of 2 to simplify the FFT computation. A typical record length for an FFT analyzer is 1024 sample points. The frequency spectrum produced by the FFT is symmetrical about the folding frequency. Thus, the first half of the output record is redundant with the second half, and the sample points numbered 0 to $N/2$ are retained. This implies that the effective length of the output record is $(N/2) + 1$. These are complex points (real + j imaginary) containing both magnitude and phase information.

Practically speaking, the output of the FFT is $(N/2) + 1$ points, extending from 0 Hz to f_f . Not all of these points are usually displayed though since the anti-alias filter begins to roll off before f_f . A common configuration is 1024 samples in the time record, producing 513 unique complex frequency domain points, with 401 of these actually displayed.

The $N/2$ (or so) frequency domain points are often referred to as *bins* and are usually numbered from 0 to $N/2$ (e.g., 0 to 512 for $N = 1024$). These bins are equivalent to the individual filter/detector outputs in the bank-of-filters analyzer. Bin 0 contains the DC level present in the input signal and is also known as the *DC bin*. The bins are spaced equally in frequency, with the frequency step, f_{step} being the reciprocal of the time record length.⁴

$$f_{\text{step}} = 1/\text{length of time record} \quad (4-10)$$

The length of the time record can be determined from the sample rate and the number of sample points in the time record.

$$f_{\text{step}} = f_s/N \quad (4-11)$$

The frequency associated with each bin is given by

$$f_n = nf_s/N \quad (4-12)$$

where

n = the bin number

The frequency of the last bin, containing the maximum frequency out of the FFT, is $f_s/2$. Therefore, the frequency range of an FFT is 0 Hz to $f_s/2$. (Note that this frequency is intentionally *not* called f_{max} , which is reserved for the upper-frequency limit of the instrument and which may not be the same as the last FFT bin.)

Suppose one cycle of a sine wave fits exactly into one time record, as shown in Figure 4-11. This sine wave will show up in bin 1 of the FFT output. If the frequency of the sine wave is doubled, then two sine waves will fit into one time record and their energy will appear in bin 2. Tripling the original sine wave frequency will cause a frequency domain response in bin 3, and so forth.

⁴ The term *frequency step* does not mean that some frequencies will be missed by the FFT. The output of the FFT is equivalent to the bank-of-filters analyzer, with contiguous band-pass filters centered at each bin.

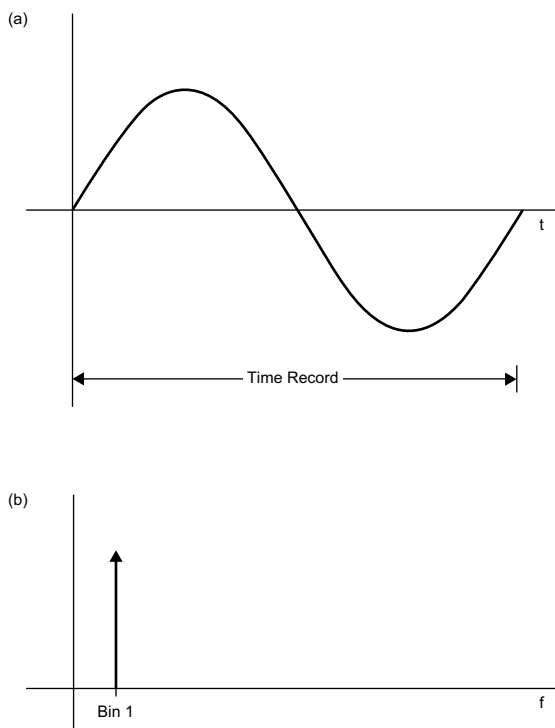


Figure 4-11 (a) A sine wave that exactly fills one time record. (b) The spectral line shows up in bin 1 of the FFT output.

4.7 Controlling the Frequency Span

The FFT is inherently a baseband transform. In other words, the frequency range of the FFT always starts at 0 Hz and extends to some maximum frequency, $f_s/2$. This can be a significant limitation in measurement situations where a small frequency band, not starting at DC, needs to be analyzed.

For example, suppose an FFT analyzer has a sample rate, $f_s = 256$ kHz. The frequency range of the FFT would be 0 Hz to 128 kHz ($f_s/2$). If $N = 1024$, the frequency resolution would be $f_s/N = 250$ Hz. Spectral lines closer than 250 Hz could not be resolved.⁵

One way to increase the frequency resolution is to increase N , the number of samples in the time record, which also increases the number of bins in the FFT output. Unfortunately, this increases the size of the arrays that the FFT has to deal with and the computation time increases accordingly. The computation time of the FFT algorithm often limits the performance of the instrument (in the form of update rate to the display), so increasing the size of the FFT is often undesirable.

⁵ This is an approximation since the frequency resolution will also depend on the window function, discussed in Section 4.9.

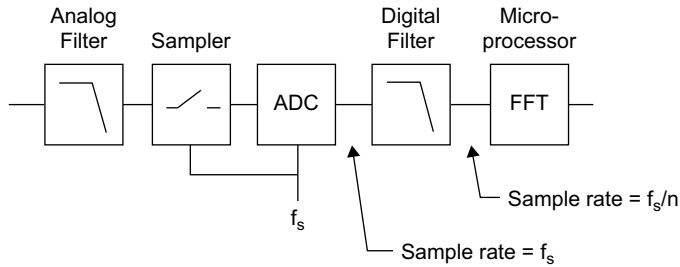


Figure 4-12 Decimating digital filters are often used to reduce the sample rate into the FFT.

Reducing f_s will also improve the frequency resolution but at the expense of reducing the upper-frequency limit of the FFT and ultimately the instrument bandwidth. This is a worthwhile trade-off that gives the user control over the frequency resolution and frequency range of the instrument. As the sample rate is lowered, the cutoff frequency of the anti-alias filter must also be lowered; otherwise, aliasing will occur. Selectable analog anti-alias filters could be provided, but it is more economical to implement the additional filters digitally. A *decimating digital filter* simultaneously decreases the bandwidth of the signal and decreases the sample rate (Figure 4-12). The sample rate into the digital filter is f_s . The sample rate out of the filter is f_s/n , where n is the *decimation factor*, which is an integer number. Similarly, the bandwidth at the input is BW , and the bandwidth at the output of the filter is BW/n .

The digital filter provides alias protection while reducing the sample rate so that the FFT frequency resolution is increased. The analog anti-alias filter is still required, since the digital filter is itself a sampled system that must be protected from aliasing. The analog filter protects the instrument at its widest frequency span, with operation at f_s . The digital filters fall in behind the protection of the analog filter and are used when narrower spans are selected by the user.

4.8 Band Selectable Analysis

By varying the sample rate, the frequency span of the analyzer can be controlled, but the start frequency of the span is always at DC. The frequency resolution of the measurement can be improved arbitrarily but at the expense of a lower maximum frequency. *Band selectable analysis* (also known as *zoom operation*) allows the user to reduce the frequency span while maintaining a constant center frequency. In other words, the displayed frequency range is not limited to starting at DC. This is useful because very narrow spans away from DC can be analyzed.

Band selectable analysis is accomplished by a change in the instrument block diagram (Figure 4-13). The output of the ADC is multiplied by a digital sinusoid, which mixes it down in frequency.⁶ Many readers will recognize this as just a digital version of the

⁶ Normally, a pair of digital mixers and a pair of digital filters are used due to the complex sinusoid and a complex FFT is required, but the operations are shown simplified here.

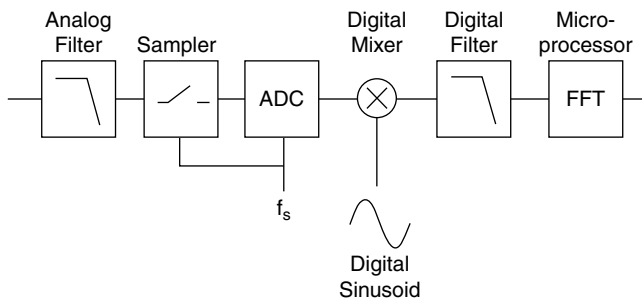


Figure 4-13 A digital mixer provides band selectable analysis in an FFT analyzer.

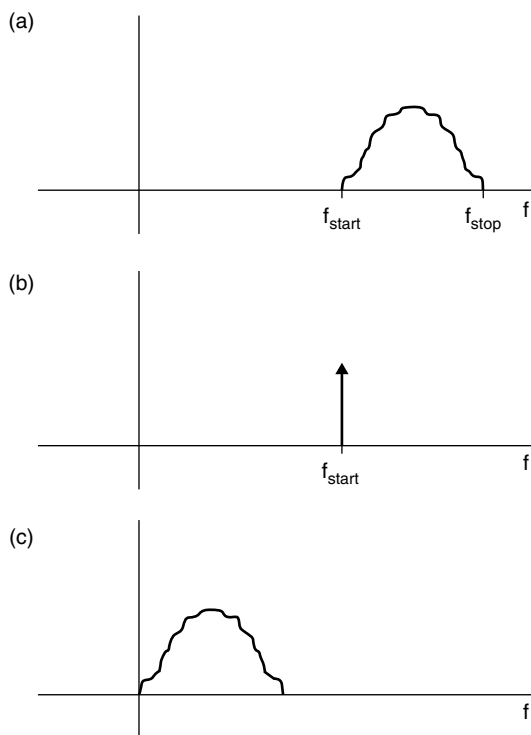


Figure 4-14 (a) The spectrum of the signal to be measured. (b) The spectrum of the digital oscillator. (c) The frequency translated version of the original spectrum.

superheterodyne technique used in radio receivers and swept spectrum analyzers. The frequency span of interest (Figure 4-14) is mixed with a complex sinusoid at the center frequency, which causes that frequency span to be mixed down to baseband. The digital filter is configured for the proper span by using the appropriate decimation factor. The FFT is used to obtain the frequency spectrum from the output of the digital filter. The bandwidth of the digital filter can be narrowed significantly, producing frequency spans as narrow as 1 Hz.

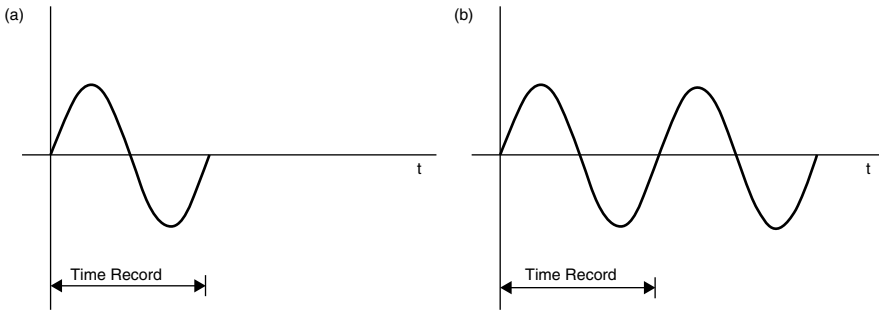


Figure 4-15 (a) A waveform that exactly fits one time record. (b) When replicated, no transients are introduced.

4.9 Leakage

The FFT operates on a finite length time record in an attempt to approximate the Fourier transform, which integrates over all time. The mathematics of the FFT operate on the finite length time record but have the effect of replicating the finite length time record over all time (Figure 4-15).⁷ With the waveform shown in Figure 4-15b, the finite length time record represents the actual waveform quite well, so the FFT result will approximate the Fourier integral very well.

However, the shape and phase of a waveform may be such that a transient is introduced when the waveform is replicated for all time, as shown in Figure 4-16. In this case, the FFT spectrum is not a good approximation for the integral form of the Fourier transform. Since the instrument user often does not have control over how the waveform fits into the time record, in general, it must be assumed that a discontinuity may exist. This effect, known as *leakage*, is very apparent in the frequency domain. Instead of the spectral line appearing thin and slender, it spreads out over a wide frequency range (Figure 4-17).

The usual solution to the problem of leakage is to force the waveform to zero at the ends of the time record; then they will always be the same, and no transient will exist when the time record is replicated. This is accomplished by multiplying the time record by a *window* function. Of course, the shape of the window is important because it will affect the data; it also must not introduce a transient of its own. Many different window functions have been developed for particular digital signal processing applications. The ones common to spectrum analyzers will be examined here.

4.10 Hanning Window

Also known as the *Hann window*, the *Hanning window* is one of the most common windows in digital signal processing. The time record samples are weighted by

$$w_n = \frac{1}{2} \{1 - \cos[2\pi n / (N - 1)]\} \quad (4-13)$$

⁷ The FFT has the effect of replicating the time record. This is a consequence of the mathematics, and there is no need for the algorithm to actually produce the replicated time record.

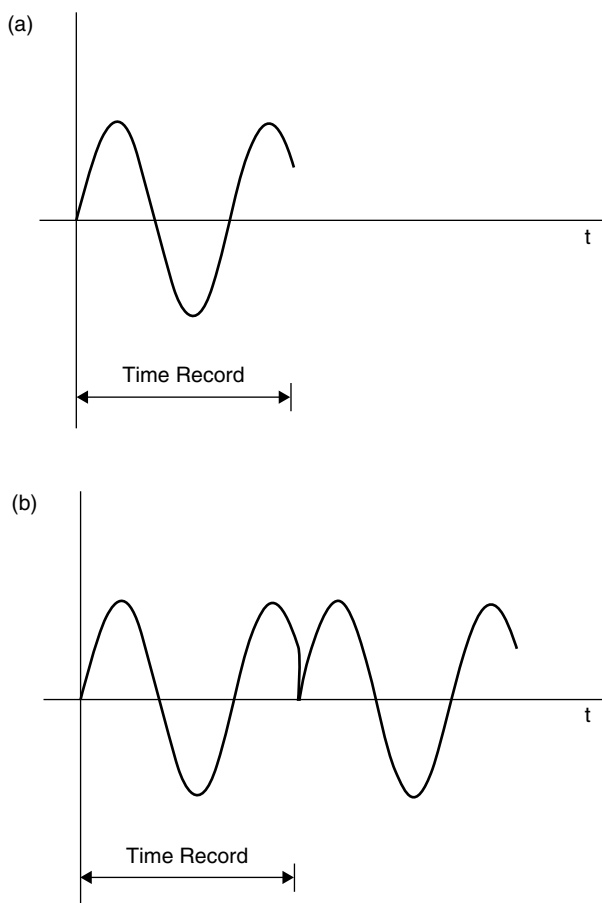


Figure 4-16 (a) A waveform that does not exactly fit into one time record. (b) When replicated, severe transients are introduced, causing leakage in the frequency domain.

where

n = bin number

N = number of bins

The Hanning window provides a smooth transition to zero as either end of the time record is approached (Figure 4-18). Therefore, the windowed time record will not produce a transient when replicated by the FFT algorithm. Clearly, the original time record has been modified and the effect in the frequency domain must be considered. The shape of the Hanning window in the frequency domain is the Fourier transform of the window function.

The frequency domain response of the window function determines the passband shape of the individual filters that the FFT produces mathematically. Figure 4-19a shows the overlapped response of several frequency bins using a Hanning window. The filter shape is rounded off, and the net response of the analyzer drops off somewhat between bins. Therefore, a spectral line falling where the two filters meet will be measured with an error

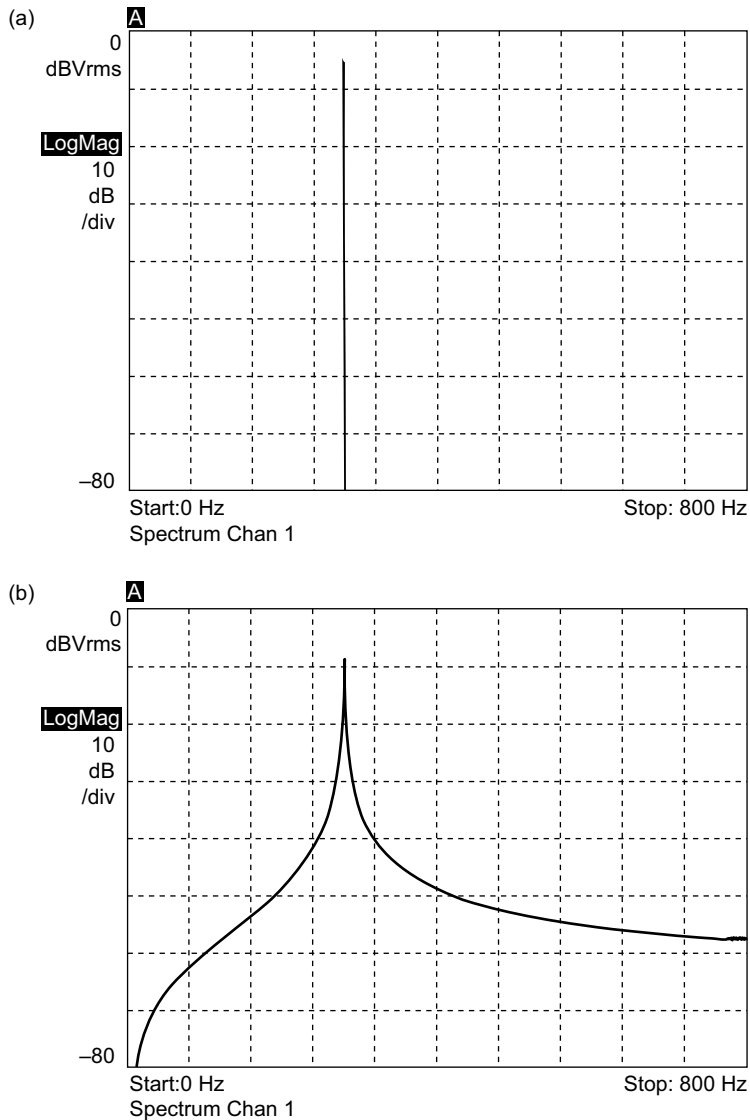


Figure 4-17 (a) Measurement of a spectral line with no leakage. (b) Measurement of a spectral line with leakage.

determined by the shape of the filter. The Hanning window introduces a maximum amplitude error of 1.5 dB (16%), which may be a significant error in some applications. The shape of a window is always a compromise between amplitude accuracy (which depends on the flatness of the filter passband) and frequency resolution (which depends on the width of the filter). The Hanning window, compared with other common windows, provides good frequency resolution at the expense of somewhat less amplitude accuracy. Figure 4-20a shows the spectrum of a sine wave measured using the Hanning window.

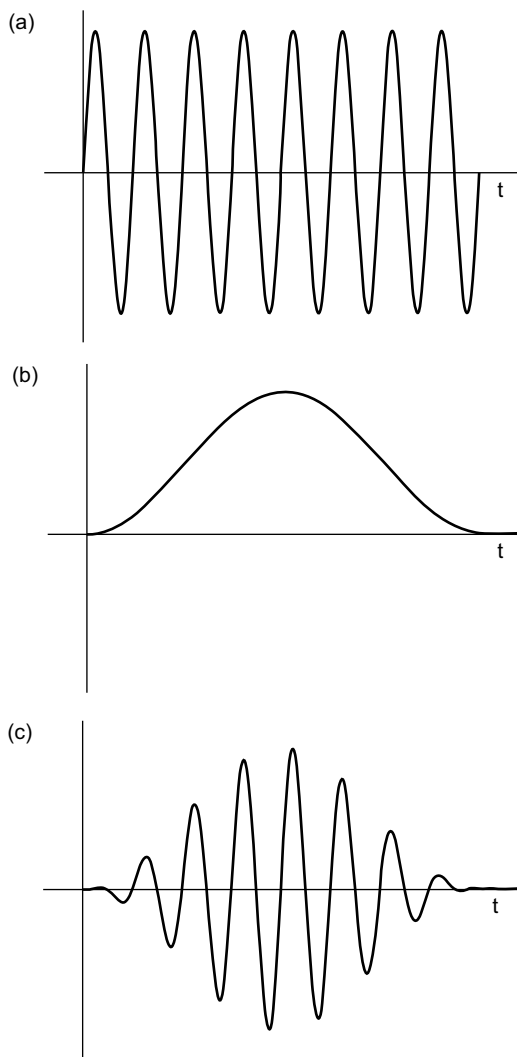


Figure 4-18 (a) The original time record. (b) The Hanning window. (c) The windowed time record.

4.11 Flattop Window

A window that has a flat passband reduces the size of the amplitude dips between bins and minimizes the amplitude error. A spectral line that falls halfway between the centers of two bins will be attenuated by a much smaller amount. The *flattop* window has such a characteristic and is shown in Figure 4-19b. Since the response of each bin overlaps considerably more than with the Hanning window, the disadvantage of the flattop window is reduced frequency resolution due to its wider profile. The spectral line will appear wider on the spectrum analyzer display, limiting the ability to resolve two closely spaced spectral lines.

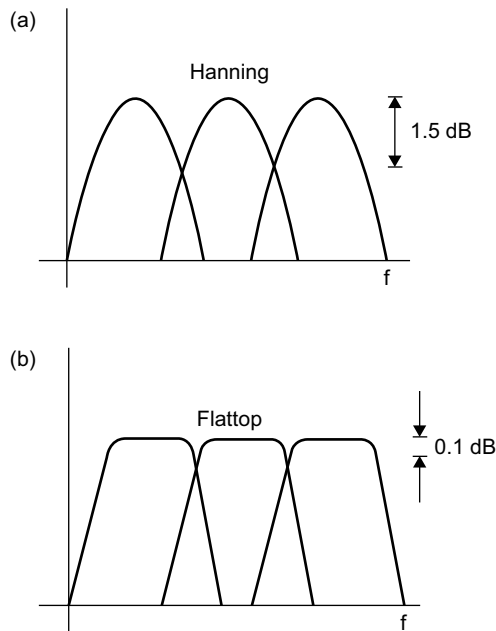


Figure 4-19 (a) The Hanning window introduces a maximum amplitude error of 1.5 dB. (b) The flattop window introduces a maximum amplitude error of up to 0.1 dB.

The flattop window is considered a high-amplitude accuracy window, having a maximum amplitude error of 0.1 dB or less, depending on the implementation. Figure 4-20b shows the spectrum of a sine wave as measured using the flattop window.

4.12 Uniform Window

The uniform window is really no window at all; all the samples are left unchanged. Although the uniform window has the potential for severe leakage problems, in some cases the waveform in the time record has the same value at both ends of the record, thereby eliminating the transient introduced by the FFT. Such waveforms are called *self-windowing*. Waveforms such as *pseudorandom noise* (PRN),⁸ sine bursts, impulses, and decaying sinusoids can all be self-windowing.

The uniform window is appropriate for making network measurements when the internal noise source of the analyzer is used. The noise source is usually a PRN generator that produces a noise waveform that is periodic within the time record of the instrument. Since the noise source and the time record are synchronized, no transients occur at the ends of the time record and leakage in the frequency domain is avoided.

⁸ PRN is not truly random noise but instead repeats at some interval.

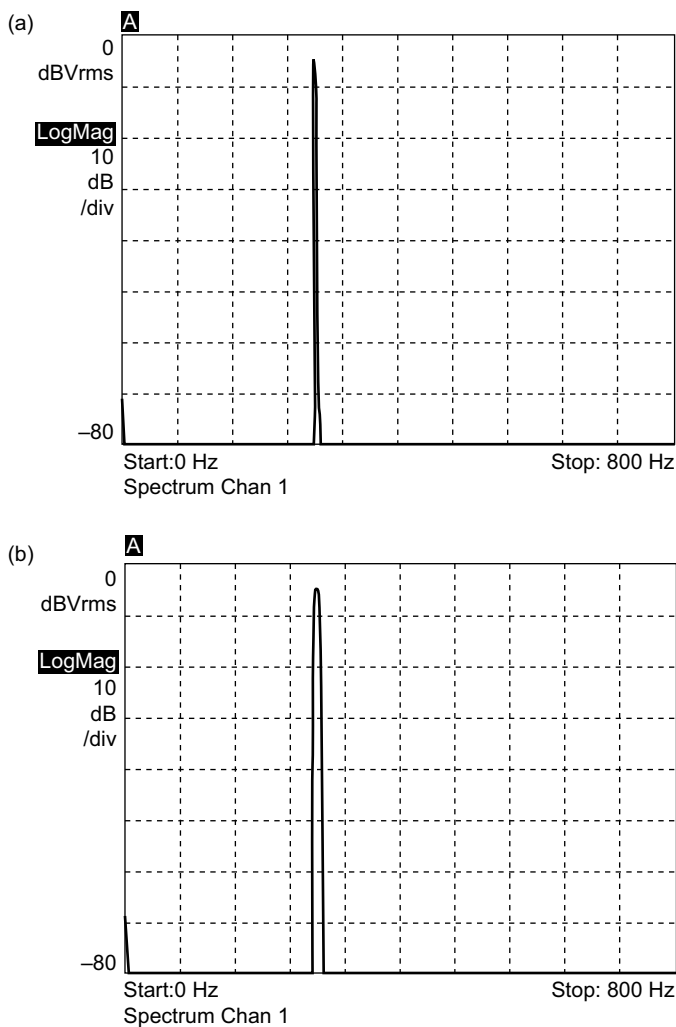


Figure 4-20 (a) Sine wave spectrum using the Hanning window. (b) Sine wave spectrum using the flattop window.

4.13 Exponential Window

One of the advantages of an FFT analyzer is that it can be used to measure the frequency content of a fast transient. (This is not usually possible in the more conventional swept analyzer since it may miss some of the transient as it is sweeping through its frequency span.) Such a transient might be the step or impulse response of an electrical network or mechanical system.

A typical transient response is shown in Figure 4-21a. As shown, the waveform is self-windowing because it dies out within the length of the time record, reducing the leakage. If the waveform does not dissipate within the time record (as shown in Figure 4-21b), then a

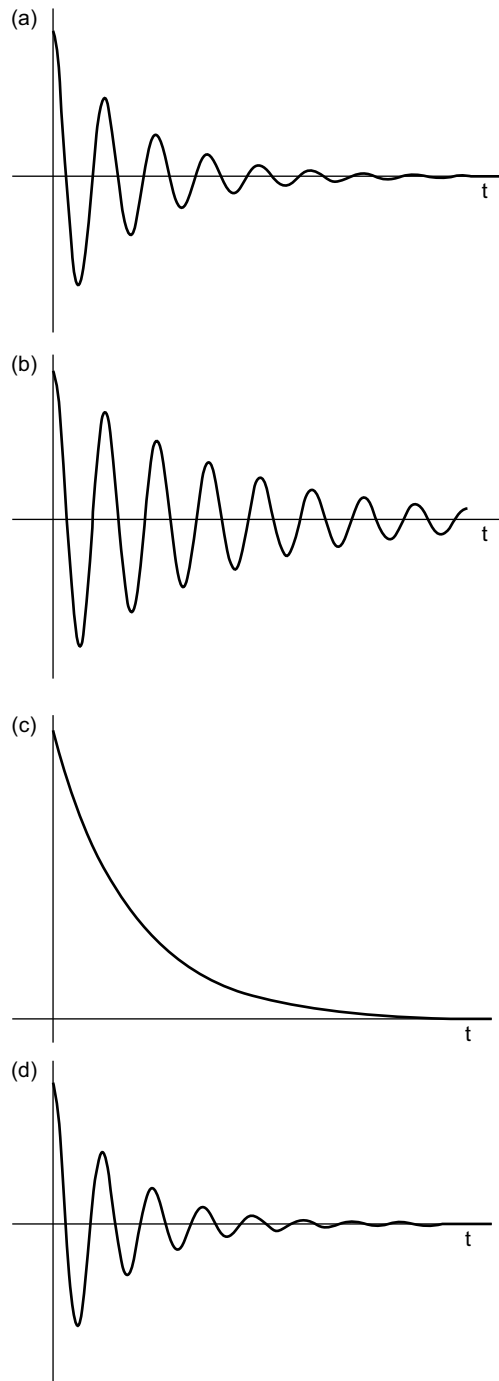


Figure 4-21 (a) A transient response that is self-windowing. (b) A transient response that requires windowing. (c) The exponential window. (d) The windowed transient response.

window function should be applied. If a window such as the Hanning window were applied to the waveform, the beginning portion of the time record would be forced to zero. This is precisely where most of the transient's energy is, so such a window would be inappropriate.

A much better choice is a window that has a decaying exponential response. The beginning portion of the waveform is not disturbed, but the end of the time record is forced to zero. There still may be a transient at the beginning of the time record, but this transient is not introduced by the FFT, it is, in fact, the transient being measured. Figure 4-21c shows the exponential window, and Figure 4-21d shows the resulting time domain function when the exponential window is applied to Figure 4-21b.

The exponential window function is given by

$$w_n = e^{-n/((N-1)k)} \quad (4-14)$$

where

n = bin number

N = number of bins

k = exponential time constant

The time constant, k , is selected by the user to provide the appropriate amount of exponential decay to prevent leakage. The exponential window is inappropriate for measuring anything but transient waveforms.

4.14 Selecting a Window Function

Selecting the appropriate window function may seem cumbersome for users familiar with swept spectrum analyzers. Most measurements will require the use of a window such as the Hanning or flattop windows. These are the appropriate windows for typical spectrum analysis measurements. Choosing between these two windows, then, involves a trade-off between frequency resolution and amplitude accuracy. Again, the Hanning window provides better frequency resolution while the flattop window has better amplitude accuracy. Having used the time domain to explain why leakage occurs, here the user should switch into frequency domain thinking. The narrower the passband of the window's frequency domain filter, the better the analyzer can discern between two closely spaced spectral lines. At the same time, the amplitude of the spectral line will be less certain. Conversely, the wider and flatter the window's frequency domain filter is, the more accurate the amplitude measurement will be and, of course, the frequency resolution will be reduced. Choosing between two such window functions is really just choosing the filter shape in the frequency domain.

The uniform and exponential windows should be considered windows for special situations. The uniform window is used where it can be guaranteed that there will be no leakage effects, such as when making network measurements using the analyzer's internal PRN source. The exponential window is for use when the input signal is a transient.

4.15 Oscillator Characterization

FFT spectrum analyzers can be used to characterize oscillators. One important specification of an oscillator is its harmonic distortion. Figure 4-22 shows the fundamental through the

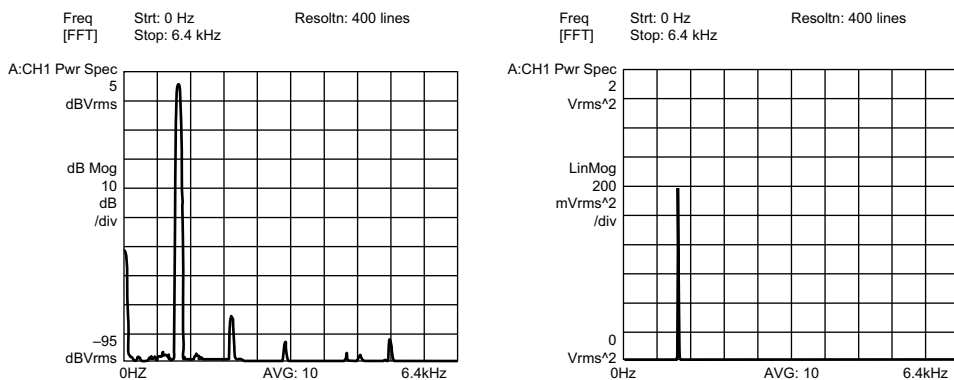


Figure 4-22 Harmonic distortion of an audio oscillator. (a) Logarithmic amplitude scale. (b) Linear amplitude scale.

sixth harmonic of a 1 kHz oscillator. Because the fundamental frequency may not exactly fit the time record of the analyzer, windowing should be used to reduce the leakage. The flattop window should be used to provide the most accurate amplitude measurements.

Notice that the input sensitivity of the analyzer is selected so that the fundamental is near the top of the display. In general, set the input sensitivity to the most sensitive range that does not overload the analyzer. Severe distortion of the input signal occurs if its peak voltage exceeds the range of the analog-to-digital converter. Therefore, FFT spectrum analyzers warn the user of this condition by some kind of overload indicator.

It is also important to make sure the analyzer is not underloaded. If the signal going into the analog-to-digital converter is too small, much of the useful information of the spectrum may be below the noise level of the analyzer. Therefore, setting the input sensitivity to the most sensitive range that does not cause an overload gives the best possible results.

Figure 4-22a is a display of the spectrum amplitude in logarithmic form to ensure that distortion products far below the fundamental can be seen. All signal amplitudes on this display are in dBV (decibels below 1 V RMS). However, since most FFT analyzers have very versatile display capabilities, this spectrum could also be displayed linearly as in Figure 4-22b. Here the units of amplitude are volts.

Another important measure of an oscillator's performance is the level of its power-line sidebands. In Figure 4-23, band selectable analysis is used to "zoom in" on the signal so that it is easy to resolve and measure sidebands that are only 60 Hz away from our 1 kHz signal. With some analyzers, it is possible to measure signals as close as 1 mHz away from the fundamental.

4.16 Spectral Maps

One feature that has been unique to the FFT analyzer but is finding its way into other analyzers is the *spectral map* (also known as a *waterfall display*). This feature displays multiple spectrums as a function of time, giving an almost three-dimensional display (Figure 4-24).

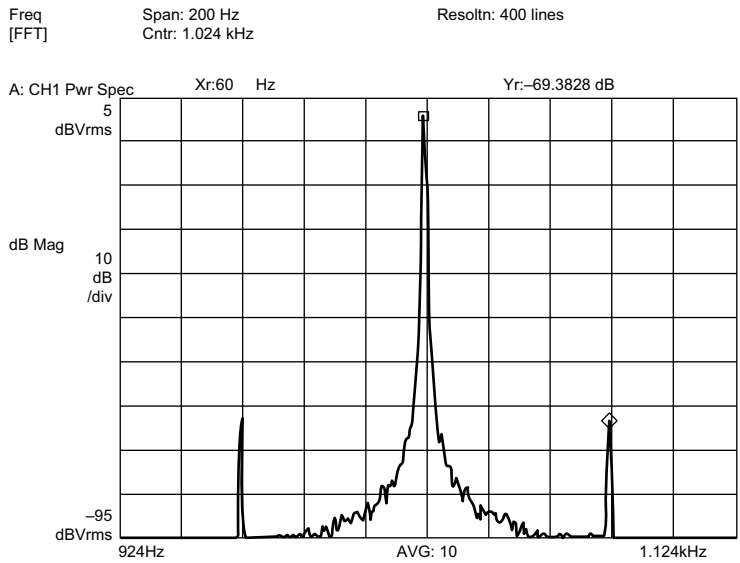


Figure 4-23 Powerline sidebands of an audio oscillator.

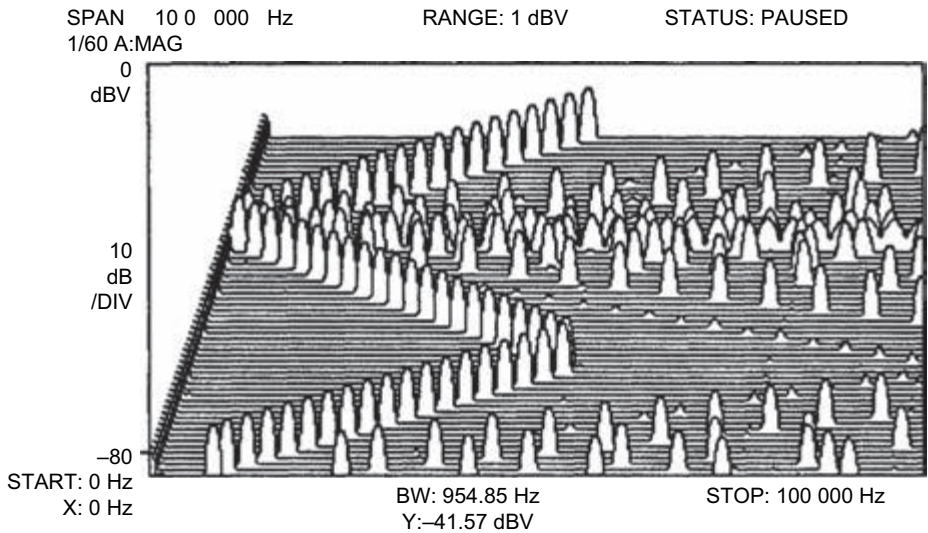


Figure 4-24 A spectral map of a swept sine wave oscillator. The largest responses are caused by the fundamental frequency of the oscillator moving up and down in frequency. The other responses are caused by harmonic distortion and other imperfections in the oscillator.

For a transient event, this frequency spectrum versus time display characterizes the signal as a function of time. This feature is particularly appropriate for an FFT analyzer since it has the ability to produce successive spectra without missing any data. Traditional swept analyzers may miss portions of the waveform while sweeping.

Spectral maps are often used to monitor vibrations in the structure of a rotating machine as its speed is steadily increased over time (called *run up*) or steadily decreased over time (called *run down*).

4.17 Time Domain Display

Many FFT analyzers provide a display of the time domain data. Although the time domain display is similar to an oscilloscope display, there are some significant differences.

First, the sample rate of the FFT analyzer has been chosen to optimize for FFT analysis results. Specifically, the sample rate must be high enough to satisfy the sampling theorem with some margin to account for the shape of the anti-alias filter. Typically, the sample rate will be 2.5 times the highest frequency. So at the highest frequency there will be between two and three samples per period of the waveform. Simply displaying so few samples per period will not produce a waveform onscreen that looks anything like an oscilloscope display. (Digital oscilloscopes normally use more samples per period and may provide additional digital signal processing to reconstruct the waveform.)

The anti-alias filter is a steep high-order filter designed to approximate an ideal low-pass filter. It abruptly limits the frequency response of the analyzer and may introduce ringing in the time domain.

In band selectable analysis, the time waveform may be displayed after it has been mixed with the complex sinusoid. The resulting waveform is complex (has a real and an imaginary part) and is often difficult to interpret.

Despite these shortcomings, the time domain display is useful for many applications. The user can monitor the input waveform that is associated with the frequency spectrum. Also, the analyzer can be used as a waveform recorder within the time domain capability of the instrument. Some analyzers provide long time buffers for capturing large amounts of time domain data. After capture, portions of the time record can be analyzed for frequency content.

4.18 Network Measurements

Traditionally, network measurements are made by supplying a sinusoid to the *device under test* (DUT) and measuring its output, repeating this at each frequency of interest. The internal source of an FFT analyzer is usually quite flexible and can output a variety of waveforms for use in network analysis. The source is connected to the input of the DUT, and the output of the DUT is connected to the input of the analyzer (Figure 4-25).

Recall that the FFT analyzer behaves the same as a bank-of-filters analyzer. To make a network measurement using a sinusoid, we would iteratively set the sinusoid's frequency to be in the center of each of the filters, recording the readings as we went. This requires as many measurements as there are bins. On the other hand, a broadband signal can be used to simultaneously produce energy in each of the FFT bins, which can be captured in one FFT measurement. Chirp sine and random noise meet such a requirement.

The chirp sine signal is a swept sine burst designed to fill the time record of the FFT analyzer. This provides energy across the frequency band of interest. The chief advantage of the chirp sine is its relatively high average power—much more than random noise (for the

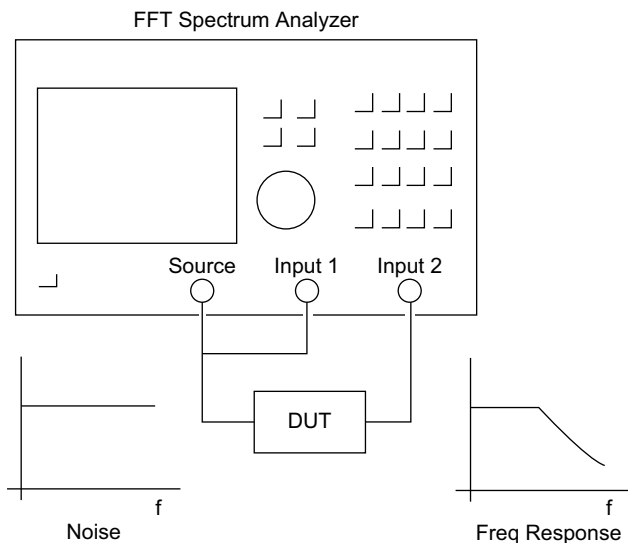


Figure 4-25 An FFT analyzer can perform a network measurement using a noise source.

same peak voltage). This produces a better signal-to-noise ratio in the measurement when compared with random noise.

Broadband random noise has equal energy in all of the FFT bins and provides a stimulus to the DUT such that the output frequency response will be the frequency response of the network. A PRN signal is often used because it is periodic within the time record of the analyzer so that it does not produce leakage. Thus, the uniform window is used when making network measurements with a PRN source.

A truly random (not pseudorandom) noise source is useful with nonlinear networks. Nonlinear networks produce considerable distortion, which corrupts the results of a network measurement. With a random noise source, these distortion effects can be averaged out since they are different for each measurement (see Chapter 10). With PRN, the noise waveform and the distortion products are the same for every measurement, and averaging will have no effect.

This points out a fundamental problem with measuring nonlinear networks: the frequency response is not a property of the network alone—it also depends on the stimulus. Each stimulus (i.e., swept sine, PRN, and random noise) will, in general, give a different result. Also, if the amplitude of the stimulus is changed, you will get a different result. To minimize this problem, consider using a test signal that closely approximates the kind of signals normally used to drive the network's inputs.

4.19 Phase

So far, we have discussed measuring only the amplitude of signals in the frequency domain. However, true network analysis requires that both magnitude and phase be measured. Earlier in the chapter, we mentioned that the output of the FFT was an array of complex points

containing both magnitude and phase information, which allows the analyzer to perform complex network measurements.

In a network measurement, phase information is the phase response of the DUT. More precisely, it is the phase difference (as a function of frequency) between the input stimulus and the measured response. Many FFT analyzers have two input channels that can be used to simultaneously measure both the input and the output of the DUT. In this case, the phase response of the DUT is the phase difference between the two channels.

In a spectrum measurement, the usefulness of the phase information is less obvious. Since phase is a relative concept, one is tempted to ask, “Phase with respect to what?” Phase displayed on an FFT analyzer depends on the relative position of the waveform in the time record. For a single sinusoid, shifting the waveform 90° in the time record causes the phase response to change by 90° . Many analyzers provide oscilloscope-like triggering capability to allow some control over the start of the time record. If this feature is used, then the phase of a particular signal can be stabilized. For example, this allows the phase of harmonics to be compared with the fundamental. If no triggering is used, the analyzer will acquire a time record when ready, which will be uncorrelated to the input signal. In this case, the phase of the signal will vary randomly from measurement to measurement. The relative phases of signals present in a single time record may be useful.

4.20 Electronic Filter Characterization

Another typical application of an FFT analyzer is to characterize a low-frequency electronic filter. One possible test setup appears in Figure 4-26. Because the filter is linear, it is possible to use pseudorandom noise as the stimulus and get very fast measurement times. The uniform window is used because the pseudorandom noise is periodic in the time record.

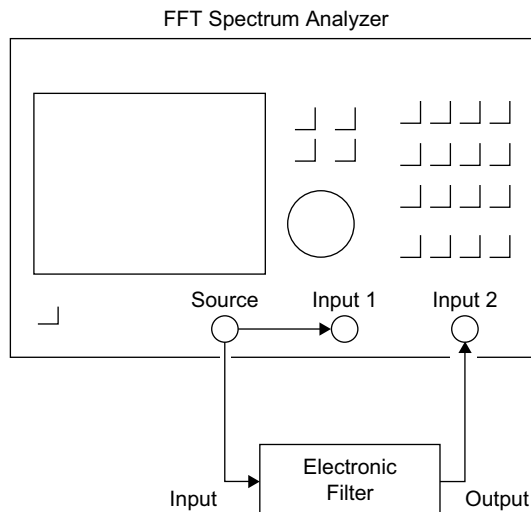


Figure 4-26 Test setup to measure frequency response of filter.

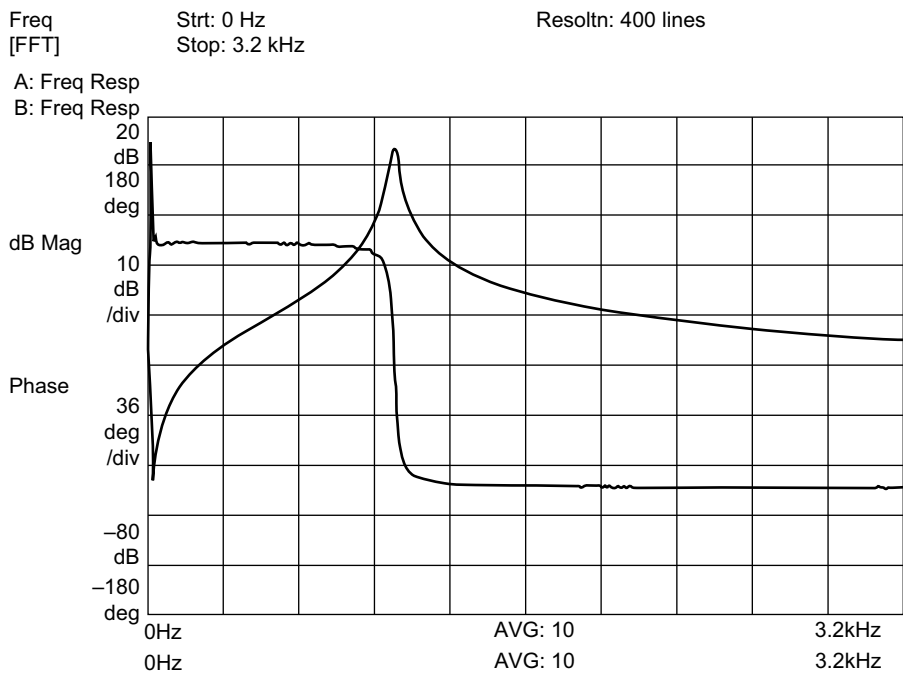


Figure 4-27 Frequency response of electronic filter using PRN and uniform window.

No averaging is needed since the signal is periodic and reasonably large. As in the single-channel case, it's important to set the input sensitivity for both channels to the most sensitive position that does not overload the analog-to-digital converters.

With these considerations in mind, the frequency response, including both magnitude and phase, is shown in Figure 4-27. The primary advantage of this measurement over traditional swept analysis techniques is speed. This measurement can be made in 1/8 second with an FFT analyzer but may take over 30 seconds with a swept network analyzer. This speed improvement is particularly important when the filter is being adjusted or when large volumes are tested on a production line.

4.21 Cross-Power Spectrum

The cross-power spectrum is not often used as a separate measurement but is used internally by FFT analyzers to compute transfer functions and coherence. The cross-power spectrum, G_{xy} , is defined as taking the FFT of two signals separately and multiplying the result together:

$$G_{xy}(f) = S_x(f)S_y^*(f) \tag{4-15}$$

where

★ = the complex conjugate of the function

With this function, we can define the transfer function, $H(f)$, using the cross-power spectrum and the spectrum of the input channel

$$H(f) = \frac{\overline{G_{xy}(f)}}{\overline{G_{xx}(f)}} \tag{4-16}$$

where the overbar denotes the average of the function in the frequency domain.

At first glance it may seem more appropriate to compute the transfer function using

$$|H(f)|^2 = \frac{\overline{G_{yy}}}{\overline{G_{xx}}} \tag{4-17}$$

This is the ratio of two single-channel, averaged measurements. Not only does this measurement fail to give any phase information, but it also will be in error when there is noise in the measurement. For example, let us solve the equations for the special case where noise is injected into the output as in Figure 4-28. The output is

$$S_y(f) = S_x(f)H(f) + S_n(f) \tag{4-18}$$

So

$$G_{yy} = S_y S_y^* = G_{xx}|H|^2 + S_x H S_n + S_x^* H^* S_n + |S_n|^2 \tag{4-19}$$

Using the RMS average of this result to try to eliminate the noise shows the $S_x S_n$ terms approaching zero because S_x and S_n are uncorrelated. That is, the expected value of $S_x S_n = 0$. However, the $|S_n|^2$ term remains as an error, giving

$$\frac{\overline{G_{yy}}}{\overline{G_{xx}}} = |H|^2 + \frac{|S_n|^2}{\overline{G_{xx}}} \tag{4-20}$$

Therefore, measuring $|H|^2$ by this single-channel technique gives a value that will be in error (too large) by the value of the noise-to-signal ratio. Averaging the cross-power spectrum eliminates this noise error. Using the example in Figure 4-28,

$$\overline{G_{yx}} = \overline{S_y S_x^*} = \overline{(S_x H + S_n) S_x^*} = \overline{G_{xx} H} + \overline{S_n S_x^*} \tag{4-21}$$

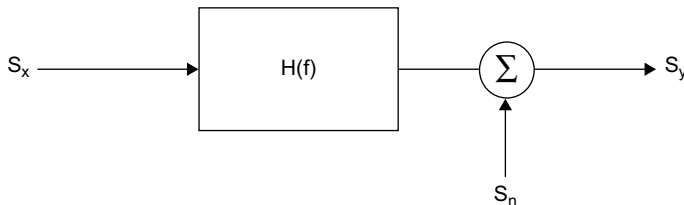


Figure 4-28 Transfer function measurements with noise present.

so

$$\frac{\overline{G_{yx}}}{\overline{G_{xx}}} = H(f) + \overline{S_n S_x^*} \quad (4-22)$$

Because S_n and S_x are uncorrelated, the second term will average to zero, making this function a much better estimate of the transfer function.

4.22 Coherence

FFT analyzers often include the ability to make *coherence measurements*, which measures the power in the response channel that is caused by power in the reference channel. Coherence is a unitless parameter that indicates how much of the output power is coherent with the input power. A coherence value of 1 implies that all the power in the output is caused by the input. A coherence value of 0 means that none of the power in the output is coherent with the input. (Care must be exercised when interpreting the coherence measurement. It does not always imply a causal relationship. For example, if two signals are caused by a third signal, they will be coherent with each other even though one is not caused by the other.)

The coherence function is often used alongside a transfer function measurement as an indicator of measurement quality. This is especially important for situations when the components to be tested cannot be isolated from outside disturbances. One example is the measurement of a switching power supply's frequency response that contains a high concentration of power at the switching frequency. Another example is measuring the frequency response of a mechanical part on one machine in the presence of strong vibration from another nearby machine.

Figure 4-29 shows one way to simulate this type of situation by adding noise and a 2 kHz signal to the output of an electronic filter. The measured frequency response is shown in the upper trace in Figure 4-30. RMS averaging has reduced the noise contribution but has not completely eliminated the 2 kHz interference. Further averaging would reduce this interference. If we did not already know the source of this interference, we would think that the filter has an additional resonance at 2 kHz. By using a coherence measurement we can eliminate the unrelated 2 kHz component.

The lower trace in Figure 4-30 shows the coherence. The coherence goes from 1 (all the output power at that frequency is caused by the input) to 0 (none of the output power at that frequency is caused by the input). The coherence function shows that the response at 2 kHz is not coherent with the input and therefore is not likely to be caused by the input but by interference. However, the filter's response near 1 kHz has excellent coherence, so the measurement here is valid.

The coherence function, $\gamma^2(f)$, is derived from the cross-power spectrum by

$$\gamma^2(f) = \frac{\overline{G_{yx}(f)} \overline{G_{xy}^*(f)}}{\overline{G_{xx}(f)} \overline{G_{yy}(f)}} \quad (4-23)$$

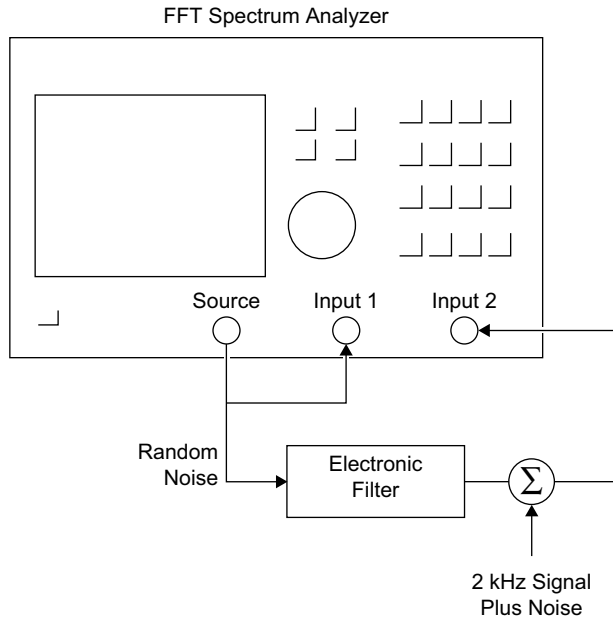


Figure 4-29 Adding noise and a 2 kHz signal to the output of a filter.

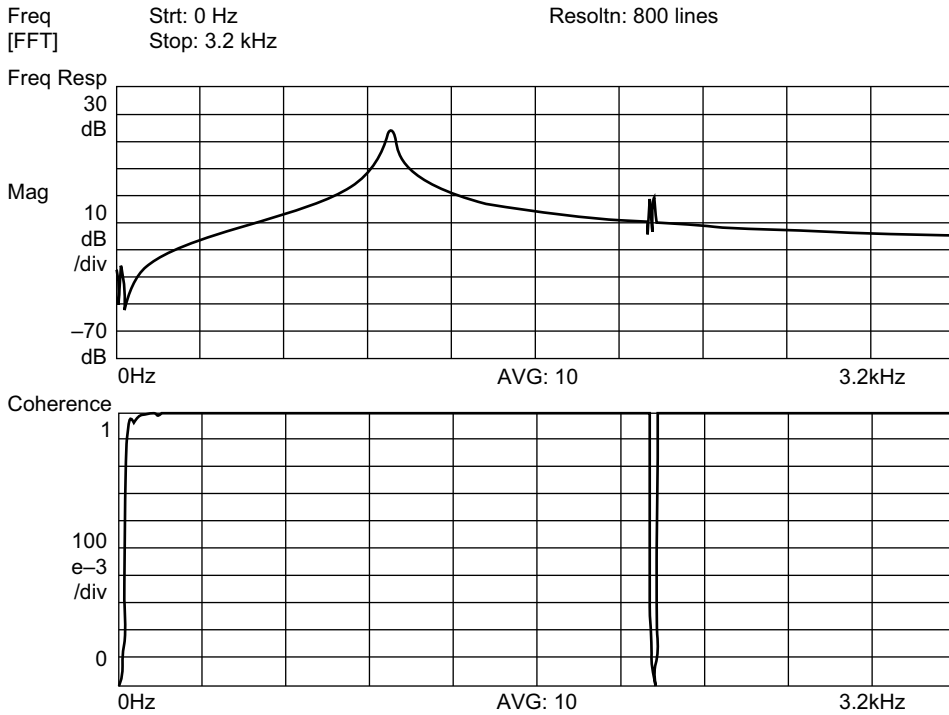


Figure 4-30 Frequency response and coherence with added noise and 2 kHz signal.

To explore the mathematics of the coherence function, we will use the example in Figure 4-28. As has been shown before,

$$\overline{G_{yy}} = \overline{G_{xx}}|H|^2 + \overline{S_x H S_n^*} + \overline{S_x^* H^* S_n} + \overline{|S_n|^2} \quad (4-24)$$

$$G_{yx} = \overline{G_{xx} H} + \overline{S_n S_x^*} \quad (4-25)$$

As the measurement is averaged, the cross-terms $S_n S_x$ approach zero, assuming that the signal and the noise are not related. So the coherence becomes

$$\gamma^2 = \frac{(H \overline{G_{xx}})^2}{\overline{G_{xx}} (|H|^2 \overline{G_{xx}} + \overline{|S_n|^2})} \quad (4-26)$$

$$\gamma^2 = \frac{|H|^2 \overline{G_{xx}}}{|H|^2 \overline{G_{xx}} + \overline{S_n}^2} \quad (4-27)$$

This shows that if there is no noise, the coherence function is unity. If there is noise, the coherence will be reduced. Note also that coherence is a function of frequency. Coherence can be unity at frequencies where there is no interference and low at frequencies where the noise is high.

4.23 Correlation

Correlation is a measure of the similarity between two functions, normally computed at a specified time offset. To understand the correlation between two waveforms, we start by multiplying the waveforms together at each instant in time and adding up all the products. If, as in Figure 4-31a, the waveforms are identical, every product is positive and the resulting sum is large. If however, as in Figure 4-31b, the two records are dissimilar, then some of the products would be positive and some would be negative. There would be a tendency for the products to cancel, so the final sum would be smaller.

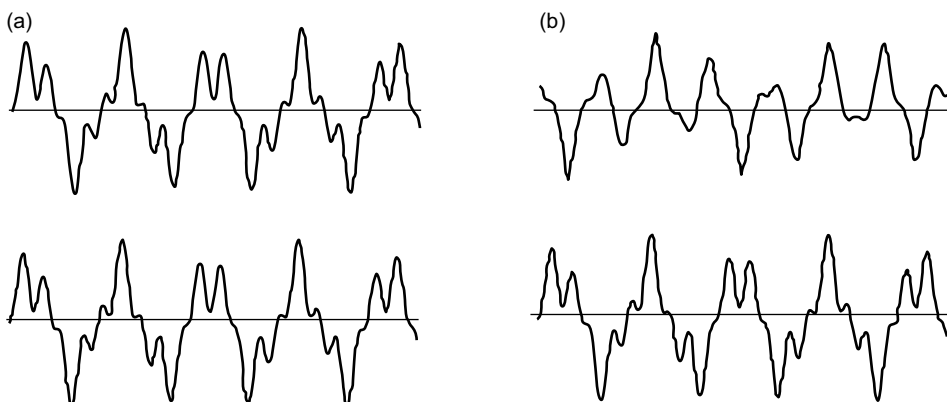


Figure 4-31 (a) Correlation of two identical signals. (b) Correlation of two different signals.

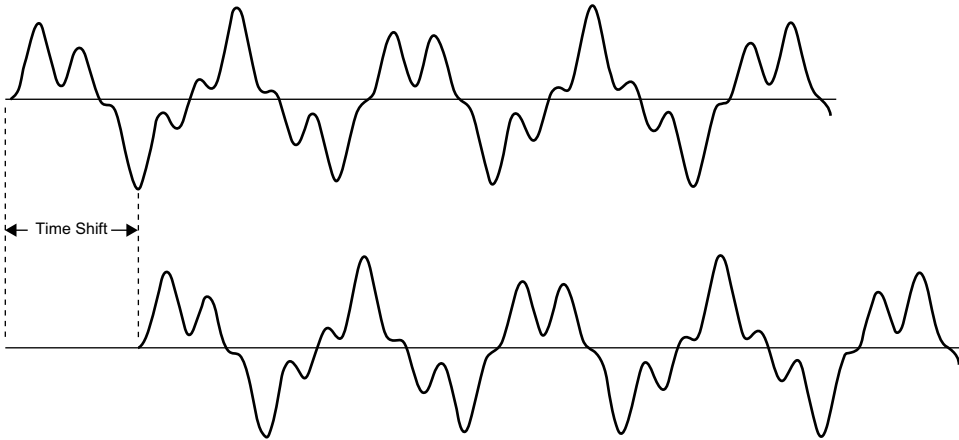


Figure 4-32 Correlation of time displaced signals.

Now consider the upper waveform in Figure 4-32, and the same waveform shifted in time shown below it. If the time shift were zero, then conditions would be the same as before; that is, the waveforms would be in phase and the final sum of the products would be large. As the time shift between the two waveforms is increased, the waveforms become increasingly dissimilar and the final sum is reduced.

4.24 Autocorrelation

Going one step further, we can find the average product for each time shift by dividing each final sum by the number of products contributing to it. By plotting the average product as a function of time shift, the resulting curve is shown to be largest when the time shift is zero and diminishes to zero as the time shift increases. This curve is called the *autocorrelation function* of the waveform. It is a graph of the similarity (or correlation) between a waveform and itself, as a function of the time shift.

The autocorrelation function, $R_{xx}(\tau)$, is a special time average defined by

$$R_{xx}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_T x(t)x(t + \tau) dt \quad (4-28)$$

That is, the autocorrelation function is found by taking a signal, multiplying it by the same signal displaced τ units in time, and averaging the product over all time.

For the sake of simplicity and speed, most FFT analyzers perform the correlation operation by taking advantage of its duality with the power spectrum. It can be shown that

$$R_{xx}(\tau) = \mathcal{F}^{-1} [S_x(f)S_x^*(f)] \quad (4-29)$$

where

$$\begin{aligned} \mathcal{F}^{-1} &= \text{the inverse Fourier transform} \\ S_x &= \text{the Fourier transform of } x(t) \end{aligned}$$

The autocorrelation function always has a maximum at $\tau = 0$ equal to the mean square value of $x(t)$. If the signal $x(t)$ is periodic, the correlation function is also periodic with the same period. Random noise, on the other hand, correlates only at $\tau = 0$.

The autocorrelation function can be used to improve the signal-to-noise ratio of periodic signals. The random noise component concentrates near $\tau = 0$, while the periodic component repeats itself with the same periodicity as the signal. Another thing to remember is that impulsive noises such as pulse trains, bearing ping, or gear chatter show up more distinctly in correlation or time record averaging than in a frequency domain analysis.

The autocorrelation function is more easily understood by looking at a few examples. The random noise shown in Figure 4-33 is not similar to itself with any amount of time shift (since it is random), so its autocorrelation has only a single spike at the point of zero time shift.

Figure 4-34 shows the autocorrelation of a sine wave and a square wave. These are both special cases of a more general statement; the autocorrelation of any periodic waveform is periodic and has the same period as the waveform itself.

This can be useful when trying to extract a signal hidden by noise. Figure 4-35a shows what looks like random noise, but there is actually a low-level sine wave buried in it. We can see this in Figure 4-35b where we have taken 100 averages of the autocorrelation of this signal. The noise has become the spike around a time shift of zero, whereas the autocorrelation of the sine wave is clearly visible, repeating itself with the period of the sine wave.

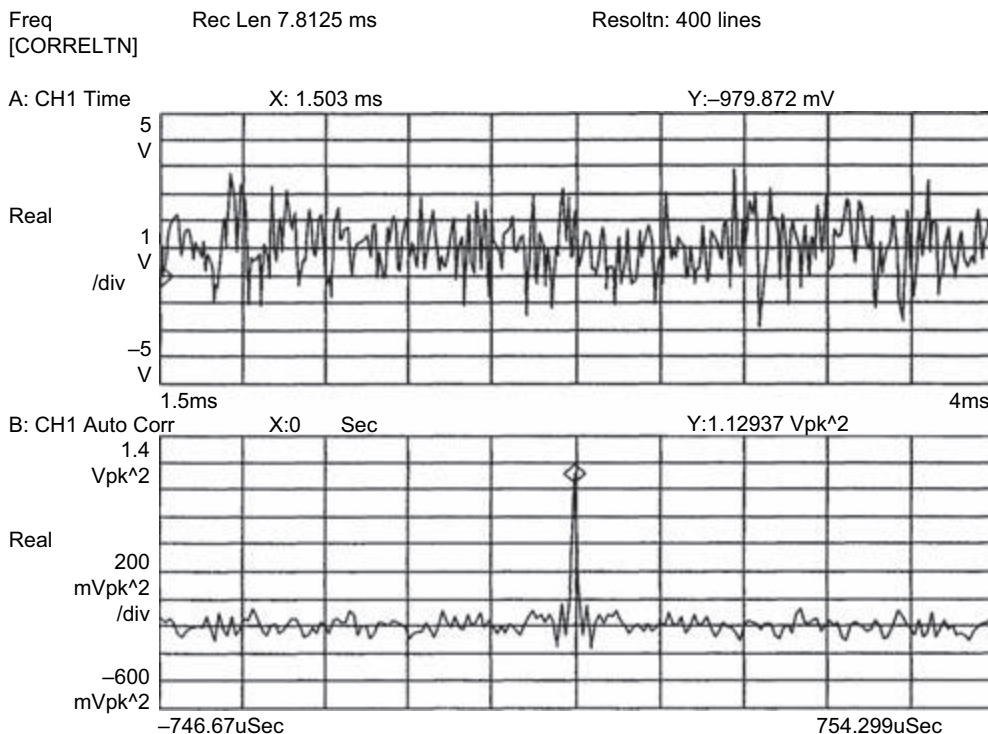


Figure 4-33 Autocorrelation of random noise.

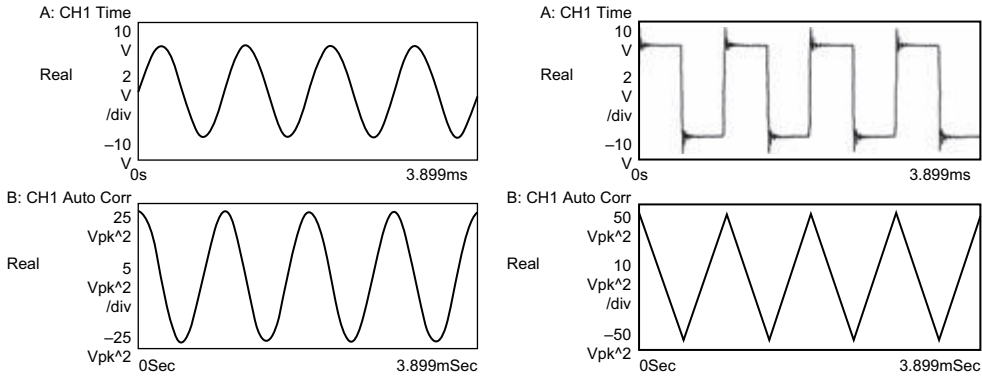


Figure 4-34 Autocorrelation of periodic waveforms.

If a trigger signal is available that is synchronous with the sine wave, it is possible to extract the signal from the noise by linear averaging as in the last example. But the important point about the autocorrelation function is that no synchronizing trigger is needed. In signal identification problems like radio astronomy and passive sonar, a synchronizing signal is not available, so autocorrelation is an important tool. The disadvantage of autocorrelation is that the input waveform is not preserved as it is in linear averaging.

Since any time domain waveform can be transformed into the frequency domain, the reader may wonder what is the frequency transform of the autocorrelation function. It is the magnitude of the input spectrum squared. Thus, there is really no new information in the autocorrelation function; the same information existed in the spectrum of the signal. But, as always, a change in perspective between these two domains often clarifies problems. In general, impulsive-type signals like pulse trains, bearing ping, or gear chatter show up better in correlation measurements. Signals containing several sine waves of different frequencies, like structural vibrations and rotating machinery, are clearer in the frequency domain.

4.25 Cross-Correlation

Cross-correlation is a measure of the similarity between two signals as a function of the time shift between them. If the same signal is present in both waveforms, it is reinforced in the cross-correlation function whereas any uncorrelated noise is reduced. In many network analysis problems, the stimulus can be cross-correlated with the response to reduce the effects of noise.

Cross-correlation is defined as

$$R_{xy}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_T x(t)y(t + \tau) dt \tag{4-30}$$

As with autocorrelation, an FFT analyzer computes this quantity indirectly. In this case it is computed from the cross-power spectrum.

$$R_{xy}(\tau) = \mathcal{F}^{-1} [G_{xy}] \tag{4-31}$$

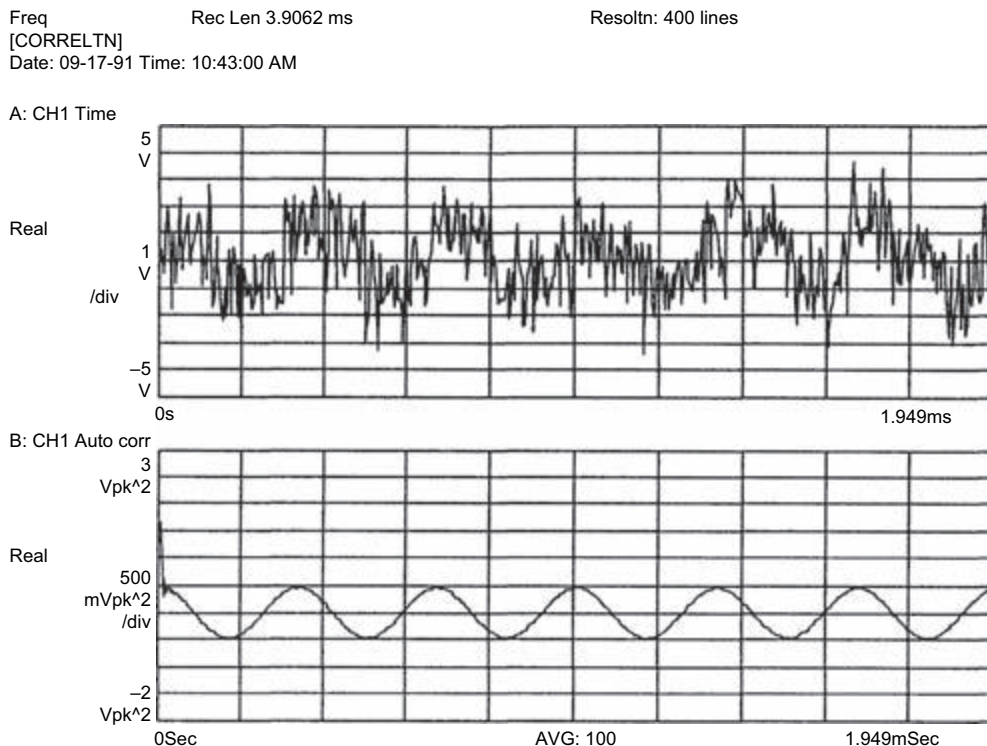


Figure 4-35 Autocorrelation of a sine wave buried by noise.

One application for cross-correlation is the determination of time delays between signals. These signals can be impulsive (e.g., radar or sonar application) or broadband random noise such as those encountered in system stimulus response measurements (transmission path delays, room acoustics, airborne noise analysis, and noise source identification).

4.26 Histogram

A histogram (Figure 4-36) shows how a signal's amplitude is distributed between its minimum and maximum values. A histogram displays number of samples versus amplitude. This measurement is useful for determining the statistical properties of noise and monitoring the performance of electromechanical positioning systems. Other measurement data derived from a histogram measurement are probability density function and cumulative density function.

The *probability density function* (PDF) (Figure 4-37) is the histogram data normalized to unit area. It is a statistical measurement of the probability that a specific level occurred. The probability of an input signal falling between two points is equal to the integral of the curve between those points. For more information see Chapter 8, section 8.1.

The *cumulative density function* (CDF) (Figure 4-38) is a measure of the probability that a level equal to (or less than) a specific level occurred. It is calculated by integrating the PDF.

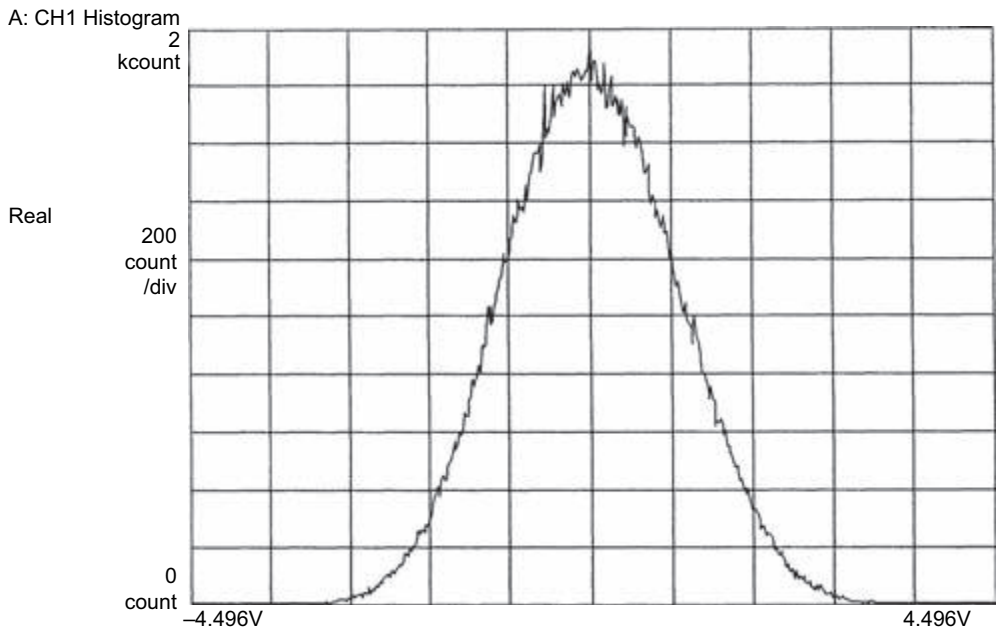


Figure 4-36 Histogram of random noise.

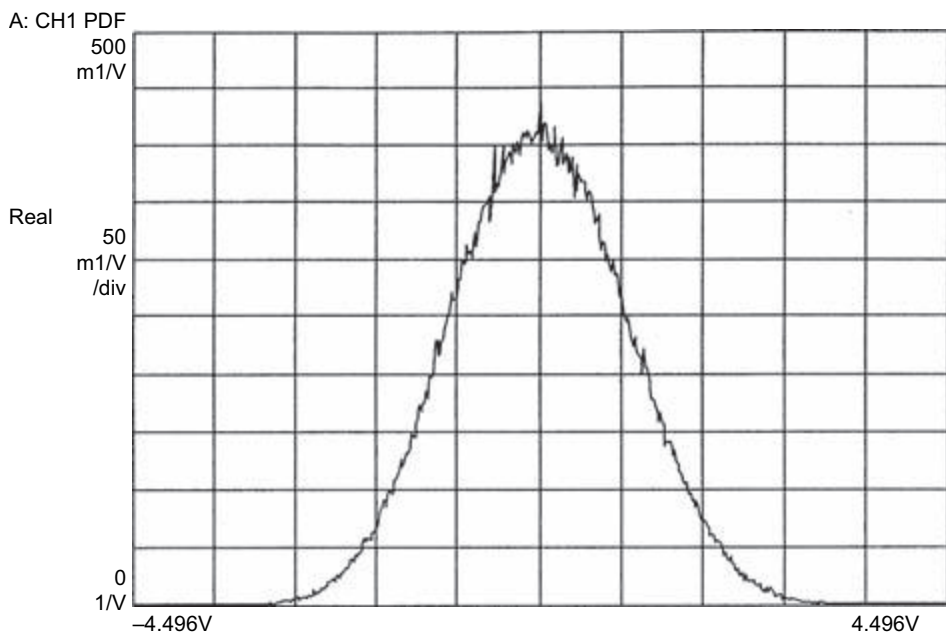


Figure 4-37 PDF of random noise.

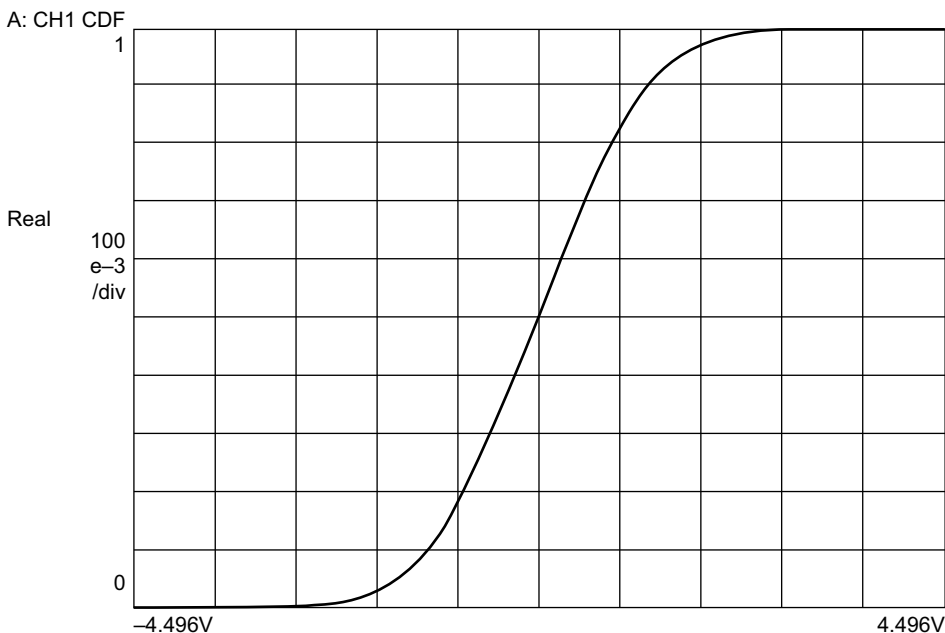


Figure 4-38 CDF of random noise.

4.27 Real-Time Bandwidth

Until now we have ignored the fact that it takes a finite time to compute the FFT of a time record. In fact, if the transform could be computed in less time than our sampling period, it could be ignored. Figure 4-39 shows that under this condition a new frequency spectrum could be obtained with every sample. As we saw from our discussion of aliasing, this could result in far more spectra every second than could be used. Because of the complexity of the FFT algorithm, it would take fast computational hardware to generate spectra this rapidly.

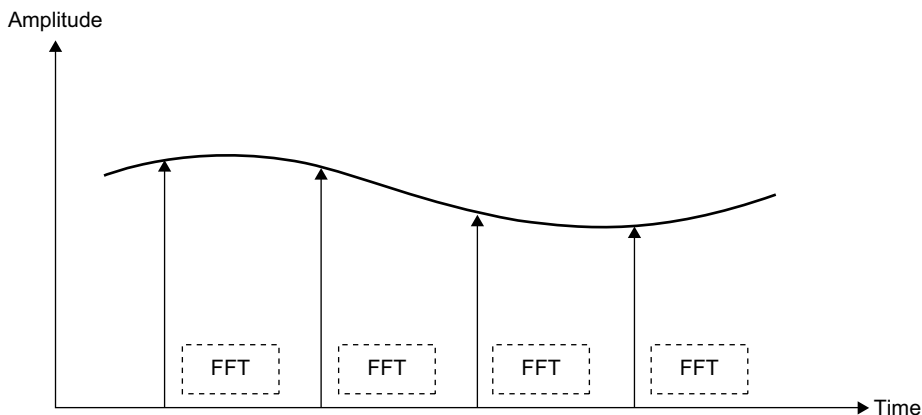


Figure 4-39 A new transform every sample.

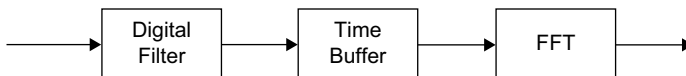


Figure 4-40 Time buffer added to block diagram.

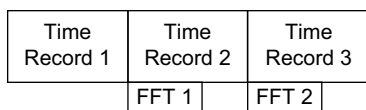


Figure 4-41 Real-time operation.

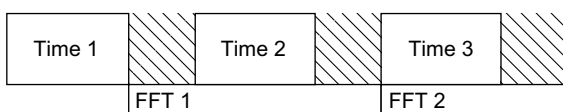


Figure 4-42 Non-real-time operation.

A reasonable alternative is to add a time-record buffer (Figure 4-40) before the FFT. Figure 4-41 shows how this arrangement allows the analyzer to compute the frequency spectrum of the previous time record while gathering the current time record. If the transform can be computed before the time-record buffer fills, then the analyzer is said to be operating in real time.

To see what this means, consider the case where the FFT computation takes longer than the time required to fill the time record. This is shown in Figure 4-42. Although the buffer is full, the last transform has not been completed, so data collection must stop. When the transform is finished, the time record can be transferred to the FFT and collection of another time record begun. Because some input data were missed, the analyzer is said to be no longer operating in real time.

Keep in mind that the time record can vary depending on the frequency span of the analyzer. For wide frequency spans, the time record is shorter, allowing less time for the FFT computation to complete. The frequency span or bandwidth setting where the FFT computation time and the time record are equal is called the *real-time bandwidth* (RTBW). For frequency spans at or below the RTBW, the analyzer does not miss any data.

4.28 Real-Time Bandwidth and RMS Averaging

There are situations when a measurement requires RMS averaging. It is important to know how the real-time specifications of the FFT analyzer affect the measurement. We might be interested in the spectral distribution of noise, or in reducing the variation of a signal contaminated by noise. There is no requirement in averaging that records must be consecutive with no gaps.⁹ In these situations, a small real-time bandwidth does not affect the accuracy of the results.

⁹ This is because averaging is useful only if the signal is periodic and the noise stationary.

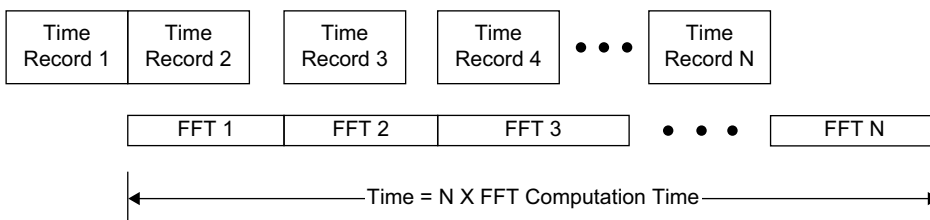


Figure 4-43 RMS averaging time.

However, the real-time bandwidth does affect the speed with which an RMS-averaged measurement can be made. Figure 4-43 shows that for frequency spans above the real-time bandwidth, the time to complete the average of N records is dependent only on the time to compute the N transforms. Rather than continually reducing the time to compute the RMS average as we increase our span, we reach a fixed time to compute N averages.

Therefore, a small real-time bandwidth is a problem only when RMS averaging large spans using a large number of averages. Under these conditions it takes longer to get the answer. Since wider real-time bandwidths require faster computations and therefore a more expensive processor, there is a straightforward trade-off of time versus money. In the case of RMS averaging, higher real-time bandwidth gives somewhat faster measurements but at increased analyzer cost.

4.29 Real-Time Bandwidth and Transients

Real-time bandwidth is an important consideration when analyzing transient events. If the entire transient fits within the time record (Figure 4-44), the FFT computation time is of little interest. The analyzer can trigger on the transient and store the event in the time-record buffer. The time to compute its spectrum is not important.

However, if a transient event contains high-frequency energy and lasts longer than the time record necessary to measure the high-frequency energy, the processing speed of the analyzer is critical. As shown in Figure 4-45, not all of the transient will be analyzed if the computation time exceeds the time-record length.

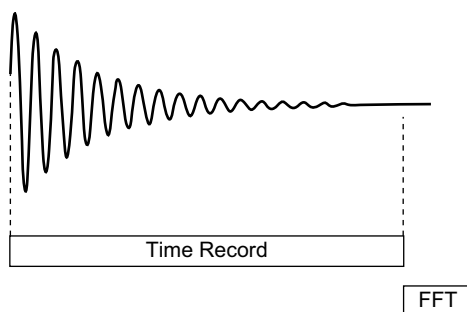


Figure 4-44 Transient fits in time record.

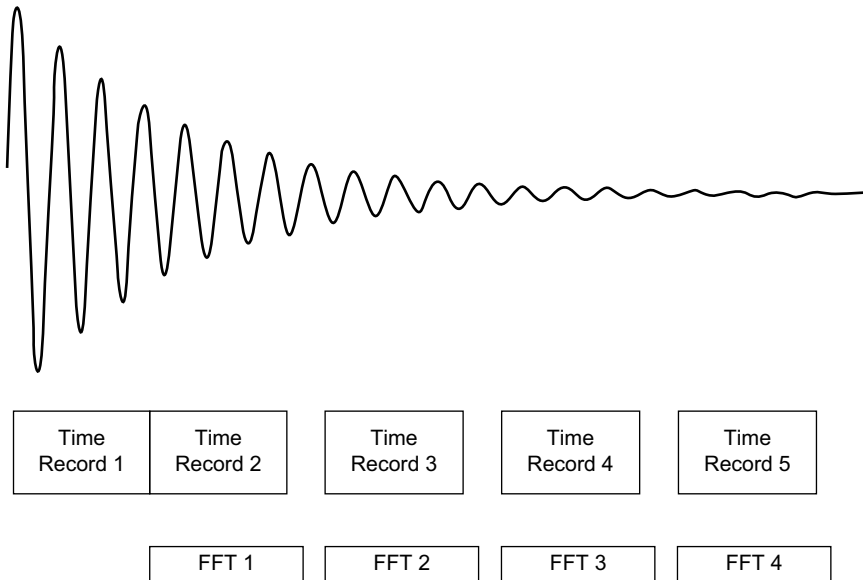


Figure 4-45 Transient longer than one time record.

When the transient is longer than the time record, it is imperative that there be some way to rapidly record the spectrum. Otherwise, the information is lost as the analyzer updates the display with the spectrum of the latest time record. A special display that can show more than one spectrum (“waterfall” display), large storage device or a high-speed link to a computer is needed. The output device must be able to record a spectrum for every time record or information will be lost.

4.30 Overlap Processing

Previously we considered the case where computing the FFT took longer than collecting the time record. In this section we will look at a technique called overlap processing. This can be used when the FFT computation takes less time than gathering the time record.

To understand overlap processing, look at Figure 4-46. This is the diagram for a situation where the time record is much longer than the FFT computation time (e.g., low-frequency analysis). Without overlap capability the FFT processor is sitting idle much of the time. If we take a snapshot of the time data each time the FFT process completes and then starts the next FFT, it is possible to do consecutive FFTs with little idle time, as shown in Figure 4-47. The data used by the current FFT process will not all be new. The snapshot of the time data will contain some of the data used in the previous FFT plus whatever new data were collected during the time required to compute the previous FFT. To understand the benefits of overlap processing, let us look at the same cases used in the last section.

Earlier we concluded that to adjust a test device effectively a new spectrum is needed every few tenths of a second. Without overlap processing, this limits our resolution to a few hertz. With overlap processing, our resolution is unlimited.

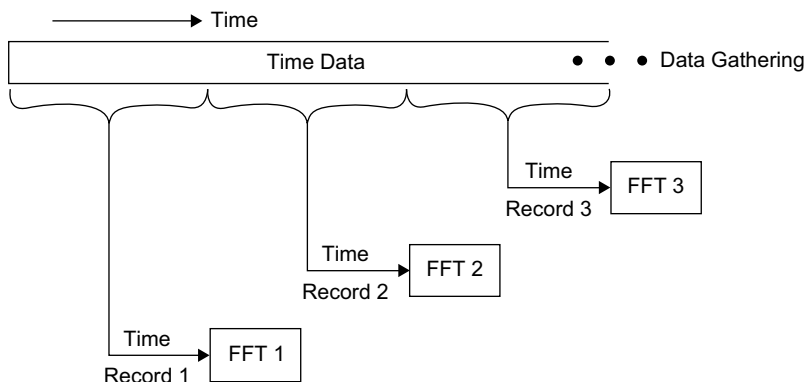


Figure 4-46 Nonoverlapped processing is performed only on completely new data (time records).

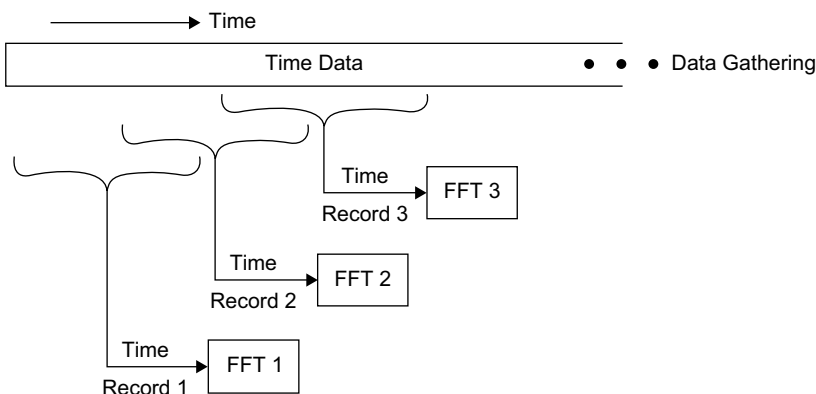


Figure 4-47 Overlapped processing is performed on data that combine old and new. The time between FFTs represents display processing.

But overlap processing does not give us something for nothing. Since the overlapped time record contains old data from before the device was adjusted, it is not completely correct. It does indicate the direction and the amount of change, but we must wait one full time record after the change for the new spectrum to be accurately displayed. Nonetheless, by indicating the direction and magnitude of the changes every few tenths of a second, overlap processing greatly increases the responsiveness of the measurement.

Overlap processing can dramatically reduce the time needed to compute RMS averages with a given variance. Recall that window functions reduce the effects of leakage by weighting the ends of the time record to zero. Overlapping eliminates most (if not all) of the time that would be wasted taking these data. Since some overlapped data are used twice, more averages must be taken to get a variance that is comparable to the nonoverlapped case. Figure 4-48 shows the improvements that can be expected by overlapping.

For transients shorter than the time record, overlap processing is useless. For transients longer than the time record, the real-time bandwidth of the analyzer and spectrum recorder is

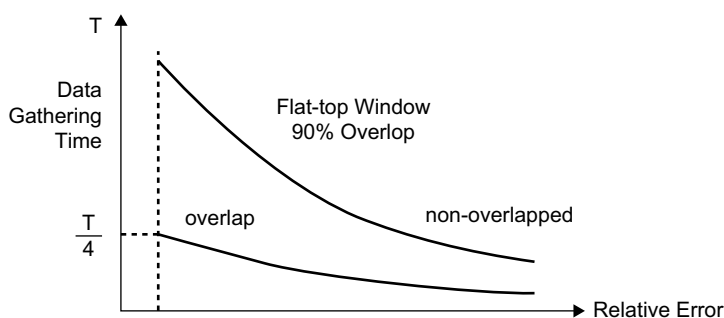


Figure 4-48 RMS averaging is faster with overlap processing.

usually a limitation. If it is not, overlap processing allows more spectra to be generated from the transient. Usually this improves the resolution of resulting plots.

4.31 Swept Sine

Some FFT analyzers also make swept sine measurements. Swept sine analysis is a very traditional measurement technique, but its implementation in an FFT analyzer is somewhat different than in a swept analyzer. FFT analyzers with swept sine capability use a swept sine wave source and a time domain integration process to emulate a tracking band-pass filter. The sine source sweeps across a user-selected frequency band, exciting the DUT. At each discrete frequency point during the sweep, the analyzer measures and displays the relative magnitudes and phases of the DUT's sinusoidal response.

Swept sine measurements provide a very good signal-to-noise ratio and can characterize nonlinear systems. At each point in the sweep, the DUT exhibits a transient response and a steady-state response. The analyzer waits for the transient response to settle out and then measures the steady-state response. The swept sine measurement allows the analyzer to characterize nonlinearities and better excite the DUT because the energy is concentrated in a narrow frequency band. To reduce the effect of noise, the analyzer integrates the input signal over several cycles.

Most FFT analyzers that provide swept sine measurements include some automatic adjustments to optimize the results. These include automatic source level adjust, automatic input range adjust, and automatic resolution adjust.

For nonlinear devices, the transfer function varies depending on the input level. With autoleveling, the analyzer adjusts the signal source level to keep the DUT output level within a specified range.

With autoranging, the analyzer adjusts the input range up or down when the DUT output level goes above or below the optimum for the current range. This can greatly increase the dynamic range.

With autoresolution, the analyzer adjusts the spacing between adjacent measurement points, taking finer or coarser steps where necessary. Coarser steps minimize measurement time, but autoresolution can narrow the steps where there are rapid changes in the response. This minimizes the sweep time while still catching fast changes in amplitude or phase.

4.32 Octave Measurements

Some FFT analyzers also make *octave measurements*, which are commonly used for performing acoustic measurements. An octave measurement computes power in bands using banks of filters covering several octaves. Each higher filter has a wider bandwidth than the previous filter, with the filter spacing being logarithmic rather than linear. The most common spacing is 1/3 octave, but some analyzers also provide full octave and 1/12 octave spacing.

Octave measurements are displayed on a logarithmic x-axis, so each band appears to be the same width. In Figure 4-49, the top trace shows a third octave with a log x-axis. The bottom trace shows the same data on a linear x-axis. Notice that the higher frequency bands are much wider than the lower frequency bands.

FFT analyzers that make octave measurements usually include an A-weight filter to simulate the frequency response of the human ear. Third-octave measurements represent how the human ear perceives the frequency content of a signal, but the frequency resolution does not reveal the exact spectral component of the signal. To diagnose the specific cause of a noise problem, the analyzer’s FFT measurement is more useful.

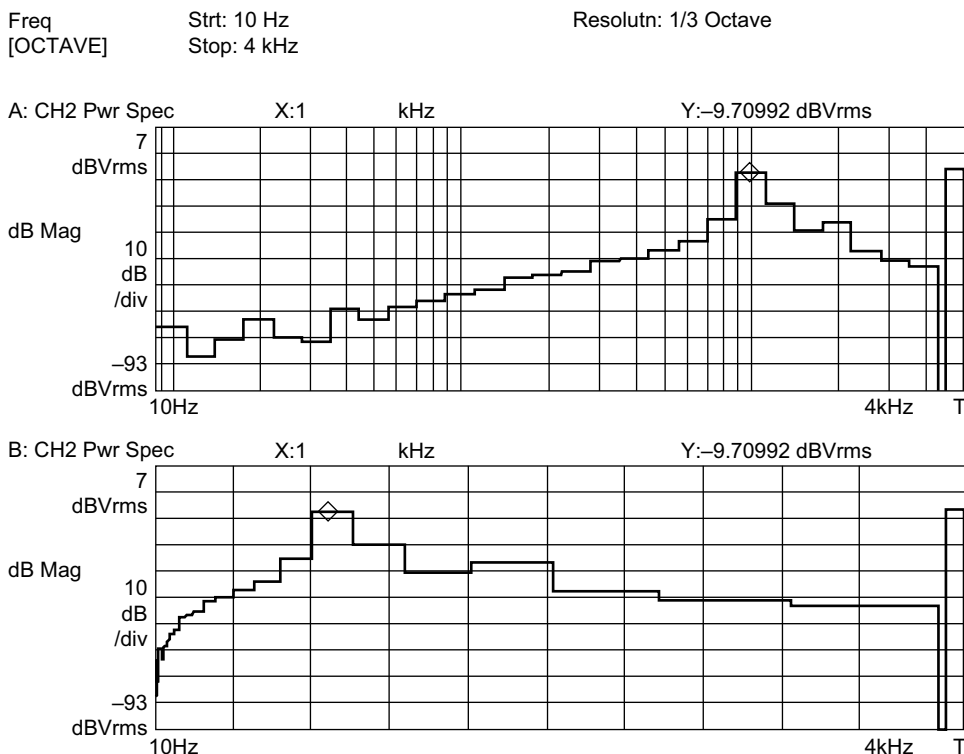


Figure 4-49 Third octave measurement results displayed on a logarithmic x-axis and a linear x-axis.

Bibliography

Agilent Technologies. "Agilent 35670A Dynamic Signal Analyzer," Publication Number 5966-3064E, Santa Clara, CA, 2009.

Agilent Technologies. "Effective Machinery Measurements Using Dynamic Signal Analyzers," Application Note 243-1, Publication Number 5962-7276E, Santa Clara, CA, 1997.

Agilent Technologies. "The Fundamentals of Modal Testing," Application Note 243-3, Publication Number 5954-7957E, Santa Clara, CA, 2000.

Agilent Technologies. "The Fundamentals of Signal Analysis," Application Note 243, Publication Number 5952-8898E, Santa Clara, CA, 2000.

Brigham, E. Oran. *The Fast Fourier Transform and Its Applications*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1988.

McGillem, Clare D., and George R. Cooper. *Continuous and Discrete Signal and System Analysis*. New York: Holt, Rinehart and Winston, Inc., 1974.

Oliver, Bernard M., and John M. Cage. *Electronic Measurements and Instrumentation*. New York: McGraw-Hill Book Company, 1971.

Oppenheim, Alan V., and Alan S. Willsky. *Signals and Systems*, 2d ed. Upper Saddle River, NJ: Prentice Hall, Inc., 1996.

Oppenheim, Alan V., and Ronald W. Schaffer. *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1975.

Ramirez, Robert W. *The FFT, Fundamentals and Concepts*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1985.

Schwartz, Mischa. *Information, Transmission, Modulation, and Noise*, 3rd ed. New York: McGraw-Hill Book Company, 1980.

Schwartz, Mischa, and Leonard Shaw. *Signal Processing*. New York: McGraw-Hill Book Company, 1975.

Ziemer, Rodger E., William H. Tranter, and D. Ronald Fannin. *Signals and Systems: Continuous and Discrete*, 4th ed. Upper Saddle River, NJ: Prentice Hall, Inc., 1998.

Swept Spectrum Analyzers

The traditional method for implementing a spectrum analyzer is the swept heterodyne block diagram. Similar to a radio receiver, the spectrum analyzer is automatically tuned (swept) over the band of interest. This type of analyzer has been gradually replaced by the fast Fourier transform (FFT) analyzer at low frequencies, but the swept analyzer remains the dominant technology in the radio frequency range and above. In recent years, the swept analyzer has been combined with the FFT analyzer to provide the advantages of both techniques.

5.1 The Wave Analyzer

The bank-of-filters analyzer, which was examined in the previous chapter, uses a large number of fixed filters to implement a spectrum analyzer. Another approach is to use one filter, but to make it tunable over the frequency range of interest (Figure 5-1). Since this technique allows only one frequency to be measured at a time, it is not a true spectrum analyzer but is called a *wave analyzer* or *wave meter*.

The user tunes the wave analyzer to the frequency of interest and reads the signal level present at that frequency. The bandwidth of the tunable filter determines the resolution bandwidth, RBW, of the wave analyzer. It is desirable for the filter to be as flattop as possible, with steep skirts so that equal amplitude signals within the passband of the filter produce the same meter reading.

This type of instrument has been used extensively for making simple “tuned voltmeter” measurements and still exists today in the form of a *selective level meter*. Selective level meters have very flattop passbands, resulting in excellent amplitude accuracy.

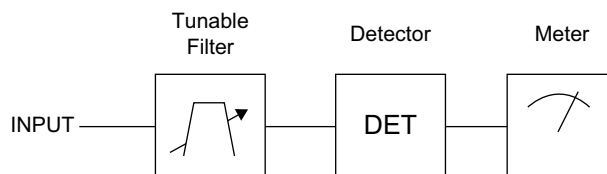


Figure 5-1 A conceptual block diagram of the wave analyzer.

5.2 Heterodyne Block Diagram

Practical tunable band-pass filters have severe restrictions on the tuning range of the filter's center frequency. Thus, the wave analyzer is rarely implemented using an actual tunable filter. Instead of moving the filter in frequency, the input signal is translated in frequency and the filter's frequency remains fixed. The band-pass filter's center frequency is called the *intermediate frequency* (IF) and the filter is called the *IF filter*.

The simplified block diagram of a practical wave analyzer is shown in Figure 5-2. The key component of this block diagram is the mixer. The mixer is a three-port device that is driven by the input signal of the analyzer (usually called the *RF signal*) and the *local oscillator* (LO) signal. The output of the mixer is at the IF.

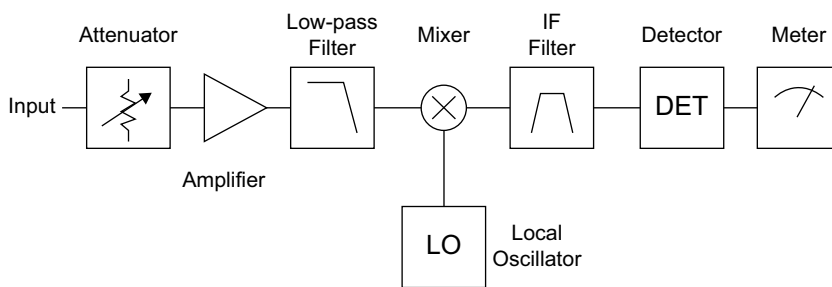


Figure 5-2 A more practical block diagram of the wave analyzer.

Ideally, the mixer functions as a multiplier. Suppose the input is a cosine:

$$v_{RF}(t) = A \cos(2\pi f_{RF}t) \quad (5-1)$$

and

$$v_{LO}(t) = \cos(2\pi f_{LO}t) \quad (5-2)$$

The output of the mixer is

$$v_{IF}(t) = A \cos(2\pi f_{RF}t) \cos(2\pi f_{LO}t) \quad (5-3)$$

$$v_{IF}(t) = \frac{A}{2} [\cos(2\pi f_{LO}t + 2\pi f_{RF}t) + \cos(2\pi f_{LO}t - 2\pi f_{RF}t)] \quad (5-4)$$

Therefore, the mixer's output is the sum and difference frequencies of the LO and RF signals.¹

This characteristic is used to implement the superheterodyne block diagram. The IF filter always remains tuned to the same center frequency, and the mixer is used to shift the input signal in frequency so that it falls on the center of the IF filter. This makes the IF filter easier to build since it does not require a tunable center frequency. Of course, the LO

¹ Practical mixers will usually have other higher-order products that are ignored here.

must be made tunable, but this is usually an easier task than building a filter that tunes over a wide range.

The mixer produces both the sum and difference frequencies of the input and LO. Only the sum or the difference frequency is used since, by design, it will fall directly on the IF. The other frequency will be rejected by the IF filter. This requires some careful choices in defining the LO and IF frequencies.

A numerical example should help explain the operation of the superheterodyne block diagram. Suppose a wave analyzer is required to measure signals from 0 to 10 MHz. The chosen IF is 20 MHz, and the LO operates between 20 MHz and 30 MHz. Now, suppose that the input frequency happens to be 5 MHz. To measure this frequency, the LO is tuned to 25 MHz, producing the sum and difference frequencies of 20 MHz and 30 MHz. The 20 MHz signal is exactly the IF (by design) and is passed through the IF filter and detected and displayed on the meter. The 30 MHz signal falls outside of the IF filter and is rejected.

If the input frequency were changed to 1 MHz, the LO would have to be tuned to 21 MHz, producing sum and difference frequencies of 20 MHz and 22 MHz. Again, the 20 MHz signal is the IF and is measured while the 22 MHz signal falls outside the IF filter and is ignored.

The low-pass filter at the input of the block diagram is known as the *image filter*. If this filter were not included, undesirable frequencies could enter the mixer and be translated down to the IF, corrupting the measurement. Suppose the wave analyzer is still tuned to 5 MHz. If a 45 MHz signal made its way into the mixer, it would mix with the LO frequency (25 MHz) and would produce sum and difference frequencies of 20 MHz and 70 MHz. The 70 MHz signal would be ignored, but the 20 MHz signal would fall directly on the IF filter and would be included in the measurement. Without an image filter the wave analyzer could not distinguish between the desired 5 MHz signal and the 45 MHz image frequency.

The image frequency (for this block diagram) causes the difference frequency (when mixed with the LO) to fall on the IF.

$$f_{\text{IF}} = f_{\text{IMAGE}} - f_{\text{LO}} \quad (5-5)$$

$$f_{\text{IMAGE}} = f_{\text{IF}} + f_{\text{LO}} \quad (5-6)$$

The LO frequency is the input frequency plus the IF or

$$f_{\text{LO}} = f_{\text{RF}} + f_{\text{IF}} \quad (5-7)$$

Thus,

$$f_{\text{IMAGE}} = f_{\text{RF}} + 2f_{\text{IF}} \quad (5-8)$$

The image frequency is twice the IF away from the desired input frequency. This holds for the case shown where the IF is higher than the input frequency.

5.3 The Swept Spectrum Analyzer

The wave analyzer can measure only one frequency at a time. An obvious enhancement is to have the analyzer automatically sweep through the frequency range of interest. In a spectrum analyzer this is accomplished by sweeping the LO. Figure 5-3 shows how the wave analyzer

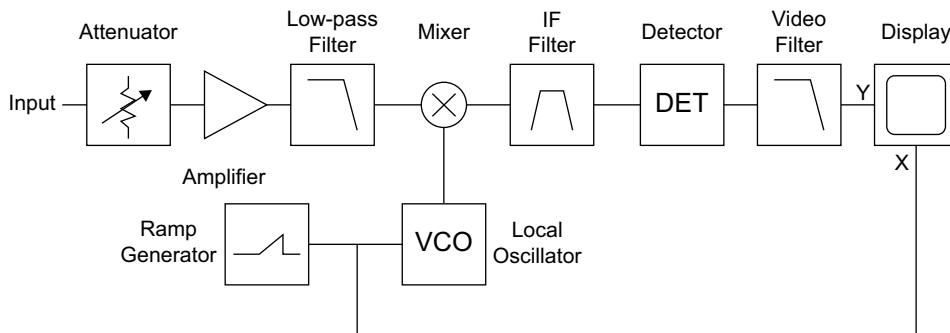


Figure 5-3 A simplified block diagram of a swept spectrum analyzer.

block diagram could be converted into a spectrum analyzer by using a voltage-controlled oscillator (VCO) as the LO. A ramp generator is used to produce a linearly increasing voltage, which drives the tuning port of the VCO. The same ramp voltage is applied to the horizontal (*x*) axis of the display, while the detector output is low-pass filtered and connected to the vertical (*y*) axis. As the LO is swept in frequency, the spectrum of the input signal is automatically plotted on the display. The low-pass filter at the output of the detector is called the *video filter* and is a postdetection filter (as discussed in Chapter 10), which serves to smooth out the response as the analyzer sweeps.

As shown, the block diagram is implemented in a totally analog fashion. Although this is a practical technique, the advent of the microprocessor and the digital display has caused the block diagram to take on a digital flavor (Figure 5-4). For example, the LO is often implemented using digital synthesis techniques that lend themselves to microprocessor control. (The LO may be stepped or swept in frequency under microprocessor control.) The output of the IF filter (or the detector) may be sampled and converted to digits by an analog-to-digital converter (ADC), which is read by the microprocessor. The display in a modern spectrum

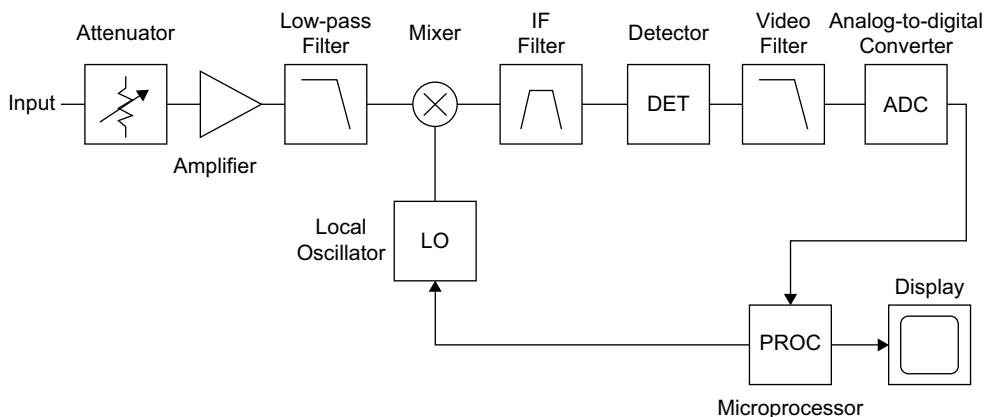


Figure 5-4 A simplified block diagram of a spectrum analyzer using microprocessor control.

analyzer is always a digital graphics display. That is, the display graphics information is written to a designated area of memory and the display is refreshed from this memory. This eliminates any problems with display refresh rate being too slow due to a slow sweep rate since the display can be refreshed much faster than the sweep rate.

Although the analog block diagram of Figure 5-3 has been largely replaced by the digital equivalent, it still represents a good conceptual basis for understanding the operation of the spectrum analyzer.

5.4 Practical Considerations

The block diagram previously discussed uses a single mixer/IF stage and is therefore called a *single conversion receiver*. This simple block diagram can be used to implement a spectrum analyzer, but its performance is limited. Modern analyzers use much more complicated block diagrams to achieve state-of-the-art performance.

Some factors in the block diagram design call for a high IF, whereas others require a low IF. A high IF makes the rejection of image frequencies easier, but narrow IF filters and detectors are more difficult to implement at high frequencies. Conversely, narrow filters and detectors are easier to build at low frequencies, but the image rejection problem is made more difficult. A compromise of sorts is often used with multiple conversion stages cascaded. Each conversion section contains a mixer, an LO, and an IF filter. (The LOs may all be derived from the same master oscillator.) Multiple conversion stages are the rule rather than the exception in spectrum analyzers.

Many of the circuit blocks in the spectrum analyzer block diagram are complex systems within themselves. For example, an LO can be made up of several oscillators or frequency synthesis loops. Each frequency synthesis loop may contain one or more mixers, a low-pass filter, and oscillator. These blocks may be configured such that the block diagram of the analyzer changes significantly depending on the frequency range that is being measured.

Regardless of the complexity of the actual spectrum analyzer block diagram, conceptually it simply implements a sweeping tuned filter.

5.5 Input Section

The input to the spectrum analyzer block diagram has a variable attenuator, often followed by an amplifier. The purpose of this input section is to control the signal level applied to the rest of the instrument. If the signal level is too large, the analyzer circuits will distort the signal, causing distortion products to appear along with the desired signal. If the signal level is too small, the signal may be masked by noise present in the analyzer. Either problem tends to reduce the dynamic range of the measurement.

Some instruments provide an autorange feature that automatically selects an appropriate input attenuation. Other instruments require the user to select the appropriate input attenuation. The input circuitry of a typical analyzer is sensitive and will not withstand much abuse. Careful attention should be paid to the allowable signal level at the input, particularly for microwave analyzers. Some instruments tolerate DC voltages at their inputs, but others require that little or no DC be applied.

5.6 Resolution Bandwidth

The bandwidth of the last IF filter usually determines the resolution bandwidth (RBW), of the instrument. If multiple IF filters are used, the composite response of the IF chain determines the resolution bandwidth. Usually, one of the IF filters will be significantly narrower than the others and alone will determine the resolution bandwidth.

Multiple resolution bandwidths are supplied by simply switching in different filters. Wider bandwidth filters settle faster, providing faster measurements. Narrow bandwidth filters take longer to settle but produce better frequency resolution and better signal-to-noise ratio (see Chapter 10).

5.7 Sweep Limitations

The swept spectrum analyzer generally provides a significant increase in measurement speed over the wave analyzer since the entire frequency range of interest can be displayed at once. This is not meant to imply that the spectrum analyzer can be swept arbitrarily fast. The IF filter (resolution bandwidth filter) must have time to respond to the changing signal level that it experiences at its input.

Consider the case where the spectrum analyzer sweeps past a sinusoidal signal (Figure 5-5). In this case, we will analyze the situation by considering the IF filter to be fixed and the signal to be moving. The signal starts well outside of the passband of the filter (Figure 5-5a). Then, the signal starts up the skirt of the filter with the filter's output level increasing accordingly (Figure 5-5b). Finally, the signal enters the passband of the filter and starts down the other side (Figure 5-5c).

If the signal is swept slowly enough, the shape of the IF filter is traced out on the spectrum analyzer display. (Normally, the IF filter bandwidth is small compared with the frequency span being swept, so the IF filter shape will appear as a spectral line on the display.) If the signal is swept too fast, the filter does not have time to respond and two types of display errors occur (Figure 5-6). The amplitude of the spectral line is smaller than the slowly swept case, and the spectral line will shift to the right slightly, causing a frequency error. Additionally, there may be filter "ringing" down the back edge of the filter shape.

How fast is too fast of a sweep rate? Ideally, the filter should be swept infinitely slow since the filter response time will always degrade the measurement. In practice, the filter can be swept at some finite rate as long as some small error can be tolerated. If this error is small compared with other errors in the analyzer, then there is no penalty for sweeping. A typical error limit due to sweep induced errors is 0.1 dB. The maximum sweep rate (with such an error limit) is proportional to the square of the resolution bandwidth.

$$\text{sweep rate (max)} = \text{RBW}^2/k \quad (5-9)$$

where

k = a factor depending on the resolution bandwidth filter characteristics

A typical value for k is 2 (for Gaussian filters), and the sweep rate has units of Hz/sec. If a steep-walled filter is used, the response time of the filter increases, causing k to be larger.

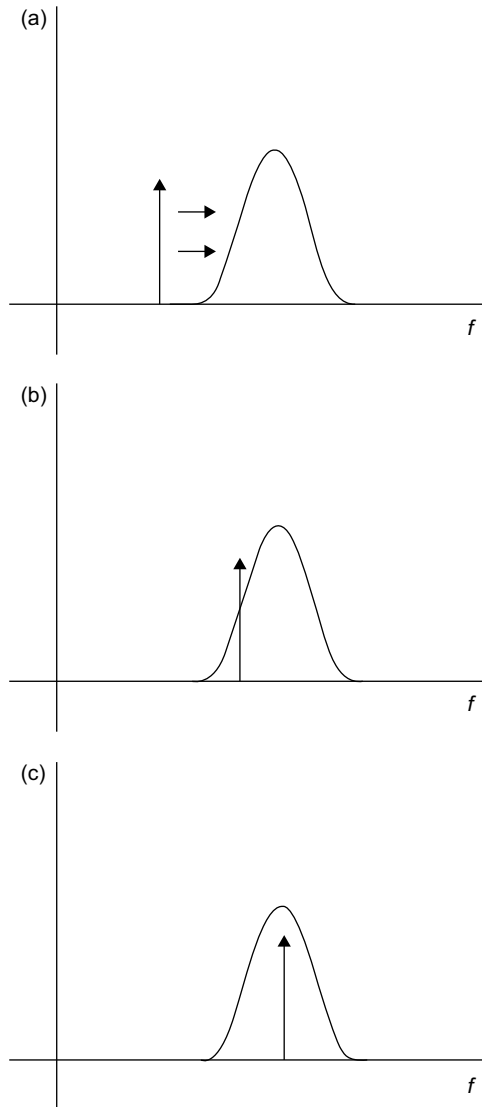


Figure 5-5 A sine wave passes through the IF filter of a spectrum analyzer.

In a wave analyzer, the shape of the IF filter is usually designed to be steep walled and as flattop as possible. However, this is inappropriate for an analyzer intended to sweep due to the increased sweep time required. In swept analyzers a more rounded filter such as a Gaussian filter is used to minimize the sweep time (Figure 5-7).

The minimum sweep time for a particular frequency span is given by

$$T_s = f_{\text{span}}k/\text{RBW}^2 \quad (5-10)$$

Other sweep limitations such as the maximum local oscillator sweep rate may be present in the instrument.

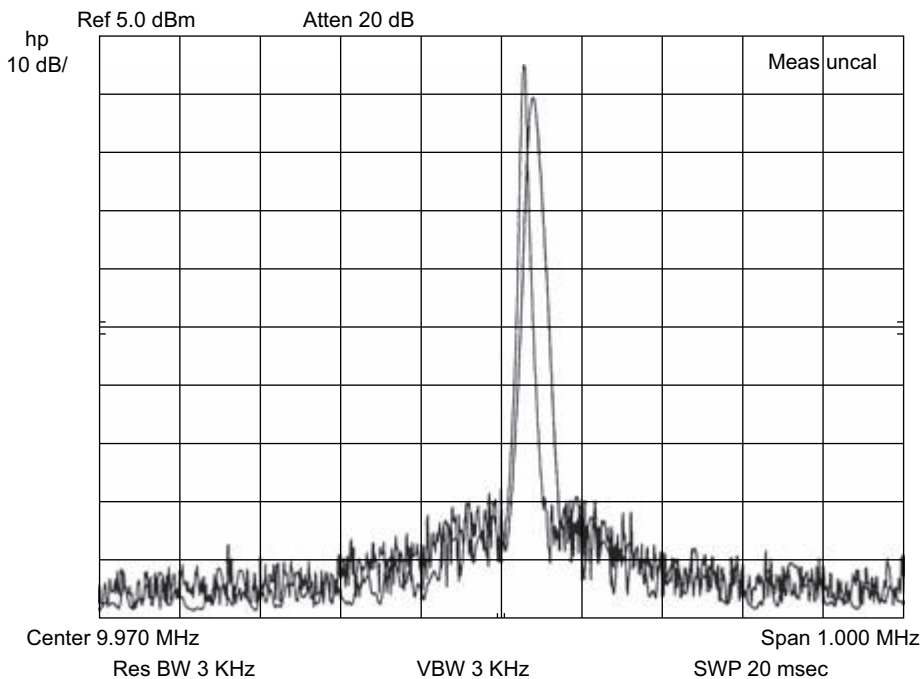


Figure 5-6 These two measurements show the effects of sweeping too fast. The leftmost spectral line was swept correctly. Sweeping too fast causes the spectral line to be smaller in amplitude and shifted to the right.

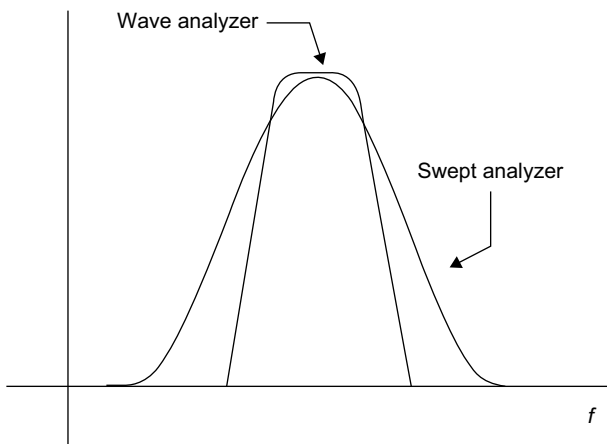


Figure 5-7 The IF (resolution bandwidth) response of a spectrum analyzer is more rounded and less selective than a wave analyzer.

Although it is important for the spectrum analyzer user to understand the sweep rate limitations to optimize the measurement, modern spectrum analyzers provide automatic selection of the sweep time. The user is protected from making an erroneous measurement as long as the autoselection feature is not overridden.

5.8 Specialized Sweep Modes

Many spectrum analyzers include a special sweep mode called *manual sweep*. This feature causes the analyzer to operate like a wavemeter, measuring the spectrum at only one frequency. The user can adjust the frequency as desired and can manually sweep the entire frequency span, if desired. In some measurement situations, the amplitude at one particular frequency is important. Manual sweep is especially useful if the required resolution bandwidth is very narrow. The user can avoid having to wait for a long sweep to occur by forcing the analyzer to measure only at the frequency of interest.

Another specialized sweep mode is *zero span* operation, which causes the analyzer to sweep while maintaining a constant measurement frequency. This mode allows amplitude variations to be displayed, as discussed in Chapter 6.

Some spectrum analyzers provide another sweep mode called *discrete sweep*, *program sweep*, or *list sweep* that lets the user specify a list of frequencies to test. The analyzer automatically hops from frequency to frequency, measuring the spectral content at each one. For measurement applications such as production test, where measurements at a small number of frequencies can adequately verify correct operation of the device under test, the total measurement time can be reduced.

5.9 Local Oscillator Feedthrough

One particularly noticeable imperfection in the mixer of a spectrum analyzer is the phenomenon known as *LO feedthrough*. The ideal mixer produces only the sum and difference frequencies at the IF port. In a real mixer, the LO and RF signals (at reduced amplitude) also appear at this port. In most cases, the LO frequency is far enough away from the center of the IF that the LO feedthrough does not appear in the measurement. However, when the LO frequency is the same as (or very near) the IF, the LO signal is passed through the IF filter and appears in the measurement. This LO frequency corresponds to an input frequency of 0 Hz (DC).

LO feedthrough is also known as the *DC response* since that is where it appears on the spectrum analyzer display. In nonsynthesized spectrum analyzers (with limited frequency accuracy in the LO), this is used as a method of locating 0 Hz on the display.

If the IF filter were infinitely narrow, the LO feedthrough would appear only at exactly 0 Hz. With a finite-width IF filter, the LO feedthrough extends from 0 Hz to approximately $RBW/2$, following the shape of the IF filter (Figure 5-8). It may be necessary to reduce the resolution bandwidth to prevent the LO feedthrough from interfering with the measurement.

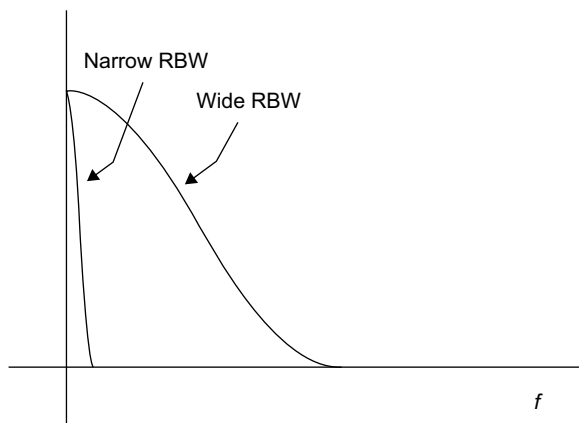


Figure 5-8 Local oscillator feedthrough appears as a response at DC whose width depends on the resolution bandwidth.

5.10 Digital IF Section

When implemented using analog circuitry, the output of the IF filter drives a log amp, which drives the detector circuit. The log amp compresses the signal level according to a logarithmic function. (For an input voltage amplitude, v , the output voltage amplitude is $\log(v)$.) This greatly reduces variation in signal level seen by the detector and simultaneously provides the user with a logarithmic vertical scale, which is calibrated to read in decibels. The logarithmic scale is desirable in a spectrum analyzer due to the large variation in signal levels. The detector produces a DC level proportional to the AC level of the signal in the IF section. When the output of the detector is sampled and converted to digital form, it is important that the output be sampled often enough so that spectral components are not missed.

In modern instruments, the resolution bandwidth filters and the detector have been implemented using digital signal processing (Figure 5-9). The signal is digitized while it is still at the last IF. A digital filter algorithm is then used to provide the resolution bandwidth function, and the filtered signal is detected digitally. This digital implementation provides a high degree of flexibility such that a variety of detector algorithms, logarithmic amplification, video filters, and averaging techniques can be employed. A digital IF section provides very stable narrow resolution bandwidths (1 Hz or even narrower), since digital filters do not

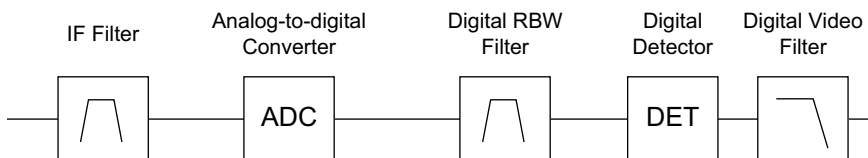


Figure 5-9 The IF detector section of a swept spectrum analyzer can be implemented using digital signal processing techniques.

exhibit any drift. Errors associated with the log amp and detector are essentially eliminated since both of these functions are performed digitally.

The use of digital RBW filters has enabled improved sweep rate and better frequency resolution in swept analyzers. The response of the digital filter can be tightly controlled and optimized for the best spectrum analyzer performance compared with their analog equivalents. A digital IF section requires an ADC with high dynamic range so that closely spaced signals can be captured accurately. That is, multiple signals may be present in the IF section such that the ADC must have enough resolution to simultaneously digitize both large and small signals.

5.11 Types of Detectors

Since the detector is implemented digitally, a variety of detector algorithms can be included, optimized for different signal types. A key point is that the number of samples coming out of the IF filter is often much larger than can be included on the display. In particular, when the sweep speed is slow there may be thousands of samples available for each displayed spectrum point.

Some common detector types are as follows:

- *Normal*: Also referred to as the Rosenfell detector,² this determines the peak of sinusoidal signals and yields alternating maximums and minimums of noise-like signals. This is the most versatile detector for general use.
- *Average*: The detector determines the average of the signal within the sweep points. Often, multiple types of averaging are available: *log power* (also called *video*), *power* (also called *RMS*), and *voltage envelope*.
- *Peak*: The detector determines the maximum of the signal within the sweep points. Peak detection is good for analyzing sinusoids but tends to overrespond to noise when sinusoids are not present.
- *Sample*: The detector indicates the instantaneous level of the signal at the center of the sweep points represented by each display point. Sample detection is good for displaying noise or noise-like signals.
- *Negative peak*: The detector determines the minimum of the signal within the sweep points. This mode is useful for distinguishing between random and impulse noise. Negative peak detection does not give the analyzer better sensitivity, although the noise floor may appear to drop.
- *Quasi-peak*: This is a fast-rise, slow-fall detector used in making electromagnetic interference (EMI) measurements, compliant with the International Special Committee on Radio Interference (CISPR) Publication 16-1-1. Quasi-peak detection displays a weighted, sample-detected amplitude using specific, charge, discharge, and meter time constants derived from the legacy behaviors of analog detectors and meters. See Chapter 16 for more information on electromagnetic compatibility (EMC) measurements.

² Rosenfell is not a person's name but rather a description of the algorithm that checks if the signal "rose and fell" within the frequency range represented by a given data point.

5.12 The Tracking Generator

A *tracking generator* is a very useful addition to the basic spectrum analyzer block diagram. A tracking generator, as the name implies, provides a sinusoidal output whose frequency is the same as the analyzer's input frequency. A tracking generator allows a spectrum analyzer to perform basic network measurements. The output of the tracking generator is connected to the input of the device under test and the response is measured with the analyzer's receiver. As the analyzer sweeps, the tracking generator is always operating at the receiver's frequency, and the transfer characteristics of the device can be measured.

While the classic tracking generator is frequency locked to the spectrum analyzer measurement frequency, some instruments support a more flexible source (which may require an external signal generator). *Power sweep* is the capability for the source to change its output power as the spectrum analyzer sweeps. This allows the analyzer to measure a device's behavior as the signal level is varied, typically used while measuring one fixed frequency.

Another powerful feature allows the source to sweep a frequency range that is different from the measurement frequency but is still synchronized with the spectrum analyzer. For example, if the source frequency is one-half of the measurement frequency, the analyzer can make a swept measurement at the second harmonic of the source frequency. The frequency of the source can be programmed to be a function of the analyzer frequency using

$$f_s = kf_a + f_{offset}$$

where

f_s = the source frequency

k = a multiplier factor, usually limited to a ratio of integers

f_{offset} = a fixed frequency offset

5.13 FFT versus Swept Measurements

Besides the inherently simpler block diagram in the FFT approach, the FFT analyzer provides a speed improvement over the swept analyzer. As previously discussed, the swept analyzer measurement speed is limited by its resolution bandwidth, with the measurement time being inversely proportional to RBW^2 . At low frequencies, very narrow resolution bandwidths are required to separate closely spaced spectral lines. Narrow resolution bandwidths require a longer sweep time so the total measurement time can get unacceptably long. On the other hand, the FFT analyzer's speed is limited by the time it takes to acquire the data and the time it takes to compute the FFT. For equivalent-frequency resolution, the FFT analyzer is much faster than the swept analyzer.

The FFT analyzer is limited in frequency range due to the need for a high-resolution ADC to sample somewhat above the Nyquist rate. As ADC technology has improved, the frequency range of FFT analyzers has increased.³ However, for high dynamic range

³ Another option for high-frequency spectrum measurements is a wide bandwidth digital oscilloscope with an FFT analysis function.

measurements at microwave frequencies and higher, the swept analyzer is still the dominant type of instrument.

As mentioned in Chapter 3, any practical measurement is limited to a finite time. For a signal that is changing, it may be desirable to measure the spectrum instantaneously so that its frequency content at an instant in time can be determined. Unfortunately, this is often not possible. Swept analyzers, in particular, may take several seconds or even minutes to perform one swept measurement. During this time, the signal may change and the swept analyzer may miss portions of the signal.

An FFT analyzer acquires a time record that contains the entire spectral content of a signal for that particular slice of time. The FFT computation transforms this time domain data into its spectrum. As long as the FFT is performed at least as fast as new time domain data are acquired, the analyzer can continue to capture and display the spectral content of the signal without ever missing any portion of the signal. Thus, an FFT analyzer is more effective at measuring dynamic signals.

5.14 Modern Spectrum Analyzer Block Diagrams

Most modern spectrum analyzers have combined the block diagrams and the benefits of the swept analyzer and the FFT analyzer. It is a natural evolution of the digital IF shown in Figure 5-9 to add the capability to compute an FFT of the IF signal. Figure 5-10 shows the combined block diagram in a simplified fashion. The front end of the spectrum analyzer employs the super heterodyne receiver of the swept analyzer shown in Figure 5-4. The later stages of the signal path uses a digital IF structure similar to what is shown in Figure 5-9.

A quadrature detector, followed by digital low-pass filters, is shown after the ADC (Figure 5-11). This quadrature detection is done digitally, transforming the IF signal into an in-phase (I) component and a quadrature (Q) component. The in-phase component is just the IF signal mixed with the digital LO, while the quadrature component uses an LO signal that is shifted by 90° . The output of these digital mixers is followed by *decimating low-pass filters*, which can be used to further reduce the bandwidth of resulting quadrature signal.

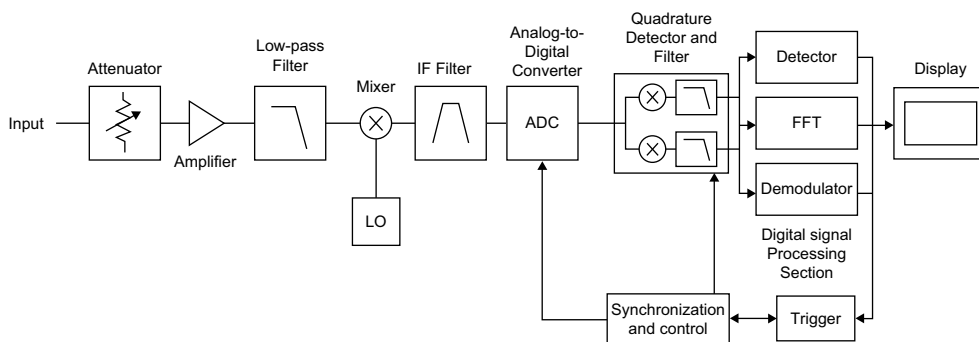


Figure 5-10 The conceptual block diagram of a spectrum analyzer that combines the swept and FFT approaches.

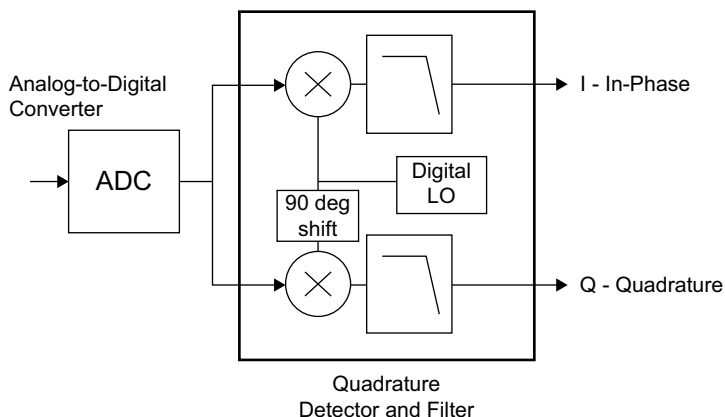


Figure 5-11 The quadrature detector splits the incoming IF signal into in-phase (I) and quadrature (Q) components.

There are actually three signal detection devices shown in the block diagram: the *swept detector*, the *FFT*, and the *I/Q demodulator*. The swept detector is the same digital detector associated with the swept analyzer block diagram and includes the actual detector itself and the video filter, if required. The FFT block includes any window function required and computes the spectrum of the signal present in the IF. The I/Q demodulator extracts and analyzes the modulation from the I/Q data and is used for measuring analog and digital modulated signals.

The block diagram is drawn to cover the most general case. Specific spectrum analyzer implementations may implement all or part of this block diagram, depending on the intended application and available technology. Although the terminology used may imply a hardware implementation, the portions of the block diagram to the right of the ADC can be implemented using hardware or software.

- *Swept analysis*: When operating in the classic swept analyzer mode, the swept detector analyzes the IF signal while the LO is swept to cover the frequency range of interest. Generally, the swept mode maintains a high dynamic range measurement while covering a wide frequency range.
- *FFT analysis*: The FFT mode is normally used to capture a specific frequency span within the frequency range of the spectrum analyzer. This span is limited to the widest IF bandwidth available in the instrument. The LO is tuned to a fixed frequency such that the IF is centered on the frequency span of interest and the FFT calculates the spectrum of the signal present in the IF. The decimating low-pass filters associated with the quadrature detector are used to narrow the frequency span. The advantage of the FFT mode is measurement speed and the ability to measure very narrow frequency spans. This mode is very useful for measuring close-in sidebands and phase noise on modulated carriers. Some analyzers use FFT analysis to cover the entire frequency span of the instrument (not just a single IF measurement). A broader frequency span can be measured by collecting multiple FFT snapshots as the LO is stepped through the frequency range of interest. This technique uses multiple stepped FFT measurements instead of a continuous frequency sweep.

- *Modulation analysis*: The digital IF approach lends itself to analyzing digital and analog modulation of communications signals. The I/Q demodulator operates similar to the FFT analysis mode in that the LO is fixed tuned to capture the communications signal in the IF and the I/Q demodulator extracts the modulation from the signal present. This mode is covered in more detail in Chapter 6.

5.15 Real-Time Spectrum Analyzer

The concept of *real-time spectrum analysis* (RTSA) first appeared in FFT spectrum analyzers, as discussed in Chapter 4. The basic idea is that the analyzer captures and processes the spectrum of a signal so fast that nothing is missed. As the FFT found its way into the modern swept spectrum analyzer, the benefits of real time followed along.⁴

The overall gap-free speed of an RTSA is often described in terms of real-time bandwidth (RTBW), which is the signal bandwidth that can be captured and processed without losing any sample points. In general, higher bandwidth requires a higher sample rate and more processing power to keep up. Alternatively, the RTSA capture ability may be specified in terms of *probability of intercept* (POI). Normally, we are interested in 100% POI. That is, we want to be sure that intermittent signals will be captured. An RTSA datasheet may specify the minimum signal duration with 100% probability of intercept (typical specification is $\sim 10 \mu\text{s}$).

The RTSA captures gapless, sequential frequency spectra at a high processing rate. This measurement capability adds in the third dimension of time: a frequency domain plot (amplitude vs. frequency) captured as a function of time. This also produces a large amount of frequency domain information that needs to be stored and displayed. The RTSA includes advanced display techniques to allow the user to make use of this information.

One display technique is to plot the density of the frequency spectrum using color-graded persistence. Frequent spectral content shows up in one color, and less frequent events are given other colors based on how often they occur. Figure 5-12a shows a swept spectrum analyzer measurement of a signal with complex time-varying spectral content. This measurement provides some idea of the spectral content but without much detail. Figure 5-12b shows a density plot of the same signal, revealing much more detail of the spectrum. The actual instrument display shows this detail in color, whereas this book is limited to a grayscale representation. (The use of color greatly enhances the displayed spectrum.)

Another common way to display the RTSA frequency content versus time is via a *spectrogram*. Figure 5-13 shows a typical spectrogram display with the conventional spectrum plot in the upper half of the display. The lower half of the display has time as the vertical axis and frequency on the horizontal axis. The amplitude of the spectrum is shown via color intensity. (Again, the figure in the book is grayscale so it does not show the spectrogram in color.)

RTSAs generally include frequency mask triggers (FMT), which provide the ability to trigger on the spectrum of a signal. This is implemented in the trigger block shown in Figure 5-10, which

⁴ The benefits of real time (not missing any part of the signal) are so compelling that it is amazing that the electronics industry has lived with non-real-time analyzers for decades.

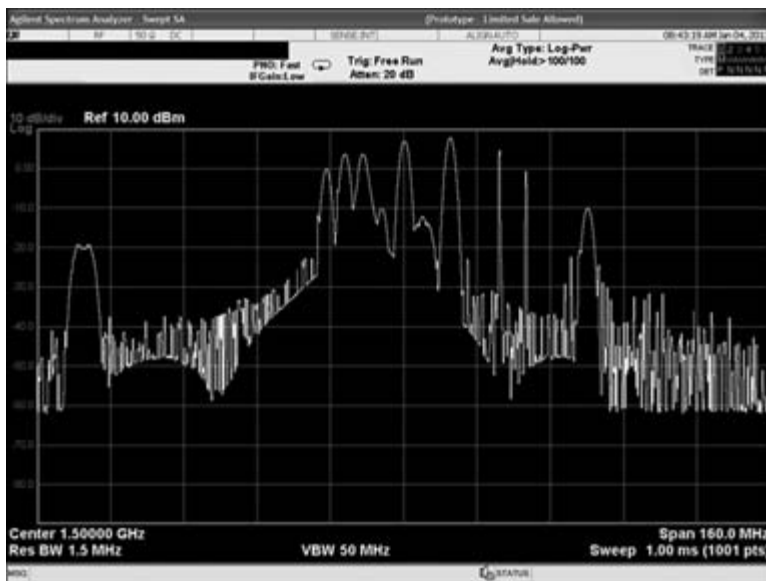


Figure 5-12a A complex, time-varying spectrum is measured using a conventional swept analyzer. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

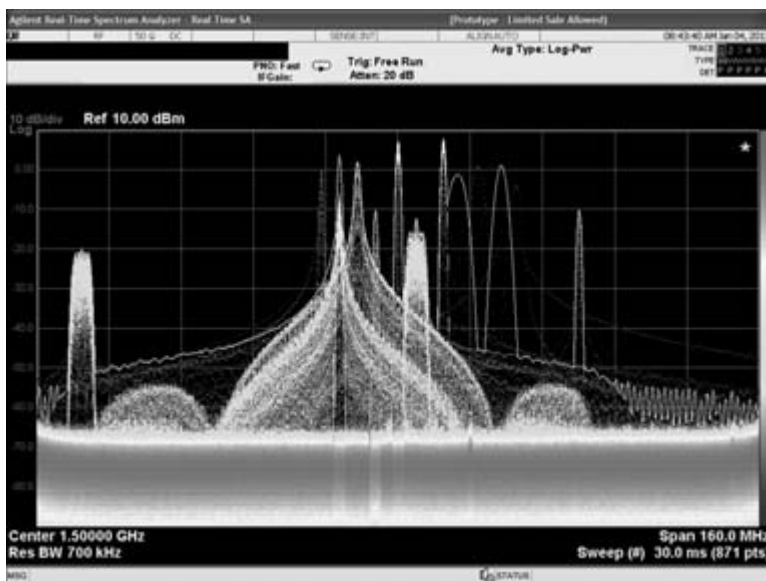


Figure 5-12b A complex, time-varying spectrum is measured using a real-time spectrum analyzer. Shown in grayscale here, spectrum analyzers use color grading to enhance the displayed spectrum. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

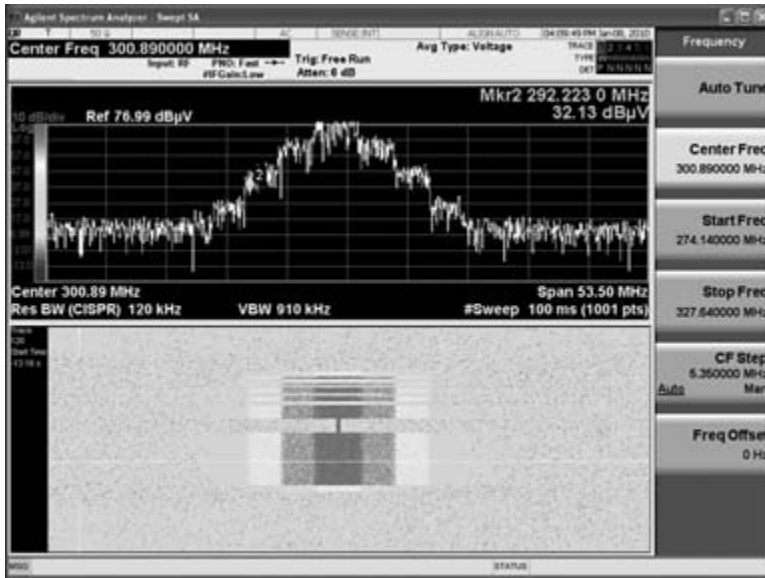


Figure 5-13 A spectrogram display (lower half of figure) shows the spectrum as a function of time using color to encode the amplitude of the signal. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

is after the FFT processing so that the spectrum information is available. Figure 5-14 shows a typical measurement using FMT. The trigger condition is specified as a limit line that the measured spectrum is compared with. Typically, the FMT feature will have a number of options on when the measurement is triggered. The trigger may be set for when the signal is inside the mask, outside the mask, enters the mask or exits the mask, depending on the type of event the user is trying to capture.

5.16 Types of Spectrum Analyzers

Spectrum analyzer block diagrams continue to evolve, with variations on the block diagram shown in Figure 5-10. *Spectrum analyzer* remains the most common name for an instrument that measures the frequency spectrum of a signal. As new measurement techniques have emerged, other names have been used to identify important attributes associated with certain categories of instruments.

- *Spectrum analyzer*: This is the most common name for an instrument that measures a signal in the frequency domain, displaying the spectral content of signals present over the frequency range of the instrument.
- *FFT spectrum analyzer*: Also called a *dynamic signal analyzer*, this type of instrument samples the input signal above the Nyquist rate and uses digital signal processing to determine the spectrum of the signal (see Chapter 4).

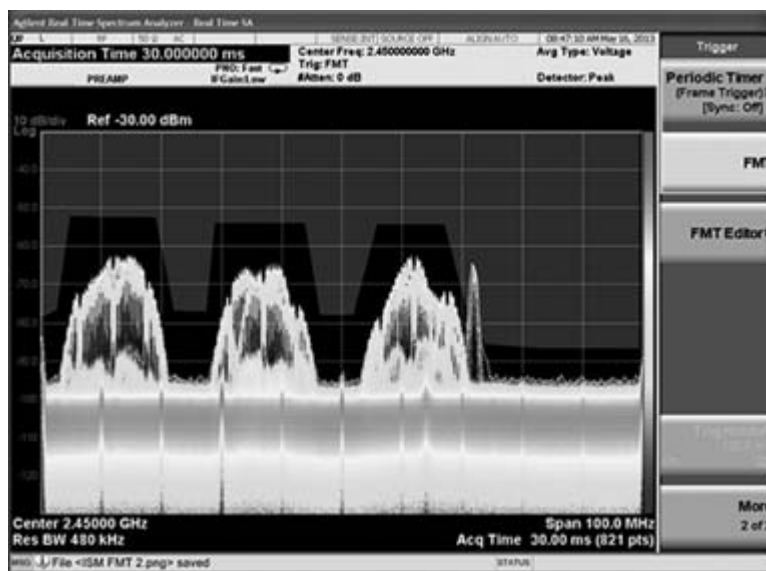


Figure 5-14 Frequency mask trigger allows the user to capture measurements based on user supplied limits in the frequency domain. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

- *Vector signal analyzer (VSA)*: This measures the magnitude and phase of a signal at a particular frequency, capturing the signal without any loss of information, which is important for many measurement applications including modulation analysis.
- *Signal analyzer*: Normally this refers to an instrument that combines the functions of a spectrum analyzer and a vector signal analyzer. That is, the instrument has general purpose spectral measurement capability plus complex modulation analysis
- *Real-time spectrum analyzer (RTSA)*: Its primary attribute is gap-free analysis so that all parts of a time-varying signal are not missed. RTSAs also usually include frequency mask trigger and advanced signal displays.
- *EMI receiver*: This spectrum analyzer includes specific features for measuring EMI (see Chapter 16).

Bibliography

Agilent Technologies, “Agilent Vector Signal Analysis Basics,” Publication Number 5990-7451EN, Santa Clara, CA, February 2011.

Agilent Technologies, “Agilent X-Series Signal Analyzer Real-Time Spectrum Analyzer Measurement Guide,” Publication Number N9030-90060, September 2013.

Agilent Technologies, “Real-Time Spectrum Analyzer (RTSA) X-Series Signal Analyzers Technical Overview,” Publication Number 5991-1748EN, December 2013.

Agilent Technologies, "Spectrum Analysis Basics," Application Note 150, Publication Number 5952-0292, December 2014.

Agilent Technologies, "Understanding and Applying Probability of Intercept in Real-Time Spectrum Analysis," Publication Number 5991-4317EN, April 2014.

Clarke, Kenneth K., and Donald T. Hess. *Communication Circuits: Analysis and Design*. Reading, MA: Addison-Wesley Publishing Company, 1971.

Engelson, Morris. *Modern Spectrum Analyzer Theory and Applications*. Dedham, MA: Artech House, 1984.

Hayward, W. H. *Introduction to Radio Frequency Design*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1982.

Oliver, Bernard M., and John M. Cage. *Electronic Measurements and Instrumentation*. New York: McGraw-Hill Book Company, 1971.

Tektronix, "Fundamentals of Real-Time Spectrum Analysis," Application Note, Beaverton, OR, 2009.

Modulation Measurements

Ever since the early days of radio, modulation techniques have played an important part in electronic communications. Typically, a low-frequency voice or data signal is used to modulate some characteristic of a carrier signal—usually the amplitude, phase, or frequency. Over time, the digital forms of modulation have become dominant. Digital modulation often is implemented using vector modulation of the carrier signal, having an in-phase (I) and a quadrature (Q) modulation component.

Communication systems represent an intentional use of modulation, but there are also incidents of unintentional modulation, such as power line sidebands on an oscillator output or residual frequency modulation on an amplitude-modulated signal. Whether the modulation is intentional or not, a spectrum analyzer can be used to characterize and measure it.

6.1 The Carrier

Analog modulation techniques start with a carrier that is a pure sinusoid.

$$v(t) = A \cos(2\pi f_c t + \theta) \quad (6-1)$$

where

A = carrier amplitude (zero-to-peak)

f_c = carrier frequency (hertz)

θ = the carrier phase

The carrier may be modulated in a variety of ways, but the various techniques fall into two main categories: *amplitude modulation* (AM) and *angle modulation*. Amplitude modulation implies that the amplitude is no longer simply a constant but is a function of time. Similarly, angle modulation occurs when the angle of the cosine term is varied. Angle modulation may take the form of *frequency modulation* or *phase modulation* depending on the particular modulation technique.

All the modulation techniques result in increasing the occupied bandwidth of the carrier by spreading out sidebands in the frequency domain. Originally, the carrier is presumed to be a single spectral line, infinitely thin, occupying only one exact frequency. When modulated, the signal bandwidth increases depending on the type of modulation and the modulating

signal. Sidebands appear beside the carrier, either in the form of discrete frequencies or, for nonperiodic modulation such as voice or music, more complex spectral shapes.

6.2 Amplitude Modulation

AM is generally considered the simplest modulation system. Although usually lumped under the general label of AM, there are several distinct variations.

An AM signal with carrier¹ is represented by

$$v(t) = A_c[1 + am(t)]\cos(2\pi f_c t) \quad (6-2)$$

where

- A_c = the constant that determines the overall signal amplitude
- a = the modulation index ($0 \leq a \leq 1$)
- $m(t)$ = the normalized modulating signal
- f_c = carrier frequency (Hz)

Note that the modulating signal is normalized, meaning that it is always within the range of -1 to $+1$. $A_c[1 + am(t)]$ defines the amplitude of the carrier envelope. With the stated restrictions on a and $m(t)$, the zero-to-peak amplitude of the carrier is always in the range of 0 to $2A_c$, inclusive. Thus, the amplitude of the carrier can be driven to zero, but it cannot go negative and change the sign of the envelope.² Figure 6-1 shows a modulating signal and the resulting modulated carrier.

Rearranging the equation for $v(t)$ allows it to be divided into the carrier portion and the modulation sidebands portion.

$$v(t) = \underbrace{A_c \cos(2\pi f_c t)}_{\text{carrier}} + \underbrace{A_c am(t) \cos(2\pi f_c t)}_{\text{sidebands}} \quad (6-3)$$

$$v(t) = v_c(t) + v_s(t) \quad (6-4)$$

where

$$v_c(t) = A_c \cos(2\pi f_c t)$$

$$v_s(t) = A_c am(t) \cos(2\pi f_c t)$$

Transforming the time domain expressions into the frequency domain,

$$V(f) = V_c(f) + V_s(f) \quad (6-5)$$

The Fourier transform of $v_c(t)$ is a pair of delta functions at $\pm f_c$.

$$V_c(f) = A_c \pi[\delta(f - f_c) + \delta(f + f_c)] \quad (6-6)$$

¹ This is the most common type of AM as it is used for standard AM radio broadcasting and in other AM voice communications systems.

² In most communication systems, if this happens the signal is overmodulated and will not be recovered properly at the receiver.

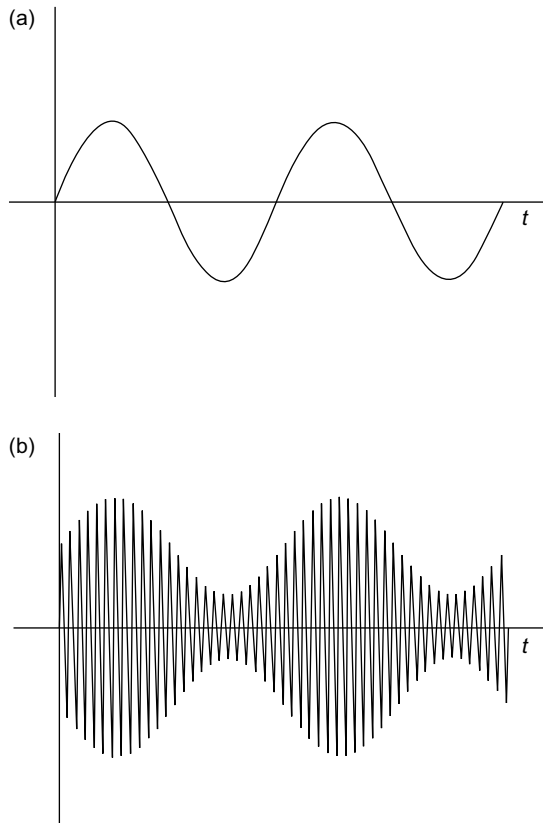


Figure 6-1 Amplitude modulation causes the amplitude of the carrier to be determined by the modulating signal. (a) The modulating signal. (b) The amplitude-modulated carrier.

The Fourier transform of $v_s(t)$ is most easily derived by using the modulation property from Table 3-3.

$$\mathcal{F}[x(t) \cos(2\pi f_0 t)] = \frac{1}{2} [X(f - f_0) + X(f + f_0)] \quad (6-7)$$

Applying this property to $v_s(t)$

$$V_s(f) = \frac{A_c a}{2} 2[M(f - f_c) + M(f + f_c)] \quad (6-8)$$

That is, the sideband term in the frequency domain is the spectrum of the original modulating signal, $M(f)$, centered on $\pm f_c$ (Figure 6-2b). Adding $V_c(f)$ and $V_s(f)$ gives $V(f)$, which is shown in Figure 6-2c.

$$V(f) = A_c \pi [\delta(f - f_c) + \delta(f + f_c)] + \frac{A_c a}{2} [M(f - f_c) + M(f + f_c)] \quad (6-9)$$

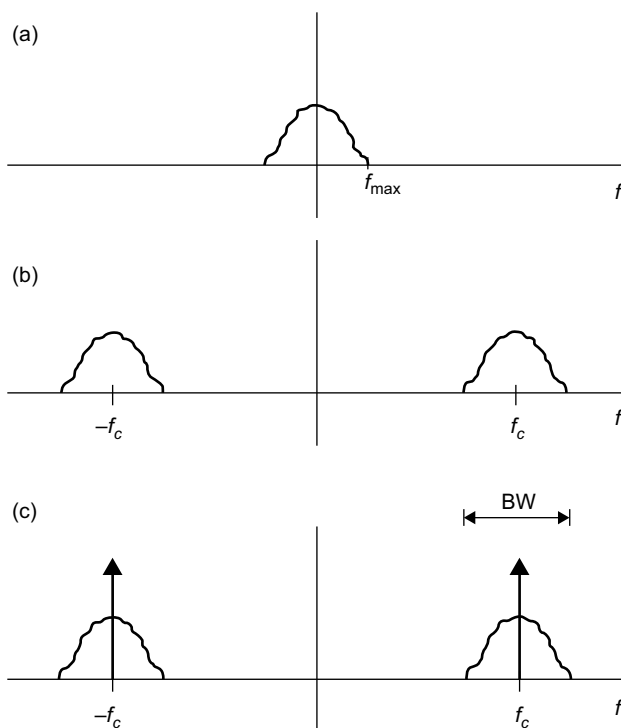


Figure 6-2 (a) The spectrum of the modulating signal. (b) The spectrum of the modulating signal centered on f_c . (c) The spectrum of the AM signal with carrier.

Consider the positive frequency portion of Figure 6-2c. We can see that the bandwidth occupied by the modulated signal is twice that of the modulating signal. This gives a simple mathematical relationship for the bandwidth of an AM signal.

$$BW = 2f_{max}$$

where

$$f_{max} = \text{the maximum frequency in the modulating signal}$$

Sinusoidal Modulation

The case where the modulating signal is a sinusoid is an important and common occurrence in electronic systems. This case can be analyzed using Fourier transforms, but it can also be easily explained using trigonometry. Since the trig approach is instructive and gives a result that is inherently one-sided in the frequency domain, we will use it here.

$$m(t) = \cos(2\pi f_m t) \tag{6-10}$$

Recall that

$$v(t) = A_c \cos(2\pi f_c t) + A_c am(t) \cos(2\pi f_c t) \tag{6-11}$$

$$v(t) = A_c \cos(2\pi f_c t) + A_c a \cos(2\pi f_m t) \cos(2\pi f_c t) \tag{6-12}$$

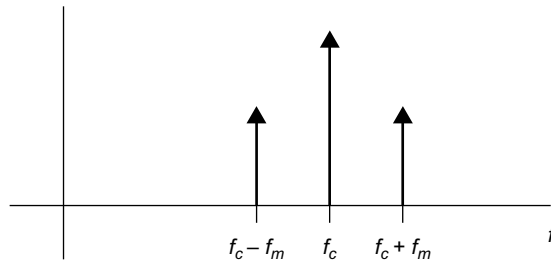


Figure 6-3 The spectrum of an AM signal with sinusoidal modulation.

Using the trig identity,

$$\cos A \cos B = 1/2[\cos(A + B) + \cos(A - B)] \quad (6-13)$$

$$v(t) = A_c \cos(2\pi f_c t) + \frac{aA_c}{2} [\cos 2\pi(f_m + f_c)t + \cos 2\pi(f_m - f_c)t] \quad (6-14)$$

Since $\cos(A - B) = \cos(B - A)$,

$$v(t) = A_c \cos(2\pi f_c t) + \frac{aA_c}{2} [\cos 2\pi(f_c + f_m)t + \cos 2\pi(f_c - f_m)t] \quad (6-15)$$

Thus, $v(t)$ consists of the carrier frequency with amplitude A_c and two sidebands, one at $f_c + f_m$ and one at $f_c - f_m$, both with amplitude $aA_c/2$ (Figure 6-3).

The modulation index, a , may vary from 0 to 100%. When a is 100%, each sideband amplitude is $A_c/2$, which is half of the carrier amplitude. Note that the carrier amplitude does not depend on the level of modulation. Table 6-1 tabulates the sideband amplitude relative to the carrier amplitude for a variety of modulation index values.

Time Domain

In the time domain, a carrier with sinusoidal amplitude modulation will appear as shown in Figure 6-4. The minimum and maximum values of the envelope of the waveform are called V_{\min} and V_{\max} . The modulation index can be computed from these two parameters.

The maximum envelope voltage occurs when the modulating sinusoid is at its most positive value, which is +1.

$$V_{\max} = 1 + a \quad (6-16)$$

The minimum envelope voltage occurs when the modulating sinusoid reaches its most negative value, which is -1.

$$V_{\min} = 1 - a \quad (6-17)$$

Solving for a ,

$$a = \frac{V_{\max} - V_{\min}}{V_{\max} + V_{\min}} \quad (6-18)$$

Table 6-1 Modulation Index and Relative Sideband Amplitude

Modulation Index (%)	Sideband Amplitude Relative to Carrier	
	(%)	dB
100	50.0	-6.02
95	47.5	-6.47
90	45.0	-6.94
85	42.5	-7.43
80	40.0	-7.96
75	37.5	-8.52
70	35.0	-9.12
65	32.5	-9.76
60	30.0	-10.46
55	27.5	-11.21
50	25.0	-12.04
45	22.5	-12.96
40	20.0	-13.98
35	17.5	-15.14
30	15.0	-16.48
25	12.5	-18.06
20	10.0	-20.00
15	7.5	-22.50
10	5.0	-26.02
9	4.5	-26.94
8	4.0	-27.96
7	3.5	-29.12
6	3.0	-30.46
5	2.5	-32.04
4	2.0	-33.98
3	1.5	-36.48
2	1.0	-40.00
1	0.5	-46.02

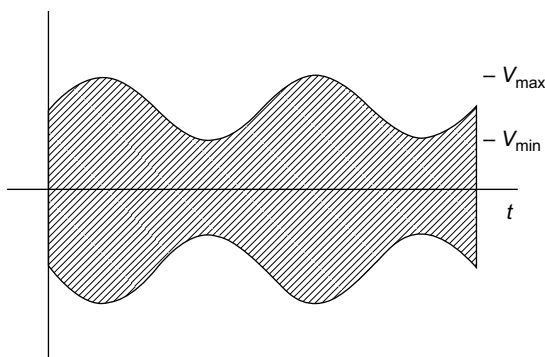


Figure 6-4 The envelope of an AM signal in the time domain can be used to determine the modulation index.

6.3 AM Measurements

The spectrum analyzer can be used to characterize an amplitude-modulated signal in the frequency domain. The parameters that can be measured are the carrier amplitude and frequency, the modulating frequency, and the modulation index.

The carrier amplitude and frequency are measured just like any other spectral component, either by reading the values using the graticule or with the help of a marker or cursor readout. The modulating frequency is the difference between the carrier frequency and one of the sidebands. (The sidebands are symmetrical around the carrier.) Measuring the difference between the carrier and the sideband is made easier by the use of a marker that has delta (offset) capability. The modulation index is determined by measuring the sideband amplitude relative to the carrier amplitude. Usually this is expressed in dB. Table 6-1 or the following equation allows the user to convert relative sideband amplitude back to modulation index.

$$a = 2 \times 10^{(A_{dB}/20)} \quad (6-19)$$

where

A_{dB} = the sideband amplitude relative to the carrier (dB)

Example 6.1

A spectrum analyzer measurement of an amplitude-modulated signal is shown in Figure 6-5. Determine the modulating frequency and modulation index of the signal.

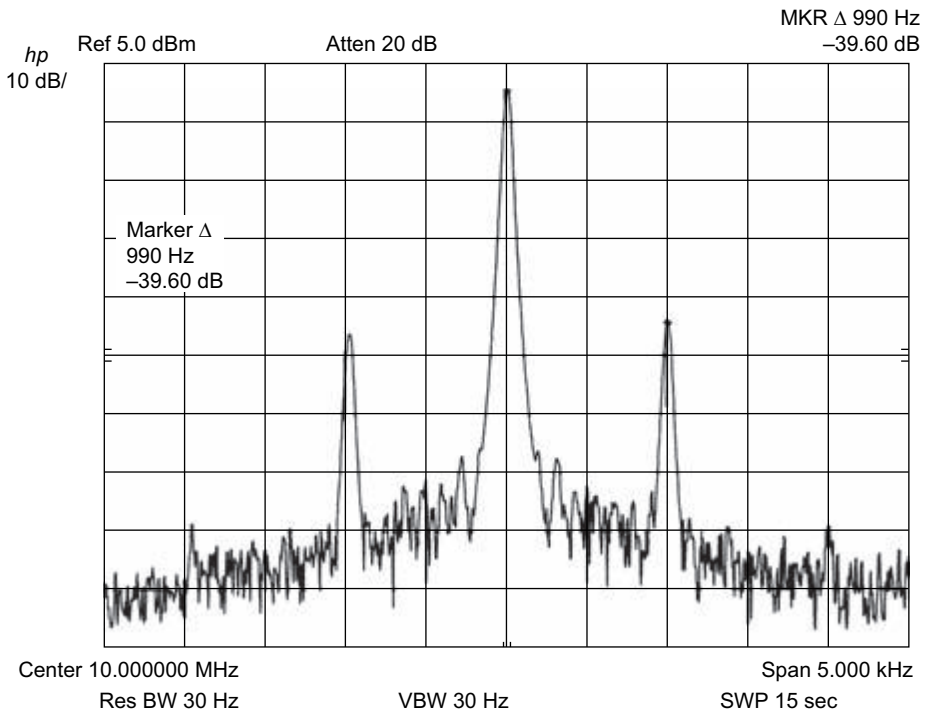


Figure 6-5 A spectrum analyzer measurement of an amplitude-modulated signal.

The delta marker feature is used to measure the amplitude and frequency differences between the carrier and the sidebands in Figure 6-5. The sidebands are -39.6 dB relative to the carrier and are offset by 990 Hz. Therefore, the modulating frequency is 990 Hz. The modulation index can be found by

$$a = 2 \times 10^{(A_{dB}/20)} = 2 \times 10^{(-39.6/20)} = 2.1\%$$

6.4 Zero-Span Operation

Most swept spectrum analyzers provide a simple yet powerful feature for observing slow amplitude variations in signals. The spectrum analyzer is set for a frequency span of zero with some nonzero sweep time. This is commonly referred to as *zero-span* or sometimes *synchronous* operation. The center frequency is set to the carrier frequency, and the resolution bandwidth must be set large enough to allow the sidebands to be included in the measurement. The analyzer will plot the amplitude of the signal versus time, within the limitations of its detector and video and resolution bandwidths. Since the minimum sweep time is about 25 msec on the fastest analyzers and may be as slow as 300 msec, this feature cannot be used for quickly varying signals. The highest modulating frequency that can be observed is roughly $1/(\text{sweep time})$, which will put only one cycle of the modulation on screen.

A spectrum analyzer measurement using zero span is shown in Figure 6-6. One can see the variation due to amplitude modulation on the signal as it is plotted across the display.

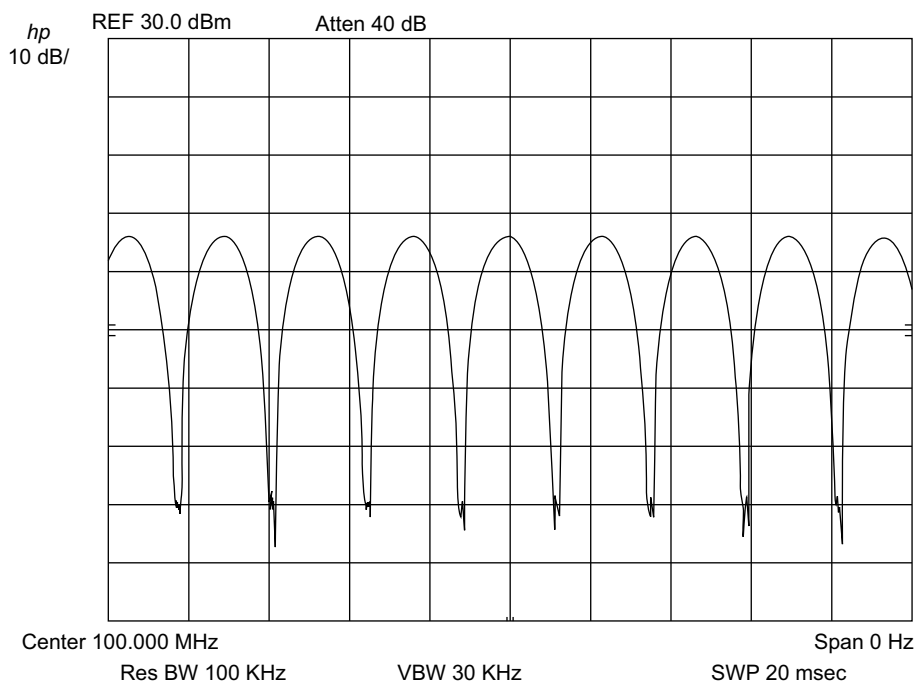


Figure 6-6 The zero-span mode of a spectrum analyzer is used here to view the amplitude variation in an AM signal.

The sweep time is 20 msec, so the horizontal axis is 2 msec/div. One cycle of the modulating signal occurs in 2.5 msec, indicating a modulating frequency of 400 Hz.

Zero-span operation is not limited to modulation measurements. It can be used to characterize any signal that is slowly varying in amplitude.

6.5 Other Forms of Amplitude Modulation

We will briefly mention some of the other varieties of amplitude-modulated signals. Note that “standard” AM includes a carrier and two sidebands that are symmetrical about the carrier. A significant amount of power is used to generate the carrier, which contains no information content since it does not vary with the modulating signal. All the information is in the sidebands. Sometimes the carrier is removed from the modulating signal, which produces *double-sideband (DSB) modulation* (also called *AM suppressed carrier* or *double-sideband suppressed carrier modulation*).

Since the modulation sidebands contain redundant information, one of them may be removed, resulting in *single-sideband (SSB) modulation*. Normally, SSB modulation also has the carrier removed (or greatly attenuated), but in some cases the carrier may be retained. If the lower sideband remains, it is called *lower-sideband (LSB) modulation* while retaining the *upper sideband* produces *upper-sideband (USB) modulation*. Compared with the other forms of AM, SSB modulation produces the narrowest occupied bandwidth, requiring half the bandwidth of standard AM and DSB modulation.

6.6 Angle Modulation

While AM modulates the amplitude of the carrier, another option is to modulate the angle or phase of the carrier. Depending on the particular implementation, this type of modulation is called frequency modulation (FM) or phase modulation (PM). The difference between FM and PM is sometimes quite subtle since either form of modulation can be derived from the other by shaping the frequency response of the modulating signal.

The equation for the carrier is modified to allow for a time-varying phase term:

$$v(t) = A_c \cos(2\pi f_c t + \theta(t)) \quad (6-20)$$

where $\theta(t)$ is the time-varying phase containing the modulation information.

For phase modulation, the phase term is directly proportional to the modulating signal:

$$\theta(t) = k_p m(t) \quad (6-21)$$

where

k_p = deviation constant
 $m(t)$ = modulating signal

The phase-modulated carrier is

$$v(t) = A_c \cos(2\pi f_c t + k_p m(t)) \quad (6-22)$$

For frequency modulation, the frequency must be proportional to the modulating signal. Since the frequency is the time derivative of phase,

$$\frac{d\theta}{dt} = k_f m(t) \quad (6-23)$$

Solving for phase,

$$\theta(t) = k_f \int_{t_0}^t m(x) dx + \theta_0 \quad (6-24)$$

where

k_f = frequency deviation constant

θ_0 = initial phase at $t = 0$

Setting the initial phase to zero, the frequency-modulated carrier is

$$v(t) = A_c \cos \left(2\pi f_c t + k_f \int m(x) dx \right) \quad (6-25)$$

Figure 6-7 shows a modulating signal, the resulting phase-modulated carrier, and the resulting frequency-modulated carrier. Notice the difference between changing the carrier phase and changing the carrier frequency.

The previous mathematical discussion centered on converting the modulating signal into a phase term to produce a frequency-modulated carrier. Thus, integrating the modulating signal in a phase-modulated system is equivalent to frequency modulating the carrier. This technique may be used in actual circuit implementations. The converse is also true. If a frequency modulator circuit was available, it could be used to produce a phase-modulated signal by taking the derivative of the modulating signal before applying it to the modulator. Since the integrator and differentiator operations can be approximated with high-pass and low-pass filters, respectively, the difference between FM and PM is often just the frequency shaping of the modulator circuits.

Sinusoidal Modulation

Consider the important case where the modulating signal is a sinusoid in an FM system:

$$m(t) = A_m \cos(2\pi f_m t) \quad (6-26)$$

The frequency-modulated carrier is

$$v(t) = A_c \cos \left(2\pi f_c t + k_f \int A_m \int \cos 2\pi f_m x dx \right) \quad (6-27)$$

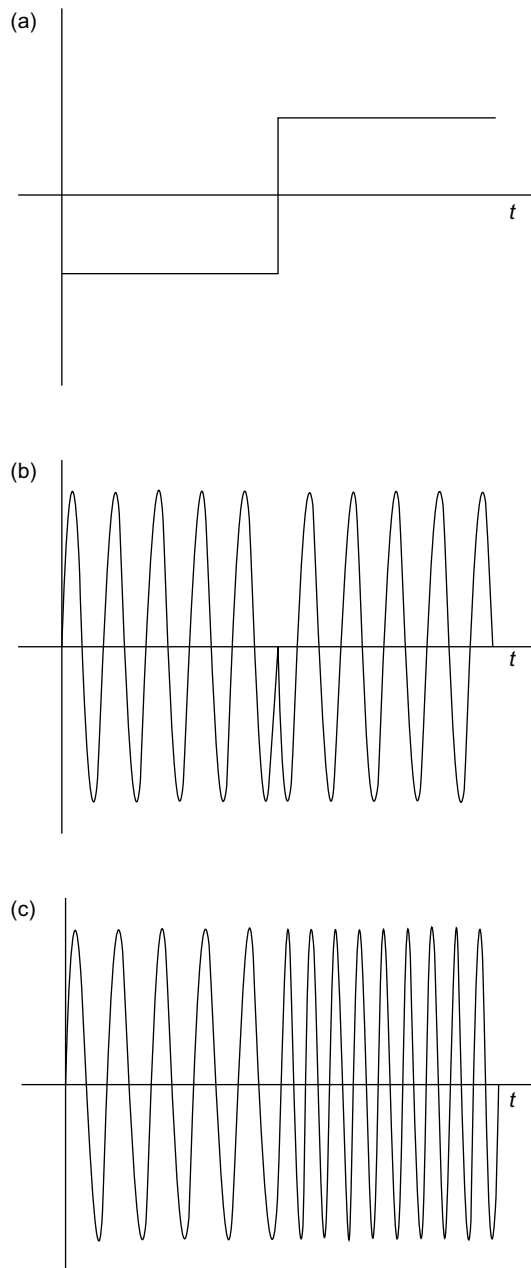


Figure 6-7 The difference between FM and PM is easily shown by considering the case where the modulating signal is a step. (a) The modulating signal. (b) The phase-modulated carrier stays at the same frequency but changes phase with the modulating signal. (c) The frequency-modulated carrier changes frequency when the step in the modulation occurs.

Taking the integral of the modulating signal,³

$$v(t) = A_c \cos\left(2\pi f_c t + \frac{k_f A_m}{2\pi f_m} \sin 2\pi f_m t\right) \quad (6-28)$$

Introducing the *modulation index*, $\beta = k_f A_m / (2\pi f_m)$

$$v(t) = A_c \cos(2\pi f_c t + \beta \sin 2\pi f_m t) \quad (6-29)$$

The modulation index is defined as

$$\beta = \frac{\Delta f}{f_m} \quad (6-30)$$

where

Δf = the frequency deviation (Hz)

The frequency deviation can also be expressed as

$$\Delta f = k_f A_m / 2\pi \quad (6-31)$$

6.7 Narrowband Angle Modulation

Angle modulation is normally divided into two cases: narrowband (small modulation index) and wideband (large modulation index). First, consider the case where β is small (i.e., less than 0.2 radians).

$$v(t) = A_c \cos(2\pi f_c t + \beta \sin 2\pi f_m t) \quad (6-32)$$

Using the identity $\cos(A + B) = \cos A \cos B - \sin A \sin B$,

$$v(t) = A_c [\cos(2\pi f_c t) \cos(\beta \sin 2\pi f_m t) - \sin(2\pi f_c t) \sin(\beta \sin 2\pi f_m t)] \quad (6-33)$$

For small β , $\cos(\beta \sin 2\pi f_m t)$ equals approximately 1 and $\sin(\beta \sin 2\pi f_m t)$ equals approximately $\beta \sin 2\pi f_m t$:

$$v(t) = A_c [\cos(2\pi f_c t) - \beta \sin(2\pi f_c t) \sin(2\pi f_m t)] \quad (6-34)$$

This can be broken down further into its spectral components using $\sin A \sin B = 1/2 [\cos(A - B) - \cos(A + B)]$:

$$v(t) = A_c \left\{ \cos(2\pi f_c t) - \frac{\beta}{2} [\cos 2\pi(f_c - f_m)t - \cos 2\pi(f_m + f_c)t] \right\} \quad (6-35)$$

$$v(t) = A_c \cos(2\pi f_c t) + \frac{A_c \beta}{2} [\cos 2\pi(f_c + f_m)t - \cos 2\pi(f_c - f_m)t] \quad (6-36)$$

³ Note that this same modulated signal would result in a phase-modulated system with $m(t) = \frac{A_m}{2\pi f_m} \sin 2\pi f_m t$.

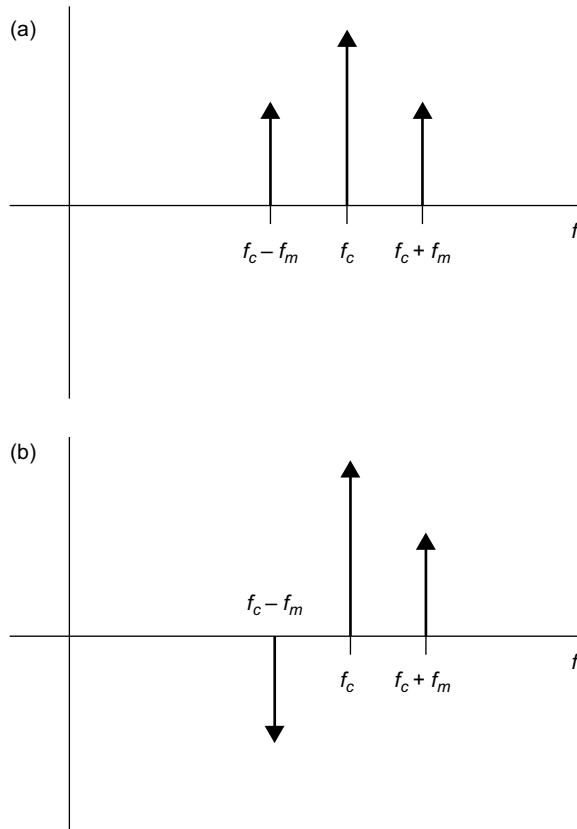


Figure 6-8 Narrowband FM is very similar to AM, except that one sideband is out of phase. (a) Spectrum of an AM signal. (b) Spectrum of a narrowband FM signal.

This result should be reminiscent of the AM formula. Just like the AM case, the narrowband FM signal has frequency components at the carrier and at $\pm f_m$ away from the carrier. The subtle difference between AM and narrowband FM is that the phase of the lower sideband ($f_c - f_m$) is changed by 180° as indicated by the minus sign in front of the term (Figure 6-8). If no phase information is available (as with most spectrum analyzers), the two types of signals are indistinguishable in the frequency domain.

6.8 Wideband Angle Modulation

For the case where the modulation index is large, wideband angle modulation will result. As the name implies, in the frequency domain the signal will occupy a much larger bandwidth than the narrowband case.

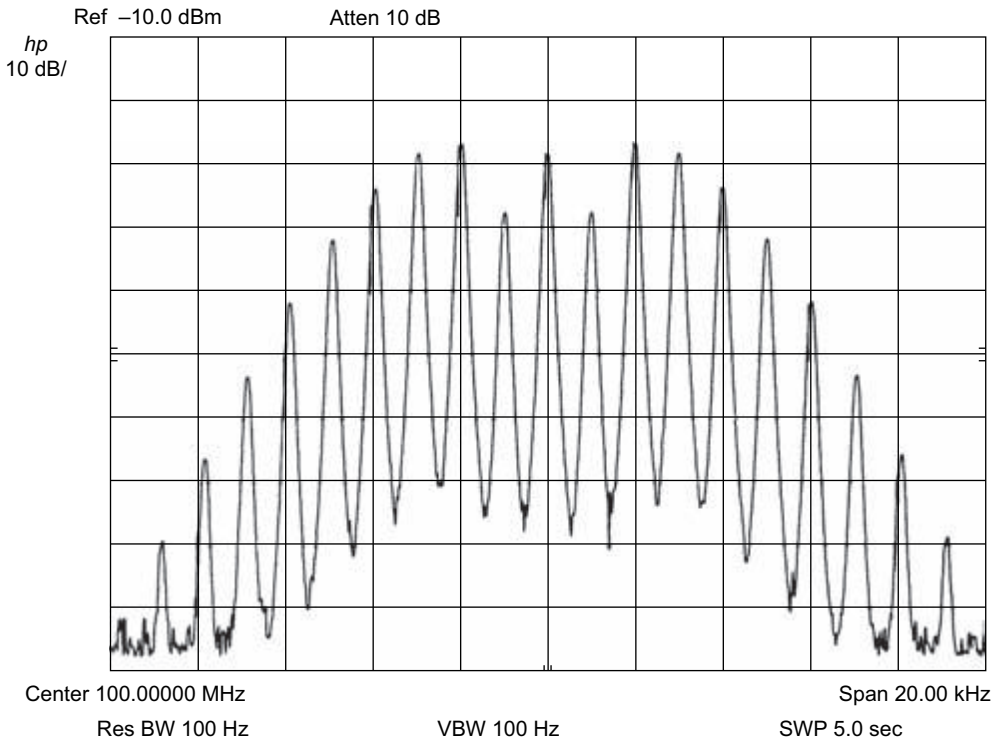


Figure 6-9 A spectrum measurement of a sine wave modulated wideband FM signal, with numerous sidebands spaced at multiples of the modulating frequency.

will be small enough to be ignored. The spectrum of a typical wideband FM signal is shown in Figure 6-9.

Carson's Rule

With a single modulating frequency, the table of Bessel functions can be used to obtain the exact spectrum of the modulated signal. The signal bandwidth can then be inferred from the spectrum. With multiple modulating frequencies (e.g., voice modulation), the analysis quickly gets unmanageable. However, *Carson's rule* can be used to estimate the bandwidth of a frequency-modulated signal.

$$BW = 2(\Delta f + f_m) \quad (6-39)$$

Recall that Δf is the peak frequency deviation, and f_m is the modulating frequency. For the single tone case, f_m retains this definition, but for multitone modulation the highest modulating frequency is substituted for f_m .

Example 6.2

Determine the spectrum of a 10 MHz carrier frequency modulated by a 5 kHz signal with a frequency deviation of 10 kHz.

The modulation index is given by $\beta = \Delta f/f_m = 10 \text{ kHz}/5 \text{ kHz} = 2$. Since the modulating signal has a frequency of 5 kHz, the sidebands will be spaced at multiples of 5 kHz relative to the carrier frequency. Table 6-2 shows the following coefficients for the sidebands (modulation index = 2):

n	Table Coefficient	Frequencies (MHz)	Amplitude Relative to Unmodulated Carrier
0	0.224	10.000	-13.0 dB
1	0.577	9.995, 10.005	-4.78 dB
2	0.353	9.990, 10.010	-9.04 dB
3	0.129	9.985, 10.015	-17.8 dB
4	0.034	9.980, 10.020	-29.4 dB
5	0.007	9.975, 10.025	-43.1 dB
6	0.001	9.970, 10.030	-60.0 dB

6.9 FM Measurements

The individual spectral components of an FM signal can be measured directly using a spectrum analyzer. Both the frequency and amplitude of the spectral lines can be determined, either absolutely or relative to the carrier. Determining the frequency deviation directly from the frequency spectrum is a more difficult task.

Table 6-3 FM Carrier Nulls

Null	Modulation index
First	2.405
Second	5.520
Third	8.654
Fourth	11.792
Fifth	14.931
Sixth	18.071

Carrier Null Method

For certain values of β , the carrier frequency of an FM signal (with sinusoidal modulation) will disappear. These carrier null points are listed in Table 6-3.

A radio transmitter or signal generator's frequency deviation can be set by using the *carrier null method*. A modulating frequency is chosen such that the desired deviation

level causes a null on the carrier frequency. The output is monitored with a spectrum analyzer or other instrument to detect the null. Since carrier nulls occur at many different values of modulation index, it is important to use the correct carrier null. Normally, the deviation level is set to zero and is then gradually increased while the carrier nulls are noted.

Example 6.3

A signal generator is to be adjusted such that its FM deviation is 5 kHz. What frequency should the modulating signal be to cause the first carrier null to occur at this frequency deviation?

The first carrier null occurs at $\beta = 2.405$. $\beta = \Delta f / f_m = \Delta f / \beta = 5000 / 2.405 = 2079$ Hz.

6.10 Combined AM and FM

In many high-frequency circuits, signals may be inadvertently amplitude modulated and frequency (or phase) modulated. When the modulation is purely amplitude or purely angle modulation, the previous sections of this chapter can be used to measure and understand it. However, when different forms of modulation appear simultaneously, the measurements may be very confusing.

Recall that the AM signal and the narrowband angle-modulated signal are identical except for the phase of the lower sideband. A carrier with simultaneous AM and narrowband FM can be described by combining the AM and narrowband FM equations (6-15) and (6-36). (We will assume that these two signals combine with negligible interaction that would produce new frequency components.)

$$v(t) = A_c \cos(2\pi f_c t) + \frac{aA_c}{2} [\cos 2\pi(f_c + f_m)t + \cos 2\pi(f_c - f_m)t] + \frac{A_c\beta}{2} [\cos 2\pi(f_c + f_m)t - \cos 2\pi(f_c - f_m)t] \quad (6-40)$$

$$v(t) = A_c \cos(2\pi f_c t) + \frac{A_c(a + \beta)}{2} [\cos 2\pi(f_c + f_m)t] + \frac{A_c(a - \beta)}{2} [\cos 2\pi(f_c - f_m)t] \quad (6-41)$$

If $a = \beta$, then cancellation of the lower sideband may occur:

$$v(t) = A_c \cos(2\pi f_c t) + \frac{A_c(a + \beta)}{2} \cos 2\pi(f_c + f_m)t \quad (6-42)$$

Several assumptions were made in this analysis. The modulation sources were assumed to be the same, and there was no phase shift between the two modulation mechanisms. The two modulation indexes must also match exactly. In practice, these conditions will not usually be met, and cancellation will not be complete. However, it is common to find some partial cancellation (Figure 6-10), causing modulation sidebands that are not symmetrical.

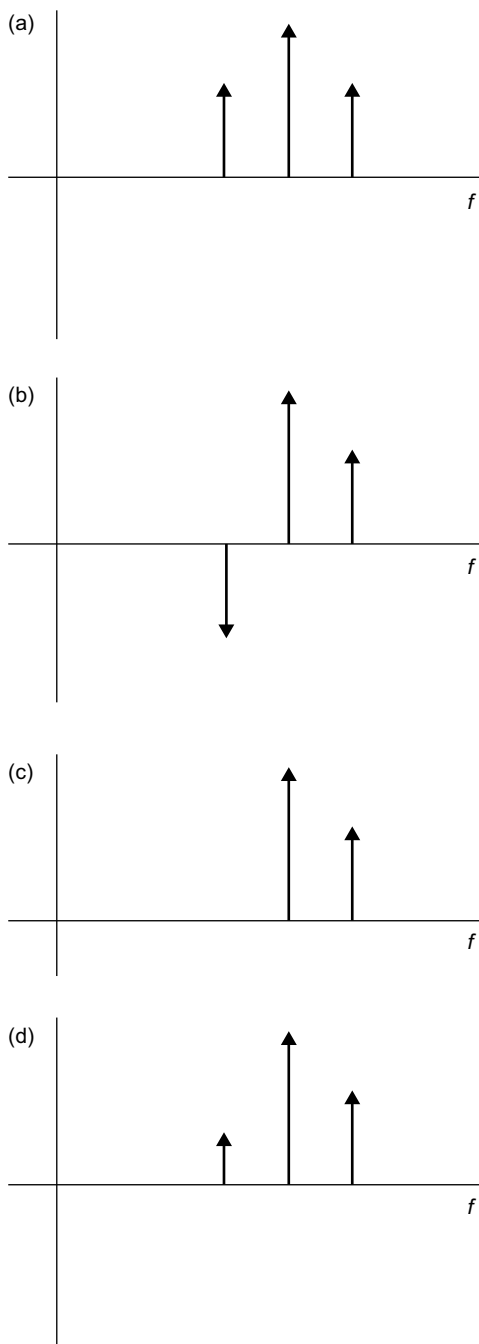


Figure 6-10 Simultaneous AM and FM can produce modulated signals that are asymmetrical in the frequency domain. (a) The spectrum of an AM signal. (b) The spectrum of a narrowband FM signal. (c) Perfect cancellation of the lower sideband due to simultaneous AM and FM. (d) Partial cancellation of the lower sideband.

The individual amounts of AM and FM can be estimated by assuming that the larger sideband is due to the AM and FM sidebands adding and that the smaller sideband is due to the subtraction of the AM and FM sidebands.

Devices that limit the amplitude of a signal (e.g., mixers and overdriven amplifiers) are notorious for converting amplitude modulation to phase modulation. Often only a portion of the AM is converted to PM, causing a combined AM/PM signal at the device's output. In the frequency domain, this may cause a single sideband spectrum or, more likely, a spectrum with asymmetrical sidebands.

Example 6.4

The carrier level of a signal with both AM and FM is 0.1 V. The upper sideband amplitude is 0.05 V, and the lower sideband amplitude is 0.02 V. Estimate the AM and FM modulation indexes.

The upper sideband is larger so it represents the addition of the AM and FM sidebands.

$$A_c(a + \beta) = 0.05, a + \beta = 0.05/0.1 = 0.5$$

The lower sideband is smaller and represents the subtraction of the AM and FM sidebands.

$$A_c(a - \beta) = 0.02, a - \beta = 0.02/0.1 = 0.2$$

Solving simultaneously, this implies that $a = 0.35$ and $\beta = 0.15$.

Modern spectrum analyzers that have an I/Q demodulator (Section 5.14) can extract and display the modulation present on a signal. Normally, this demodulation feature is quite flexible and is able to independently extract amplitude, phase, and frequency modulation. The modulation index and frequency deviation of the signal can be measured, and the frequency spectrum of the modulating signal can be displayed.

6.11 Digital Modulation

The widespread adoption of digital technology has resulted in the need for digital modulation techniques. Instead of an analog signal modulating the carrier frequency, the modulating signal is a stream of digital bits. These digital bits may, in fact, represent an analog signal, or they may have originated from a digital data source. Either way, the communications system must be designed to transfer logical 1's and 0's from transmitter to receiver.

The top row of Figure 6-11 shows the digital data to be transmitted, which are converted into the digital waveform labeled *modulation*. This digital signal is used to modulate the frequency, amplitude, or phase of the carrier frequency. More advanced modulation techniques use more than one of these attributes of the carrier simultaneously. For example, the bottom waveform of Figure 6-11 shows both the amplitude and phase of the carrier being modulated.

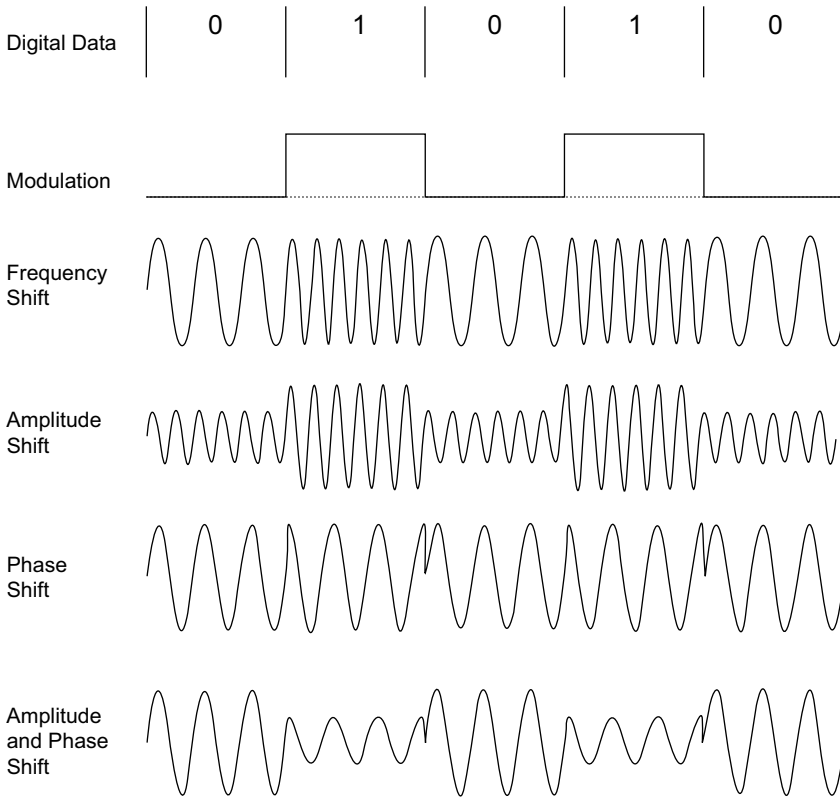


Figure 6-11 Digital modulation techniques use the digital signal to control the amplitude, frequency, or phase of the carrier.

6.12 Quadrature Modulation

Most digital modulation schemes are defined and generated in terms of *quadrature modulation*. Starting with the generalized case with both amplitude and phase modulation:

$$v(t) = A_c(t)\cos(2\pi f_c t + \theta(t)) \tag{6-43}$$

Using $\cos(x + y) = \cos x \cos y - \sin x \sin y$

$$v(t) = A_c(t)[\cos(\theta(t)) \cos(2\pi f_c t) - \sin(\theta(t)) \sin(2\pi f_c t)] \tag{6-44}$$

$$v(t) = A_c(t) \cos(\theta(t)) \cos(2\pi f_c t) - A_c(t) \sin(\theta(t)) \sin(2\pi f_c t) \tag{6-45}$$

This produces two components to the modulated waveform: one multiplies the $\cos(2\pi f_c t)$ carrier frequency; the other multiplies the $\sin(2\pi f_c t)$ term, which is 90° out of phase with the main carrier.

Defining the two modulation terms as

$$\text{In-phase modulation : } v_i(t) = A_c(t) \cos(\theta(t))$$

$$\text{Quadrature modulation : } v_q(t) = -A_c(t) \sin(\theta(t))$$

$$v(t) = v_i(t) \cos(2\pi f_c t) + v_q(t) \sin(2\pi f_c t) \tag{6-46}$$

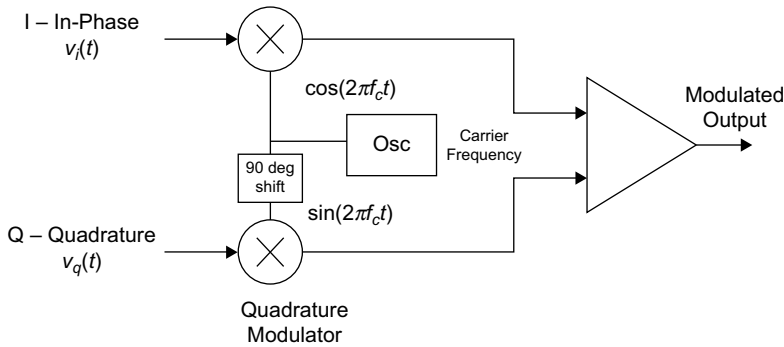


Figure 6-12 A quadrature modulator is capable of generating digital modulation using *in-phase* and *quadrature* components of the signal.

Figure 6-12 shows a block diagram for a quadrature modulator. The in-phase component and the quadrature component are multiplied by the carrier frequency (with a 90° phase shift for the quadrature component) and summed to obtain the desired modulation vector.

A common tool used for vector modulation is to plot the in-phase and quadrature components as a vector (Figure 6-13a). The magnitude of the vector corresponds to the amplitude of the signal, and the phase of the vector is the angle, θ , shown in the diagram. In many measurement applications, only the point of the vector is drawn (Figure 6-13b). There are a number of terms used to describe this vector representation: *vector*, *phasor*, and *quadrature modulation*.

6.13 Common Digital Modulation Formats

Binary phase-shift keying (BPSK) is one of the simplest forms of digital modulation. BPSK shifts the phase of the carrier between 0° and 180° based on the digital modulation signal. The constellation diagram for BPSK has two dots, as shown in Figure 6-14. You can imagine a vector that points to the right-hand dot when the digital modulation is a 0, flopping over to the left-hand dot when the digital value changes to a 1.

Quadrature phase-shift keying (QPSK) uses four equally spaced phases to represent four distinct states of the vector modulation. While the amplitude remains constant for all states, the phase angle of the vector takes on the values of 45° , 135° , -135° , and -45° . As shown in Figure 6-15, each of the four states represents two bits of information. To describe this, we introduce the concept of a *symbol*, which is defined as a unique state of the modulated waveform that persists for a fixed period of time. In the case of QPSK, a symbol represents 2 bits, so the number of bits transmitted per second (*bit rate*) is twice the number of symbols per second (*symbol rate*). Compare this with BPSK, which has one bit per symbol, resulting in the bit rate being the same as the symbol rate.

$$\text{Bit rate} = \text{symbol rate} \times \text{number of bits/symbol} \quad (6-47)$$

In general, as the number of bits per symbol is increased, the error rate at the receiver increases since it has to distinguish the different symbols from each other, which may be difficult and cause bit errors.

Quadrature amplitude modulation (QAM) uses both amplitude and phase change to represent the digital signal. One version of QAM is 16QAM, which has 16 logical states

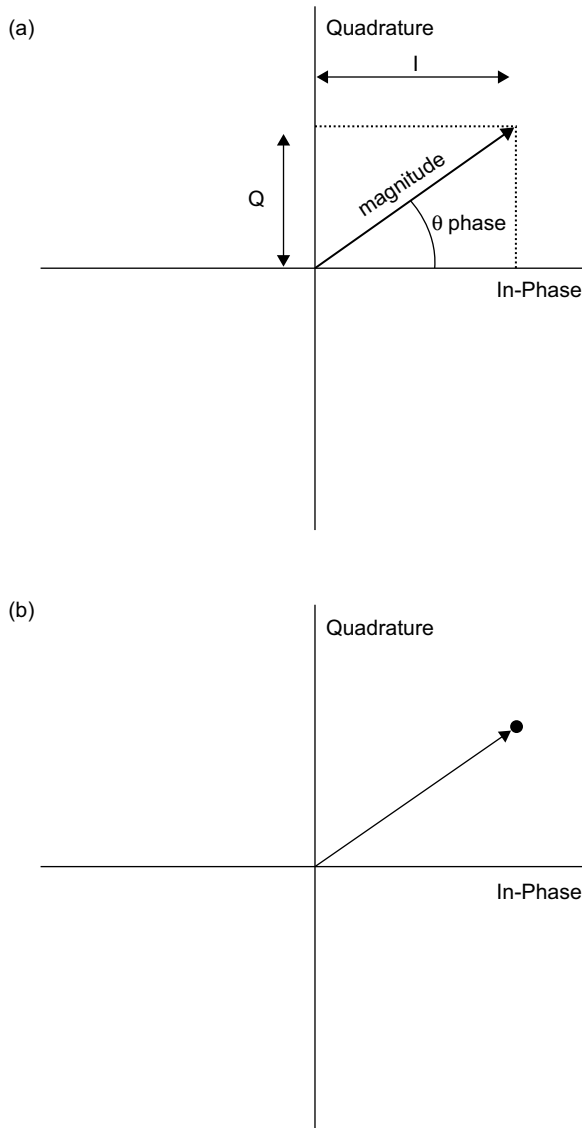


Figure 6-13 (a) Digital modulation is often expressed in terms of a vector diagram, with in-phase and quadrature components. (b) For constellation diagrams, only the end of the vector is shown as a point.

represented by the vector modulation. Figure 6-16 shows how these 16 states are plotted in the vector space. Each of the states represents four bits of information, so the bit rate is 4 times the symbol rate.

Figure 6-17 shows a practical measurement of a 16QAM signal. The plot on the left shows the transitions of the vector modulation as it changes between the various states, resulting in a busy plot. The plot on the right shows the vector modulation only at the appropriate sample times, emulating how a receiver will evaluate the modulated signal.

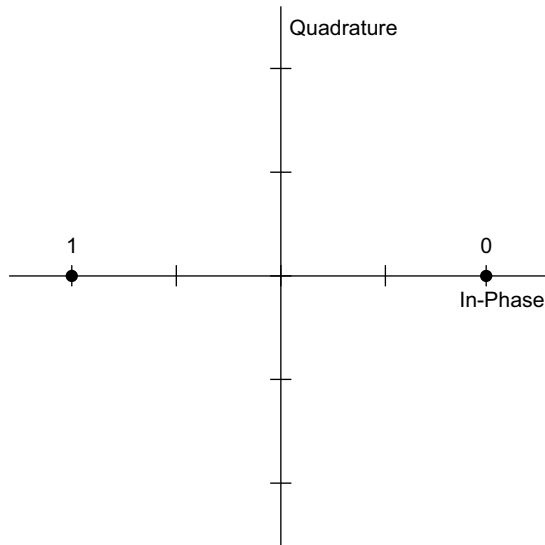


Figure 6-14 The constellation diagram of a BPSK signal shows a dot at 0° and a dot at 180° , corresponding to the two states of the vector modulation.

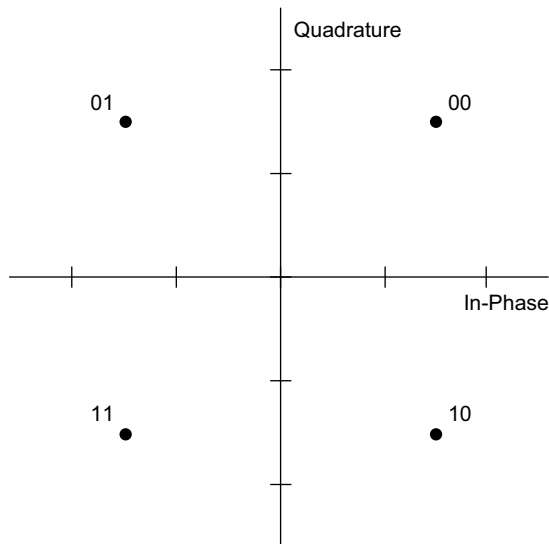


Figure 6-15 The QPSK constellation diagram shows four equally spaced constant amplitude states, with a phase angle of 45° , 135° , -135° , and -45° . Each position of the vector represents two bits of information.

The vector modulation states are represented by the 16 dots in a constellation diagram. The circles around the dots represent a nominal deviation from the center of the ideal vector position. The signal shown has very little noise in it, so the constellation diagram shows small dots with small statistical variation.

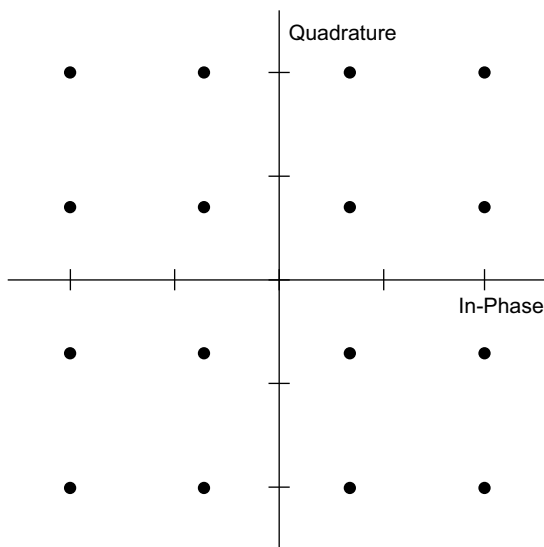


Figure 6-16 The 16QAM constellation diagram shows 16 states, with produced by varying the amplitude and the phase of the vector modulation. Each position of the vector represents four bits of information.

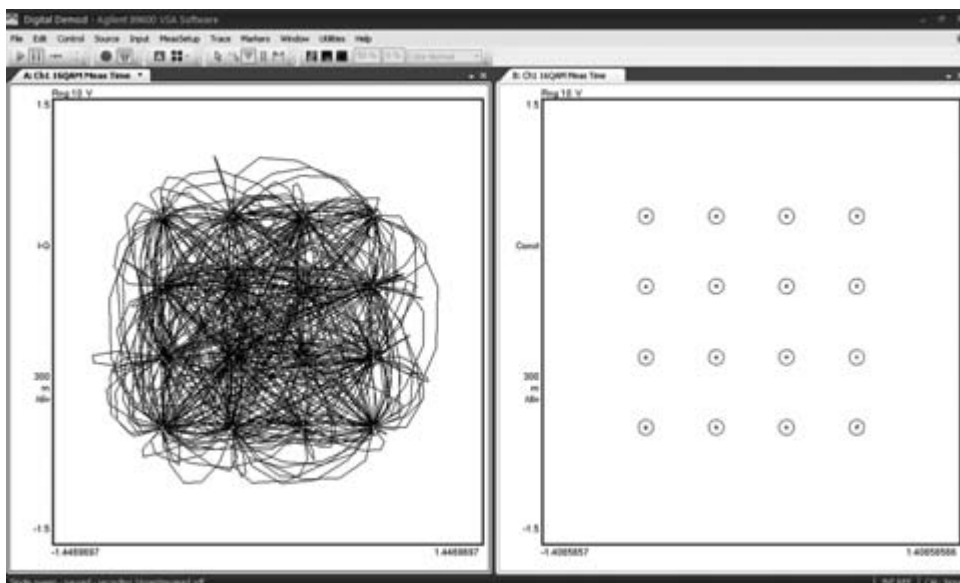


Figure 6-17 A practical measurement of a 16QAM signal using vector signal analysis software. The display on the left shows all of the transitions of the signal, while the display on the right displays only the recovered I/Q symbols.

6.14 Error Vector Magnitude

The move to digital modulation in vector form created the need for a standard method to describe the error in the modulation vector. *Error vector magnitude* (EVM), also referred to as *relative constellation error* (RCE), is a quantitative measure of the quality of a vector-modulated signal. As shown in Figure 6-18, the *error vector* represents the vector difference between the ideal signal and the actual measured signal. EVM is the root mean square (RMS) value of the error vector over some time interval (evaluated at the valid symbol times).

EVM may be reported as a percentage or in decibels, referenced to the square root of the mean power of the ideal signal, the square root of the average symbol power, or the peak signal level.

Modulation error ratio (MER) is another way to represent the quality of a vector signal, often expressed in dB.

$$MER_{(dB)} = 10 \log \left(\frac{P_{\text{signal}}}{P_{\text{error}}} \right) \quad (6-48)$$

where

P_{signal} = the average power of the signal

P_{error} = the average power of the error vector

A practical digital modulation measurement is shown in Figure 6-19. QPSK is used in this measurement example for simplicity and enables understanding of the various measurements. Four different views of the signal provide a comprehensive look at this vector modulated signal. The state diagram in the upper left of the figure shows the four QPSK states along with the transitions made by the vector signal. The lower left chart is just the frequency spectrum of the baseband signal, with the vertical axis in dBm and the horizontal axis in frequency. The upper right view is the time domain signal, which shows a relatively constant amplitude with phase shift keying. The lower right chart is a tabular view of the measurement parameters and a list of the decoded QPSK symbols.

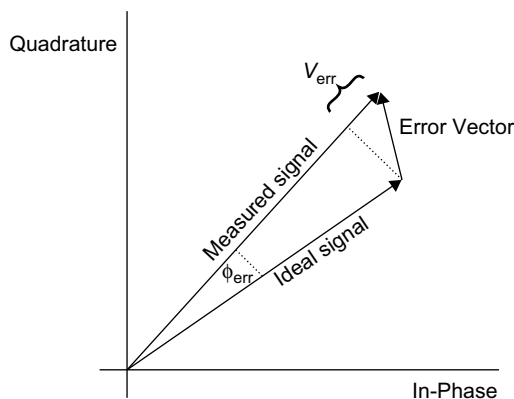


Figure 6-18 Error vector magnitude (EVM) is the magnitude of the error vector, drawn from the ideal signal to the measured signal.

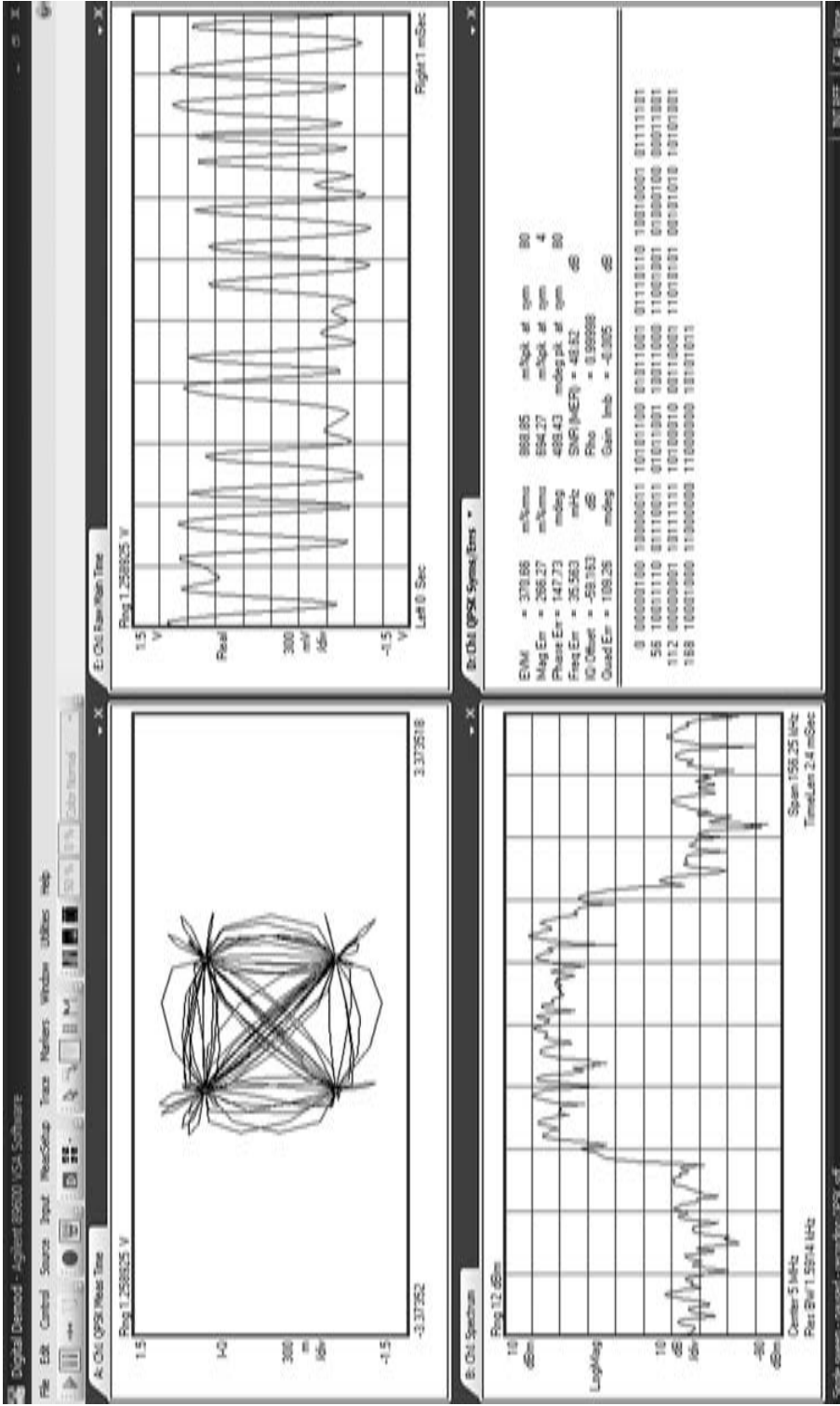


Figure 6-19 A comprehensive measurement of a QPSK signal shows four different views. Upper left: The modulation states with transitions. Lower left: The frequency spectrum of the modulated signal. Upper right: The time domain view of the modulated signal (the real portion of the complex I/Q waveform before synchronization). Lower right: Textual readout of EVM and other measured parameters.

6.15 Channel Measurements

Many communication systems are based on channels that divide up the available frequency spectrum such that each signal is assigned a particular channel. It is important to understand and measure the bandwidth that a particular signal is occupying and whether it is spilling over into adjacent channels. Digital modulation produces signals that spread out in frequency, causing signal and channel measurements to be more complex.

The *occupied bandwidth* (OBW) of a signal is a measure of how wide a signal is in frequency, usually defined as the bandwidth that contains 99% of the signal's power. This can be a tedious measurement to do manually but spectrum analyzers often perform this measurement automatically (Figure 6-20). The analyzer integrates the power from all frequencies in the signal and determines the bandwidth that corresponds to 99% of the signal's power.

Another important measurement is *channel power*, which measures the total power in the specified channel. Again, modern spectrum analyzers provide an automated measurement that integrates the total power in the channel, based on the center frequency and bandwidth provided by the user.

It is important that a signal in a channel not interfere with adjacent channels by spewing energy into those channels. Thus, an important measurement is *adjacent channel power* (ACP). As shown in Figure 6-21, a spectrum analyzer can automatically measure the power in the desired channel along with any power that is present in the adjacent channels.

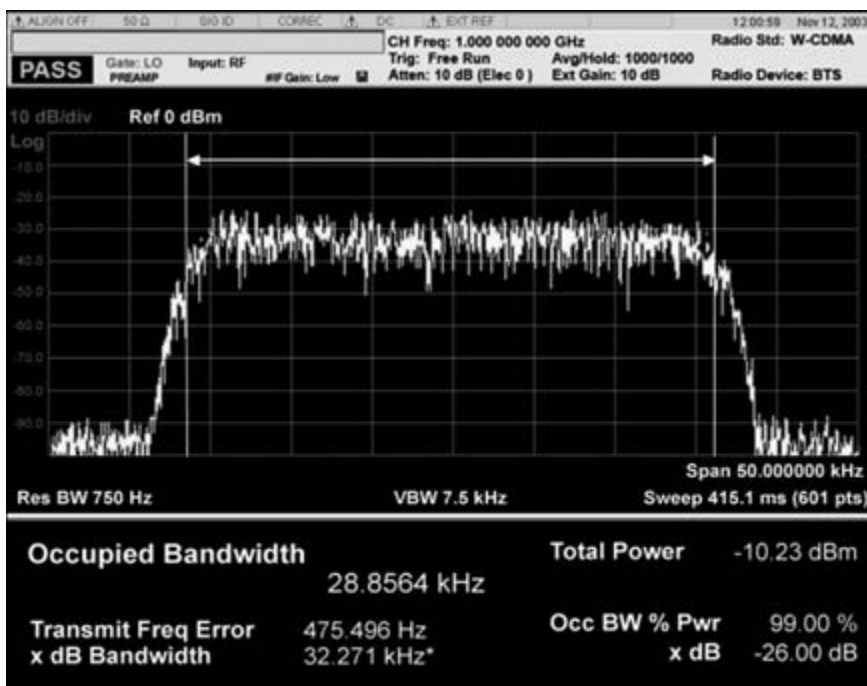


Figure 6-20 An automated measurement of occupied bandwidth also shows the total power in the signal. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)



Figure 6-21 Adjacent channel power measures the impact of a signal on channels that are adjacent to it. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

The *adjacent channel power ratio* (ACPR) expresses the ACP relative to the power in the main channel:

$$ACPR = \frac{P_{adj}}{P_{main}} \tag{6-49}$$

where

P_{adj} = power in the adjacent channel

P_{main} = power in the main channel

ACPR is often expressed in dB. Some wireless communication standards use the term *adjacent channel leakage ratio* (ACLR), which is the same concept as ACPR. Some specific ACPR measurement techniques may be required by a wireless communications standard.

Poor ACPR can be caused by improper filtering in the transmitter (too wide of a passband or poor stop-band rejection). More commonly, ACPR problems are caused by intermodulation distortion, which causes energy to spill over into adjacent communication channels.

Bibliography

Adam, Stephen F. *Microwave Theory and Applications*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1969.

Agilent Technologies, “8 Hints for Making and Interpreting EVM Measurements,” Publication Number 5989-3144EN, May 2005.

Agilent Technologies, “Agilent Vector Signal Analyzer Basics,” Application Note, Publication Number 5990-7451EN, February 2011.

Agilent Technologies, “Digital Modulation in Communications Systems—An Introduction,” Application Note 1298, Publication Number 5965-7160E, March 2001.

Agilent Technologies, “Spectrum Analyzer Mode User’s and Programmer’s Reference, Agilent X-Series Signal Analyzer,” Publication Number N9060-90027, February 2012.

Agilent Technologies, “Using Error Vector Magnitude Measurements to Analyze and Troubleshoot Vector-Modulated Signals,” Product Note 89400-14, Publication Number 5965-2898E, 2000.

Engelson, Morris. *Modern Spectrum Analyzer Theory and Applications*. Dedham, MA: Artech House, 1984.

Kinley, R. Harold. *Standard Radio Communications Manual*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1985.

Malvino, Albert Paul. *Electronic Principles*, 3rd ed. New York: McGraw-Hill Book Company, 1984.

Schwartz, Mischa. *Information, Transmission, Modulation, and Noise*, 3rd ed. New York: McGraw-Hill Book Company, 1980.

Scott, Allan W., and Frobenius, Rex. *RF Measurements for Cellular Phones and Wireless Data Systems*. Hoboken, NJ: John Wiley & Sons, Inc., 2008.

Ziemer, R. E., and W. H. Tranter. *Principles of Communications*. Boston: Houghton Mifflin Company, 1976.

Distortion Measurements

Many electronic circuits and systems are considered to be *linear time invariant* (LTI), as introduced in Chapter 1. For a sinusoidal input the output of an LTI system is also sinusoidal with perhaps a different amplitude and phase. In the time domain, the output waveform is a sinusoid, the exact same shape as the input waveform. In the frequency domain, we expect to see at the output the same frequency that was at the input (and only that frequency). Any other frequencies that are generated due to the input signal are considered to be distortion.¹

7.1 The Distortion Model

It is difficult to create systems that are purely linear since some distortion of the signal is normally present. Many of the distortion mechanisms measured with spectrum analyzers are low level. That is, the devices producing the distortion are mostly linear and have only a slight nonlinear behavior. Such a weakly nonlinear system can be modeled with a power series.

$$V_{\text{out}} = k_0 + k_1 V_{\text{in}} + k_2 V_{\text{in}}^2 + k_3 V_{\text{in}}^3 + k_4 V_{\text{in}}^4 + \dots \quad (7-1)$$

The first coefficient, k_0 , represents the DC offset in the system. The second coefficient, k_1 , is the gain of the circuit associated with linear circuit theory. The remaining coefficients, k_2 and above, represent the nonlinear behavior of the circuit. If the circuit were completely linear, all of the coefficients except k_1 would be zero.

The model can be simplified by ignoring the terms that come after the k_3 term. For gradual nonlinearities, the size of k_n decreases rapidly as n gets larger. For many applications the reduced model is sufficient, since the second-order and third-order effects dominate. (Expansion of the model to higher order is discussed later.)

$$V_{\text{out}} = k_0 + k_1 V_{\text{in}} + k_2 V_{\text{in}}^2 + k_3 V_{\text{in}}^3 \quad (7-2)$$

¹ This does not include frequency components that are generated in the circuit independent of the input signal, such as spurious responses.

7.2 Single-Tone Input

The simplest distortion test of a system is to input a pure sinusoid and measure the frequency content of the output signal:

$$V_{\text{in}} = A \cos \omega t \quad (7-3)$$

The angular frequency, $\omega = 2\pi f$
where

f = frequency (Hz)

Inserting this into the distortion model gives

$$V_{\text{out}} = k_0 + k_1 A \cos \omega t + k_2 A^2 \cos^2 \omega t + k_3 A^3 \cos^3 \omega t \quad (7-4)$$

$$\begin{aligned} V_{\text{out}} = k_0 + k_1 A \cos \omega t + (k_2 A^2 / 2)(1 + \cos 2\omega t) \\ + k_3 A^3 (3/4 \cos \omega t + 1/4 \cos 3\omega t) \end{aligned} \quad (7-5)$$

Collecting terms,

$$\begin{aligned} V_{\text{out}} = k_0 + k_2 A^2 / 2 + (k_1 A + 3k_3 A^3 / 4) \cos \omega t \\ + (k_2 A^2 / 2) \cos 2\omega t + (k_3 A^3 / 4) \cos 3\omega t \end{aligned} \quad (7-6)$$

This leaves us with an output voltage containing a DC component, the original (fundamental) frequency, and its second and third harmonics. These distortion products are called *harmonic distortion* since they occur at multiples of the original frequency. Had we used a higher-order model, the analysis would have shown even higher-order harmonics present at the output. Note that the fundamental amplitude is affected by the nonlinear third-order coefficient of the model, k_3 . Similarly, the DC component of the equation is affected by the second-order coefficient. The fundamental is mostly proportional to A , the second harmonic is proportional to A^2 , and the third harmonic is proportional to A^3 .

The model is somewhat limited since we do not usually know the values of k_0 , k_1 , k_2 , and k_3 for a particular device. However, we can infer some useful information from the model anyway. Consider what happens when the signal level, A , is reduced. The fundamental will be reduced almost in direct proportion to the signal amplitude. We might say that the fundamental is reduced 1 dB per dB of change in the signal level. The second harmonic will go down as the square of A , or converting to dB

$$20 \log (A^2) = 2(20 \log A) = 2 A_{\text{dB}} \quad (7-7)$$

This means that the second harmonic will be changed 2 dB per dB of signal level change. Similarly, the third harmonic term amplitude is proportional to A^3 . Converting to dB,

$$20 \log (A^3) = 3(20 \log A) = 3 A_{\text{dB}} \quad (7-8)$$

which means that the third harmonic will be reduced by 3 dB per dB of signal level reduction.

Figure 7-1 shows the spectrum of a typical signal having harmonic distortion. (Ideally, a pure sine wave would have no harmonics.) Note that the odd harmonics, particularly the third harmonic, is larger than the even harmonics. Distortion that maintains the 50% duty

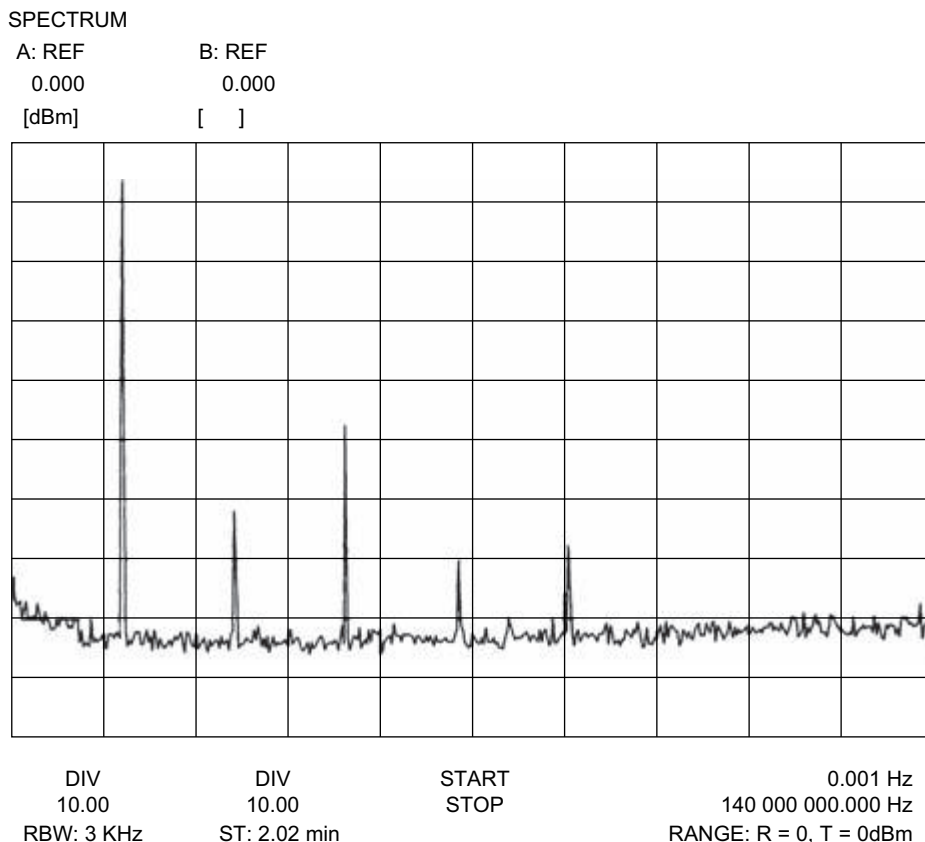


Figure 7-1 Measurement of harmonic distortion in a signal.

cycle of the ideal waveform will create only odd harmonics. (Recall the case of the square wave from Chapter 3.) Distortion mechanisms that upset the symmetry of the signal produce even harmonics.

When making this kind of measurement, one must get accustomed to the fact that there are very few pure sine waves. For example, a good signal or function generator may have a third harmonic that is 30 or 40 dB lower than the fundamental. When viewed on an oscilloscope, this signal will appear to be a pure sine wave since the distortion is not discernable. When measured with even a moderate performance spectrum analyzer, the harmonics will be easily visible. This illustrates one advantage of a narrowband receiver (the spectrum analyzer) versus a wideband receiver (the oscilloscope).

7.3 Two-Tone Input

Another input signal commonly used for distortion tests is the two-tone signal.

$$V_{\text{in}} = A_1 \cos \omega_1 t + A_2 \cos \omega_2 t \quad (7-9)$$

Using our distortion model,

$$V_{\text{out}} = k_0 + k_1 V_{\text{in}} + k_2 V_{\text{in}}^2 + k_3 V_{\text{in}}^3 \quad (7-10)$$

The result is in the form

$$\begin{aligned} V_{\text{out}} = & c_0 + c_1 \cos \omega_1 t + c_2 \cos \omega_2 t + c_3 \cos 2\omega_1 t \\ & + c_4 \cos 2\omega_2 t + c_5 \cos 3\omega_1 t + c_6 \cos 3\omega_2 t \\ & + c_7 \cos(\omega_1 t + \omega_2 t) + c_8 \cos(\omega_1 t - \omega_2 t) \\ & + c_9 \cos(2\omega_1 t + \omega_2 t) + c_{10} \cos(2\omega_1 t - \omega_2 t) \\ & + c_{11} \cos(2\omega_2 t + \omega_1 t) + c_{12} \cos(2\omega_2 t - \omega_1 t) \end{aligned} \quad (7-11)$$

where

c_0, \dots, c_{12} = coefficients determined by k_0, \dots, k_3, A_1 , and A_2 .

Besides the harmonics of the two tones (as in the single-tone case), there are also sum and difference frequencies. These new frequency components are called *intermodulation distortion* (IMD) because they result from the two tones modulating together. The frequencies present in the output satisfy the following criterion:

$$\omega_{nm} = |n\omega_1 \pm m\omega_2| \quad (7-12)$$

where

n and m = positive integers such that $n + m \leq 3$

With the frequency expressed in Hz,

$$f_{nm} = |nf_1 \pm mf_2| \quad (7-13)$$

If the distortion model is expanded from the third-order model to a higher-order model, the limit on the sum of $n + m$ is raised accordingly.

The order of a particular frequency component is the sum of the n and m values used to obtain that frequency (e.g., f_{12} and f_{21} are third-order terms, and f_{20} and f_{11} are second-order terms). As in the single-tone case, second-order terms will be reduced 2 dB in amplitude when the input tones are reduced by 1 dB. Equivalently, second-order terms are reduced 2 dB/dB of input signal reduction. Third-order terms are reduced 3 dB/dB of signal reduction and so on for higher-order terms, if present.

Example 7.1

Assuming a third-order distortion model, what frequencies will be present at the output with a two-tone input signal with frequencies of 10.7 MHz and 10.8 MHz?

The output frequencies are given by $f = |nf_1 \pm mf_2|$. For $n = 1$ and $m = 0$,

$$f_{10} = |10.7 \text{ MHz} \pm 0| = 10.7 \text{ MHz}$$

For $n = 2$ and $m = 0$,

$$f_{20} = |2(10.7 \text{ MHz}) \pm 0| = 21.4 \text{ MHz}$$

For $n = 3$ and $m = 0$,

$$f_{30} = |3(10.7 \text{ MHz}) \pm 0| = 32.1 \text{ MHz}$$

For $n = 0$ and $m = 1$,

$$f_{01} = |0 \pm 10.8 \text{ MHz}| = 10.8 \text{ MHz}$$

For $n = 0$ and $m = 2$,

$$f_{02} = |0 \pm 2(10.8 \text{ MHz})| = 21.6 \text{ MHz}$$

For $n = 0$ and $m = 3$,

$$f_{03} = |0 \pm 3(10.8 \text{ MHz})| = 32.4 \text{ MHz}$$

So far, these frequencies are simply the first three harmonics of the two input tones. Now the sum and difference frequencies will be calculated.

For $n = 1$ and $m = 1$,

$$f_{11} = |10.7 \text{ MHz} \pm 10.8 \text{ MHz}| = 0.1 \text{ MHz}, 21.5 \text{ MHz}$$

For $n = 2$ and $m = 1$,

$$f_{21} = |2(10.7 \text{ MHz}) \pm 10.8 \text{ MHz}| = 10.6 \text{ MHz}, 32.2 \text{ MHz}$$

For $n = 1$ and $m = 2$,

$$f_{12} = |10.7 \text{ MHz} \pm 2(10.8 \text{ MHz})| = 10.9 \text{ MHz}, 32.5 \text{ MHz}$$

The spectrum of the output signal is shown in Figure 7-2. The amplitudes of the frequency components will depend on the levels of the input tones and the coefficients of the distortion model. However, the amplitudes that are shown in the figure are typical of a distorted signal.

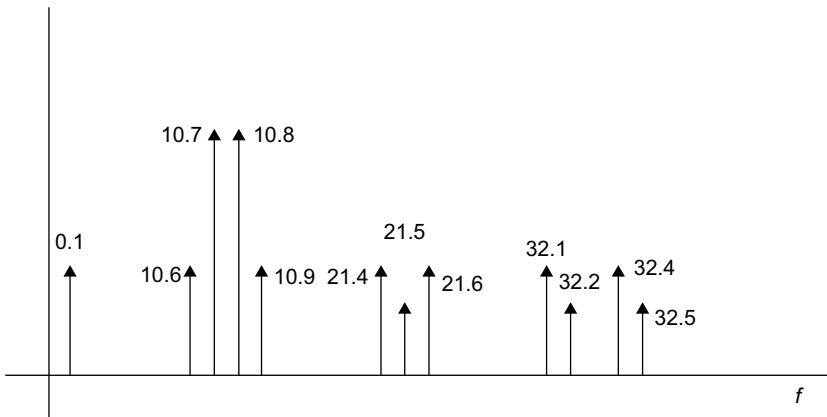


Figure 7-2 The spectrum of a two-tone signal with third-order intermodulation distortion products.

A few comments are in order now that a numerical example has been given. The two input tones were chosen to be close to each other in frequency, as is usually the case for two-tone testing. An examination of Figure 7-2 will reveal that the spectral lines tend to fall in four groupings. The $f_1 - f_2$ frequency (0.1 MHz) will fall down near DC. The other frequencies fall in groups near the fundamentals (near 10.7 MHz), the second harmonics (near 21.5 MHz), and the third harmonics (near 32.4 MHz) of the original two tones. Depending on the system involved, some of these distortion components can be neglected since they will be filtered out at some point. For instance, an intermediate frequency (IF) amplifier stage will usually be narrowband, centered on the two input tones. Spectral components out at the second and third harmonics can be easily filtered out. The distortion components close to the original tones (f_{21} and f_{12}) will be more troublesome since they fall near the desired frequencies. In many cases, odd-order intermodulation products are of particular concern to radio frequency (RF) designers since the distortion products fall in band.

7.4 Higher-Order Models

We have chosen to limit the number of terms in the distortion model to produce a third-order behavior. Even with such a simple model, the derivation of the output signal frequency components is lengthy and expanding the model to a higher order makes the situation only worse. Fortunately, for many situations a third-order model is sufficient.

But what if the third-order model is insufficient? For instance, it is common to have significant energy in the fifth, sixth, or seventh harmonic of a single tone, yet the third-order model does not show this effect. The analytical approach used previously can simply be expanded to include the higher-order terms, with the penalty of the mathematics getting more difficult. Another approach is to simply expand on the concepts demonstrated by the third-order model, even though they have not been proven rigorously. As stated previously, the frequencies generated by the distortion model obey the $n f_1 \pm m f_2$ rule, where the maximum value of $m + n$ is the order of the model. So it is possible to predict the frequency components of higher-order systems without extensive mathematics.

The example shown in Figure 7-2 is for the simple case of two tones. As shown in Chapter 6, modern digital modulation techniques use signals that have complex spectrums. When these signals are subject to intermodulation distortion, the frequency content spreads out, causing interference to nearby communication channels (Figure 7-3).

7.5 The Intercept Concept

Increasing the signal level at the input to a weakly nonlinear device will cause the distortion products to increase at the output. The distortion products increase in amplitude and they do so *faster than the input signal increases*. Figure 7-4 shows a plot of the output power versus the input power for the fundamental, second-order frequency components, and third-order frequency components. For increasing fundamental input power, the fundamental output power increases in a linear manner, consistent with the gain or loss of the device. At some point, gain compression may occur, and the fundamental output power no longer increases

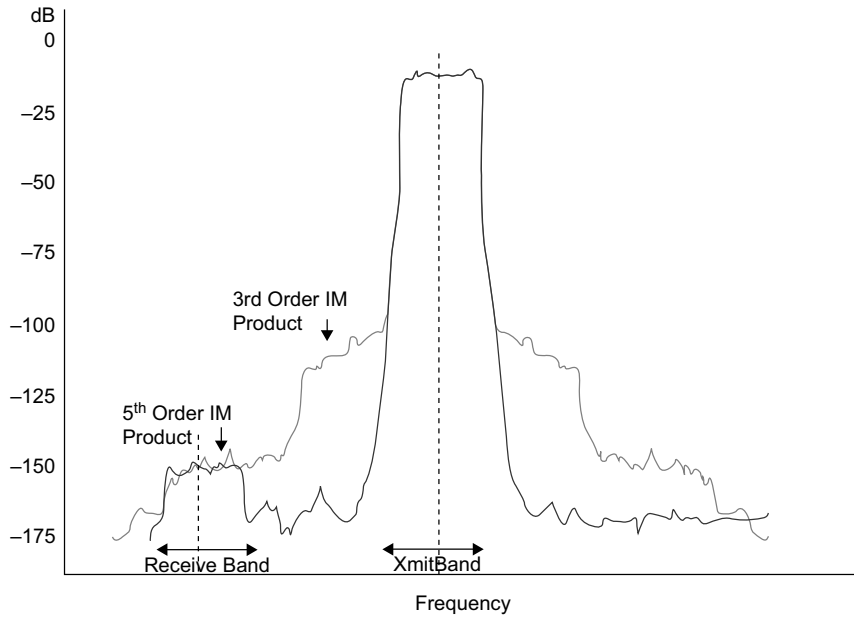


Figure 7-3 A digitally modulated signal with wide spectral content spills over into adjacent channels when subject to intermodulation distortion.

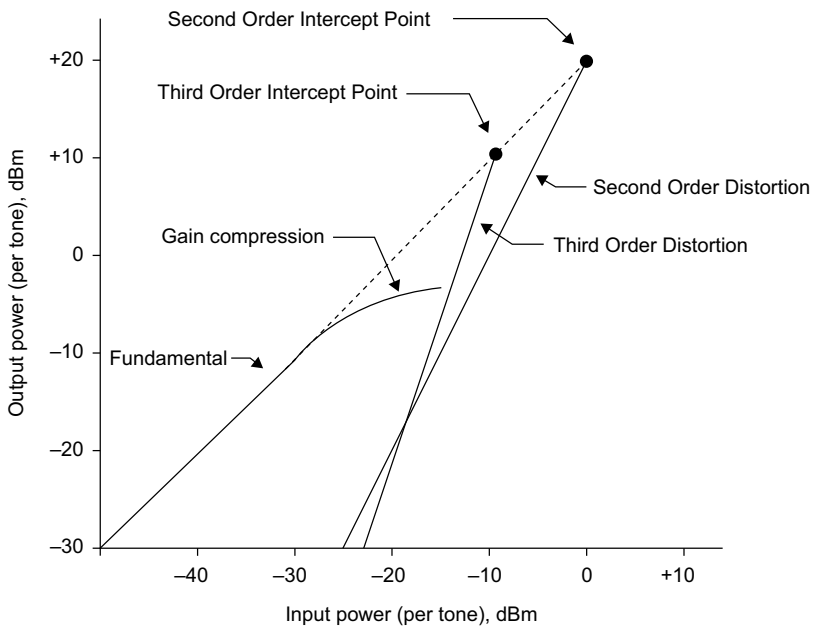


Figure 7-4 Plot of fundamental, second-order distortion product, and third-order distortion product power levels illustrates the concept of second- and third-order intercept points.

with input power. The output power of the second-order distortion products also increases with fundamental input power, but at a faster rate. Recall that the distortion model shows that second-order terms change 2 dB per 1 dB of change in the fundamental. Thus, on a decibel plot the line representing the second-order output power has twice the slope of the fundamental line. Similarly, the third-order distortion products change 3 dB per 1 dB of change in the fundamental so that line has a slope that is three times the slope of the fundamental line.

If there is no gain compression, the fundamental input power could be increased until the second-order distortion products would eventually catch up with it and the two output power levels would be equal. This point is referred to as the *second-order intercept point*. The third-order distortion products also increase faster than the fundamental, and those two lines will intersect at the *third-order intercept point*. Rarely can either of these two points be measured directly due to the gain compression of the fundamental. Instead, the intercept points are extrapolated from measurements of the fundamental and distortion products at power levels below where gain compression occurs. The intercept points are usually specified in dBm and may refer either to the output or the input. (It is important to always specify whether the intercept point refers to the output power or the input power. The two points will differ by the gain of the device.)

The utility of the intercept concept is in specifying and predicting the distortion level in a system. One might be tempted to specify the distortion of a circuit or system directly by stating the level of the distortion products in decibels relative to the signal level. This can be done but is not very meaningful unless the signal level is also specified. One circuit's distortion might be -80 dB relative to the signal while another circuit might achieve only -40 dB. However, these two values are not a fair comparison unless the same signal level is used. The second- and third-order intercept points are figures of merit that are independent of signal level. Therefore, the distortion performance of two different circuits can be compared quite easily if their intercept points are known.

Most often, an engineer is interested in the level of the distortion products relative to the signal level. The intercept points do not indicate this directly and may seem cumbersome to use, but a few observations will show how the relative distortion level can be easily determined from the intercept point. The difference between the level of the second-order distortion products and the fundamental signal level is the same as the difference between the fundamental signal level and the intercept point. Suppose the second-order intercept point is $+15$ dBm and the fundamental signal level is -10 dBm (both referred to the output of the device). The difference between these two values is 25 dB. Therefore, the second-order distortion products will be 25 dB below the fundamental, or -35 dBm. So the intercept point allows easy conversion between fundamental signal level and distortion level. Often the distortion level is specified relative to the fundamental power level, and the conversion to absolute power (dBm) is not necessary.

The difference between the level of the third-order distortion products and the fundamental signal level is *twice* the difference between the fundamental signal level and the third-order intercept point. (Note that the second-order intercept point is *not* the same as the third-order intercept point.) Suppose that the third-order intercept point is $+5$ dBm and the fundamental signal level is -25 dBm, both referred to the output of the device. The

difference between the intercept and the fundamental is 30 dB, so the third-order distortion products will be two times 30 dB down from the fundamental. The relative distortion level is -60 dB and the absolute power level of the distortion products is -85 dBm.

Example 7.2

What is the maximum allowable power level of the input signal if the third-order distortion products are to be less than -70 dB relative to the fundamental? The third-order intercept point is $+10$ dBm, referred to the input.

The third-order distortion products are to be 70 dB below the fundamental, so the fundamental must be $70/2$ dB or 35 dB below the intercept point. The intercept point is $+10$ dBm, so the signal level should be -25 dBm at the input.

7.6 Harmonic Distortion Measurements

Harmonic distortion measurements can easily be made with a spectrally pure signal source and a spectrum analyzer. The quality of the measurement is limited by the harmonic distortion of both the signal source and spectrum analyzer. The signal source is most often the limiting factor, with harmonic distortion performance often not much better than 40 dB below the fundamental.

The source provides a signal to the device under test and the spectrum analyzer is used to monitor the output. Figure 7-5 shows a typical harmonic distortion measurement, with the distortion level specified using the largest harmonic level expressed in dB relative to the fundamental.

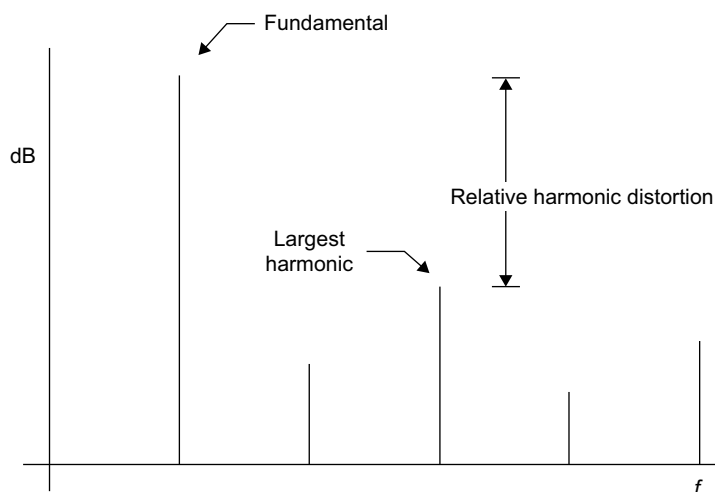


Figure 7-5 The harmonic distortion of a signal is often specified by stating the amplitude of the largest harmonic in dB relative to the fundamental.

Alternatively, the distortion may be specified as *total harmonic distortion* (THD), usually as a percent of the fundamental. THD takes into account the power in all the harmonics:

$$\text{THD} = \sqrt{V_2^2 + V_3^2 + \dots} / V_1 \quad (7-14)$$

where V_1 is the RMS voltage of the fundamental and V_2, V_3, \dots are the RMS voltages of the harmonics.

All harmonics of the fundamental are summed in a RMS manner and are divided by the fundamental RMS voltage. Since an infinite number of harmonics cannot be measured, a finite number will have to suffice. Fortunately, the harmonic amplitudes tend to decrease with higher harmonic numbers. The calculation is somewhat tedious for a large number of harmonics, but some spectrum analyzers include an automatic THD function. If not, the user must determine each harmonic amplitude and compute the THD.

Example 7.3

Determine the total harmonic distortion of a signal with the following spectral components: 1 MHz, 3.5 V RMS; 2 MHz, 0.1 V RMS; 3 MHz, 0.2 V RMS; 4 MHz, 0.05 V RMS. Express the largest harmonic in decibels relative to the fundamental.

The fundamental frequency is 1 MHz.

$$\begin{aligned} \text{THD} &= \sqrt{(0.1)^2 + (0.2)^2 + (0.05)^2} / 3.5 = 0.229 / 3.5 \\ &= 0.065 \text{ or } 6.5\% \end{aligned}$$

The largest harmonic is the third harmonic (3 MHz). In decibels, this harmonic is $20 \log(0.2/3.5) = -24.9$ dB relative to the fundamental.

7.7 Use of Low-Pass Filter on Source

The signal source is often the limiting factor in a harmonic distortion measurement due to its own harmonic distortion. A typical signal generator may have harmonic distortion around -40 dB relative to the fundamental,² whereas a typical spectrum analyzer may have a dynamic range of over 80 dB.

A low-pass filter can be used to improve the source's effective harmonic distortion, as shown in Figure 7-6. The cutoff frequency of the low-pass filter is chosen such that the fundamental frequency is passed largely intact, while the harmonics are attenuated significantly. The performance of the source/filter combination can be verified directly by the spectrum analyzer. The passband attenuation of the filter should be kept to a minimum, but the exact value is not critical. If the loss through the filter at the fundamental frequency is significant, it should be accounted for when setting the source output level. The spectrum analyzer can be used to check directly the amplitude of the fundamental at the output of the filter.

² Sources that are designed with distortion measurements in mind may have considerably better harmonic distortion, but are usually restricted to frequencies below 10 MHz with best distortion performance in the audio range.

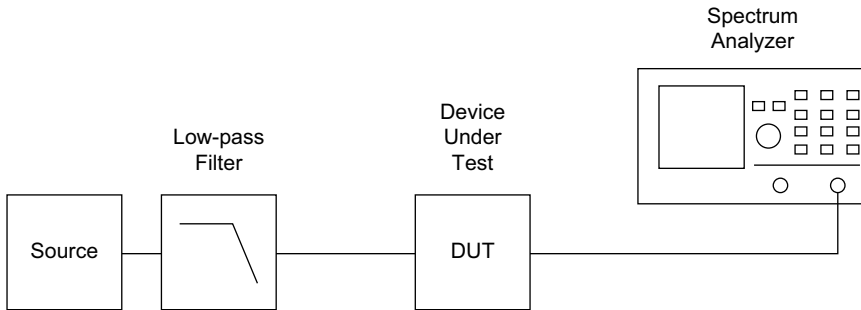


Figure 7-6 The harmonic distortion of a signal source can be improved by installing a low-pass filter at the source's output.

7.8 Intermodulation Distortion Measurements

To test for intermodulation distortion, two stimulus sine waves are required. The test setup shown in Figure 7-7 has two independent signal sources connected with a power splitter (used as a combiner) to drive the device under test. The sources are set at the same output level, but at different frequencies. The 6 dB loss of the combiner should be accounted for when setting the output amplitudes of the sources. A typical spectrum analyzer display of the two-tone distortion test is shown in Figure 7-8. As shown, the third-order products (f_{21} and f_{12}) that fall close to the original two tones are being measured. This is a common measurement since the two distortion products fall close to the original two tones and are difficult to remove by filtering.

In some cases, the two sources may interact and produce intermodulation distortion. This problem can be detected with the spectrum analyzer and can be cured by inserting fixed attenuators at the outputs of the sources. These attenuators increase the isolation between the sources and prevent internally generated intermodulation distortion. The output levels of the sources should be increased to compensate for the signal loss in the attenuators. The automatic leveling circuits in the sources can also be a source of intermodulation as each

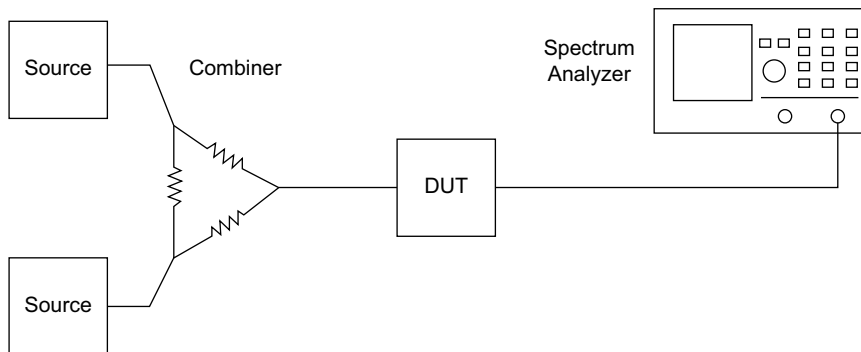


Figure 7-7 The outputs of two signal sources can be combined to a two-tone signal for intermodulation distortion tests.

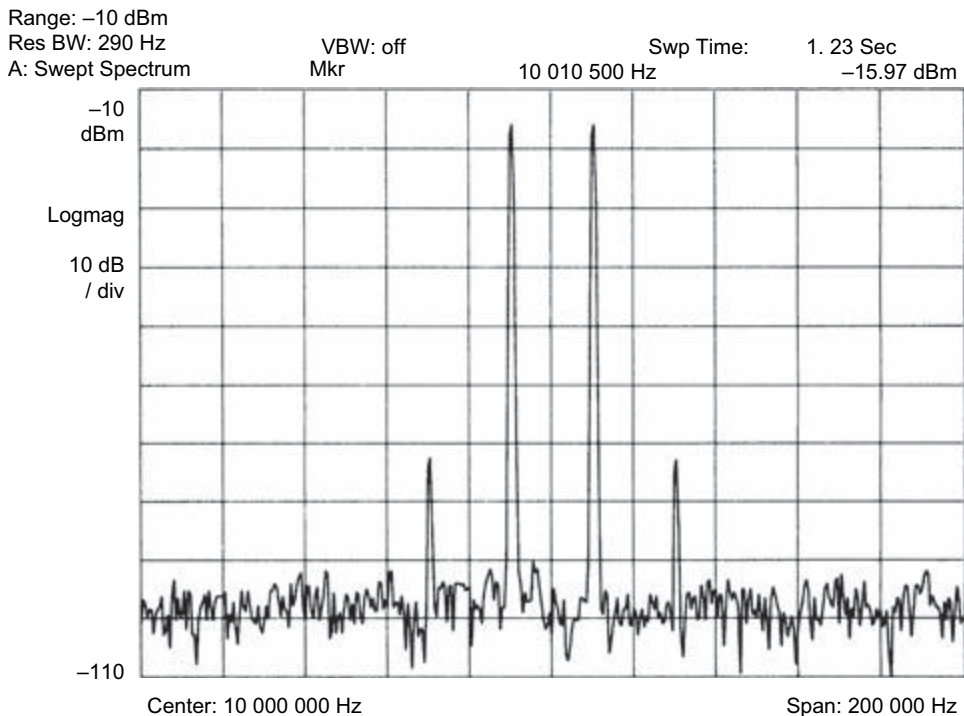


Figure 7-8 A typical two-tone intermodulation distortion measurement that measures the third-order products close to the original two tones.

source attempts to compensate for the other sources signal. Disabling automatic leveling can correct this problem.

It should be kept in mind that the two sine waves will combine to create a signal 6 dB larger than the individual tones (after accounting for the combiner loss). The device under test is often sensitive to the peak instantaneous voltage applied to it, and the user may inadvertently supply twice the desired peak input voltage.

7.9 Distortion Internal to the Analyzer

The preceding discussion was oriented toward understanding and measuring distortion in the device under test. However, the internal circuits of the analyzer are imperfect and will also produce distortion products. The distortion performance of the analyzer is specified by the manufacturer, either specified by a third-order intercept or lumped into a dynamic range specification. The instrument user can stretch the performance of the analyzer by understanding the nature of these distortion products.

As shown in this chapter, distortion products can be reduced in amplitude by reducing the signal level. Not only do the absolute levels of the distortion products decrease, they also decrease more than the decrease in signal level. So as the signal level decreases, the relative

distortion level also decreases, depending on the order of the distortion product. Higher-order distortion products decrease the fastest. This implies that the distortion products internal to the analyzer can be reduced by reducing the signal level into the analyzer.³ The internal input attenuators of the analyzer may be used or an external attenuator may be attached, improving the distortion measurement range of the analyzer. The most obvious disadvantage of reduced signal level is reduced signal-to-noise ratio. The user may find that the low-level distortion products are buried in the noise. Reducing the resolution bandwidth of the analyzer will reduce the measured noise, but at the expense of a slower sweep rate. See Section 17.3 for information on optimizing the dynamic range of a spectrum analyzer.

In some measurement situations, the amount of distortion is not of concern, and the signal level at the input of the analyzer can be increased to provide a better signal-to-noise ratio. For many measurements, the distortion products are known to occur at frequencies that are not of interest. For example, a narrowband measurement around the fundamental frequency of a sine wave will not be degraded by the presence of harmonic distortion since the harmonics will fall far away from the frequency range of interest. The instrument user must always be careful not to apply too large a signal to the input of an analyzer so that the damage level is not exceeded.

Bibliography

Bartz, Manfred. "Designing Effective Two-Tone Intermodulation Distortion Test Systems." *RF Design*, November 1987.

Hardy, James K. *High Frequency Circuit Design*. Reston, VA: Reston Publishing Company, Inc., 1979.

Hayward, W. H. *Introduction to Radio Frequency Design*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1982.

³ This may not be true of some spectrum analyzers that use digital IF sections (see Chapter 5). The analog-to-digital conversion process can introduce low-level distortion products that do not decrease in amplitude in response to decreasing signal level.

Noise and Noise Measurements

In frequency domain measurements, electronic noise shows up in two distinctly different ways. The first case is when the measurement is affected by the presence of unwanted noise, with noise being a nuisance. For example, we could be measuring the distortion of an amplifier with the amplifier's noise degrading the measurement. The second case occurs when the noise present in the system is the parameter to be measured. In that same amplifier, we may want to measure the noise at the output. Many of the same principles apply to both cases, but it is important to know whether the noise *is* the measurement or whether it *degrades* the measurement.

The electronic noise present in our measurements may come from the device under test (DUT) that is being measured or may be generated internally by the analyzer. In the general case, the analyzer internal noise must be significantly lower than the noise of the DUT. However, techniques that compensate for the noise in the analyzer can lower the measurement floor of the analyzer.

8.1 Statistical Nature of Random Noise

Many waveforms that we measure can be reliably characterized in the time domain. For instance, a sine wave can be completely described by its amplitude, frequency, and phase. Once we know these values, the instantaneous voltage of the waveform can be predicted for any arbitrary instant in time. Such a waveform is said to be *deterministic*. Noise, on the other hand, is often random in nature such that the instantaneous voltage cannot be predicted for arbitrary points in time.¹ Thus, random noise is *nondeterministic*.

Noise cannot be characterized in the time domain by simple parameters such as amplitude and phase since the voltage at any point in time is a random function. However, we can describe noise with a statistical approach by tabulating how often a certain voltage appears. In a continuous form, this results in the *probability density function* (PDF) of a random waveform. Figure 8-1 shows the probability density function of a particular waveform. The PDF shown happens to have a Gaussian shape, which is very common, but other PDF shapes

¹ The definition of noise is restricted in this chapter to include only truly random noise. Other noise processes exist which are not completely random.

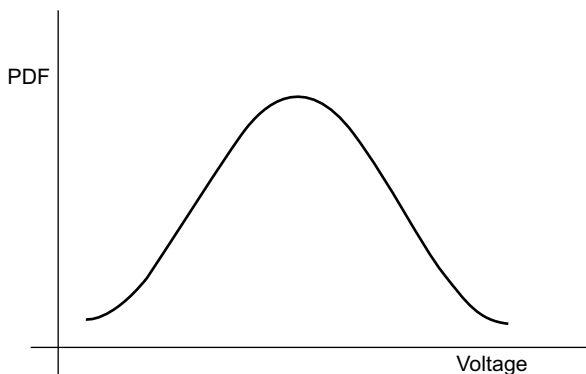


Figure 8-1 The probability density function shows the probability of a particular voltage occurring in a waveform.

are possible. The PDF does not define the shape of the time domain waveform, but tells us the probability of a certain voltage occurring.

8.2 Mean, Variance, and Standard Deviation

The statistical characteristics of a random waveform can be described by a few simple parameters. First, the waveform will have an average or *mean* value given by

$$\bar{x} = E(x) = \int_{-\infty}^{\infty} xp(x) dx \quad (8-1)$$

where

$E(x)$ = expected value of x

$p(x)$ = probability density function of $x(t)$

Of course, the mean value has a less mathematical and more intuitive definition of being the average value of the waveform.

With the mean value of the waveform defined, a measure of how much the voltage of the waveform varies is in order. The *variance* of a waveform is given by

$$\sigma^2 = E[(x - \bar{x})^2] = \overline{x^2} - (\bar{x})^2 \quad (8-2)$$

The variance is a measure of how far the instantaneous value of x strays from the mean value of x . If the variance is zero, the waveform is a DC level that never changes from its mean value. Closely related to the variance is the *standard deviation*, σ . Because the square of the standard deviation is equal to the variance, the two quantities are redundant. The variance is proportional to the power in the random waveform, while the standard deviation is proportional to the voltage.

8.3 Power Spectral Density

A random waveform can also be characterized in the frequency domain. One is tempted to simply compute the Fourier transform of the waveform, but this is not possible since the waveform is random and is not easily defined in terms of a time domain function. This problem is sidestepped mathematically by using the expected value of the Fourier transform of the random waveform.² A slightly modified form of frequency domain representation is produced, namely, the *power spectral density* (PSD). The PSD of a random signal is given by

$$S_x(f) = \lim_{T \rightarrow \infty} \frac{E[|X_T(f)|^2]}{2T} \quad (8-3)$$

where

$E(x)$ = expected value of x

$X_T(f)$ = Fourier transform of the random waveform, $x(t)$, evaluated over the time interval, $-T < t < T$

A less rigorous but more useful definition is that the PSD gives the density of power in a signal as a function of frequency. The power over a particular frequency range is given by

$$P_{12} = \int_{f_1}^{f_2} S_x(f) df \quad (8-4)$$

The total power in the signal is found by integrating over all frequencies:

$$P_T = \int_{-\infty}^{\infty} S_x(f) df = \overline{x^2(t)} \quad (8-5)$$

The power spectral density is a two-sided function, having values for both positive and negative frequencies. The PSD gives the power in the signal referenced to 1 Ω . That is, since no resistance is specified, $x(t)$ is interpreted as a voltage (or current) across (through) a 1 Ω resistor, with the power in the resistor equal to $x^2(t)$. Figure 8-2 shows the power spectral density of a particular random signal.

The spectrum or frequency domain representation of a signal has been discussed previously in this book. Here, the emphasis should be placed on the word *density* that appears in power spectral density. The frequency domain representation of noise does not have discrete spectral lines but instead is a continuous function of frequency that represents the density per unit frequency. The basic units of PSD are V^2/Hz , so to determine the voltage or power of a noise waveform the measurement bandwidth must be specified.

8.4 Frequency Distribution of Noise

In general, noise may have any arbitrary frequency content, resulting in a variety of possible PSD shapes. Noise that has equal power density at all frequencies is called *white noise*

² This concept is described in more detail in Chapter 6 of McGillem and Cooper (1974).

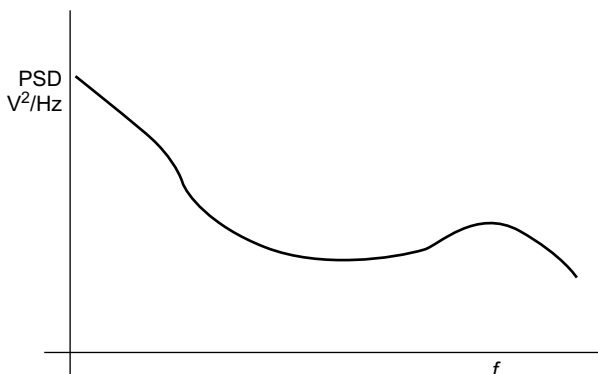


Figure 8-2 The power spectral density function shows the power density of a signal as a function of frequency.

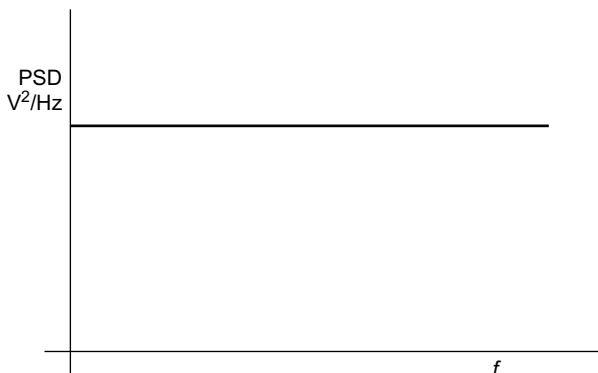


Figure 8-3 White noise has a power spectral density which is constant with frequency.

(Figure 8-3). A strict definition of white noise requires that the flat power density characteristic extend out for an infinite bandwidth. A more practical definition requires the noise to have a flat PSD over some frequency range.

Another common type of noise spectrum is $1/f$ noise (also called *flicker noise*), as shown in Figure 8-4. This type of noise spectrum is found in many physical systems, including electronic circuits. The contribution of $1/f$ noise is usually significant only at low frequency and becomes less important at higher frequencies. As the name implies, the amplitude of this type of noise is inversely proportional to frequency.³

Other PSD shapes are possible, since they may result from a combination of electronic noise sources. In addition, the noise PSD will be affected by the frequency response of the system. In many cases, we can consider the noise power density to be constant over a small frequency range, which simplifies the mathematical complexity involved.

³ This model has the annoying property that the noise density approaches infinity near 0 Hz.

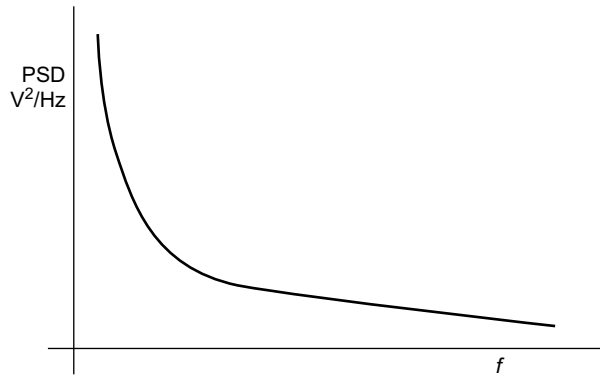


Figure 8-4 Another common type of noise is $1/f$ noise.

8.5 Equivalent Noise Bandwidth

One problem that presents itself in designing and understanding spectrum analyzers is how a filter with some arbitrary shape will respond to noise. More specifically, how much noise will be present at the output of a filter with a known power density of noise at its input? Consider the filter shape shown in Figure 8-5. The filter is a band-pass filter with a nominal gain of G_0 at its center frequency, f_0 . If the input noise is constant across the filter shape, the output noise power can be determined by integrating the gain of the filter:

$$P_N = N_0 \int_0^x G(f) df \tag{8-6}$$

where

- N_0 = power density of the input noise (V^2/Hz)
- $G(f)$ = power gain of the filter

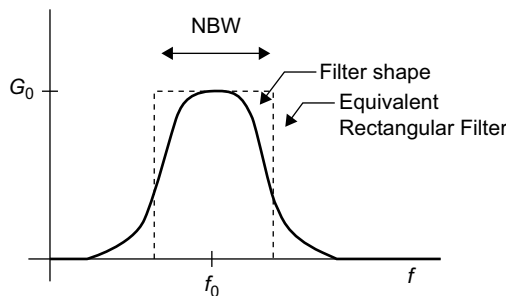


Figure 8-5 The equivalent noise bandwidth of a filter is defined by a rectangular filter that passes the same amount of white noise as the original filter.

Now suppose that an ideal rectangular filter with gain G_0 is used instead. What bandwidth will this filter need to be to produce the same noise power at the output? The output power with such a filter is given by

$$P_N = N_0 G_0 NBW \quad (8-7)$$

where

NBW = bandwidth of the ideal rectangular filter

Therefore,

$$NBW = \frac{1}{G_0} \int_0^{\infty} G(f) df \quad (8-8)$$

The rectangular filter bandwidth is called the *equivalent noise bandwidth* (NBW) of the filter (also called the *noise equivalent bandwidth*). If the equivalent noise bandwidth of a filter is known, the exact filter shape is not needed to perform noise calculations as long as the input noise is constant over the bandwidth of the filter. Note that this definition of bandwidth is *not* the same as some of the other classical definitions such as the 3 dB and 6 dB bandwidth.

Example 8.1

What is the equivalent noise bandwidth of a single-pole low-pass filter? The filter is shown in Figure 8-6 with cutoff frequency, f_c .

A single-pole low-pass filter has the voltage transfer function

$$H(f) = \frac{f_c}{f_c + jf}$$

Taking the magnitude of $H(f)$ and squaring to get the power gain gives

$$G(f) = |H(f)|^2 = \frac{f_c^2}{f_c^2 + f^2}$$

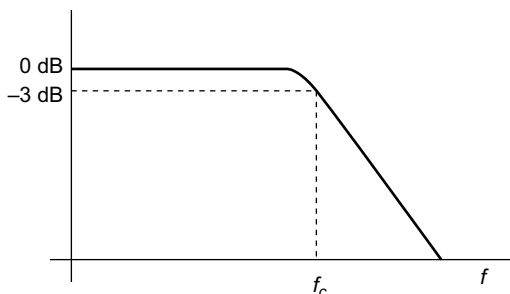


Figure 8-6 The transfer function of a single-pole low-pass filter.

The nominal gain $G_0 = 1$, and the equivalent noise bandwidth is

$$NBW = \int_0^{\infty} \frac{f_c^2}{f_c^2 + f^2} df = \frac{\pi}{2} f_c = 1.57 f_c$$

Therefore, the noise equivalent bandwidth is 1.57 times the 3-dB bandwidth. This relationship is valid only for a single-pole filter.

8.6 Noise Units and Decibel Relationships

As previously stated, the power spectral density of noise has the units of V^2/Hz . The V^2 implies that this is a measure of power. The PSD can be expressed as a voltage by taking the square root, with the units of $V/\sqrt{\text{Hz}}$. Spectrum analyzers may display the measured results this way, which is convenient when the user is working in terms of voltage.

Alternatively, it is often convenient to refer to the noise level as normalized to 1 Hz and expressed in decibel form. At this point, we will introduce the possibility of a resistance other than 1 Ω .

$$\text{Noise (dBm, 1 Hz)} = 10 \log[N_0/(Z_0 \times 1 \text{ mW})] \quad (8-9)$$

where

N_0 = power spectral density (in V^2/Hz)

Z_0 = impedance of the system

The 1 Hz noise level can be converted to other equivalent noise bandwidths, assuming that the noise density remains constant across all bandwidths of interest, using

$$\text{Noise (dBm)} = 10 \log[\text{NBW } N_0/(Z_0 \times 0.001)] \quad (8-10)$$

$$= 10 \log(\text{NBW}) + 10 \log[N_0/(Z_0 \times 0.001)] \quad (8-11)$$

The dBm (1 Hz) value is increased by $10 \log(\text{NBW})$ to obtain the new noise power adjusted to the new bandwidth. To go from one bandwidth directly to another, a correction factor can be computed:

$$K_{\text{dB}} = 10 \log(BW_2/BW_1) \quad (8-12)$$

To convert from BW_1 to BW_2 , add K_{dB} to the noise value associated with BW_1 . Note that the bandwidth is treated with a factor of 10 in decibel form similar to power (and not voltage). This is because noise power is proportional to bandwidth, given the foregoing assumptions. The reader is urged to be careful in applying these equations, making sure that the assumption of constant noise power density across the bandwidths of interest is true.

Example 8.2

Given a noise power density of $2 \times 10^{-12} V^2/\text{Hz}$ and a resistance of 50 Ω , what is the noise power present in a noise equivalent bandwidth of 1 kHz? What is the noise voltage (in the same bandwidth)? Express the noise level in dBm (1 Hz). Convert the dBm (1 Hz) value to a 1 kHz bandwidth.

The noise power in 1000 Hz is

$$\begin{aligned} P_N(1 \text{ kHz}) &= (1000 \text{ Hz}) (2 \times 10^{-12} \text{ V}^2/\text{Hz}) / (50 \Omega) \\ &= 40 \text{ pW} \end{aligned}$$

In terms of voltage,

$$\begin{aligned} V_N &= \sqrt{2 \times 10^{-12} \text{ V}^2/\text{Hz} \times 1000 \text{ Hz}} \\ &= 44.7 \mu\text{V} \end{aligned}$$

$$\begin{aligned} P_N(1 \text{ Hz}) &= 10 \log(2 \times 10^{-12} \text{ V}^2/\text{Hz} / (50 \Omega \times 0.001 \text{ mW})) \\ &= -104 \text{ dBm (1 Hz)} \end{aligned}$$

To convert to a 1 kHz bandwidth, add $10 \log(1000/1) = 30 \text{ dB}$:

$$P_N(1 \text{ kHz}) = -104 \text{ dBm(1 Hz)} + 30 \text{ dB} = -74 \text{ dBm}$$

8.7 Noise Measurement

Since the level of noise at the analyzer detector is affected by the resolution bandwidth (RBW), the noise level on the analyzer display depends on the RBW setting. Narrowing the RBW reduces the displayed noise level, and widening the bandwidth increases the noise level. In general, such a measurement may be uncalibrated due to the unknown noise equivalent bandwidth of the RBW filter and the unknown characteristics of the detector. Because spectrum analyzer RBW filters are designed to be swept quickly, the filter shape is not very steep. The noise equivalent bandwidth of an ideal Gaussian filter is 6.4% wider than the half-power bandwidth; modern spectrum analyzer filters are very close to Gaussian with a 5.5% wider behavior. Some older spectrum analyzers with analog filters have an NBW that is 11–12% wider than the half-power bandwidth.

The classic analog spectrum analyzer block diagram uses an amplifier with logarithmic gain followed by a detector to determine the measured value. These instruments are calibrated to detect sinusoids accurately and exhibit a significant error when measuring random noise. The log amplifier introduces error by compressing the noise peaks and expanding the smaller noise values downward. For a particular detector–amplifier combination, correction factors to account for the error in random noise measurements are known. For a spectrum analyzer with an envelope detector and a logarithmic amplifier before the detector, the error correction factor is 2.5 dB. That is, the spectrum analyzer trace will show noise as 2.5 dB lower than the actual value. The correctly calibrated noise reading is given by

$$\text{Noise (dBm, 1 Hz)} = \text{spectrum analyzer reading (dBm)} + K_{\text{det}} - 10 \log(\text{NBW}) \quad (8-13)$$

where

$$K_{\text{det}} = \text{error correction factor for the detector and log amp}$$

Because the measurement is inherently noisy, it is often desirable to use video filtering or averaging to smooth out the noise reading.

If relative (not absolute) noise measurements are required, the correction factor can be eliminated. Also, except for the widest ones, the ratio of the noise equivalent bandwidth to

the 3 dB bandwidth of the RBW filter is generally constant. Thus, a factor of two change in RBW implies the NBW also changes by a factor of two.

8.8 Automatic Noise Level Measurement

The level of random noise can be measured with most any spectrum analyzer, but the calculations required to produce a calibrated measurement may be difficult. The noise equivalent bandwidth of the analyzer RBW filter and the characteristics of the detector must be known, and the measured result must be adjusted accordingly. Modern spectrum analyzers provide an automatic and calibrated means of measuring the spectral density of random noise, often implemented as a special trace marker feature called *noise marker*. When this feature is invoked, the analyzer combines many neighboring trace elements and averages them together to reduce the variation in the noise measurement. The result is automatically corrected for the noise equivalent bandwidth of the RBW filter, the logarithmic intermediate frequency (IF) gain, and the characteristics of the detector. Finally, the measurement is normalized to a 1 Hz bandwidth. Modern spectrum analyzers with digital IFs can be configured to average signal envelopes on a power scale even while displaying them on a log (decibel) scale and thus avoid the need for the 2.5 dB correction factor.⁴

The noise measurement can be performed using any RBW, and the analyzer will still normalize the result to a 1 Hz bandwidth. The user must be careful not to pick too wide a bandwidth since the noise is assumed to be white over the RBW. The analyzer cannot discern a noise spectral density shape that varies over its RBW. Also, any discrete spectral lines that fall inside the RBW will be lumped into the measurement and treated as noise in the calculations, introducing an error into the noise measurement.

8.9 Noise Floor

To the user of an analyzer, the noise present in a measurement (either from the circuit or signal under test or the internal noise of the analyzer) shows up as a noise trace on the analyzer display. The noise is usually relatively constant with frequency but may be worse at certain frequencies (particularly low frequencies due to $1/f$ noise). Although the use of a narrower RBW forces the measured noise power to be lower, the spectral density of the noise remains unchanged.

When measuring spectral lines, the instrument user may narrow the RBW as needed to lower the noise level. Thus, small signals can still be measured in the presence of noise. The situation is different when measuring random noise—the internal noise of the analyzer must be lower than the noise being measured. Reducing the RBW does not help since it will reduce the noise being measured along with the internal noise of the analyzer. Again, in terms of spectral density, the analyzer noise must be less than the noise being measured.

Obviously, reducing the bandwidth of the RBW filter reduces the amount of noise power in the measurement system. A narrow RBW filter will remove as much of the noise as

⁴ For a more detailed look at these issues, see Agilent Technologies (2012).

Table 8-1 Error Due to Analyzer Internal Noise

Measured Noise Level in dB Relative to Internal Noise	Error in Measured Noise, dB
20	0.04
15	0.14
10	0.46
9	0.58
8	0.75
7	0.97
6	1.26
5	1.65
4	2.20
3	3.02
2	4.33
1	6.87

possible from the measurement. However, narrow RBW filters slow down the measurement, so the trade-off is increased measurement time.

8.10 Correction for Noise Floor

If the measured noise is much larger than the internal noise of the analyzer, no significant error will be introduced. However, as the external noise level approaches the internal noise level, the measurement will be in error. Table 8-1 summarizes this effect. The left column corresponds to the noise level as measured by the analyzer, whereas the right column indicates the amount of error in that measurement due to the analyzer internal noise. The error is always positive (i.e., the measured value is larger than the actual noise). To obtain the actual noise level, the error (in dB) should be subtracted from the measured value. Note that even with measured noise levels as large as 10 dB above the analyzer noise floor, an error of 0.46 dB is introduced. When the actual noise equals the analyzer noise, the measurement will appear to be 3 dB above the noise floor (with an error of 3 dB).

Example 8.3

A spectrum analyzer with a noise floor of -140 dBm (1 Hz) shows a measured value of -135 dBm (1 Hz). What is the actual noise level being measured?

The measured value is 5 dB above the internal noise floor. From Table 8-1, this produces an error of 1.65 dB. The actual noise level being measured is

$$-135 \text{ dBm} - 1.65 \text{ dB} = -136.65 \text{ dBm (1 Hz)}$$

Some modern spectrum analyzers include a special function called *noise floor extension*, which automatically performs a calculation to remove the analyzer noise from the measurement.

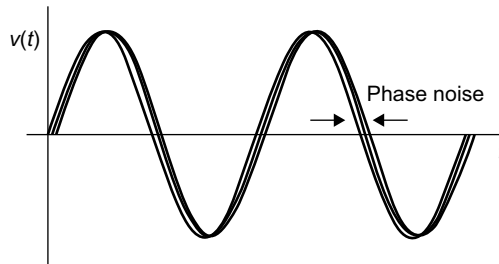


Figure 8-7 In the time domain, phase noise causes a jitter in the zero crossings of a waveform.

Correcting for the noise floor can be a complex algorithm to achieve maximum improvement.⁵ The actual noise improvement depends on the specific measurement setup but may result in up to 12 dB of noise reduction.

8.11 Phase Noise

Phase noise is an important measure of the spectral purity of a sine wave, often associated with synthesized (phase-locked) oscillators. In the time domain, phase noise is exhibited as a jitter in the zero crossings of the waveform (Figure 8-7). For a high-quality oscillator design, the phase noise will usually not be discernible in the time domain. In the frequency domain, the phase noise shows up as noise sidebands on the carrier (Figure 8-8).

A pure sine wave can be represented by

$$v(t) = V_0 \sin 2\pi f_0 t \quad (8-14)$$

where

V_0 = zero-to-peak amplitude

f_0 = carrier frequency

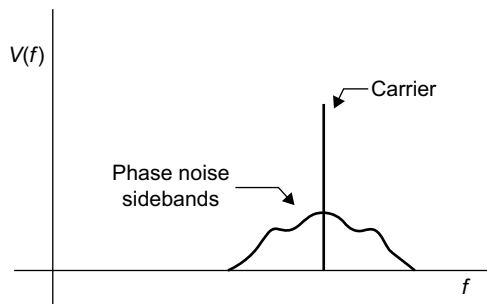


Figure 8-8 In the frequency domain, phase noise appears as noise sidebands on both sides of the carrier.

⁵ See Agilent Technologies (2010) for a detailed explanation of how *noise floor extension* is implemented in a high-performance spectrum analyzer.

A sine wave that exhibits both amplitude and frequency fluctuations is given by

$$v(t) = [V_0 + a(t)] \sin [2\pi f_0 t + \phi(t)] \quad (8-15)$$

where

$a(t)$ = amplitude noise

$\phi(t)$ = phase noise

Notice that this noise process resembles the amplitude and angle modulation processes, but with the modulation “source” being random noise mechanisms in the system. The amplitude noise may be significant, even in a high-quality oscillator design. However, in many systems, the amplitude noise may be removed when the signal passes through an amplitude limiting device such as a mixer.

Phase noise in the frequency domain can be expressed as

$$\mathcal{L}(f) = \frac{V_N(1 \text{ Hz BW})}{V_c} \quad (8-16)$$

where

$V_N(1 \text{ Hz BW})$ = RMS noise level in a 1 Hz bandwidth at a frequency f Hz
away from the carrier

V_c = RMS voltage of the carrier

$\mathcal{L}(f)$ is often expressed in terms of decibels,

$$\mathcal{L}(f)_{\text{dBc}} = 20 \log[\mathcal{L}(f)] \quad (8-17)$$

The resulting plot of $\mathcal{L}(f)$ shows the phase noise level relative to the carrier as a function of frequency away from the carrier (Figure 8-9). If the phase noise sidebands are within the measurement range of the spectrum analyzer, $\mathcal{L}(f)$ can be measured directly. On most analyzers, a marker function can be used to read the noise level at an offset, often expressed relative to the carrier amplitude.

The phase term, $\phi(t)$, could include both long-term and short-term phase or frequency fluctuations. Long-term effects are usually specified in terms of frequency drift while the short-term effects are characterized as phase noise. As the frequency offset, f , approaches

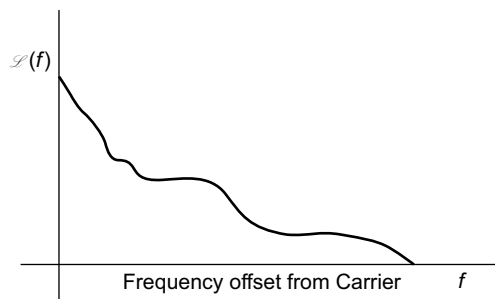


Figure 8-9 Phase noise is usually plotted as a function of frequency away from the carrier.

zero the period of the frequency offset gets arbitrarily large. For example, a frequency offset of 0.001 Hz corresponds to a period of 17 min, which is usually categorized as frequency drift. Our ability to measure the noise at such a small frequency offset becomes increasingly difficult. The resolving power of the analyzer must be fine enough to reject the large carrier power while still measuring the phase noise.

For the spectrum analyzer to measure the phase noise of a sine wave directly, a few conditions must be met. The noise floor of the spectrum analyzer must be significantly lower than the measured phase noise. A more subtle phenomenon occurs due to the spectrum analyzer local oscillator purity. Recall from Chapter 5 that the analyzer local oscillator mixes with the input signal to produce a new signal at the analyzer IF. We usually think of this process as mixing the input signal down to the IF. Certainly, the phase noise of the input signal will also be mixed down, but any phase noise present on the local oscillator will also appear at the IF. This means that a close-in measurement of phase noise on an input signal is actually a measurement of the combined phase noise of the input signal and the local oscillator. If the input signal were completely free of phase noise, it could be used to measure the internal phase noise of the spectrum analyzer. To obtain a precise measurement of an input signal phase noise, the analyzer local oscillator must have lower phase noise than the signal being measured. Figure 8-10 shows the phase noise of an oscillator as directly measured by a spectrum analyzer.

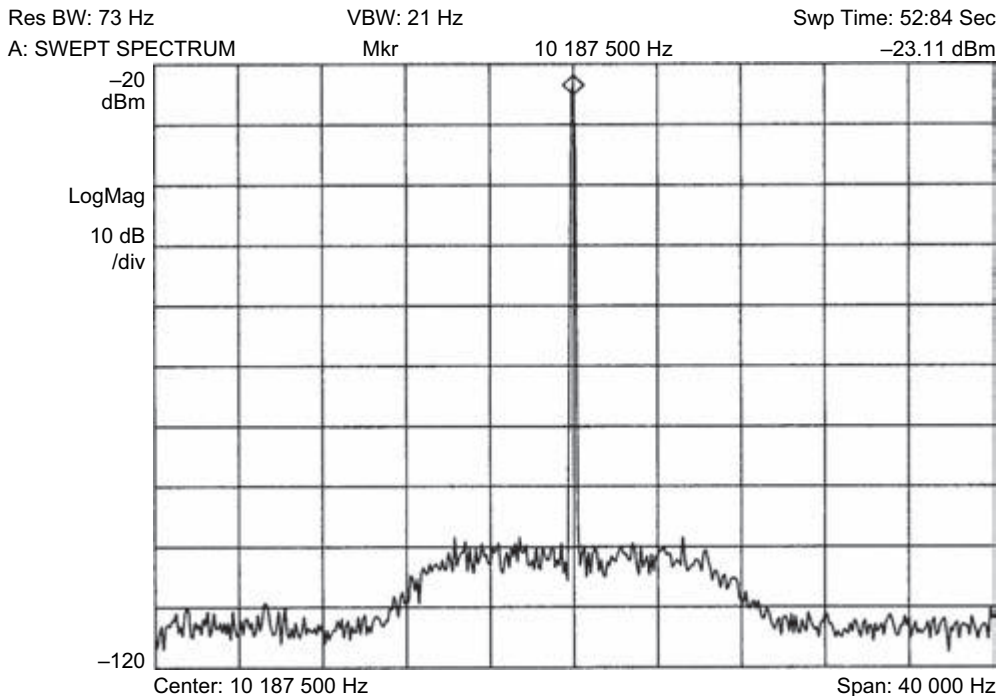


Figure 8-10 A spectrum analyzer measurement showing the close-in phase noise of an oscillator that appears as a noise pedestal.



Figure 8-11 A spectrum analyzer measurement showing the single-sideband phase noise characteristics of a signal. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

The most common way to display the phase noise of a signal is to show only one side of the spectrum plotted versus frequency offset from the carrier. Figure 8-11 shows a spectrum analyzer measurement using a software application specifically designed for phase noise measurements. The frequency offset is shown using a log frequency plot so that a wide range of offsets can be shown on one chart while still showing the close-in phase noise performance.

Using a spectrum analyzer to measure phase noise is referred to as the *direct spectrum* technique. While we have just shown that the phase noise of an oscillator can sometimes be measured directly with a spectrum analyzer, in many cases the internal phase noise and

broadband noise floor of the spectrum analyzer is not good enough to perform the measurement accurately. More advanced methods for phase noise measurement include multiple *phase detector* techniques and the *two-channel cross-correlation* technique.⁶

Bibliography

Agilent Technologies. “Agilent’s Phase Noise Measurement Solutions,” Selection Guide, Publication Number 5990-5729EN, September 2011.

Agilent Technologies. “Spectrum and Signal Analyzer Measurements and Noise,” Application Note, Publication Number 5966-4008E, May 2012.

Agilent Technologies. “Using Noise Floor Extension in the PXA Signal Analyzer,” Application Note, Publication Number 5990-5340EN, February 2010.

Cooper, George R., and McGillem, Clare D. *Probabilistic Methods of Signal and System Analysis*. New York: Holt, Rinehart and Winston, Inc., 1971.

Engelson, Morris. *Modern Spectrum Analyzer Theory and Applications*. Dedham, MA: Artech House, 1984.

Feher, Kamilo. *Telecommunications Measurements, Analysis, and Instrumentation*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1987.

McGillem, Clare D., and George R. Cooper. *Continuous and Discrete Signal and System Analysis*. New York: Holt, Rinehart and Winston, Inc., 1974.

Oliver, Bernard M., and Cage, John M. *Electronic Measurements and Instrumentation*. New York: McGraw-Hill Book Company, 1971.

Pettai, Raoul. *Noise in Receiving Systems*. New York: John Wiley & Sons, Inc., 1984.

Ziemer, R. E., and Tranter, W. H. *Principles of Communications*. Boston: Houghton Mifflin Company, 1976.

⁶ For a description of advanced phase noise measurements, see Agilent Technologies (2011).

Pulse Measurements

Pulsed waveforms are an important class of signals in systems such as radar and digital radio. Pulsed signals can present a more difficult measurement problem than continuous waveforms. The resolution bandwidth used in a measurement can affect the displayed spectrum. With a small-resolution bandwidth, the displayed spectrum has discrete spectral lines, but with wider wide-resolution bandwidths these line spectra are smeared together and the spectrum appears to be continuous.

The principles associated with the pulsed waveform are also applicable to pulsed radio frequency signals. The envelope of the spectrum is the same and depends on the pulse width, but the spectrum is centered on the radio carrier frequency.

9.1 Spectrum of a Pulsed Waveform

As shown in Chapter 3, the Fourier transform of a single pulse has a $(\sin x)/x$ shape (Figure 9-1):

$$V(f) = \tau \frac{\sin[2\pi f(\tau/2)]}{2\pi f(\tau/2)} \quad (9-1)$$

The nulls of the spectrum occur at multiples of $1/\tau$. The amplitude of the spectrum is proportional to the pulse width—the wider the pulse, the more energy present in the signal.

The classic swept spectrum analyzer is not capable of measuring a transient event such as a single pulse. However, a fast Fourier transform (FFT) spectrum analyzer can produce the spectrum of such a signal as long as it is within the bandwidth of the analyzer.

A pulse train is produced by repeating the pulse periodically (Figure 9-2a). Since the waveform is periodic, it can be expanded into a Fourier series to determine the harmonic content of the waveform. As listed in Table 3-1, the Fourier series for this waveform is

$$x(t) = \frac{\tau}{T} + \frac{2\tau}{T} \sum_{n=1}^{\infty} \frac{\sin(\pi n\tau/T)}{(\pi n\tau/T)} \cos(2\pi n\tau/T) \quad (9-2)$$

The waveform has a DC component of τ/T , which is just the average value of the waveform. The harmonics of the signal will fall at multiples of the waveform frequency, which is $1/T$ (Figure 9-2b). The period of the waveform is also known as the *pulse repetition*

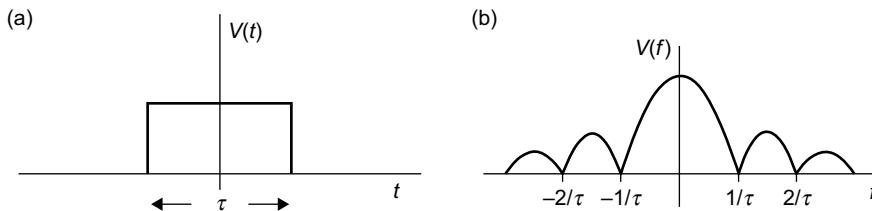


Figure 9-1 (a) A pulse in the time domain. (b) The corresponding spectrum in the frequency domain.

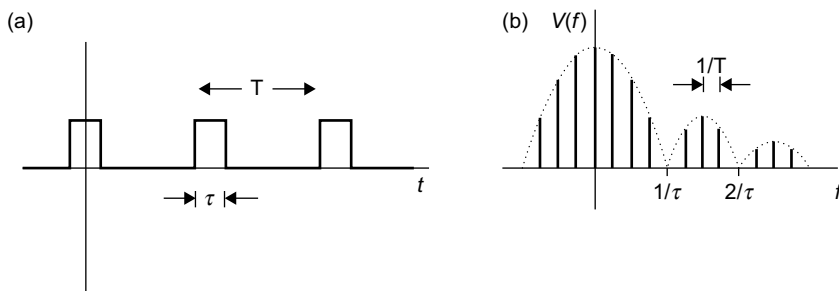


Figure 9-2 (a) A repetitive pulse train in the time domain. (b) The pulse train in the frequency domain.

frequency (PRF). The overall shape or envelope of the harmonics takes on the $(\sin x)/x$ characteristic, which is the same shape as the Fourier transform of the single pulse. As shown in Figure 9-2b, the envelope of the spectrum has nulls at integer multiples of $1/\tau$.

The amplitude of the spectrum of the pulse train is proportional to the *duty cycle* of the waveform, which is the ratio of the pulse width to the waveform period. The greater amount of time that the pulse is at its high voltage, the greater the power in the waveform.

$$\text{duty cycle} = \frac{\tau}{T} \tag{9-3}$$

The overall shape of the spectrum is determined by the width of the pulse, while the PRF determines the spacing of the spectral lines. Figure 9-3 illustrates the phenomenon. In Figure 9-3a, with $\tau/T = 1/4$, the spectral lines are widely spaced. If the PRF is decreased with τ remaining constant (Figure 9-3b), the spectral lines move closer together while the shape of the spectrum remains the same. Note that the amplitude of the spectrum decreases, consistent with the decrease in duty cycle of the waveform. (A factor of 2 decrease in duty cycle corresponds to a factor of 2 decrease in the amplitude of the spectrum.)

If the PRF is made very small, the spectral lines get very close together and begin to approximate a continuous spectrum (Figure 9-3c). In reality, the spectral lines are always distinct for repetitive waveforms, but as the spacing between the harmonics gets smaller than the resolution bandwidth of the spectrum analyzer the spectrum will appear to be continuous. The amplitude of the spectrum continues to be proportional to the duty cycle of the waveform. Note that as the PRF approaches zero, corresponding to the waveform period

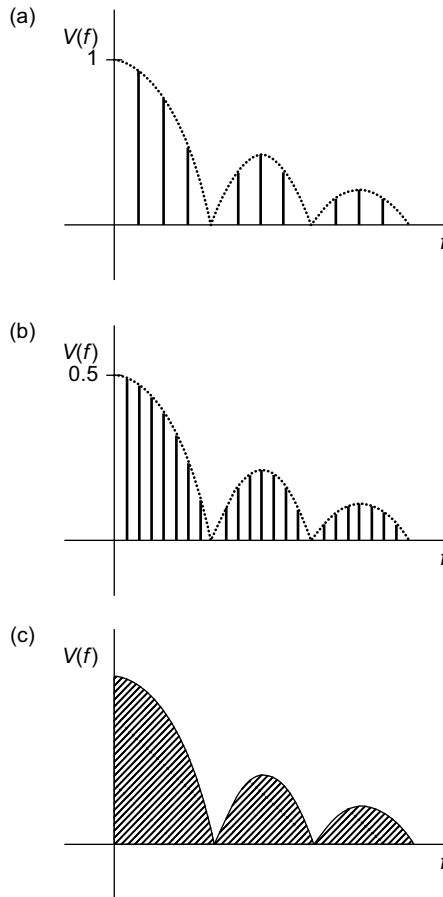


Figure 9-3 The PRF determines the spacing of the spectral lines within the $(\sin x)/x$ envelope. (a) Widely spaced spectral lines (high PRF). (b) Closely spaced spectral lines (moderate PRF). (c) Continuous spectrum (low PRF).

approaching infinity, the time domain and frequency domain representations of the signal revert back to being those of a single pulse.

9.2 Effective Pulse Width

Many pulsed waveforms are not exactly the ideal shape as shown in Figure 9-2a. An example of such a waveform is pictured in Figure 9-4. For this type of waveform we define an *effective pulse width* to be used in the calculations relating to frequency spectrum.

$$\tau_{\text{eff}} = \frac{1}{V_{\text{max}}} \int_{-T/2}^{T/2} v(t) dt \quad (9-4)$$

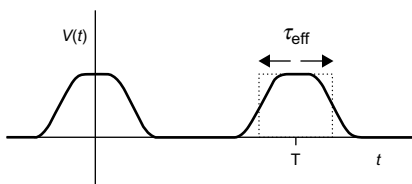


Figure 9-4 An effective pulse width can be determined for nonideal pulse shapes.

The effective pulse width is the width of an ideal rectangular pulse, which would have the same maximum voltage and energy as the original pulse.

9.3 Line Spectrum

When the resolution bandwidth of the spectrum analyzer is narrow enough, each of the spectral lines will be shown distinctly on the display. Although not a hard limit, the general requirement for a *line spectrum* display of a pulsed waveform is

$$RBW < 0.3 PRF \quad (9-5)$$

With the resolution bandwidth narrow enough to resolve the individual spectral lines, the spectrum measurement is fairly conventional, with the display being a close representation of the signal's spectrum. (Compare this with the pulse spectrum case discussed later.) Changing the measurement span widens or narrows the displayed spectrum as appropriate and changing the sweep time does not affect the shape of the spectrum.¹

Example 9.1

A pulse waveform has a period of 10 μsec and a duty cycle of 10%. What is the maximum resolution bandwidth that will cause a line spectrum to be displayed?

The waveform has a period of 10 μsec , so

$$PRF = 1/T = 1/10 \mu\text{sec} = 100 \text{ kHz}$$

The maximum resolution bandwidth is determined by

$$RBW < 0.3PRF = (0.3)(100 \text{ kHz}) = 30 \text{ kHz}$$

A typical spectrum analyzer measurement resulting in a line spectrum is shown in Figure 9-5.

¹ As long as the sweep limitations of the resolution bandwidth are not violated.

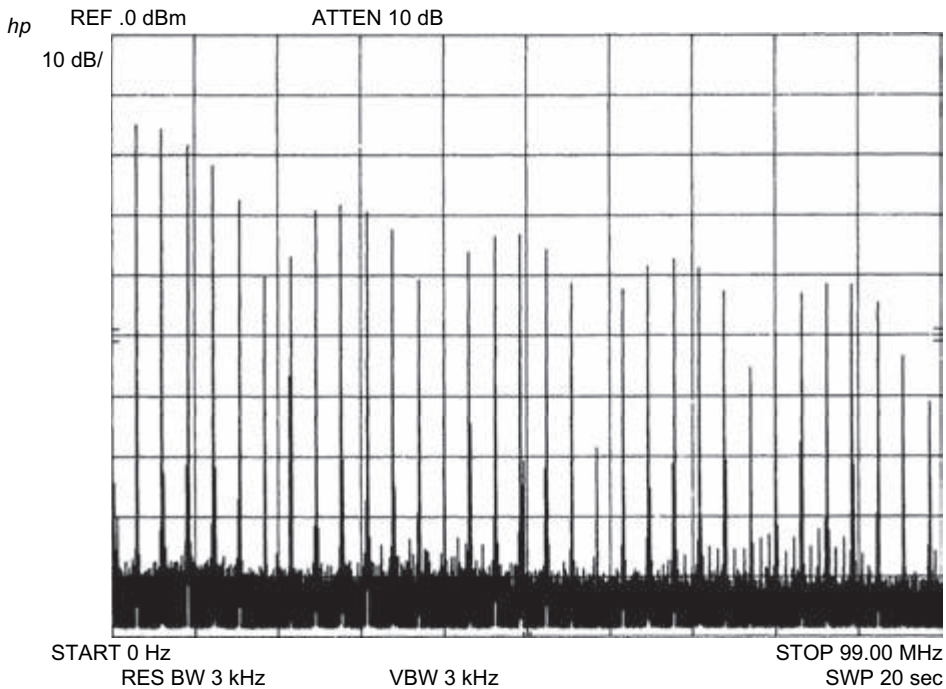


Figure 9-5 A line spectrum measurement of a pulsed waveform.

Example 9.2

Estimate the pulse repetition frequency and the effective pulse width of the signal shown in Figure 9-5.

The spectral lines are spaced approximately every 3 MHz, so $PRF = 3$ MHz. The first null of the spectrum envelope occurs on the sixth spectral line from the left, which is approximately 18 MHz.

$$\tau = 1/18 \text{ MHz} = 56 \text{ nsec}$$

9.4 Pulse Spectrum

It may seem desirable to always measure a pulsed waveform using spectrum analyzer settings that cause a line spectrum to be displayed. However, such a display may not always be possible or desirable. When the PRF is very small, the spacing of the spectral lines is also small, and a suitably small resolution bandwidth may not be available. Even if such a bandwidth setting is available, the required sweep time may result in an unacceptably slow measurement.

Supposing that the required bandwidth is available and the sweep time is not prohibitive, the user may still not want to view the individual spectral lines of the pulsed spectrum. Often the user is interested in the spectrum associated with the pulse and not the PRF. In such a

case, viewing the individual spectral lines is an unnecessary inconvenience. By using a wide resolution bandwidth, the envelope of the pulsed waveform's spectrum can be shown without revealing the details of the individual spectral lines. This type of spectrum display is called a *pulse spectrum*. The requirement for a pulse spectrum type of display is $RBW > 1.7PRF$. (Again, this is not a hard limit but a rule of thumb.) With a bandwidth significantly wider than the PRF, more than one spectral line will be inside the measurement bandwidth at one time. The wider the bandwidth, the more spectral lines are included in the measurement and the measured amplitude of the pulse spectrum is larger. Increasing the bandwidth by a factor of 2 roughly doubles the number of spectral lines included in the measurement, causing a 6 dB increase in displayed amplitude. Thus, the measured amplitude depends on the resolution bandwidth.

The previous statement should cause some concern on the part of the reader! Having the measured amplitude be a strong function of the resolution bandwidth is not a desirable feature.² One might expect this type of behavior with random noise, but not when the signal has discrete spectral lines. On closer examination, we see that the case is similar to random noise when the spectral lines are very closely spaced. The wider the bandwidth, the more "noise" (spectral lines) is let in. This type of signal is sometimes categorized as "impulse noise," which implies a large number of closely spaced spectral lines.

The resolution bandwidth must not be too large; otherwise, the envelope of the pulsed spectrum may become washed out. The resolution bandwidth must remain small compared with $1/\tau$, which defines the spacing of the nulls in the spectrum envelope. To summarize both constraints, the resolution bandwidth must be larger than the PRF but significantly smaller than $1/\tau$.

$$1.7PRF < RBW < 0.1/\tau \quad (9-6)$$

With a swept spectrum analyzer, the sweep time can interact with the PRF to produce discrete spectral lines. If the sweep time is set much greater than the period of the pulse train, the pulse spectrum is continuous. With faster sweep times, the on/off rate of the pulse train can show up as spectral lines. For example, if the sweep time is 100 msec and the pulse waveform repeats every 5 msec, spectral lines will occur at every 5 msec during the sweep, which corresponds to every 1/20 of the frequency span. If the frequency span is 200 MHz wide, these spectral lines would appear every 10 MHz. Changing the sweep time to 50 msec will cause the responses to appear at every 1/10 of the frequency span, which would appear to be spaced every 20 MHz. Clearly, both cases are misleading as to the actual spectral content of the signal. When individual spectral lines are visible on the display, they no longer represent the actual spacing of the harmonics in frequency but instead occur every $1/PRF$ seconds during the analyzer's sweep. Increasing the sweep time to much greater than $1/PRF$ will eliminate this effect and cause the spectrum to appear as a continuous $(\sin x)/x$ function. A useful guideline is

$$\text{sweep time} \geq \frac{100}{PRF} \quad (9-7)$$

which will produce at least 100 spectral lines in the spectrum.

² Unless the user likes the freedom of adjusting the measuring instrument until the desired result appears.

Example 9.3

A pulse waveform has a period of 1 msec and a pulse width of 500 nsec. Determine the limitations on sweep time and resolution bandwidth for a pulse spectrum measurement.

$$PRF = 1/(1 \text{ msec}) = 1 \text{ kHz}, \tau = 500 \text{ nsec}$$

$$1.7PRF < RBW < 0.1/\tau$$

$$1.7(1 \text{ kHz}) < RBW < 0.1/(500 \text{ nsec})$$

$$1.7 \text{ kHz} < RBW < 200 \text{ kHz}$$

$$\text{sweep time} \geq \frac{100}{PRF} = 100/1 \text{ kHz} = 100 \text{ msec}$$

Figure 9-6 shows a typical measurement of a pulsed waveform resulting in a pulse spectrum.

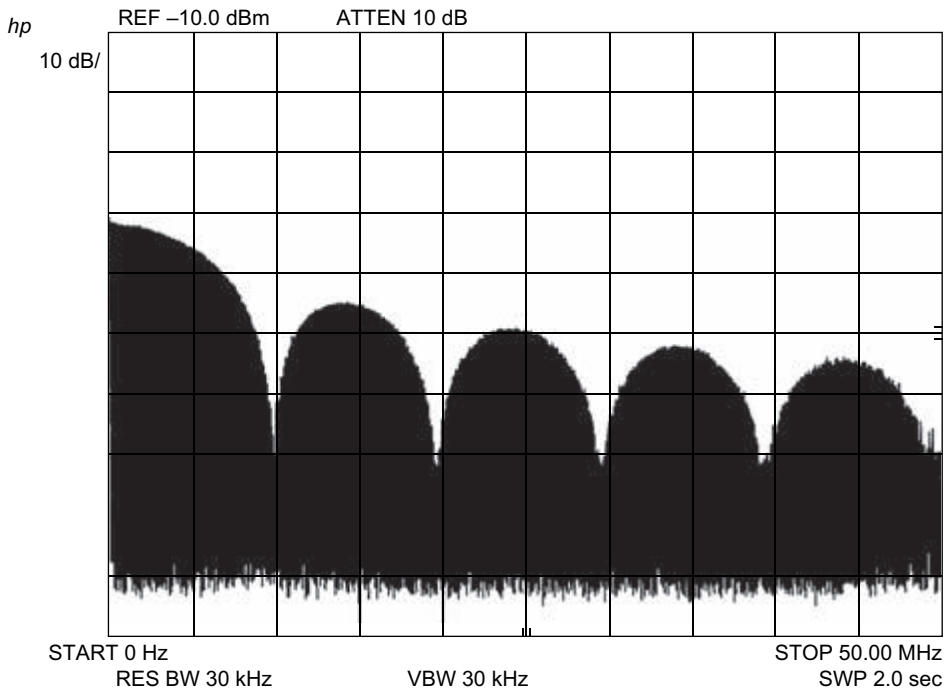


Figure 9-6 A pulse spectrum measurement of a pulsed waveform.

Example 9.4

Determine the effective pulse width of the signal shown in Figure 9-6.

The first null of the pulse spectrum occurs at approximately 10 MHz, so the effective pulse width is $\tau = 1/10 \text{ MHz} = 100 \text{ nsec}$.

9.5 Pulsed RF

Another signal that is closely related to the pulse train is the *pulsed sinusoid* or *pulsed RF* (Figure 9-7a). This type of signal can be derived by pulse modulating a radio frequency carrier (i.e., using the pulse train to turn the carrier on and off.) Radar signals are one common example of pulsed RF. The modulation property from Table 3-3 can be used to derive the pulsed RF spectrum from our previous results. The modulation property is described by the following transform pair:

$$x(t) \cos(2\pi f_0 t) \leftrightarrow 1/2[X(f - f_0) + X(f + f_0)] \quad (9-8)$$

$X(f)$, the spectrum of the modulating signal $x(t)$, appears centered on the carrier frequency, f_0 . Since the two-sided transform is used, the modulating spectrum occurs at $\pm f_0$. In the case of pulsed RF, the modulating signal is the pulse train, so the $(\sin x)/x$ shaped spectrum is no longer centered on the origin but is centered on the carrier frequency (Figure 9-7b).

Since the pulsed RF case can be directly related to the baseband pulse train, the principles derived relative to the pulse train are also valid for pulsed RF. For example, when measuring a pulsed RF signal with a spectrum analyzer, the display may show discrete spectral lines or may show a continuous pulse spectrum, depending on the pulse repetition frequency.

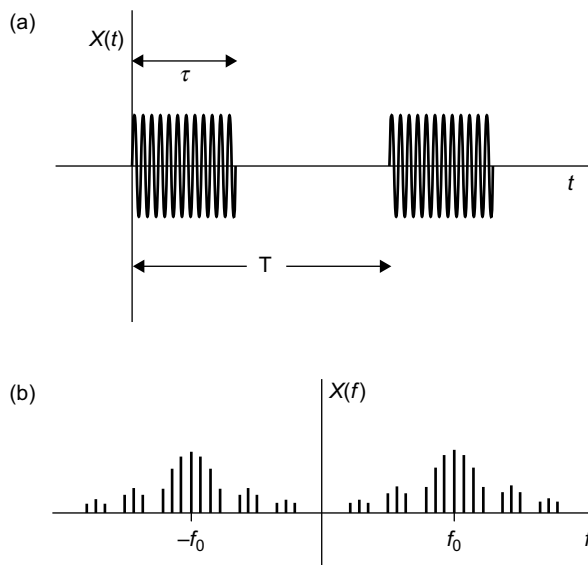


Figure 9-7 (a) A pulsed RF signal in the time domain. (b) The frequency spectrum of a pulsed RF signal.

9.6 Pulse Desensitization

The root mean square (RMS) value of a pulsed RF signal is proportional to the duty cycle of the waveform. The extreme cases are when the RF carrier is left on all of the time and when the RF carrier is always off. In between, the RMS value depends on the duty cycle of the waveform, τ/T . The RMS value of the pulsed RF waveform is given by

$$V_{\text{RMS}} = V_C \tau/T = V_C \tau PRF \quad (9-9)$$

where

V_C = RMS voltage of a constant carrier

Equation (9-9) can be expressed in decibel form as

$$V_{\text{RMS(dB)}} = V_{C(\text{dB})} + 20 \log(\tau PRF) \quad (9-10)$$

It is customary to define the *pulse desensitization factor*

$$\alpha_L = 20 \log(\tau PRF) \quad (9-11)$$

which represents the difference in amplitude between the continuous carrier and the pulsed RF signal (in decibels). This equation is valid only for a line spectrum—the pulse spectrum case will be discussed shortly. The term *desensitization* may be a poor choice since it may imply a loss of sensitivity in the measuring instrument. The instrument is not really any less sensitive—the average power in the waveform decreases, which should be reflected in a measurement of it. The range (or attenuator setting) of the spectrum analyzer should be set according to the power level of the continuous carrier. Otherwise, the peak power in the signal may overload the input circuitry of the analyzer. For small duty cycle signals, the measured amplitude will be much smaller than the peak signal power, forcing the measured response to be considerably lower than the full-scale analyzer response. This effect cuts into the dynamic range of the analyzer that is available for making the measurement—hence the term *pulse desensitization*.

Example 9.5

A pulsed RF signal has a peak power of -10 dBm, a PRF of 1 kHz, and pulse width of 10 μsec . What is the amplitude of the main lobe of the spectrum? If the spectrum is to be measured with 40 dB of dynamic range, what is the required dynamic range of the spectrum analyzer (assuming that full scale on the spectrum analyzer corresponds to -10 dBm)?

The power of a continuous carrier is -10 dBm.

$$\begin{aligned} V_{\text{RMS(dB)}} &= V_{C(\text{dB})} + 20 \log(\tau PRF) \\ &= -10 \text{ dBm} + 20 \log(10 \mu\text{sec} \cdot 1 \text{ kHz}) \\ &= -10 \text{ dBm} - 40 \text{ dB} = -50 \text{ dBm} \end{aligned}$$

The amplitude of the main lobe is -50 dBm.

Another 40 dB below the main lobe is -90 dBm. Therefore, the spectrum analyzer must measure -90 dBm with full scale equal to -10 dBm, which requires 80 dB of dynamic range.

In the case of the pulse spectrum, the situation is different. In addition to the pulse width, the measured amplitude also depends on the resolution bandwidth of the spectrum analyzer. The pulse desensitization factor for the pulse spectrum case is defined as

$$\alpha_p = 20 \log(\tau IBW) \quad (9-12)$$

The resolution bandwidth of a spectrum analyzer is normally specified in terms of a 3 dB bandwidth, which is useful in most cases but is not appropriate when analyzing pulsed signals. Instead, we introduce a new concept of bandwidth called the *effective impulse bandwidth* (IBW). This bandwidth is the bandwidth of an ideal rectangular filter that has a pulse response equivalent to the actual resolution bandwidth filter. Alternatively, we can say that the typical RBW filter behaves like it is wider than the normal 3 dB bandwidth when the input signal is a pulse. Mathematically, we can state

$$IBW = k \times RBW \quad (9-13)$$

where

k = factor relating the resolution bandwidth and the impulse bandwidth; typical value is ~ 1.5 , assuming a typical synchronously tuned RBW filter

The pulse desensitization factor can also be expressed as

$$\alpha_p = 20 \log(\tau k RBW) \quad (9-14)$$

For a pulse spectrum, reduced τ or IBW decreases the amplitude of the measured response. Reduced τ causes the average power in the signal to decrease, while reduced IBW leaves the average power in the signal unchanged but reduces the amount of energy present in the analyzer's RBW. Either way, the measurement is desensitized in that the measured reading will be lower.

For more information on pulsed RF measurements, see Agilent Technologies (2012).

Bibliography

Adam, Stephen F. *Microwave Theory and Applications*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1969.

Agilent Technologies. "Spectrum and Signal Analysis...Pulsed RF," Application Note 150-2, Publication Number 5952-1039, July 2012.

Engelson, Morris. *Modern Spectrum Analyzer Theory and Applications*. Dedham, MA: Artech House, 1984.

Averaging and Filtering

Many analyzer measurements have considerable amounts of noise present in them. The noise is often undesirable but may actually be a desired component of the measurement. Two basic techniques are used to reduce the noise: filtering and averaging. Filtering usually takes the form of an analog filter. However, it can also be implemented in digital form, whereas averaging is always done digitally. The two concepts are closely related and are treated here in a unified manner. Both filtering and averaging can be classified as either *predetection* (before the detector) or *postdetection* (after the detector). Predetection averaging/filtering reduces the noise present in a measurement, while postdetection averaging/filtering reduces the amount of fluctuation in the noise.

10.1 Predetection Filtering

In the spectrum or network analyzer block diagram, filtering can be broken down into two types—predetection and postdetection—depending on whether the filter resides before or after the detector, as shown in Figure 10-1. In the classic swept analyzer, the detector is implemented as a distinct circuit block, but modern implementations use digital signal processing to implement it. There is no detector circuit in a fast Fourier transform (FFT) analyzer; instead, the magnitude detection is done by computing the magnitude of the complex frequency domain data provided by the FFT algorithm. Whether there is or is not an actual detector circuit, the concept remains the same.

Noise

There is always some noise present in the front end and intermediate frequency (IF) sections of an analyzer. This noise may come from the signal or network being measured or may be generated internal to the analyzer. Noise present at the input of the detector degrades the measurement depending on the relative amplitudes of the signal and noise. As shown in Figure 10-2, the wider the predetection bandwidth, the more noise that gets included in the measurement at the detector. Because the resolution bandwidth of the analyzer is relatively narrow, the noise can often be considered constant or white across its passband. The noise power is

$$P_N = N_0 NBW \quad (10-1)$$

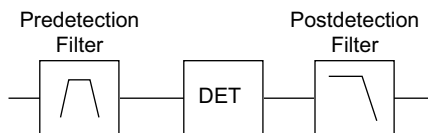


Figure 10-1 Predetection filtering takes place in front of the detector, while postdetection filtering is performed behind the detector.

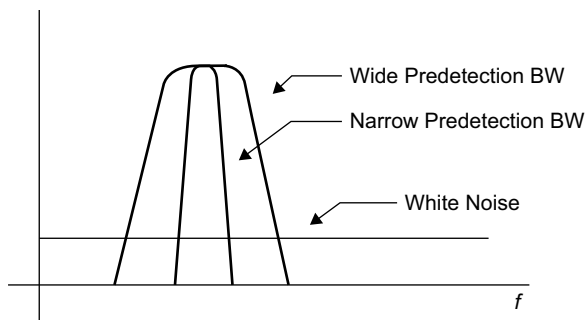


Figure 10-2 A wide predetection filter lets more noise into the measurement than does a narrow predetection filter.

where

N_0 = noise spectral density (watts/Hz)

NBW = the noise equivalent bandwidth of the predetection filter (Hz)

The noise power may be expressed in dBm (dB relative to 1 mW).

$$P_{N(\text{dBm})} = 10 \log(P_N/0.001) = 10 \log(N_0 NBW/0.001) \tag{10-2}$$

Note that a factor of 2 change in bandwidth results in a 3 dB change in noise level, assuming the spectral density of the noise is constant across the bandwidth.

$$\Delta P_{N(\text{dB})} = 10 \log(k_B) \tag{10-3}$$

where

ΔP_N = change in noise power

k_B = ratio of the two noise equivalent bandwidths

The following table shows the change in noise power (dB) corresponding to cardinal values of k_B .

k_B	ΔP_N
2	3.01 dB
5	6.99 dB
10	10.00 dB

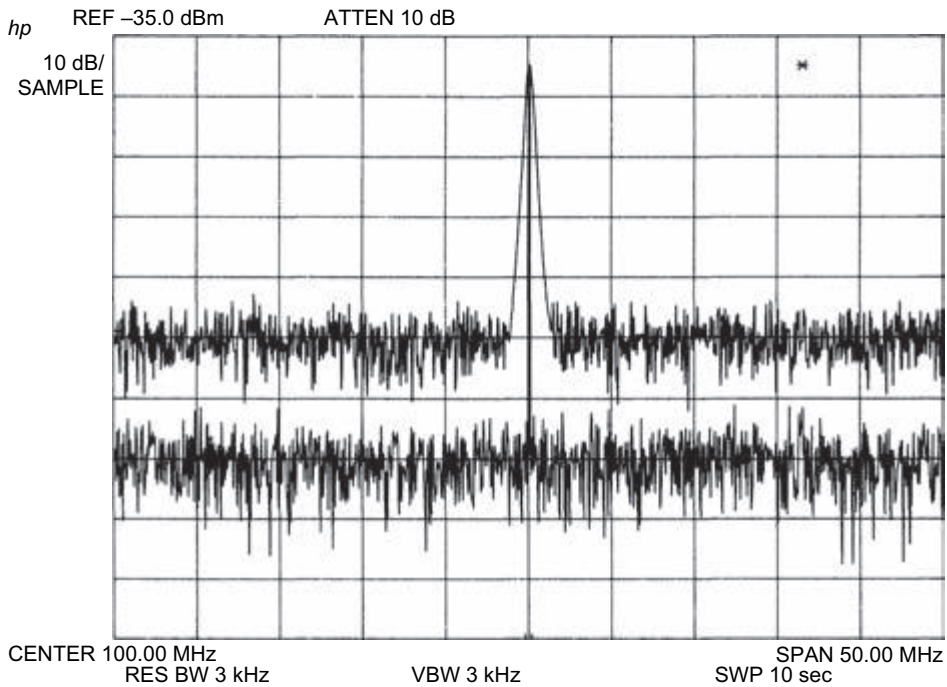


Figure 10-3 The upper trace was measured with a 300 kHz resolution bandwidth, and the lower trace was measured with a 3 kHz resolution bandwidth. The displayed noise level is 20 dB lower with the narrower bandwidth.

Figure 10-3 shows the effect on the analyzer display due to changing the resolution bandwidth. With a wider bandwidth, the noise on the display is higher while a narrower bandwidth reduces the noise.

Example 10.1

A spectrum analyzer measures the noise at a particular frequency to be -70 dBm using a resolution bandwidth of 1 kHz. What will the noise reading be using a resolution bandwidth of 300 Hz? Assume that the noise is white noise and that the bandwidths given are noise equivalent bandwidths.

The noise will be reduced by the ratio of the bandwidths, expressed in dB.

$$P_{N(\text{dBm})} = -70 \text{ dBm} + 10 \log(300/1000) = -75.2 \text{ dBm}$$

10.2 Predetection Filters

Predetection filters are easily identified in the block diagrams of traditional swept analyzers. The narrowest filter in the signal path before the detector is effectively the predetection filter. This filter's bandwidth is usually selectable and is indicated on the front panel as resolution

bandwidth or IF bandwidth. Modern analyzers use a digital IF to implement these filters, providing much more control of the filter shape and bandwidth.

In the case of the FFT analyzer, the predetection bandwidth is not a distinct filter in the block diagram. Instead, it is the effective bandwidth resulting from the use of the FFT. To a rough approximation, the predetection bandwidth will be the frequency span divided by the number of displayed points. To be more exact, this number must be adjusted depending on the time window function, with the actual bandwidth usually being somewhat larger. For a given number of displayed points, the selected frequency span will determine the predetection bandwidth. Most FFT analyzers will provide the effective predetection bandwidth for a given measurement setting, either via the operator's manual or via the display. For noise measurements it is important to always use the noise equivalent bandwidth, which is not necessarily the same as the 3-dB bandwidth (see Chapter 8).

10.3 Postdetection Filtering

Filters that reside after the detector in the signal processing chain are called postdetection filters. On swept analyzers, postdetection filters are usually called *video filters*. Postdetection filtering is not capable of reducing the noise level since the noise has already been detected. However, it can reduce the variation in the noise, exposing previously obscured signals that are near the noise floor. Also, if noise is being measured, postdetection filtering helps stabilize the measurement. Notice that in Figure 10-3 there is considerable variation in the amplitude of the noise, independent of which resolution bandwidth is used.

The output of the detector (over a short period of time) can be thought of as being a constant DC value with some noise superimposed on it (Figure 10-4a). The DC level represents the amount of energy present within the predetection bandwidth in front of the detector. This energy could be made up of discrete spectral lines or noise or both. The noise on the DC level is the statistical variation in the predetection energy, which is caused by the noise in the measurement. A low-pass filter applied to the detector output can reduce the variation in the detector output (Figure 10-4b), giving a more stable, noise-free output. This does not, however, reduce the DC level. Thus, postdetection filtering can reduce the variation in the detector's output but does not affect the average output level.

It is important to distinguish between predetection and postdetection noise. Predetection noise can be reduced by narrower predetection filtering, thereby reducing the output of the detector. The predetection noise will be detected and will contribute to the absolute level at the output of the detector. Postdetection noise is the variation in the predetection noise. This variation can be reduced by appropriate filtering, but the DC level that represents predetection noise cannot be reduced by postdetection filtering.

To understand the effect of postdetection filtering on a typical measurement, consider the analyzer display shown in Figure 10-5a and Figure 10-5b. With a wide postdetection filter, the variance in the noise is quite large. With a narrower postdetection filter, the variance is reduced considerably. Note that the average value of the measured noise remains the same, and only the variation in the noise is different. While postdetection filtering does not lower the average noise level, the reduction in the variance does reduce the peak noise level and may expose low-level signals that cannot be observed with a wider postdetection bandwidth.

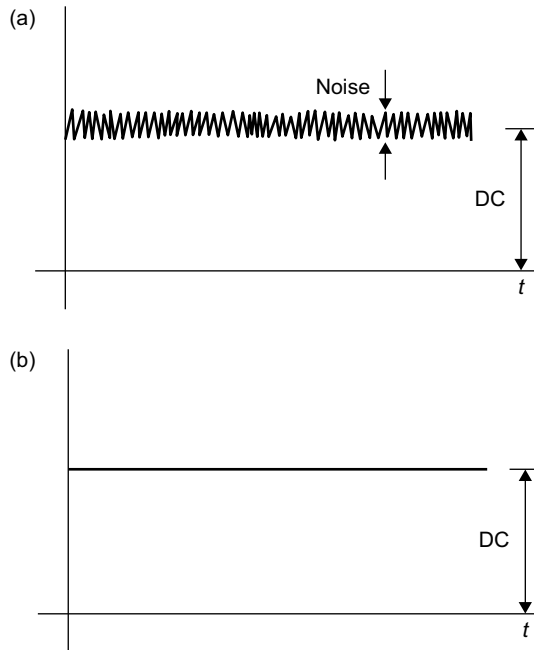


Figure 10-4 (a) The output of the detector consists of a constant DC level plus some noise.
 (b) Low-pass filtering the detector output removes the noise without altering the DC level.

Noise in a measurement is not always undesirable. The signal being measured may consist only of noise or some combination of noise and spectral lines. In such a case, the noise is intended to be part of the measurement. With no postdetection filtering, the measurement will tend to vary due to the effect of the noise. Postdetection filtering can be used to smooth out these variations and cause the measurement to converge on a single smooth trace.

10.4 Postdetection Filters

In a classic swept analyzer, the postdetection filter is implemented using a low-pass analog filter following the detector. This filter is usually a single-pole filter, often a simple resistor-capacitor (RC) network. The filter can also be implemented by digital techniques provided that the signal has been digitized at some point in the block diagram. Since the video filter slows down the response of the analyzer's receiver, the sweep rate must be increased for smaller video bandwidths, and most analyzers have mechanisms for automatically selecting a suitable sweep speed.

In an FFT analyzer, there is no exact equivalent to the postdetection filter, but post-detection averaging can produce a similar effect.

This discussion of predetection and postdetection filters may leave the user wondering how to choose the appropriate bandwidths. Fortunately, most modern spectrum analyzers

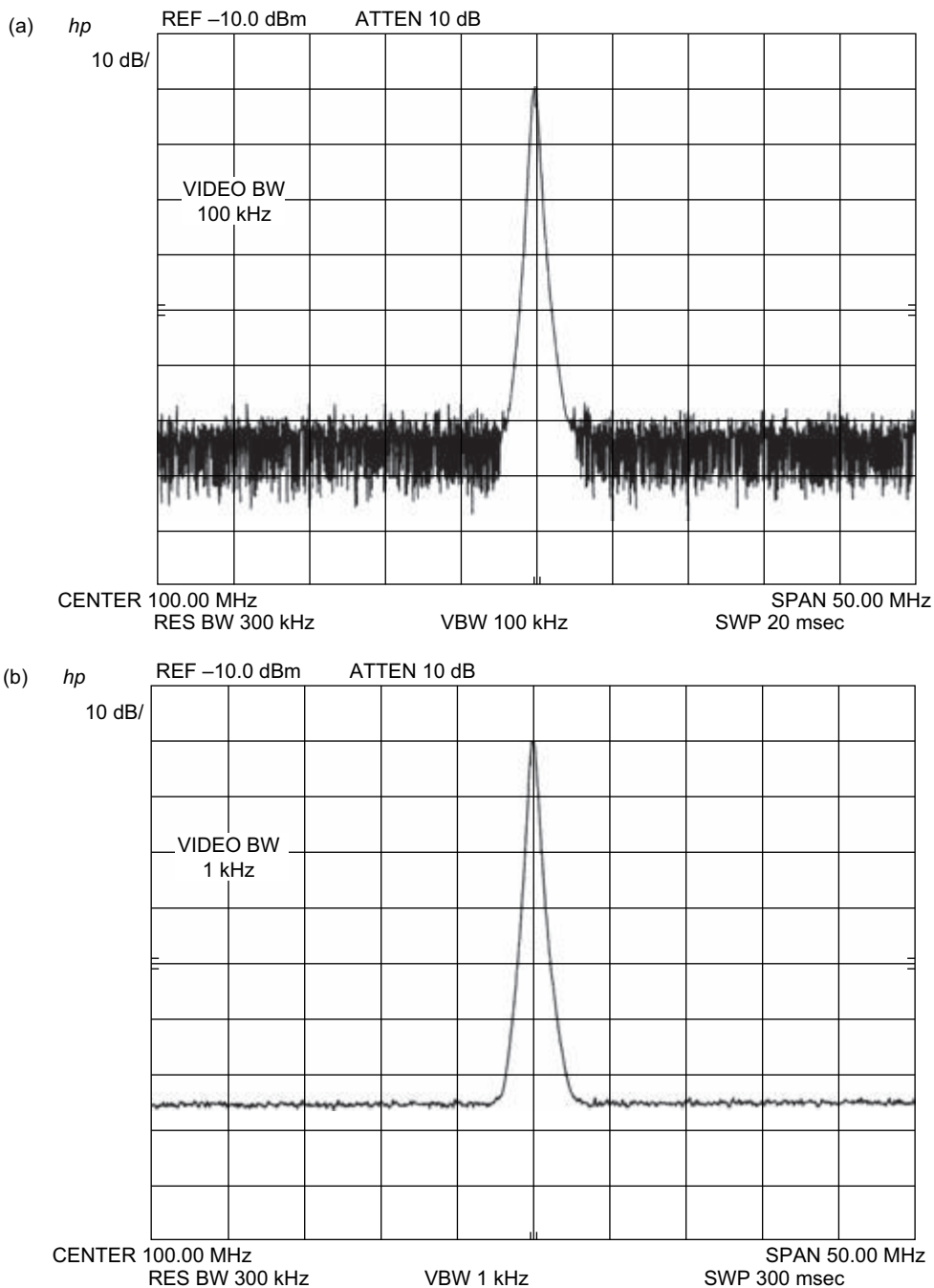


Figure 10-5 (a) The noise variance is relatively high with a wide video bandwidth. (b) A narrow video bandwidth causes the noise variance to be significantly reduced.

have built-in algorithms for choosing the two bandwidths automatically. In more critical applications, these choices can be overridden to optimize the measurement.

10.5 Averaging

Averaging was originally used in FFT analyzers to provide a method of reducing the noise. With the increased use of digital IFs techniques in swept analyzers, averaging has found its way into those instruments. Averaging techniques can be divided into predetection and postdetection types, similar to predetection and postdetection filtering. Again, filtering and averaging are very similar operations, so predetection filtering and predetection averaging have similar effects. The same can be said for postdetection filtering and averaging.

First, the process of averaging will be discussed in a general way, without reference to analyzer applications. Many electrical parameters can be thought of as being made up of two parts:

$$x(t) = s(t) + n(t) \quad (10-4)$$

where

$x(t)$ = the measured value

$s(t)$ = the desired signal to be measured

$n(t)$ = the noise present in the measured value

Noise and signals contaminated by noise must be treated on a statistical basis. The variance, σ^2 , is defined as

$$\sigma^2 = E[x^2] - E^2[x] \quad (10-5)$$

where

$E[x]$ = expected value of x

The *variance* is the square of the *standard deviation*, σ . As the name implies, the variance is a measure of how much a noisy parameter varies away from its average value. If the noise has zero mean, then the average value of $x(t)$ equals $s(t)$, the desired signal.

Usually when a measured parameter is averaged, the signal portion of $x(t)$ will be retained while the noise portion, $n(t)$, will be decreased. This assumes that the signal portion is consistent, producing the same value on each sample. Similarly, the noise is assumed to be uncorrelated to the sample rate and will vary in value with each sample. Any portion of $x(t)$ that is correlated to the sample rate will tend to be retained after averaging. Any portion that is uncorrelated will tend to be averaged out.

10.6 Variance Ratio

Averaging can be considered as a process with an input, $x(t)$, and an output, $y(t)$, as shown in Figure 10-6. Both the input and the output have corresponding variances, σ_x^2 and σ_y^2 . By averaging, the variance of the measured signal is reduced and $y(t)$ is a better approximation

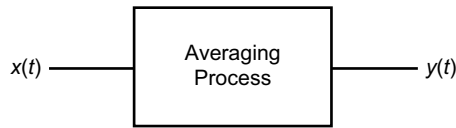


Figure 10-6 Averaging produces an output signal, $y(t)$, which has a lower variance than the input signal, $x(t)$.

to the desired signal, $s(t)$. The *variance ratio* (VR) is used as a figure of merit for the averaging process.

$$VR = \frac{\sigma_y^2}{\sigma_x^2} \quad (10-6)$$

where

$$\begin{aligned} \sigma_x^2 &= \text{variance of the unaveraged signal} \\ \sigma_y^2 &= \text{variance of the averaged signal} \end{aligned}$$

The variance of a signal is associated with its noise power (not its voltage). The *standard deviation*, which is the square root of the variance, should be used to analyze noise in terms of voltage. Since the VR is power related, it can be converted to decibel form by

$$VR_{(\text{dB})} = 10 \log(VR) \quad (10-7)$$

Example 10.2

For a given averaging process, the variance of the averaged output is 0.2 times the variance of the input. If the noise power at the input is -45 dBm, what is the noise power at the output of the averaging process?

$$\begin{aligned} \sigma_y^2 &= 0.2\sigma_x^2 \\ VR &= 0.2 \\ VR_{(\text{dB})} &= 10 \log(0.2) = -6.99 \text{ dB} \\ P_{N(\text{dBm})} &= -45 \text{ dBm} - 6.99 \text{ dB} = -51.99 \text{ dBm} \end{aligned}$$

10.7 General Averaging

In general, averaging is accomplished by weighting a set of data samples according to some algorithm and summing them together. Mathematically, this can be expressed as

$$y_N = w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_Nx_N \quad (10-8)$$

where

$$\begin{aligned} w_1, w_2, w_3, \dots, w_N &= \text{weighting factors} \\ x_1, x_2, x_3, \dots, x_N &= \text{last } N \text{ samples of } x(t) \\ y_N &= \text{current averaged output} \end{aligned}$$

The VR for the averaging process can be shown to be just the sum of the squares of the weighting factors.

$$VR = \sum_{n=1}^N w_n^2 \quad (10-9)$$

Such a general approach to averaging requires that N samples of $x(t)$ be always stored with the averaged output, and y_N computed by summing the weighted x_n values. As a new x_n sample is acquired, the oldest x_n is discarded and the new sample saved. (This requires some type of memory buffer to store the sample points.)

10.8 Linear Weighting

The most obvious way to weight the data is to weight them all equally. After all, what makes one sample point any more valid than any other point? This type of averaging is probably what most engineers think of when the term *averaging* is used.

The VR for linear averaging is

$$VR = \frac{1}{N} \quad (10-10)$$

where

N = number of samples averaged

Thus, for N measurements averaged together (N averages), the noise power is reduced by a factor N , and the noise voltage is reduced by a factor of the square root of N .

When linear weighting is used in instrumentation averaging, the final averaged result (with all N measurements averaged together) cannot be displayed until all N measurements have been acquired. Many instruments will display the intermediate results of the averaging process so the user has some measurement information without having to wait for all N acquisitions.

For a given number of samples, linear weighting provides the best possible VR . The more samples that are averaged together, the better the noise reduction at the expense of longer measurement time.

10.9 Exponential Weighting

The weighting function may be an exponential function with the most recent sample weighted the highest and previous samples weighted exponentially less. Although an exponential weighting function would seem to be computationally complex, it can be implemented with a simple algorithm. The averaged output is computed by summing the input sample multiplied by a factor of $1/k$ and the previous result multiplied by $1 - (1/k)$. A single accumulation register (memory location) is used to hold the y_{n-1} (previous) value. In a typical analyzer application, an accumulation register is required for each displayed frequency point.

$$y_n = (1/k)x_n + (1 - (1/k))y_{n-1} \quad (10-11)$$

Exponential averaging does not have a fixed number of samples, and the averaging algorithm can continue to run indefinitely. A newly acquired sample is initially weighted heavily, and then the weighting factor for that sample gradually decreases as additional samples are taken. This type of algorithm has the ability to track changes in the measured value.

The VR , assuming that the averaging process has been running for a very long time (much more than k samples), is

$$VR = \frac{1}{2k - 1} \quad (10-12)$$

The exponential weighting function produces a step response very similar to a single-pole low-pass filter (Figure 10-7). If the input of the averaging process starts at zero and abruptly changes to a constant value, the output of the process rises exponentially and asymptotically approaches the final value of the input. For k 's of interest, the time constant of this system, T is given by

$$T = k + \frac{1}{2} \quad (10-13)$$

Thus, the time constant of the system is approximately k . After k samples the step response will reach 63% of the final value, just as one would expect in a single-pole analog system. In a purely exponential average, the user should keep this behavior in mind and wait several time constants for the measurement to settle out. Selecting a large value for k will provide the maximum amount of noise reduction, but at the expense of increased measurement time and slower response to changes.

The exponential average has an initialization problem, in that if the y_{n-1} accumulation register starts out set to zero it will take k samples to get 63% of the way to the final value. It will take even longer before the averaging process produces an output close to the true answer. One solution is to load the first sample into the y_{n-1} register and then let the averaging algorithm run. This immediately gets the averaged output close to the correct answer (depending on how “good” the first sample is). Unfortunately, this also causes the first sample to be weighted much heavier than the others, which is most noticeable with large k 's since the subsequent samples are weighted very lightly.

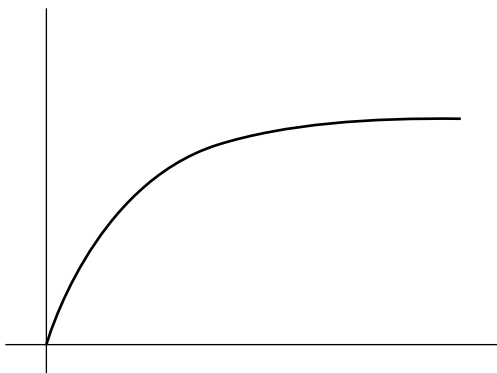


Figure 10-7 The step response of an averaging process with exponential weighting is an exponential function.

A better solution to the initialization problem is to start the averaging algorithm with k small and load the first sample into the y_{n-1} register. As the averaging algorithm progresses, the k value is automatically increased, eventually stopping at the value selected by the user. This technique has the effect of producing a linear or near-linear weighting on the early samples. Later, when the k value reaches the maximum imposed by the user, the averaging algorithm reverts to a pure exponential. This type of weighting combines the advantages of linear weighting and exponential weighting. It provides the good variance improvement of a linear average in the early samples but with the exponential average advantage of tracking changes.

10.10 Averaging in Spectrum and Network Analyzers

The previous discussion has centered on the general principles of averaging along with the common weighting functions. Now we will focus on how these weighting functions are applied to the measured data in a spectrum or network analyzer.

First, it must be made clear in what dimension the averaging is taking place. *Trace-to-trace averaging* is accomplished by taking a sample at one particular frequency or bin and averaging it with samples *at the same frequency* from other traces or sweeps. This represents most of the averaging algorithms found in spectrum and network analyzers. *Adjacent-point averaging* is accomplished by averaging several data points together from the same sweep. For example, the n -th bin might be computed by the linear average of the n -th $- 1$, n -th, and n -th $+ 1$ bins. This type of averaging is used to implement *smoothing* functions.

For the classic swept spectrum analyzer with an analog detector, there is only one type of data to be averaged: scalar magnitude data produced by sampling the output of the detector. Since these data exist after the detector, only postdetection averaging is possible. However, modern spectrum and network analyzers have digital IFs that provide access to the vector data samples before the detector. Vector data are represented by complex numbers that contain the magnitude and phase information necessary for vector signal analysis or vector network measurements. During detection, the vector data may be converted into scalar magnitude information.

Log Detector Problem

As mentioned in Chapter 8, the combination of an envelope detector and log amp can understate the noise level in a measurement by 2.5 dB. This is because the statistical variation in noise (or a noise-like signal) is distorted by the log amp: positive noise excursions are reduced in relative amplitude while negative noise blips end up passing through with relatively larger amplitude.

For many measurements, it is very useful to have a logarithmic scale (in dB) for the vertical axis, so we don't want to give that up. The problem occurs when we apply the averaging algorithm *after* the log amp. The average of the log output is not the same as the log of the average.¹ Fortunately, modern analyzers can use the digital IF to advantage here by averaging the IF signal before the log function is applied. The *power average* (also

¹ See Agilent Technologies (2011) for a detailed analysis of the effects of averaging and detection on noise.

called *root mean square (RMS) average*) feature captures the voltage of the IF signal, squares it to get the signal power, and performs the trace-to-trace averaging to produce the measured result. This averaging approach does not affect continuous wave signals but provides a much more accurate measurement of noise and complex signals.

10.11 RMS Average

RMS averaging is a scalar trace-to-trace average commonly used in FFT analyzers and is now also applied to swept analyzers. Several V^2 of each bin from several traces are averaged together, and then the square root of the averaged data is displayed. Since the square of the voltage corresponds to the power in a signal, this averaging technique is also known as a *power average*.

$$y_n = \sqrt{(x_{n1}^2 + x_{n2}^2 + x_{n3}^2 + \cdots + x_{nN}^2)/N} \quad (10-14)$$

where

$x_{n1}, x_{n2}, \dots, x_{nN}$ = unaveraged data

y_n = RMS averaged output

N = number of averages

Linear weighting of the RMS data is assumed here, but exponential weighting can also be used. RMS averaging is a postdetection process and therefore reduces the variance of the noise, but not the absolute noise level (Figure 10-8).

10.12 Vector Averaging

Many types of spectrum and network analyzers capture the magnitude and phase of the signal being measured. A vector representing a complex signal can be plotted on the complex plane (Figure 10-9a). The vector has a real part and an imaginary part, which can be converted into magnitude and phase form, if necessary.

Noise present in the measurement will add vectorally to the signal vector (Figure 10-9b), producing a single-plus-noise vector that varies in amplitude and phase (Figure 10-9c). Applying an averaging algorithm independently to the real part and to the imaginary part of the noisy vector tends to average out the noise portion while retaining the signal. (This assumes that the noise vector is not correlated to the signal vector.) With sufficient averaging, the noise portion will approach zero, leaving only the signal. Thus, vector averaging removes the uncorrelated noise and leaves the mean of the desired signal. This type of averaging lowers the noise floor of the measurement, similar to a reduction in resolution bandwidth.

Vector averaging is predetection averaging, since it operates on the complex data before the detector and reduces the amount of noise seen by the detector. For vector averaging to be effective, the phase of the signal must be consistent from trace to trace; otherwise, the signal will tend to be averaged away.

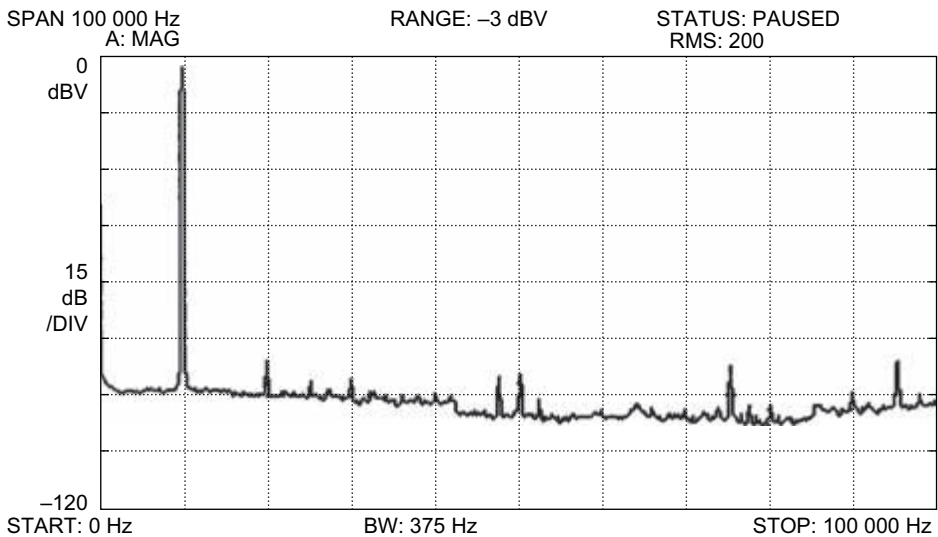
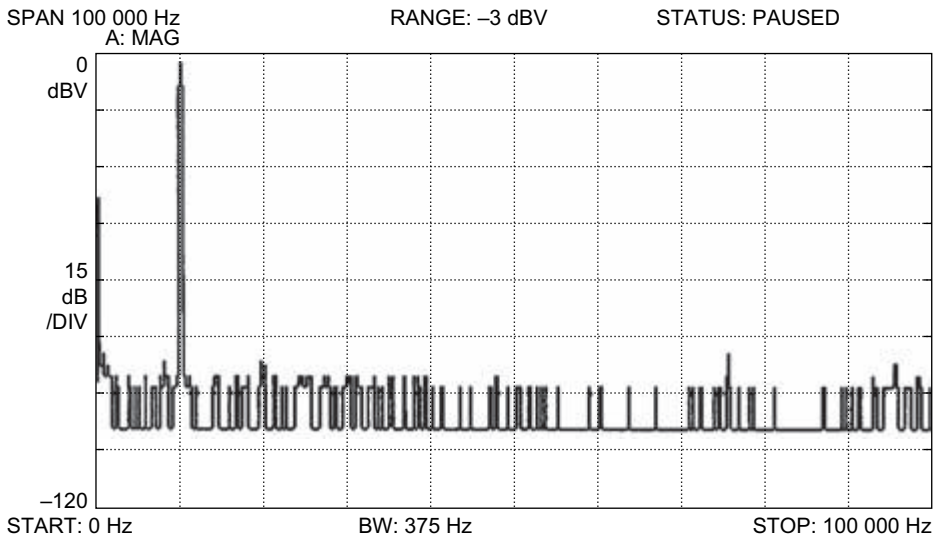


Figure 10-8 (a) FFT spectrum measurement with no averaging. (b) The same measurement with RMS averaging.

In the case of an FFT analyzer, the measured data are captured in the time domain. If the sampled waveform in the time record is repeatable from record to record, the frequency domain data will have a repeatable phase from trace to trace and vector averaging can be used. A triggering circuit, similar to the triggering circuit found in an oscilloscope, is often employed to start the collection of the time domain data. Assuming that the trigger always

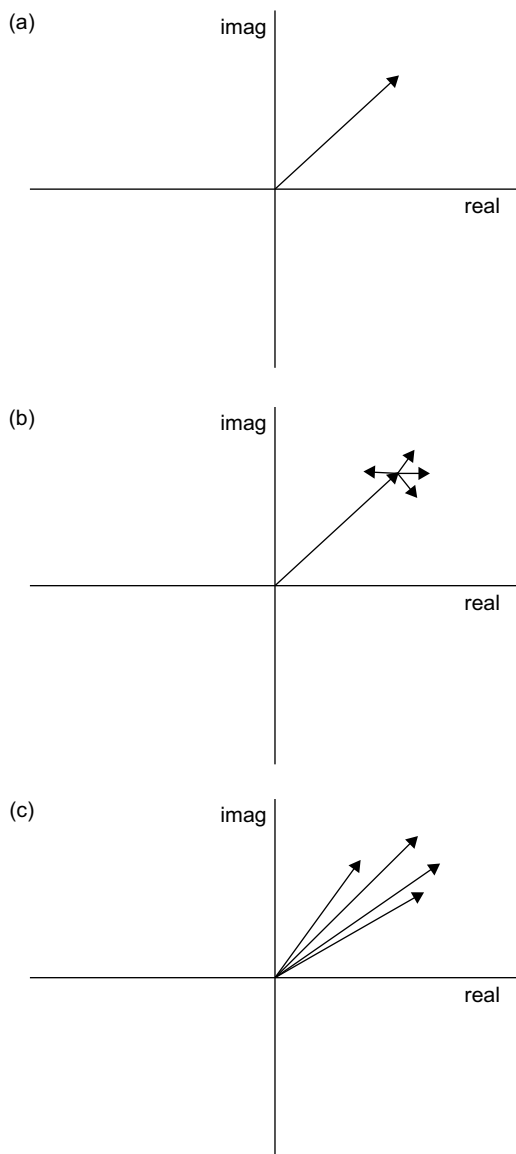


Figure 10-9 (a) The vector representation of a complex signal. (b) Random noise adds vectorially with the signal. (c) The original signal is varied in both magnitude and phase.

occurs at the same point on the waveform, the phase of the waveform will be the same for each time record acquired. Actually, the averaging can be performed in the time domain to produce the same effect as vector averaging in the frequency domain. (Sometimes vector averaging is called *time average* in FFT analyzers.) Signals that are repeatable in the time domain will remain in the final measurement while noise tends to be averaged out.

10.13 Smoothing

Smoothing functions use adjacent-point averaging to reduce the amount of fluctuation in the measured trace due to noise. This is different from other averaging techniques that combine data points from different measurements to produce the final result. N points of the trace are averaged together to produce one smoothed point. (N is odd.) The $(N-1)/2$ previous points, the $(N-1)/2$ subsequent points, and the current point are summed together with appropriate weighting. The general formula for smoothing is

$$y_n = w_{-(N-1)/2}x_{n-(N-1)/2} + \cdots + w_{-1}x_{n-1} + w_0x_n + w_1x_{n+1} + \cdots + w_{(N-1)/2}x_{n+(N-1)/2} \quad (10-15)$$

where

y_n = smoothed output data for bin n

x_k = unsmoothed input data for bin k

$w_{-(N-1)/2}$ through $w_{(N-1)/2}$ = the weighting coefficients

N = the number of points used in the smoothing algorithm (N is odd)

The data points are all taken from the same trace of data.

A simple implementation of a smoothing algorithm is to use just three points ($N = 3$) in the smoothing of the data.

$$y_n = 0.25x_{n-1} + 0.5x_n + 0.25x_{n+1} \quad (10-16)$$

Since the user has control over the amount of smoothing applied to a displayed trace, good judgment must be applied. It is possible to smooth a trace to the point where it provides little or no useful information. The user is generally required to select the amount of smoothing that reduces the noise without significantly changing the shape of the trace.² Figure 10-10 shows the transfer function of a network with varying amounts of smoothing.

Since the smoothing algorithm operates on the data after the detector, it is a type of postdetection averaging. Its effect is similar to video filtering except for two things. First, it uses the data from bins on both sides of the bin of interest, while video filtering averages only frequency bins that it has swept through (usually to the left of the bin of interest). Second, it does not impact the allowable sweep rate in a swept analyzer, although excessive smoothing can distort the trace similar to sweeping too fast for a video filter.

10.14 Averaging versus Filtering

One important difference between averaging and filtering in analyzers is the dimension in which the averaging/filtering takes place. Since filtering occurs in the IF section, the analyzer is filtering as it sweeps. Thus, filtering is done across the frequency axis of the display

² In other words, select the amount of smoothing that makes the trace look good.

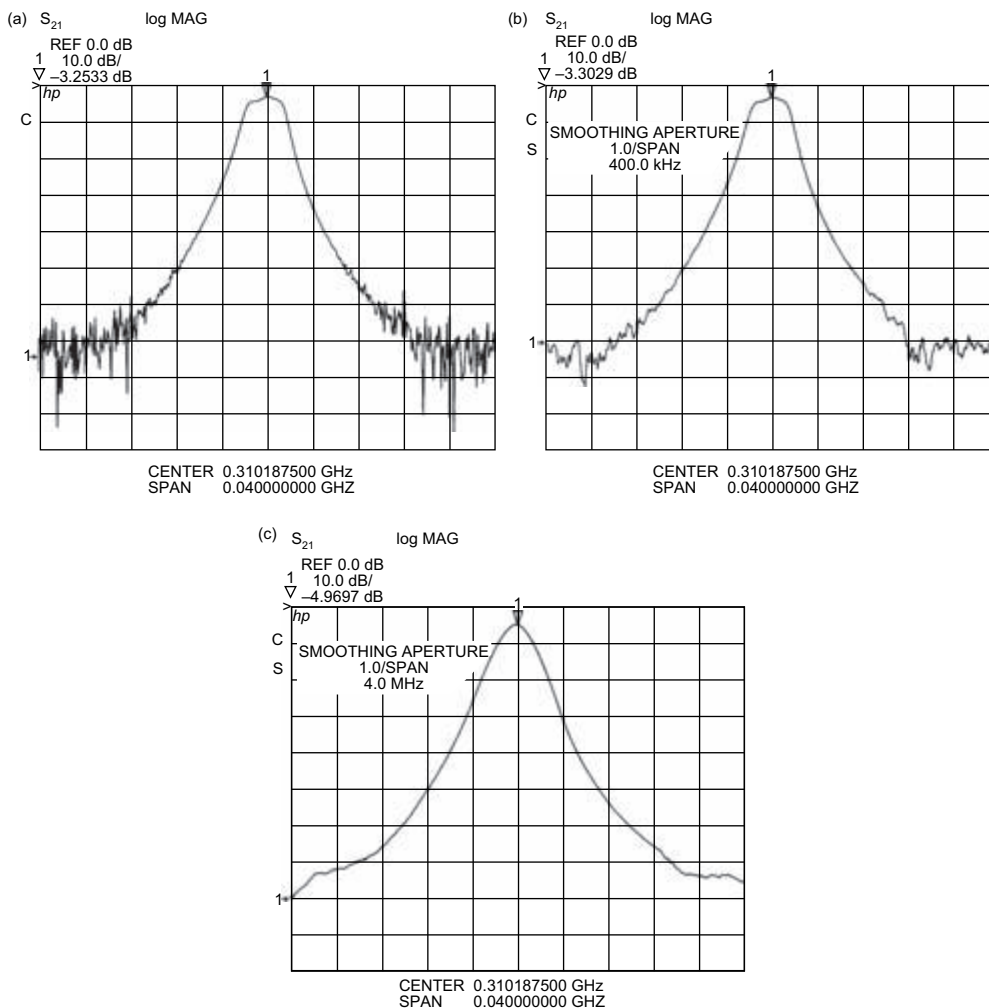


Figure 10-10 (a) A network measurement with no smoothing. (b) The same measurement with some smoothing. (c) The same measurement with excessive smoothing.

(similar to adjacent-point averaging). For the filtering to not distort the measurement by smearing the trace, the sweep rate must not be too fast. The narrower the resolution bandwidth or video filter the slower the sweep must be.

Trace-to-trace averaging, on the other hand, averages data from different sweeps together. This does not cause smearing in the direction of the frequency axis and does *not* affect the required sweep rate. Of course, multiple sweeps must be acquired that slow down the rate at which an averaged measurement can be completed.

Bibliography

Agilent Technologies. "Fundamentals of Signal Analysis," Application Note 243, Publication Number 5952-8898E, 2000.

Agilent Technologies. "Spectrum and Signal Analyzer Measurements and Noise," Publication Number 5966-4008E, October 2011.

Schwartz, Mischa, and Shaw, Leonard. *Signal Processing*. New York: McGraw-Hill Book Company, 1975.

Stanley, William D., Dougherty, Gary R., and Dougherty, Ray. *Digital Signal Processing*, 2nd ed. Reston, VA: Reston Publishing Company, Inc., 1984.

Witte, Robert A. "Averaging Techniques Reduce Test Noise, Improve Accuracy." *Microwaves & RF*, February 1988.

Transmission Lines

Transmission lines are commonly used to connect test and measurement instruments to the device under test. Transmission lines are used to control the effects of inductance and capacitance, which are unavoidable in high-frequency systems. Coaxial cables are the most common transmission lines, providing shielding of the signals being measured.

Measurement error can be introduced due to impedance mismatch at either end of a transmission line. These errors must be understood and minimized to ensure an accurate measurement.

11.1 The Need for Transmission Lines

When connecting DC circuits, the major issue is the resistance of the wires. According to Ohm's law, a drop in voltage will occur when a current flows through a wire with nonzero resistance. Inductance and capacitance are not usually a big concern for DC voltages and currents.

For circuits with AC voltages and currents, the inductance and capacitance of wires come into play. A typical wire exhibits self-inductance and has some capacitance to other nearby conductors. The higher the frequency, the more significant the effect of this inductive and capacitive reactance. Uncontrolled, these reactive effects can distort signals by loading the driving circuit and causing reflections on the wire. Transmission lines avoid these problems by controlling the inevitable inductance and capacitance of the cable and the electromagnetic fields associated with them.

Signals do not travel down a wire infinitely fast but require a finite amount of time to propagate from one place to another. For circuits and systems that have short connections (relative to the wavelength of the signal), these effects are usually ignored. As the frequency of the signal or the length of the wire is increased, the delays along the wire become significant. As the signal propagates down the wire, it may encounter variations in the impedance that it sees. The signal will be fully or partially reflected at each of these impedance discontinuities.

Reflections on a wire can cause the impedance looking into the wire to be uncontrolled, which can present an unknown or undesirable impedance to the driving circuit. When terminated properly, transmission lines provide a controlled impedance at each end of the line. This allows the system to be designed for maximum power transfer, with the signal source loaded by an impedance equal to its output impedance.

11.2 Distributed Model

The inductance and capacitance associated with lengthy wires are used to advantage in a transmission line. The reactances are controlled such that a signal traveling down the line sees a constant impedance. A circuit model for an arbitrarily small section of transmission line is shown in Figure 11-1a. The transmission line contains some series inductance, L , and some capacitance, C , between the two conductors. Also included in the circuit model is a series resistance, R , and a shunt conductance, G , associated with losses incurred in the transmission line.

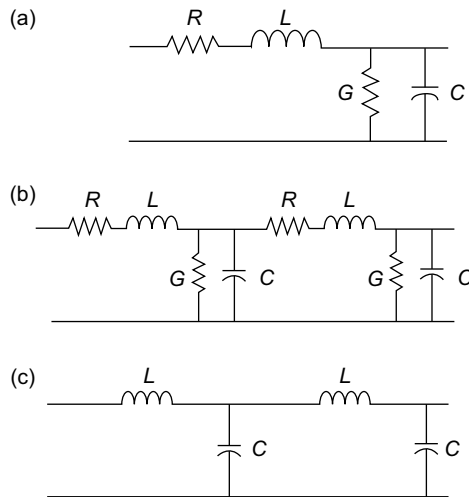


Figure 11-1 (a) The circuit model for a small section of transmission line. (b) The circuit model for a transmission line. (c) The circuit model for a lossless transmission line.

The circuit model represents an extremely small section of transmission line. A finite-length line is modeled as a large number of these sections cascaded end to end (Figure 11-1b). If the line is assumed to be lossless ($R = G = 0$), the resistive components are removed from the model (Figure 11-1c).¹

The capacitance per unit length and the inductance per unit length depend on the physical construction of the transmission line. The dielectric constant of the material between the conductors and the physical geometry of the transmission line both affect the electrical properties of the line.

11.3 Characteristic Impedance

The impedance looking into the end of an infinitely long lossless transmission line is called the *characteristic impedance*, Z_0 . (The transmission line is specified here as infinitely long to avoid any impedance changes due to reflections from the other end of the line.)

$$Z_0 = \sqrt{L/C} \quad (\text{lossless line}) \quad (11-1)$$

¹ Transmission line losses will be ignored in this chapter unless otherwise specified.

11.4 Propagation Velocity

Electromagnetic waves in free space propagate at the speed of light. Inside a transmission line there is usually a dielectric material that lowers the propagation velocity. Thus, the *propagation velocity* of a transmission line is given by

$$v_p = k_v c \quad (11-2)$$

where

k_v = velocity factor

c = velocity of light in free space (approximately 3×10^8 m/sec)

The velocity factor simply expresses the propagation factor as a percent of free space light velocity. The velocity factor has a value between 0 and 1 depending on the dielectric material in the transmission line and is specified by the cable manufacturer. Typically, k_v ranges from 60% to 90%.

11.5 Generator, Line, and Load

First, consider the generator and load shown in Figure 11-2. The generator produces a 1 V step and has an output impedance equal to Z_0 . The generator is connected to the Z_0 load by very short wires, and therefore there are no transmission line effects. At the same instant the generator voltage changes from 0 V to 1 V, the voltage across the load resistor changes from 0 V to 0.5 V. Note that the load voltage is one-half of the generator voltage due to the voltage divider effect.

Z_0 Load

If a transmission line is inserted between the generator and the load, the situation changes (Figure 11-3a). When the generator voltage changes from 0 V to 1 V, a forward-going (incident) voltage is created at the generator end of the transmission line. Since the generator sees the Z_0 impedance of the line, this incident voltage is equal to one-half of the generator voltage. This voltage moves down the transmission line at the propagation velocity until it meets the load. Since the load impedance is equal to the characteristic impedance of the line, no reflections occur. The incident voltage is “absorbed” by the load.

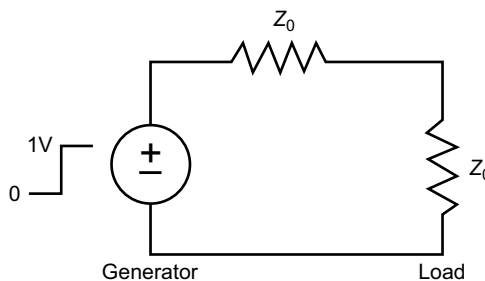


Figure 11-2 A Z_0 generator drives a Z_0 load.

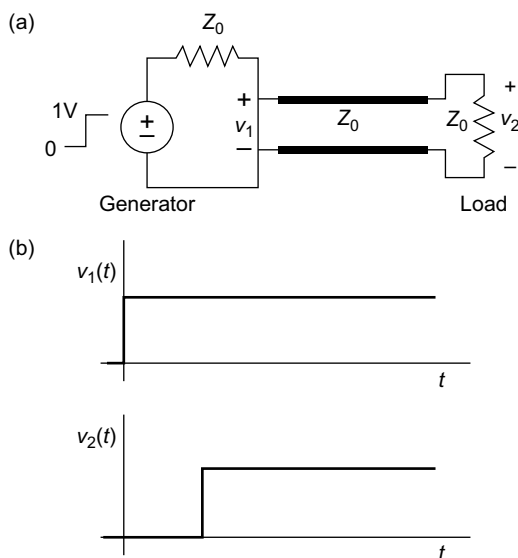


Figure 11-3 (a) A Z_0 generator drives a Z_0 load using a transmission line. (b) It takes a finite amount of time for the voltage to travel down the transmission line.

There is a time delay in the system as the voltage wave travels down the transmission line (Figure 11-3b). This is unlike the previous example, where the wires are so short that the load voltage instantaneously follows the generator voltage. Notice that the final value of the load voltage is the same in both cases. After the transmission line effects settle out, the DC voltages will be the same.

The system shown in Figure 11-3 has the transmission line matched at both ends. That is, the impedances that the transmission line sees at the generator end and the load end are both Z_0 . This eliminates any possible reflections and is usually the desirable case in instrumentation use. However, the generator and the load impedances may not be Z_0 , so other cases must be considered.

Non- Z_0 Load

Suppose the Z_0 load is replaced by a load that is some other value (Figure 11-4a). As in the Z_0 load case, the incident voltage of 0.5 V initially appears at the generator end of the transmission line. The incident voltage is not affected by the change in load impedance since the generator initially sees only the Z_0 impedance of the line. The 0.5 V step propagates down the line and eventually reaches the load. The load is *not* matched to the Z_0 line, so some of the forward-going voltage is reflected back toward the generator. The reflected voltage is given by

$$V_R = \Gamma V_I \quad (11-3)$$

where

V_R = reflected voltage

V_I = incident voltage

Γ = reflection coefficient

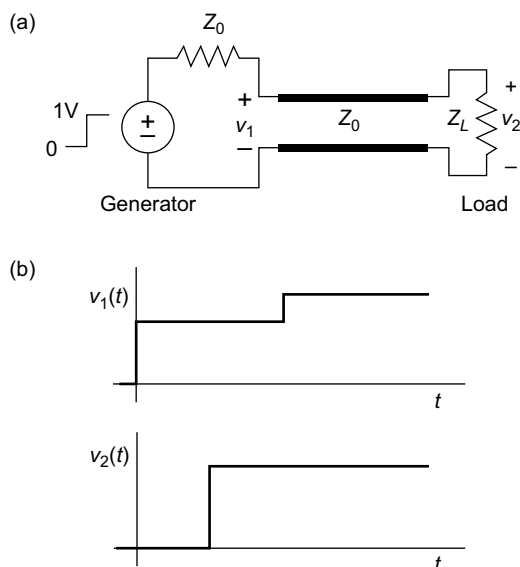


Figure 11-4 (a) A Z_0 generator driving a non- Z_0 load. (b) The incident wave appears immediately at v_1 and travels down the line to v_2 , and a portion is reflected back toward the source. After the reflected wave travels back to the source, it appears at v_1 .

The absolute value of Γ cannot exceed 1 since the reflected voltage cannot be larger than the incident voltage. The value of Γ can vary between -1 and $+1$, inclusive.²

For the case shown, Γ can be computed by

$$\Gamma = \frac{Z_L - Z_0}{Z_L + Z_0} \quad (11-4)$$

Also,

$$Z_L = Z_0 \frac{1 + \Gamma}{1 - \Gamma} \quad (11-5)$$

The reflected voltage, V_R , propagates back down the line toward the generator. The voltage at any point on the line is the sum of the incident and reflected voltages, taking into account how far the two waves have traveled. The line is initially at 0 V (because the generator has presumably been at 0 V for some time). As the incident wave travels down the line, the line becomes charged to V_I . Then the reflected wave starts back down the line moving from the load toward the generator. As the wave passes any given point, the voltage on the line at that point goes from V_I to $V_I + V_R$. When the reflected wave reaches the generator, it sees a Z_0 impedance (of the generator) and no additional reflections occur. Had the generator impedance been other than Z_0 , additional reflections would occur.

² The reflection coefficient is introduced here as a scalar quantity, but the definition will be expanded to include complex values.

Example 11.1

Determine the incident and reflected voltages for the case shown in Figure 11-5. What is the final value of the load voltage?

The incident wave $V_I = 4(50)/(50 + 50) = 2$ V, the reflected wave $V_R = \Gamma V_I$, and $\Gamma = [(30 - 50)/(30 + 50)] = -0.25$. So $V_R = (-0.25)(2) = -0.5$ V.

The final value of the load voltage is

$$V_L = V_I + V_R = 1.5 \text{ V}$$

Note that this answer agrees with a simple DC analysis, ignoring the transmission line:

$$V_L = (4)[30/(50 + 30)] = 1.5 \text{ V}$$

Let's describe what happens. The line is initially at 0 V. When the voltage source steps to 4 V, a 2 V incident wave propagates down the line. When the incident voltage reaches the load, a -0.5 V reflected wave starts its way back. As the reflected wave propagates back, the transmission line voltage becomes 1.5 V. Finally, the reflected wave is absorbed when it reaches the source, since the source is matched to Z_0 .

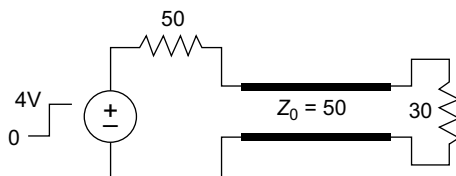


Figure 11-5 A 50 Ω source drives a 30 Ω load via a 50 Ω line.

Open Load

A special case of load impedance is when there is no load at all (i.e., an infinite impedance) as shown in Figure 11-6a. The reflection coefficient can be calculated for this case:

$$\Gamma = \frac{Z_L - Z_0}{Z_L + Z_0} \Big|_{Z_L=\infty} = 1 \quad (11-6)$$

Thus, all of the incident voltage is reflected back toward the generator. The incident voltage is once again 0.5 V, which propagates down the line until it encounters the load.

$$V_R = \Gamma V_I = (1)(0.5) = 0.5 \text{ V} \quad (11-7)$$

So 0.5 V is reflected back down the line toward the generator. After the reflected voltage propagates down the line, the voltage on the transmission line is $V_I + V_R = 1$ V. This value agrees with a simple DC analysis.

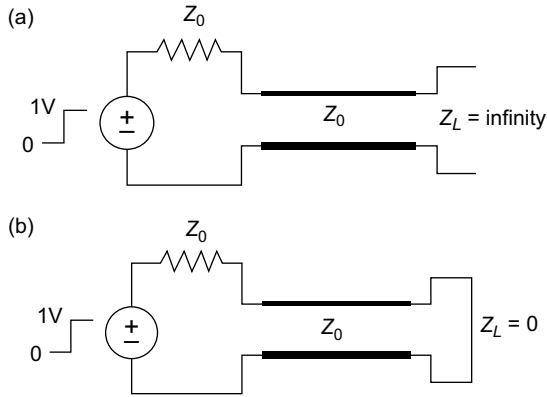


Figure 11-6 (a) When the line is terminated in an open circuit, the reflection coefficient is 1. (b) When the line is terminated in a short circuit, the reflection coefficient is -1 .

Short Load

Another special case is when the load impedance is a short circuit ($Z_L = 0$) as shown in Figure 11-6b. For this case,

$$\Gamma = \left. \frac{Z_L - Z_0}{Z_L + Z_0} \right|_{Z_L=0} = -1 \tag{11-8}$$

The incident voltage is again 0.5 V. The reflected voltage is

$$V_R = \Gamma V_I = (-1)(0.5) = -0.5 \text{ V}$$

The incident voltage of 0.5 V propagates down the line to the load where the negative of it is reflected back toward the generator. Since $V_I + V_R = 0$, the net result is that the voltage returns to zero since the incident and reflected voltages cancel. This is required for the result to make any sense: the final DC voltage across a short circuit must be zero.

11.6 Impedance Changes

So far, we have given examples where a generator drives a line that is connected to a load impedance. Now we will expand the concept of reflection coefficient to include the case where a voltage is incident at the junction of two different impedances. When the incident voltage encounters an impedance change, part of the incident voltage is reflected back and part of it travels on through (Figure 11-7), analogous to a lightwave encountering a partially reflective lens.

The two impedances involved, Z_1 and Z_2 , might be two transmission lines with differing characteristic impedance or perhaps a slight impedance mismatch due to connector imperfections. Whatever the cause of the impedance change, it will result in a reflected wave.

Rewriting the expression for the reflection coefficient

$$\Gamma = \frac{Z_2 - Z_1}{Z_2 + Z_1} \tag{11-9}$$

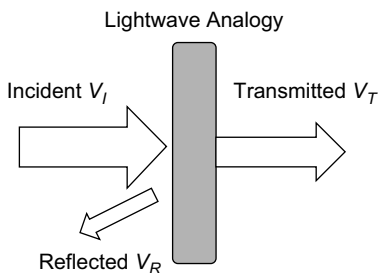


Figure 11-7 When a traveling wave, V_I , encounters an impedance change, a portion of it is reflected, V_R , while a portion of it is transmitted through the mismatch, V_T .

the portion of the voltage wave that is reflected is

$$V_R = \Gamma V_I \quad (11-10)$$

A portion of the incident voltage may be transmitted through the impedance mismatch but is modified by the amount reflected. The transmitted coefficient, T , is defined by

$$T = \frac{V_T}{V_I} = 1 + \Gamma = \frac{2 Z_2}{Z_1 + Z_2} \quad (11-11)$$

11.7 Sinusoidal Voltages

The sinusoid is a common electrical signal, so we will expand our discussion of transmission lines to include this type of waveform.

Wavelength

The wavelength of a sinusoidal electromagnetic wave in free space is given by

$$\lambda = c/f \quad (11-12)$$

where

- c = velocity of light in free space
- f = frequency of the sinusoid

However, in a transmission line the velocity of propagation must be taken into account. Thus, the wavelength of a sinusoidal voltage propagating down a transmission line is

$$\lambda = v_p/f = k_v c/f \quad (11-13)$$

where

- v_p = propagation velocity
- f = frequency of the sinusoid
- k_v = velocity factor
- c = velocity of light in free space

The slower the propagation velocity, the shorter the wavelength.

Example 11.2

What is the wavelength of a sine wave that has a frequency of 146.52 MHz in a transmission line with a 66% velocity factor?

$$\begin{aligned}\lambda &= k_v c/f = 0.66(3 \times 10^8)/(146.52 \times 10^6) \\ &= 1.35 \text{ m}\end{aligned}$$

11.8 Complex Reflection Coefficient

Sine waves are normally characterized by their magnitude and phase. To accommodate this representation, the concept of reflection coefficient is expanded to allow for the reflection coefficient as a complex number. The reflection coefficient is often shown as a magnitude and a phase angle.

$$\Gamma = \rho \angle \theta \quad (11-14)$$

Both ρ and Γ are referred to as the reflection coefficient. However, ρ is a scalar quantity, whereas Γ is a complex number.

The previously introduced definition of the reflection coefficient (reflected voltage over the incident voltage) is modified to allow complex (vector) voltages.

$$\Gamma = \frac{|V_R| \angle \theta_R}{|V_I| \angle \theta_I} = \frac{|V_R|}{|V_I|} \angle (\theta_R - \theta_I) \quad (11-15)$$

$$\Gamma = \rho \angle (\theta_R - \theta_I) \quad (11-16)$$

The magnitudes of the incident and reflected voltages do not change with position on the transmission line, and therefore ρ does not change with position. The phase of the complex reflection coefficient does change as the position changes.

The complex reflection coefficient due to an impedance mismatch (as shown in Figure 11-7) is computed using the values of the complex impedances.

$$\Gamma = \frac{Z_2 - Z_1}{Z_2 + Z_1} \quad (11-17)$$

11.9 Return Loss

Another commonly measured quantity in radio frequency systems is *return loss*. The return loss of a particular system is the scalar reflection coefficient expressed in decibels.

$$RL_{(\text{dB})} = -20 \log(\rho) \quad (11-18)$$

The minus sign in the equation causes the decibel form to indicate the amount of loss from the incident wave to the reflected wave—hence the name return loss. It is a measure of how large the reflected wave is relative to the incident wave. For example, if the return loss

is 30 dB, then a 0 dBm incident wave causes a –30 dBm reflected wave. The return loss of a system can range from 0 to ∞ dB, with 0 dB being the case where the entire incident wave is reflected and ∞ occurring when none of the incident wave is reflected.

It is common to see return loss expressed with the sign reversed. That is, measured data of return loss may be shown as negative (e.g., –30 dB). This is technically incorrect but naturally follows from how the measurement is usually made. As long as the user remembers that return loss is defined as a positive number, confusion is avoided.³

Example 11.3

A sine wave generator with an output impedance of 50 Ω drives a load impedance of 30 + j20 Ω through a 50 Ω transmission line. What is the value of the reflection coefficient and return loss at the load?

$$\Gamma = \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{(30 + j20) - 50}{(30 + j20) + 50} = \frac{28.28 \angle 135^\circ}{82.46 \angle 14.0^\circ}$$

$$\Gamma = 0.342 \angle 121^\circ \text{ or } -0.176 + j0.293$$

$$\rho = |\Gamma| = 0.342$$

$$RL_{(\text{dB})} = -20 \log(\rho) = 9.32 \text{ dB}$$

11.10 Standing Waves

When a sinusoidal signal first propagates down a transmission line, the sinusoidal voltage moves toward the load similar to how the step waveform moves down the line. When the incident voltage encounters the load, a portion of the incident wave is reflected, according to the complex reflection coefficient (Figure 11-8). This reflected wave travels back toward the generator and may also be reflected again when the generator is reached, depending on the generator's impedance. All of this is very similar to the behavior of the step waveform.

Something different happens with a sinusoidal signal. The incident voltage and the reflected voltage are both sine waves. When they intersect, going up and down the transmission line, an interference pattern is set up. The envelope of the sinusoidal voltage will remain in a constant shape called a *standing wave*, as shown in the Figure 11-9. The envelope (or magnitude) of the voltage varies with distance down the line but the voltage at each point on the line varies sinusoidally.

The *voltage standing wave ratio* (VSWR) or simply the *standing wave ratio* (SWR) is the ratio of the maximum and minimum of the envelope.

$$SWR = V_{\max}/V_{\min} \quad (11-19)$$

The *SWR* is always greater than or equal to one, with 1.0 being the case where no mismatch occurs. In this case, V_{\max} is equal to V_{\min} since no reflections occur. A properly loaded transmission line is often called a “flat” line, referring to the lack of standing waves.

³ See Bird (2009) for more information.

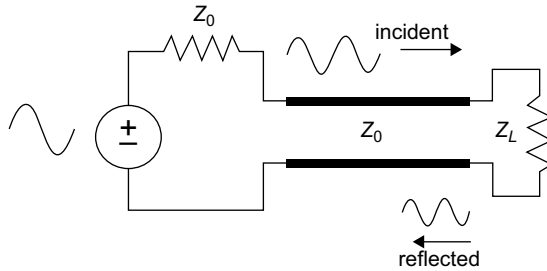


Figure 11-8 Sinusoidal signals also experience reflections on a transmission line.

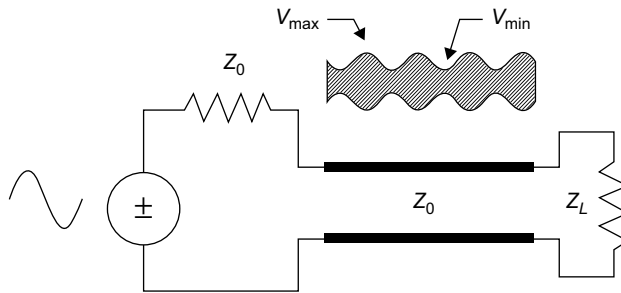


Figure 11-9 The envelope of the voltage on a transmission line will form standing waves.

The maximum of the envelope occurs when the incident and reflected voltages add constructively. Similarly, the minimum of the envelope occurs where the incident and reflected voltages add destructively.

$$V_{\max} = |V_I| + |V_R| \quad (11-20)$$

and

$$V_{\min} = |V_I| - |V_R| \quad (11-21)$$

Thus, the *SWR* can be determined from the incident and reflected voltages as well as the scalar reflection coefficient.

$$SWR = \frac{|V_I| + |V_R|}{|V_I| - |V_R|} \quad (11-22)$$

$$SWR = \frac{1 + \rho}{1 - \rho} \quad (11-23)$$

Also,

$$\rho = \frac{SWR - 1}{SWR + 1} \quad (11-24)$$

Example 11.4

A 50 Ω sine wave generator drives a 100 Ω load through a 50 Ω transmission line. What are the values of the reflection coefficient and standing wave ratio? If the incident wave has a peak voltage of 4 V, determine the minimum and maximum envelope voltages occurring on the line.

$$\begin{aligned}\Gamma &= \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{100 - 50}{100 + 50} = 0.33 \\ \rho &= |\Gamma| = 0.33 \\ SWR &= \frac{1 + \rho}{1 - \rho} = 2 \\ &= \frac{|V_I| + |V_R|}{|V_I| - |V_R|}\end{aligned}$$

so

$$|V_R| = \frac{|V_I|(SWR + 1)}{(SWR + 1)} = \frac{4(2 - 1)}{(2 + 1)} = 1.33 \text{ V}$$

The maximum and minimum envelope voltages are

$$\begin{aligned}V_{\max} &= |V_I| + |V_R| = 4 + 1.33 = 5.33 \text{ V} \\ V_{\min} &= |V_I| - |V_R| = 4 - 1.33 = 2.67 \text{ V}\end{aligned}$$

Example 11.5

What are the values of reflection coefficient, return loss, and SWR for the special cases of a shorted load and an open load?

Shorted Load:

$$\begin{aligned}\Gamma &= \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{0 - Z_0}{0 + Z_0} = -1 \\ \rho &= 1, RL = -20 \log(1) = 0 \text{ dB} \\ SWR &= \frac{1 + \rho}{1 - \rho} = \infty\end{aligned}$$

Open Load:

$$\begin{aligned}\Gamma &= \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{\infty - Z_0}{\infty + Z_0} = 1 \\ \rho &= 1, RL = -20 \log(1) = 0 \text{ dB} \\ SWR &= \frac{1 + \rho}{1 - \rho} = \infty\end{aligned}$$

Note that both load conditions result in an infinite SWR and 0 dB return loss, although the sign of the complex reflection coefficient is different.

Table 11-1 Reflection Coefficient, Return Loss, and Standing Wave Ratio

Reflection Coefficient	Return Loss	Standing Wave Ratio
1.00	0.00	∞
0.90	0.92	19.00
0.80	1.94	9.00
0.70	3.10	5.67
0.60	4.44	4.00
0.50	6.02	3.00
0.40	7.96	2.33
0.30	10.46	1.86
0.20	13.98	1.50
0.10	20.00	1.22
0.09	20.92	1.20
0.08	21.94	1.17
0.07	23.10	1.15
0.06	24.44	1.13
0.05	26.02	1.11
0.04	27.96	1.08
0.03	30.46	1.06
0.02	33.98	1.04
0.01	40.00	1.02
0.00	∞	1.00

Table 11-1 gives values for reflection coefficients, return losses, and standing wave ratios.

For the special case where both Z_0 and Z_L are real, we can calculate SWR easily by taking the ratio of the two impedances. If $Z_L > Z_0$, then

$$\rho = |\Gamma| = \left| \frac{Z_L - Z_0}{Z_L + Z_0} \right| = \frac{|Z_L - Z_0|}{|Z_L + Z_0|} \tag{11-25}$$

$$SWR = \frac{1 + \rho}{1 - \rho} = \frac{Z_L}{Z_0} \quad \text{for } Z_L > Z_0, \text{ both real} \tag{11-26}$$

Also

$$SWR = \frac{Z_0}{Z_L} \quad \text{for } Z_L < Z_0, \text{ both real} \tag{11-27}$$

11.11 Input Impedance of a Transmission Line

When an incident wave first encounters a transmission line, it sees an impedance of Z_0 . As it propagates down the line, a portion of the incident wave may be reflected back toward the generator end of the line. When this wave encounters the generator end, it will affect the voltage at that end of the line. This also means that the impedance seen looking into the end

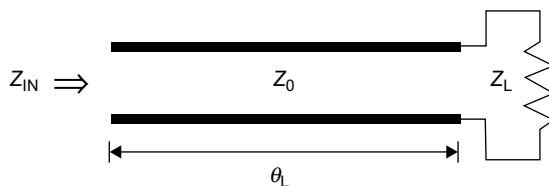


Figure 11-10 Z_{IN} is the impedance looking into the end of a transmission line with a load Z_L at the end of it.

of the line will depend on the magnitude and phase of the reflections, and it will no longer be simply Z_0 . For the situation shown in Figure 11-10, the input impedance looking into the line is⁴

$$Z_{IN} = Z_0 \frac{Z_L + j Z_0 \tan \theta_L}{Z_0 + j Z_L \tan \theta_L} \quad (11-28)$$

where

θ_L = distance from the load, expressed as an angle (degrees or radians)

The angle representing the distance from the load may be found from the physical distance.

$$\theta_L = \frac{360 d}{\lambda} \quad (11-29)$$

where

d = distance from the load

λ = wavelength

θ_L = in units of degrees

The velocity factor must be accounted for when computing the wavelength.

Matched System

For the special case of a perfectly matched system, $Z_L = Z_0$, the Z_{IN} equation reduces to

$$Z_{IN} = Z_0 \frac{Z_L + j Z_0 \tan \theta_L}{Z_0 + j Z_L \tan \theta_L} \Big|_{Z_L=Z_0} = Z_0 \quad (11-30)$$

Since there are no reflections in a perfectly matched system, the impedance looking into the end of the line just equals Z_0 .

⁴ This equation is adapted from Hayt (1974).

Example 11.6

For $f = 50$ MHz, what is the impedance looking into the end of a 1 m length of 50Ω transmission line with a 70% velocity factor, terminated in 25Ω ?

$$\begin{aligned} \lambda &= k_v c/f = 0.70(3 \times 10^8)/(50 \times 10^6) = 4.2 \text{ m} \\ \theta_L &= 360 d/\lambda = 360(1)/4.2 = 85.7^\circ \\ Z_{IN} &= Z_0 \frac{Z_L + j Z_0 \tan \theta_L}{Z_0 + j Z_L \tan \theta_L} \\ &= 50 \frac{25 + j 50 \tan(85.7^\circ)}{50 + j 25 \tan(85.7^\circ)} \\ &= 98.9 \Omega \angle 6.4^\circ \text{ or } 98.3 + j 11.0 \Omega \end{aligned}$$

Not only is the input impedance larger than either Z_0 or Z_L , the impedance has an imaginary component while Z_0 and Z_L are both real. (This illustrates the transformer action possible with a transmission line, as the 25Ω load impedance is transformed up to approximately 100Ω .)

11.12 Measurement Error Due to Impedance Mismatch

In a measurement situation, the generator or source may be an instrument (such as a signal generator) or the device under test. The load is the input impedance of a measuring instrument such as a power meter, spectrum analyzer, or network analyzer. Often the output impedance of the source, transmission line, and analyzer are all nominally Z_0 (Figure 11-11). However, the exact value of each of these impedances may vary somewhat, causing slight mismatches in the system and errors in the measurement.

CASE 1. PERFECT SOURCE, IMPERFECT LOAD

As shown in previous examples, the incident voltage from the source will propagate to the load and a portion of it will be reflected back. The reflected voltage will travel back to the source and will be absorbed there if the source is perfectly matched. The major portion of the incident wave is transmitted into the load, which is to say it is measured by the instrument.

This mismatch effect will be examined from a power transfer point of view. The incident wave traveling toward the load has a power associated with it.

$$P_I = V_I^2/Z_0 \tag{11-31}$$

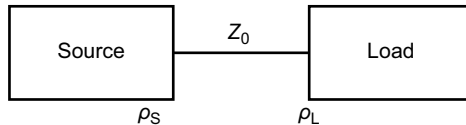


Figure 11-11 Measurement diagram for discussing mismatch loss and mismatch uncertainty.

The power reflected back from the load is

$$P_R = V_R^2/Z_0 = (\rho_L V_I)^2/Z_0 \quad (11-32)$$

The power delivered to the load must be the difference between these two powers.

$$P_L = P_I - P_R = \frac{V_I^2}{Z_0}(1 - \rho_L^2) = P_I(1 - \rho_L^2) \quad (11-33)$$

Ideally, all the incident power would be delivered to the load, but the mismatch causes some amount of power loss. The *mismatch loss*⁵ is the power transfer to the load, relative to the incident power, expressed in dB.

$$\text{mismatch loss} = -10 \log(P_L/P_I) = -10 \log(1 - \rho_L^2) \quad (11-34)$$

Example 11.7

A power meter with an *SWR* of 1.2 is used to measure the power at the end of a transmission line. How much error will be introduced in the measurement due to the mismatch at the power meter?

First, we need to compute the reflection coefficient from the *SWR*.

$$\rho = \frac{SWR - 1}{SWR + 1} = \frac{1.2 - 1}{1.2 + 1} = 0.091$$

The mismatch loss is given by

$$\begin{aligned} \text{mismatch loss} &= -10 \log(1 - \rho_L^2) \\ &= -10 \log(1 - 0.091^2) = 0.036 \text{ dB} \end{aligned}$$

The power meter will read too low by 0.036 dB.

The mismatch loss has been derived in terms of a perfect source driving an imperfect load. The same effect exists for the case where the load impedance is exactly Z_0 and the source has a non- Z_0 impedance. The mismatch loss is computed the same way but using the source's reflection coefficient (not the load's).

CASE 2. IMPERFECT SOURCE, IMPERFECT LOAD

Now consider the case where the source also has an output impedance that is not exactly Z_0 . In addition to the error described in Case 1, an additional error will be introduced due to source mismatch. The mismatch loss at the source is described mathematically as $(1 - \rho_S^2)$,

⁵ Beatty (1964) further refines this definition, calling it Z_0 mismatch loss to differentiate it from other possible definitions.

where ρ_S is the source reflection coefficient. Note the similarity of this term to the load mismatch equation of Case 1.

There is another source of error when both the load and source are not perfectly matched. When the reflection from the load travels back to the source, instead of being absorbed at the source it is reflected back again to the load. It adds constructively or destructively at the load, depending on the phase of the signal. This double reflection adds another term to the equation, which will be derived shortly. Additional reflections also occur, with each reflected wave being smaller than the previous one. For source and load impedances reasonably close to Z_0 , the additional reflections are much smaller than the first load/source round-trip reflection. Since the traveling waves in the system may add vectorally, they will be analyzed as voltage waveforms and then converted to power.

The incident wave, V_S , leaves the source and travels to the load. At the load end, an incident voltage, V_I , gets reflected back to the source, which reflects it again to the load, which produces a second incident wave, $\rho_S\rho_L V_I$. There are really two waves incident at the load: the first direct wave from the source and the doubly reflected wave. Mathematically, we can express this as

$$V_I = V_S \pm \rho_S\rho_L V_I \quad (11-35)$$

Solving for V_I/V_S ,

$$\frac{V_I}{V_S} = \frac{1}{1 \pm \rho_S\rho_L} \quad (11-36)$$

The scalar reflection coefficient is used here since we usually don't know the phase of the reflection and after the signal travels an unknown length of cable the phase relationship is lost anyway. The uncertainty in the sign of the reflected term indicates that we don't know whether the reflected wave will add constructively or destructively.

Taking the square of equation (11-33) and combining it with the mismatch losses at the source and load gives the complete power transfer function.

$$\frac{P_L}{P_S} = \frac{(1 - \rho_S^2)(1 - \rho_L^2)}{(1 \pm \rho_S\rho_L)^2} \quad (11-37)$$

The numerator of the equation indicates the effect of the mismatch loss. These two terms are deterministic in the sense that a given reflection coefficient will cause a corresponding loss in power to the load. On the other hand, the denominator represents a *mismatch uncertainty*, with the uncertainty in the power transfer bounded by taking the + or – sign in the denominator. The actual power transfer can fall anywhere in between these two extremes. From a measurement point of view, we concentrate on the maximum and minimum power transfer, which represents the maximum and minimum error that can be incurred due to impedance matching problems.

Taking the decibel form of the equation allows it to be easily broken up into the individual error mechanisms.

$$10 \log\left(\frac{P_L}{P_S}\right) = 10 \log(1 - \rho_S^2) + 10 \log(1 - \rho_L^2) + 20 \log(1 \pm \rho_S\rho_L) \quad (11-38)$$

Example 11.8

Determine the worst-case error, expressed in decibels, due to mismatches when a source with a 10 dB return loss drives a lossless transmission line connected to a power meter having a return loss of 20 dB.

$$\rho_S = 10^{(RL/-20)} = 10^{(10/-20)} = 0.32$$

$$\rho_L = 10^{(RL/-20)} = 10^{(20/-20)} = 0.1$$

The mismatch loss due to the source is

$$10 \log(1 - \rho_S^2) = 10 \log(1 - 0.32^2) = -0.469 \text{ dB}$$

The mismatch loss due to the load is

$$10 \log(1 - \rho_L^2) = 10 \log(1 - 0.1^2) = -0.0436 \text{ dB}$$

The mismatch uncertainty due to the double reflection is

$$20 \log(1 \pm \rho_S \rho_L) = 20 \log[1 \pm (0.32)(0.1)] = -0.282 \text{ dB}, +0.274 \text{ dB}$$

The total error is bounded by

$$-0.469 - 0.0436 + 0.274 = -0.239 \text{ dB}$$

$$-0.469 - 0.0436 - 0.282 = -0.795 \text{ dB}$$

11.13 Insertion Gain and Loss

Measurement of insertion gain or loss is shown in Figure 11-12. The output level of a signal generator is measured by a power meter. The device under test (DUT) is then inserted between the generator and the power meter. The gain or loss of the device is determined by taking the ratio of the output power to the generator power or, in decibels, the input power is subtracted from the output power.

$$\text{insertion loss (dB)} = 10 \log(P_{\text{REF}}/P_{\text{MEAS}}) \tag{11-39}$$

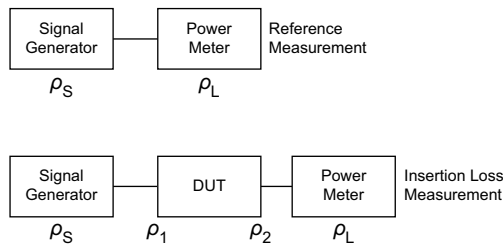


Figure 11-12 Measurement of insertion gain (or loss) can be accomplished with a signal generator and a power meter.

where

P_{REF} = reference power

P_{MEAS} = power measured at the DUT output

The insertion loss equation can be inverted to produce insertion gain:

$$\text{insertion gain (dB)} = 10 \log(P_{\text{MEAS}}/P_{\text{REF}}) \quad (11-40)$$

The reference power is determined by

$$P_{\text{REF}} = K_M P_S \quad (11-41)$$

where

K_M = constant corresponding to the mismatch error in the power meter

P_S = power out of the source

$$P_{\text{MEAS}} = K_M P_S K_{\text{DUT}} \quad (11-42)$$

where

K_{DUT} = power transfer through the DUT

The insertion gain can be expressed in terms of these two equations:

$$\begin{aligned} \text{insertion gain (dB)} &= 10 \log(K_M P_S K_{\text{DUT}}/K_M P_S) \\ &= 10 \log(K_{\text{DUT}}) \end{aligned} \quad (11-43)$$

The result is that the insertion gain measurement depends only on the DUT's transfer characteristics and not on the mismatch error at the meter or the source's power level. This is an idealized view of the insertion gain/loss measurement, appropriate when there is a good Z_0 match at both ports of the DUT.

Example 11.9

Determine the insertion loss and insertion gain in decibels for the following measurement. The reference power was measured at 120 mW, and the DUT power was measured as 20 mW.

$$\begin{aligned} \text{insertion loss (dB)} &= 10 \log(P_{\text{REF}}/P_{\text{MEAS}}) \\ &= 10 \log(0.120/0.020) \\ &= 7.78 \text{ dB} \end{aligned}$$

The insertion gain is just the negative of the insertion loss, or -7.78 dB.

Errors Due to Impedance Mismatch

Mismatch errors will contribute to the error in insertion gain measurements. First, consider the reference measurement. A mismatch loss occurs at the output of the signal generator and

the input to the power meter, both due to their imperfect Z_0 impedances. The mismatch loss at the signal generator is consistent between the reference and insertion loss measurement and is considered to be included in P_S , which has been shown to be removed from the insertion loss calculation. Similarly, the mismatch loss at the power meter is included in K_M and has already been shown to be removed from the insertion loss calculation.

Another mismatch error occurs, namely, the mismatch uncertainty due to the reflection off the power meter being reflected again at the signal generator. The doubly reflected wave ends up again at the power meter and introduces an error into the reference measurement. Since this same error mechanism does *not* occur during the insertion gain measurement, it will introduce an error into the insertion gain calculation. Including this error in the reference measurement gives a more accurate equation for P_{REF} . (Note that the mismatch uncertainty has the same form as equation (11-36) except that it is squared to be consistent with the use of power instead of voltage.)

$$P_{\text{REF}} = \frac{K_M P_S}{(1 \pm \rho_S \rho_L)^2} \quad (11-44)$$

The insertion gain measurement has a different set of reflections due to the imperfect Z_0 impedances. There are double reflections between the signal generator and the DUT (ρ_S, ρ_1) and between the DUT and the power meter (ρ_2, ρ_L). These two sets of reflections introduce two more sources of measurement uncertainty, and the resulting measured power is

$$P_{\text{MEAS}} = \frac{K_M P_S K_{\text{DUT}}}{(1 \pm \rho_S \rho_1)^2 (1 \pm \rho_2 \rho_L)^2} \quad (11-45)$$

There is one more source of mismatch uncertainty. The reflected wave from the power meter can pass through the DUT and be reflected back from the signal generator. This reflection passes through the DUT again and is finally incident on the power meter, introducing a mismatch uncertainty. This mechanism will be ignored here—if the round-trip loss through the DUT is greater than 10 dB, its effect is negligible.⁶

Now that the mismatch uncertainties have been included in the power measurements, consider again the insertion gain measurement.

$$\begin{aligned} \text{insertion gain (dB)} &= 10 \log(P_{\text{MEAS}}/P_{\text{REF}}) \\ &= 10 \log \left[\frac{K_M P_S K_{\text{DUT}} (1 \pm \rho_S \rho_L)^2}{K_M P_S (1 \pm \rho_S \rho_L)^2 (1 \pm \rho_2 \rho_L)^2} \right] \end{aligned} \quad (11-46)$$

$$\begin{aligned} &= 10 \log(K_{\text{DUT}}) + 20 \log(1 \pm \rho_S \rho_L) \\ &\quad - 20 \log(1 \pm \rho_S \rho_L) - 20 \log(1 \pm \rho_2 \rho_L) \end{aligned} \quad (11-47)$$

The first term of the equation is the ideal result for insertion gain, and the remaining terms represent the mismatch uncertainty in the measurement.

$$\text{mismatch uncertainty} = 20 \log(1 \pm \rho_S \rho_L) - 20 \log(1 \pm \rho_S \rho_L) - 20 \log(1 \pm \rho_2 \rho_L) \quad (11-48)$$

⁶ See Adam (1969) and Hewlett-Packard (1978) for a discussion of this effect.

Example 11.10

The insertion loss of an attenuator is measured using a signal generator and power meter. Determine the mismatch uncertainty in measuring the 10 dB attenuator with an SWR at each port of 1.5. The signal generator and power meter have return losses of 20 dB and 30 dB, respectively.

First, compute the reflection coefficient of each device.

$$\rho = \frac{SWR - 1}{SWR + 1}$$

$$\rho_1 = \rho_2 = \frac{1.5 - 1}{1.5 + 1} = 0.20$$

$$\rho_S 10^{(-RL/20)} = 10^{(-20/20)} = 0.10$$

$$\rho_L = 10^{(-RL/20)} = 10^{(-30/20)} = 0.0316$$

Now compute the mismatch uncertainty.

$$\begin{aligned} \text{mismatch uncertainty} &= 20 \log(1 \pm \rho_S \rho_L) - 20 \log(1 \pm \rho_S \rho_1) \\ &\quad - 20 \log(1 \pm \rho_S \rho_L) \\ &= 20 \log[1 \pm (0.1)(0.0316)] - 20 \log[1 \pm (0.10)(0.20)] \\ &\quad - 20 \log[1 \pm (0.20)(0.0316)] \\ &= (+0.0274, -0.0275) + (+0.175, -0.172) \\ &\quad + (+0.0551, -0.0547) \\ \text{mismatch uncertainty} &= +0.258 \text{ dB}, -0.254 \text{ dB} \end{aligned}$$

Therefore, the errors due to mismatch uncertainty are bounded by +0.258 dB and -0.254 dB, for a total uncertainty of 0.512 dB. Note that the mismatch uncertainty is not the same in both directions, but with small mismatch uncertainty the two limits have approximately the same magnitude.

Now consider the errors internal to the power meter (instrument accuracy). The *absolute* accuracy of the power meter is not important as long as the *relative* accuracy of the meter is good. In other words, the meter does not need to be able to determine whether the signal power is exactly a certain power level, but it does need to measure accurately *changes* in power level. In the insertion gain measurement, the change in power level that must be measured is the change caused by the insertion of the device under test into the measurement system. The measurement is accurate to the extent that the meter correctly measures this change.

The signal generator may introduce an error due to instrument drift. If the signal generator is slightly off in signal level, it will not affect the measurement accuracy since the power meter measures the signal level anyway. However, the signal generator's output power must be stable so that it does not change between the time the reference measurement

and the insertion loss measurement takes place. Such a change in output power would introduce an error classified as instrument drift.

11.14 Line Losses

Lossless transmission lines have been assumed thus far, which is a good approximation for many situations. If high-quality cables are used, frequencies are low and cable length is short, other error mechanisms in the measurement system will dominate. The longer the cable becomes and the higher the frequency, the more attention needs to be paid to the cable loss. At microwave frequencies, the loss of even high quality cables can be significant.

Manufacturers normally specify the loss in their cables in dB, often in dB per hundred feet.

11.15 Coaxial Lines

Transmission lines are available in a variety of physical configurations, but coaxial transmission lines are most commonly used in measurement applications (Figure 11-13). The center conductor is surrounded by a dielectric material, which is surrounded by the outer conductor shield. The characteristic impedance is given by

$$Z_0 = \frac{138}{\sqrt{\epsilon}} \log(D/d) \quad (11-49)$$

where

- D = inner diameter of the shield
- d = outer diameter of the center conductor
- ϵ = dielectric constant of the dielectric material

For air, the dielectric constant is equal to 1.

The coaxial structure inherently provides shielding from external electromagnetic fields and can result in transmission lines that can be moved and flexed somewhat without causing the characteristic impedance to change. Not all coaxial lines are flexible since some of the

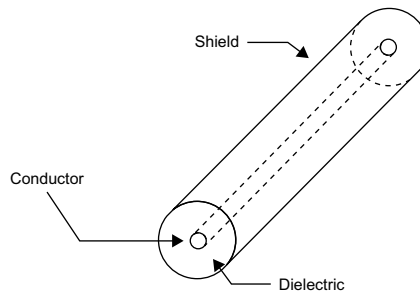


Figure 11-13 The coaxial transmission line is commonly used in measurement applications.

highest-quality lines, often with air dielectric, are fabricated with rigid or semirigid outer conductors.

The most common impedances in high-frequency measurement applications are 50 and 75 Ω . The 50 Ω impedance is most common due to its good compromise between power handling capability and loss. The 75 Ω impedance is found more in telecommunications applications where its low-loss characteristics are important.

Bibliography

Adam, Stephen F. *Microwave Theory and Applications*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1969.

Beatty, Robert W. "Insertion Loss Concepts," *Proceedings of the IEEE*, June 1964.

Bird, Trevor S. "Definition and Misuse of Return Loss," *IEEE Antennas and Propagation Magazine*, Vol. 51, No. 2, April 2009.

Hayt, William H., Jr. *Engineering Electromagnetics*, 3rd ed. New York: McGraw-Hill Book Company, 1974.

Hayward, W. H. *Introduction to Radio Frequency Design*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1982.

Hewlett-Packard Company. "High Frequency Swept Measurements," Application Note 183, December 1978.

Laverghetta, Thomas S. *Practical Microwaves*. Indianapolis, IN: Howard W. Sams & Co., Inc., 1984.

Van Valkenburg, M., and W. Middleton (ed.), *Reference Data for Radio Engineers*, 9th ed. Woburn, MA: Butterworth-Heinemann, 2002.

Measurement Connections

Connecting an instrument to a device under test (DUT) invariably involves disturbing that device. When making precision measurements, it is desirable to minimize loading and other effects so that the measurement is not corrupted by the measuring instrument. Probes, attenuators, impedance matching devices, and filters are used to couple the signal of interest into the instrument in the most efficient and accurate manner.

12.1 The Loading Effect

Any attempt to measure a voltage in a circuit will change that voltage. Consider the circuit shown in Figure 12-1. The circuit under test is modeled as a voltage source with some internal impedance, Z_S . The open circuit voltage of the circuit is V_S since no current can flow through Z_S under open circuit conditions. When load impedance, Z_L , is connected to the circuit, the situation changes. By the voltage divider relationship

$$V_L = \frac{V_S Z_L}{(Z_S + Z_L)} \quad (12-1)$$

Unless Z_S is equal to zero or Z_L is equal to infinity, V_L will always be less than V_S .

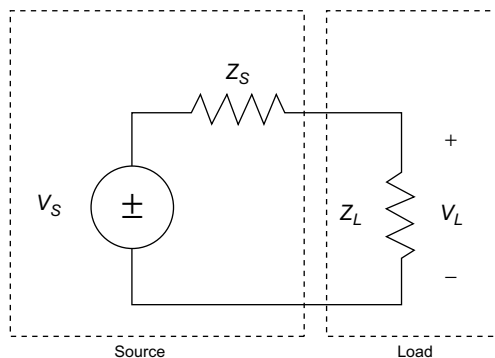


Figure 12-1 A finite impedance voltage source is attached to a resistive load.

12.2 Maximum Voltage and Power Transfer

Some electronic systems are designed to maximize the voltage transfer in the system. To maximize the voltage transfer, Z_L should be much larger than Z_S . If Z_L is infinite, the desired result of $V_L = V_S$ occurs.

Other electronic systems are designed to maximize the power transfer in the system. The power dissipated in the load impedance of Figure 12-1 is given by

$$P_L = \frac{V_S^2 Z_L}{(Z_S + Z_L)^2} \quad (12-2)$$

It can be shown that the power in the load is maximized when¹

$$Z_L = Z_S^\star \quad (12-3)$$

(★ indicates complex conjugate)

If the impedances are real, then the complex conjugate designation can be dropped and power is maximized when both of the impedances (resistances) are the same.

As discussed in Chapter 11, transmission lines are often used to transfer the measured signals, especially at high frequencies. By keeping all input and output impedances the same value as the characteristic impedance of the transmission line, we can avoid reflections and deliver maximum power transfer.

12.3 High-Impedance Inputs

If our measuring instrument has a high impedance relative to the DUT, we can measure a given voltage with minimal loading. In most measurement situations, Z_S is predetermined since it is a function of the circuit being measured and Z_L must be much larger than Z_S .

Spectrum and network analyzers that cover frequencies below 10 MHz sometimes provide high-impedance inputs. Typically, these inputs can be modeled by a 1 MΩ resistor in parallel with a small capacitor (typically 30 pF). This type of input is very similar to the high-Z inputs of the conventional oscilloscope. At low frequencies, the input impedance is 1 MΩ, which is sufficiently large for most applications. As the frequency increases, the parallel capacitance becomes dominant and reduces the equivalent input impedance of the instrument. The instrument user must be careful to not assume that a “high-impedance” input is high impedance for all frequencies. For example, at 10 MHz, the impedance of a 30 pF capacitor is only 530 Ω. Besides causing a reduction in amplitude, the high-impedance input can introduce a phase shift due to the parallel capacitance.

High-Impedance Probes

Standard oscilloscope probes can be used with high-impedance analyzer inputs to provide convenient probing of circuit nodes (Figure 12-2). A 1X or 1:1 probe has no attenuation and

¹ This assumes that Z_S is nonzero and Z_L is to be chosen. Otherwise $Z_S = 0$ would be a good choice.

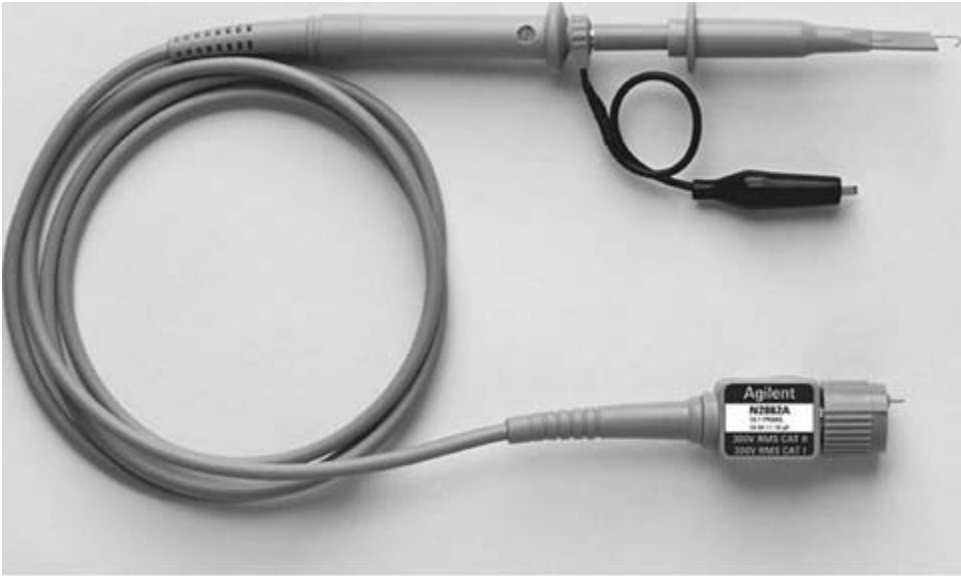


Figure 12-2 A typical 10:1 high-impedance oscilloscope probe. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

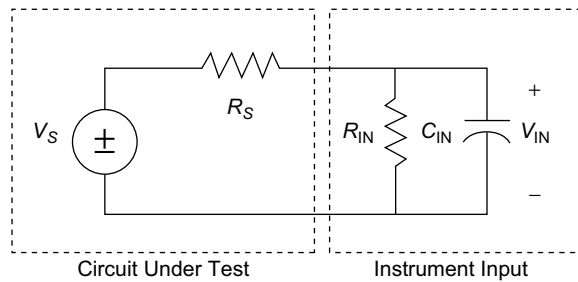


Figure 12-3 The circuit under test and a high-impedance instrument input produce a single-pole low-pass transfer function.

is roughly equivalent to connecting the instrument input to the circuit under test with a shielded cable. The circuit model is shown in Figure 12-3.

The measured voltage is

$$V_{IN} = V_S \frac{R_{IN}}{R_{IN} + R_S} \frac{1}{1 + j2\pi f C_{IN}(R_{IN} \parallel R_S)} \quad (12-4)$$

The input capacitance, C_{IN} creates a pole in the transfer function, causing V_{IN} to decrease at high frequencies. The magnitude of the transfer function is reduced by 3 dB at $f = 1/[2\pi(R_{IN} \parallel R_S)C_{IN}]$. Note that this frequency (essentially the bandwidth of the system) depends on R_{IN} , C_{IN} , and R_S . Normally, R_{IN} is much larger than R_S , so R_S dominates.

While C_{IN} is part of the measuring instrument, R_S is the equivalent output impedance of the circuit under test. Thus, the impedance of the node being measured will influence the bandwidth of the measurement.

Attenuating Probes

The bandwidth-limiting effect of the analyzer's input capacitance can be compensated for at the price of some attenuation of the input signal. An attenuating probe includes a resistor and capacitor in the signal path (Figure 12-4).² The voltage delivered to the analyzer input is

$$V_{IN} = V_S \frac{R_{IN}(j2\pi f R_P C_P + 1)}{R_{IN}(j2\pi f R_P C_P + 1) + R_P(j2\pi f R_{IN} C_{IN} + 1)} \quad (12-5)$$

If $R_P C_P = R_{IN} C_{IN}$ then the equation reduces to

$$V_{IN} = V_S \frac{R_{IN}}{R_{IN} + R_P} \quad (12-6)$$

Under this condition, the effect of the input capacitance is canceled and other parameters such as cable capacitance will limit the probe bandwidth. The loading on the device under test is decreased, since the DUT sees a higher probe impedance and smaller loading capacitance. For a 10X or 10:1 probe, R_P is chosen to be 9 times R_{IN} ; V_{IN} is one-tenth of V_S . Any particular model of probe is designed for a certain range of input capacitance and since input capacitance will vary with the design of the instrument, the probe must be selected to match the input.

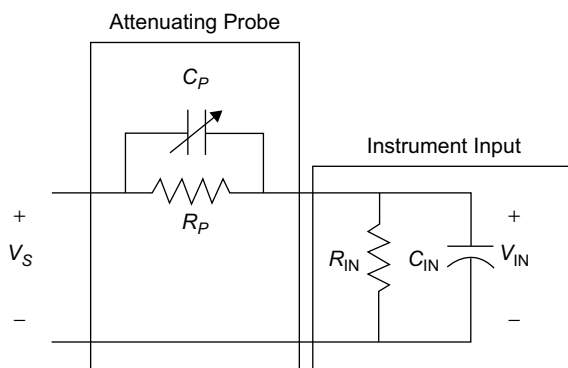


Figure 12-4 An attenuating probe will cancel out the effect of the instrument's input capacitance, producing a higher measurement bandwidth.

C_P is made variable to allow the user to match precisely the probe to the input. When used with an oscilloscope, the probe is compensated (tweaked) by optimizing the pulse response of the system. In spectrum and network analyzer applications, a probe can be

² This is a simplified probe circuit. Practical probe circuits may be arranged differently and may have additional circuit components.

compensated by adjusting it for the flattest possible frequency response, using a tracking generator or signal generator with flat frequency response.

The 10:1 probe is the most common attenuating probe, supplying 20 dB of attenuation. Other attenuation factors are possible, with each trading off increased signal attenuation for increased system bandwidth.

12.4 Active High-Impedance Probes

Most attenuating oscilloscope probes have a bandwidth of less than 1 GHz. To probe signals with higher frequency content, an *active probe* can be used. An active probe is designed to drive the Z_0 input of the instrument. It uses an amplifier with low input capacitance and wide bandwidth to sense wideband signals with minimal loading. For example, the probe shown in Figure 12-5 has a bandwidth of 12 GHz and an input impedance of 0.35 pF in parallel with 25 K Ω . Active probes may be *single ended* (one probe connection goes to ground) or *differential* (both probe connections can measure away from ground). A differential active probe allows measurement of a differential signal, although the analyzer has a single grounded input.



Figure 12-5 An active probe is designed with an amplifier having very low input capacitance so that wide bandwidth signals can be probed with minimal loading. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

12.5 Z_0 Impedance Inputs

At higher frequencies (above ~10 MHz), stray capacitance and other effects seriously degrade the performance of high-impedance inputs. Although high-impedance inputs may be present on high-frequency analyzers, for quality measurements a Z_0 input impedance is used.

The purpose of these Z_0 inputs is not so much to provide maximum power to the analyzer, but to provide a Z_0 match for systems that need to be matched during the measurement. Many circuits such as filters, amplifiers, attenuators, and oscillators need to see a Z_0 load to function properly. The analyzer is usually connected to the device under test via a Z_0 impedance transmission line, so a Z_0 input properly terminates the line.

Although the analyzer may function as a Z_0 load, it usually is not capable of handling large power levels. The input voltage or power to the analyzer must not exceed its recommended rating. A *power attenuator* or *attenuating coupler* is designed to handle large signal levels and may be used to reduce the power present at the analyzer's input.

12.6 Input Connectors

A variety of connector types are found on network and spectrum analyzer inputs, depending on the accuracy and frequency range of the instrument. A connector must introduce very little impedance mismatch to enable an accurate measurement. A connector's impedance varies depending on the frequency of operation. Connectors that are perfectly acceptable at low frequencies may perform miserably at 1 GHz. Connector repeatability is also important since it will limit the repeatability of the measurement. This also has implications about the quality of the instrument calibration since connector repeatability errors will occur during the calibration procedure.

For analyzers with upper frequency limits less than 40 MHz, the bayonet Neill Concelman (BNC)³ connector is often used (Figure 12-6). The bayonet-style locking mechanism provides a quick and convenient means of attaching and removing the connectors. The BNC

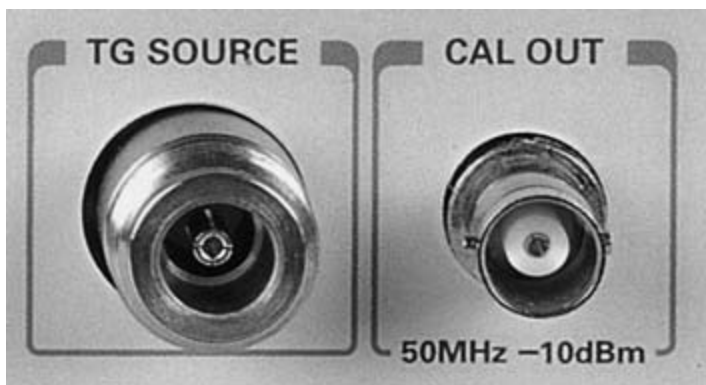


Figure 12-6 The two most common connectors used on spectrum and network analyzers are the Type N (left) and BNC (right). (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

³ The BNC connector is named for its inventors, Paul Neill and Carl Concelman.

connector's return loss degrades at higher frequencies, but the BNC fills the role of general-purpose connector for noncritical inputs and outputs. The BNC is considered usable to 11 GHz and is available in both 50 and 75 Ω versions.

For precision radio frequency measurements, the type N (Neill) connector is widely used. Found in both 50 and 75 Ω versions, this connector is much larger than the BNC and works well beyond 10 GHz. The threaded coupling mechanism provides good repeatability and is often used below microwave frequencies for this reason. The 50 and 75 Ω versions are not the same, and accidentally mixing them can cause damage to the connectors.

At microwave frequencies, the choice of connector is even more important. Examples of connectors used in this frequency range are the APC-7 (7 mm), the APC-3.5, the SMA, and the SMB.

12.7 Z_0 Terminations

In many measurement situations, it is important that all ports of the DUT are properly terminated. This means that the device must see the correct (usually Z_0) impedance at the port. A Z_0 termination is basically just a high-quality resistor conveniently packaged with the appropriate connector. A *feedthrough termination* is one that has a connector at both ends. Such a connector may be used to connect a high-impedance input to a DUT (Figure 12-7). The instrument input impedance is presumably much higher than Z_0 , and the device under test sees roughly a Z_0 impedance.

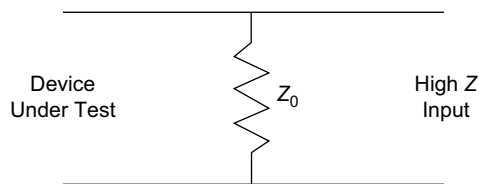


Figure 12-7 A feedthrough termination can be used to connect a Z_0 device with a high-impedance input.

12.8 Power Dividers and Splitters

Power dividers are used to provide a common signal to multiple ports, such as multiple DUTs or instrument inputs. Most power dividers are two-way dividers (providing two outputs), but dividers with additional outputs are also available.

Three-resistor power dividers can be configured in one of two ways (Figure 12-8). These two circuits are totally equivalent from an external point of view. If each port is loaded by a Z_0 resistor to ground, then the impedance looking into each port of the power splitter is also Z_0 . (This is a simple example of a circuit that must be properly terminated at each port.)

This power divider is a symmetrical design, so any port can be considered an input or output. In addition, the divider can also be used to combine two signal sources into one output (may be referred to as *power combiner*). This power divider introduces a nominal loss of 6 dB between any two ports.

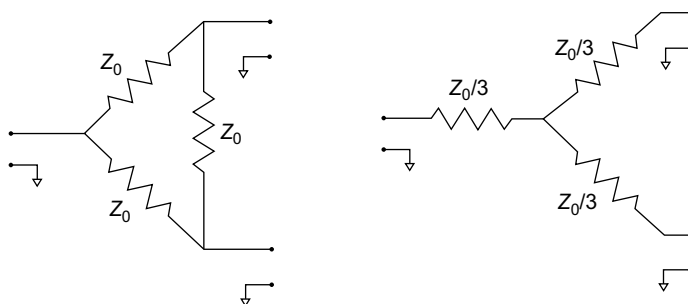


Figure 12-8 Circuit diagrams for the two types of three-resistor power dividers.

Power Splitters

The two-resistor power divider, also called a *power splitter*, is shown in Figure 12-9.⁴ This power splitter has only two resistors and is *not* symmetrical, so the input and two outputs are labeled in the circuit diagram. This type of splitter should be used in leveling and ratio applications. The nominal loss through the splitter is 6 dB.

Figure 12-10 shows the two-resistor splitter driven by a voltage source with Z_0 impedance, with a Z_0 load on each output. The impedance looking into the input of the splitter is Z_0 in this configuration. However, the impedance looking back into the output ports of the splitter is $1.67 Z_0$, not a perfect Z_0 match. However, the splitter circuit has a desirable characteristic that enables more accurate ratio measurements. The voltage labeled V_x is connected to both output ports via a Z_0 resistor. While V_x may vary with changing load impedances on either output, this divider always maintains a common voltage fed to both outputs through a Z_0 resistor.⁵ Ratio measurements are discussed further in Chapter 14.

Figure 12-11 shows examples of a power splitter and a power divider.

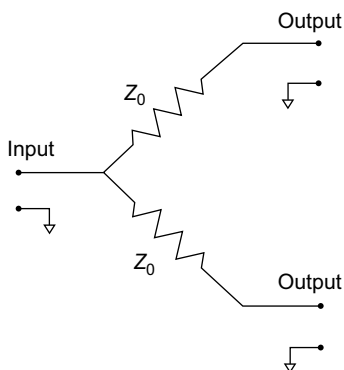


Figure 12-9 The circuit diagram for the two-resistor power splitter.

⁴ The terms *power divider* and *power splitter* are easily confused—hence the reference to the number of resistors in the circuit to keep it straight.

⁵ See Johnson (1975) for a more detailed analysis of this effect.

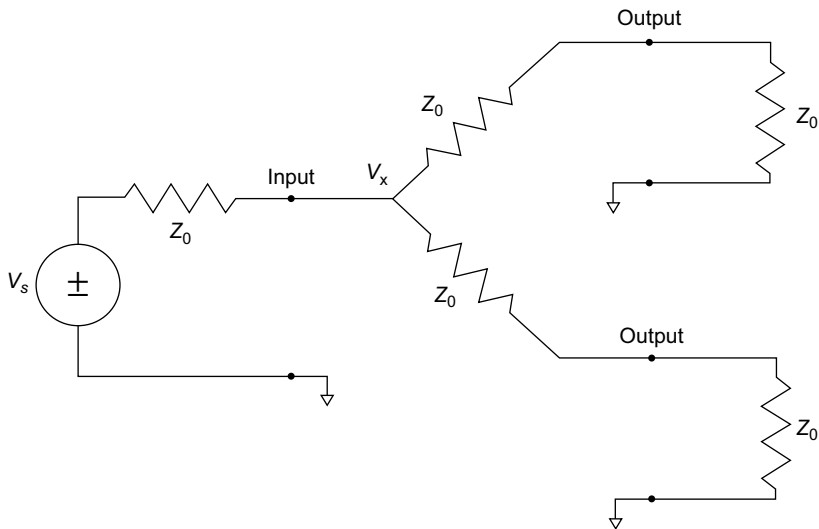


Figure 12-10 The two-resistor splitter drives the two output ports with the common voltage, V_x , via two Z_0 resistors. This provides a more consistent output voltage level for ratio measurements.



Figure 12-11 Examples of a power splitter and power divider. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

12.9 Attenuators

Attenuators (sometimes called *pads*) are used to reduce the signal level in a measurement system. This may be required to bring a large signal within the measuring range of an instrument or to control distortion by reducing the signal level. Also, improvement in the impedance match (return loss) may be achieved by using attenuators. Fixed attenuators provide only one, fixed amount of attenuation, while variable attenuators can be adjusted (often in discrete steps). High-quality attenuators are commercially available, or they can be constructed by the instrument user.

High-Impedance Attenuators

If an instrument with a high input impedance is used and the device driving the attenuator has a low impedance, a simple voltage divider can be used as an attenuator (Figure 12-12a). The high-impedance input of the instrument ensures that little or no loading will occur on the output of the divider. However, the effect of the input capacitance should be considered as the frequency increases. A compensated high-impedance attenuator can compensate for any stray capacitance across R_2 or instrument input capacitance by placing a capacitor in parallel with R_1 (Figure 12-12b). This is the same technique used in an attenuating probe, described earlier in the chapter. The capacitor value must satisfy the equation.

$$R_1 C_1 = C_{IN}(R_2 || R_{IN})$$

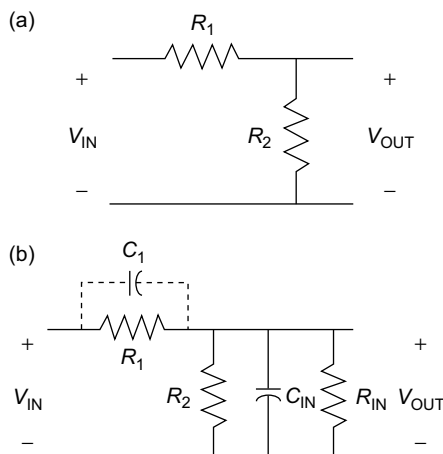


Figure 12-12 (a) A voltage divider circuit can be used as an attenuator for high-impedance systems. (b) A capacitor can be used to compensate the attenuator, improving the frequency response.

The lack of a Z_0 or other reference impedance and the usage of the common voltage divider suggest that the attenuator be specified in terms of voltage gain or loss. The voltage gain⁶ of the circuit is given by

$$G_V = V_{OUT}/V_{IN} = R_2/(R_1 + R_2) \quad (12-7)$$

⁶ Of course, the attenuator will always be lossy ($V_{OUT} < V_{IN}$), and the gain will be less than unity.

The sum of R_1 and R_2 should be chosen so that it is much larger than the output impedance of the device driving the voltage divider. Then, R_1 and R_2 can be computed.

$$R_2 = G_V(R_1 + R_2) \quad (12-8)$$

$$R_1 = (1 - G_V)(R_1 + R_2) \quad (12-9)$$

Example 12.1

Design a high-impedance attenuator to produce a 1 V root mean square (RMS) signal at the output when the input signal is 5 V RMS. The loading on the driving circuitry must be at least 1 k Ω .

The load on the driving circuitry is $R_1 + R_2$, so choose $R_1 + R_2$ equal to 1 k Ω . The voltage gain of the attenuator is $G_V = V_{\text{OUT}}/V_{\text{IN}} = 1/5 = 0.2$.

$$R_2 = G_V(R_1 + R_2) = 0.2(1000) = 200 \Omega$$

$$R_1 = (1 - G_V)(R_1 + R_2) = (1 - 0.2)(1000) = 800 \Omega$$

Z_0 Attenuators

Devices that have input and output impedances of Z_0 require attenuators designed to match the Z_0 impedance. In such systems it is customary to specify the loss of the attenuator as a power ratio, usually expressed in decibels.

$$K = P_{\text{IN}}/P_{\text{OUT}} \quad (12-10)$$

$$K_{\text{dB}} = 10 \log(P_{\text{IN}}/P_{\text{OUT}}) \quad (12-11)$$

The T attenuator circuit is shown in Figure 12-13. The resistor values can be determined from

$$R_1 = \frac{Z_0(\sqrt{K} - 1)}{\sqrt{K} + 1} \quad (12-12)$$

$$R_2 = \frac{2Z_0\sqrt{K}}{K - 1} \quad (12-13)$$

Another Z_0 attenuator configuration, the π attenuator, is shown in Figure 12-14.

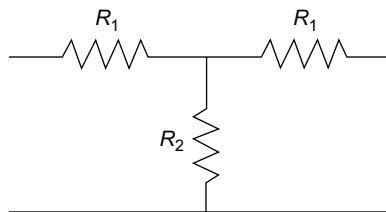


Figure 12-13 The T attenuator circuit.

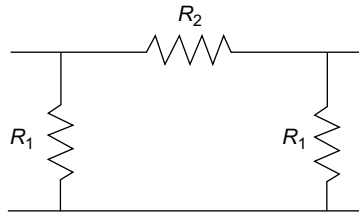


Figure 12-14 The π attenuator circuit.

The resistor values are determined from

$$R_1 = Z_0 \frac{\sqrt{K} + 1}{\sqrt{K} - 1} \quad (12-14)$$

$$R_2 = Z_0 \frac{K - 1}{2\sqrt{K}} \quad (12-15)$$

The two attenuator configurations shown are equivalent, but for a particular application, one configuration may result in more reasonable component values. Note that both Z_0 attenuator configurations are symmetrical from one port to another. Therefore, they provide the same attenuation in both directions.

Example 12.2

In a particular $50\ \Omega$ system, a $-10\ \text{dBm}$ signal must be attenuated to $-30\ \text{dBm}$. Design an attenuator to accomplish this.

First compute the loss required. In dB,

$$K_{\text{dB}} = -10\ \text{dBm} - (-30\ \text{dBm}) = 20\ \text{dB}$$

In power ratio form,

$$K = 10^{(K_{\text{dB}}/10)} = 100$$

Using the T attenuator configuration,

$$R_1 = \frac{Z_0(\sqrt{K} - 1)}{\sqrt{K} + 1} = \frac{50(\sqrt{100} - 1)}{\sqrt{100} + 1} = 40.9\ \Omega$$

$$R_2 = \frac{2Z_0\sqrt{K}}{K - 1} = \frac{2(50)\sqrt{100}}{100 - 1} = 10.1\ \Omega$$

12.10 Return Loss Improvement

Attenuators can be used to improve the return loss of a device, at the expense of reduced signal level. Consider the situation shown in Figure 12-15. An attenuator with power loss, K , is connected to a load having a reflection coefficient of ρ_L . The attenuator input and output

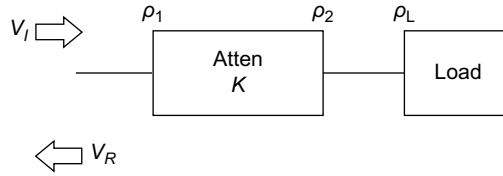


Figure 12-15 The return loss of a device can be improved by adding an attenuator.

impedances are not perfect, so they each have a reflection coefficient associated with them, ρ_1 and ρ_2 .

Without the attenuator in the system, an incident voltage, V_I , produces a reflected voltage,

$$V_R = \rho_L V_I \quad (12-16)$$

With the attenuator connected, V_I is attenuated by \sqrt{K} , the loss in the attenuator. (Since K is the ratio of input and output power, the voltage is reduced by \sqrt{K} .) This makes the voltage incident on the device equal to V_I/\sqrt{K} . A reflected voltage equal to $\rho_L V_I/\sqrt{K}$ is produced. Assuming a symmetrical attenuator, the reflected voltage is attenuated by \sqrt{K} on its way back. The reflected voltage seen at the attenuator port is

$$V_R = \frac{\rho_L V_I}{K} \quad (12-17)$$

The attenuator input and output impedances also cause another set of reflections. A reflected voltage is caused by ρ_1 as V_I is incident on the attenuator input. Also, when the incident voltage makes its way to the load and a reflected voltage is produced due to ρ_L , that reflected voltage may be reflected again by ρ_2 . This reflection will be ignored in the analysis, since with a good-quality attenuator ρ_2 will be small and any subsequent reflection from the load will be even smaller. Also, its effect will be reduced by \sqrt{K} as it passes back through the attenuator. So combining the main reflection from the load and the attenuator input reflection,

$$V_R = V_I \left(\rho_I + \frac{\rho_L}{K} \right) \quad (12-18)$$

Since we don't know whether the reflections will add in phase or out of phase, we simply added them together to produce a worst-case reflection.

The reflection coefficient looking into the input of the attenuator is

$$\rho_a = \frac{V_R}{V_I} = \rho_I + \frac{\rho_L}{K} \quad (12-19)$$

Thus, the reflection coefficient equals the reflection coefficient of the attenuator plus the load reflection coefficient reduced by the power loss factor, K . With a perfect attenuator, ρ_1 is zero, and the equation reduces to

$$\rho_a = \frac{\rho_L}{K} \quad (12-20)$$

This is a good approximation for estimating the improvement in return loss achievable with an attenuator. For a high-quality attenuator and a load with a poor reflection coefficient,

this is a reasonable approximation. The ρ_1 term in equation (12-19) serves to remind us that the improved return loss will never be better than the inherent return loss of the attenuator.

By expressing the reflection coefficient in decibel form, the return loss can be determined.

$$\begin{aligned} RL_a &= -20 \log(\rho_a) = -20 \log(\rho_L/K) \\ &= -20 \log(\rho_a) + 20 \log(K) \end{aligned} \quad (12-21)$$

Since K is a power ratio, we will rewrite the equation as

$$\begin{aligned} RL_a &= -20 \log(\rho_L) + 2[10 \log(K)] \\ RL_a &= RL_L + 2K_{dB} \end{aligned} \quad (12-22)$$

Thus, the return loss is improved by twice the attenuation (expressed in dB). The penalty for such an improvement is that the signal level to the load is reduced.

Although the previous analysis used an attenuator connected to a load, the same principles apply for improving the return loss of a source. The source's return loss will be improved by twice the loss of the attenuator, except where limited by the attenuator match. Again, the disadvantage of such an improvement is the reduced signal level available at the attenuator output.

Example 12.3

What are the return loss, RL , and the reflection coefficient, ρ , at the output of a perfect 10 dB attenuator connected to signal generator having a return loss of 8 dB? Does the answer change if the attenuator has a return loss of 20 dB?

Perfect Attenuator:

$$RL = 8 + 2(10) = 28 \text{ dB}$$

The reflection coefficient is $\rho = 10^{(-RL/20)} = 0.040$.

Attenuator with 20 dB return loss:

Since the return loss predicted by the ideal analysis is 28 dB, the attenuator's 20 dB return will clearly limit the overall return loss. Adding the reflection coefficient of the attenuator to the reflection coefficient of the ideal case will produce the overall reflection coefficient.

The reflection coefficient of the attenuator is $\rho_1 = 10^{(-20/20)} = 0.1$. Since

$$\rho_a = \rho_1 + \rho_{a(\text{IDEAL})} = 0.1 + 0.04 = 0.14$$

then

$$RL_a = -20 \log(0.14) = 17 \text{ dB}$$

Note that the overall return loss is somewhat worse than the attenuator return loss.

12.11 The Classical Attenuator Problem

When a device under test is placed between a single-ended source and a single-ended detector such as a spectrum or network analyzer, a significant error can be introduced. This

effect is due to nonzero cable shield impedance and occurs only at low frequency (less than 100 kHz). As the frequency increases, the cable acts more like a transmission line, and the shield impedance is less critical.

Consider the circuit model shown in Figure 12-16. A signal source is connected to an attenuator via a coaxial cable. The output of the attenuator is connected to the input of an analyzer with another cable. R_{C1} and R_{C2} represent the cable ground impedances. The input of the analyzer is initially assumed floating, with impedance R_G between its input ground and chassis ground. To illustrate the problem, the attenuation of the attenuator is infinite and no signal should be present at the input to the analyzer.

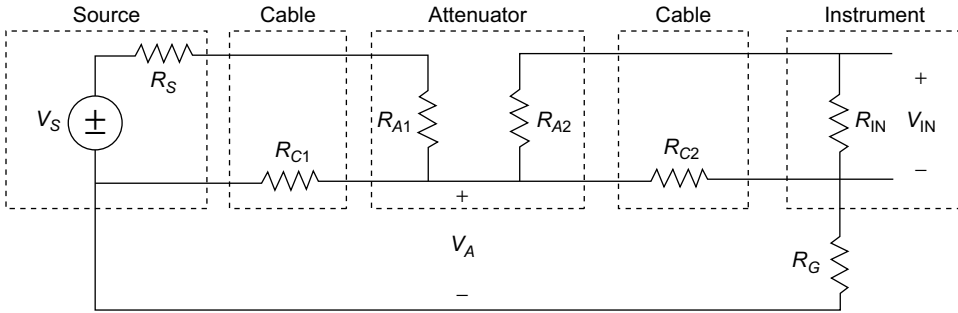


Figure 12-16 Circuit model for demonstrating the classical attenuator problem.

A voltage is generated across the shield impedance of the first cable. Assuming R_{IN} is large compared with R_{C2}

$$V_A = \frac{V_S [R_{C1} \parallel (R_{C2} + R_G)]}{R_S + R_{A1} + [R_{C1} \parallel (R_{C2} + R_G)]} \tag{12-23}$$

This voltage is in turn transferred onto R_{C2} and the instrument input. Again, if R_{IN} is much larger than R_{C2}

$$V_{IN} = \frac{V_A R_{C2}}{R_{C2} + R_G} \tag{12-24}$$

With infinite attenuation, V_{IN} should be zero. But as shown, a small voltage is present at the instrument input. To drive V_{IN} to zero and reduce this error, without corrupting the measurement, R_G can be made large.

To put the situation in perspective, a plot of the measured loss in an attenuator in a typical measurement setting is shown in Figure 12-17. With $R_G = 0$ (the input is grounded), the attenuator loss is measured as low as 90 dB. With R_G large (the input is isolated or floating), the attenuator loss is measured correctly as 120 dB.

The classical attenuator problem applies to any low-frequency situation where a large amount of attenuation is encountered, including measurements of devices such as filters. Analyzers that measure in this frequency range often supply two forms of defense against the problem. One is to isolate or float the input relative to chassis ground at all frequencies. In this case, a switch is often supplied to allow the user to conveniently select a grounded or floated input. The other technique is to isolate the input from the chassis, but only at low frequencies. For high frequencies, where the classical attenuator problem doesn't exist, the input is grounded.

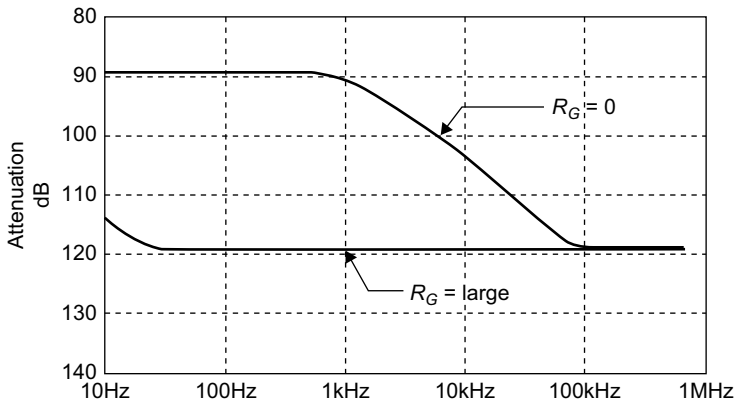


Figure 12-17 The measured attenuation of a 120 dB attenuator with ($R_G = 0$) and without ($R_G = \text{large}$) the classical attenuator problem.

12.12 Impedance Matching Devices

Spectrum and network analyzers are normally offered with standard input impedances, typically 50 or 75 Ω . Sometimes the analyzer does not have the same impedance as the device under test, and the user may need a matching network to eliminate mismatch problems.

Minimum Loss Pads

An attenuator can be used to provide an impedance change, at the expense of some signal loss. Such an attenuator is called an impedance matching attenuator or impedance matching pad. A class of impedance matching pads called *minimum loss pads* incurs the theoretical minimum amount of loss.

The circuit for a minimum loss pad that matches impedances Z_1 and Z_2 is shown in Figure 12-18. Z_1 must be greater than Z_2 . The resistor values can be computed from

$$R_1 = Z_1 \sqrt{1 - (Z_2/Z_1)} \tag{12-25}$$

$$R_2 = \frac{Z_2}{\sqrt{1 - (Z_2/Z_1)}} \tag{12-26}$$

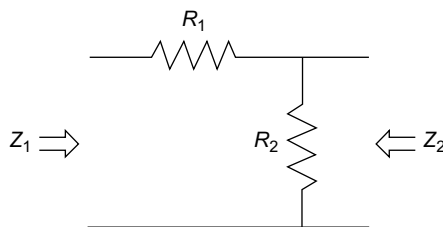


Figure 12-18 A minimum loss pad is used to match unequal impedances. Z_1 must be greater than Z_2 .

The loss (power ratio) is given by

$$K = \frac{2Z_1}{Z_2} - 1 + 2\sqrt{(Z_1/Z_2)(Z_1/Z_2 - 1)} \quad (12-27)$$

This is an example of a system with different input and output impedances, which means that care should be taken when computing gain or loss in dB. The loss factor, K , is a power ratio so to find K in dB use

$$K_{\text{dB}} = 10 \log(K) \quad (12-28)$$

Problems often occur when K_{dB} is used to determine the voltage gain or loss of the minimum loss pad. Unless the unequal impedances are explicitly accounted for, the results will be in error.

Example 12.4

Compute the values for a minimum loss pad that will convert 50Ω to 75Ω . What is the (power) loss of the pad? What is the ratio of the input and output voltage when the pad is correctly terminated?

$$\begin{aligned} Z_1 &= 7 \text{ and } Z_2 = 50 \\ R_1 &= Z_1 \sqrt{1 - (Z_2/Z_1)} = 75 \sqrt{1 - (50/75)} = 43.3 \Omega \\ R_2 &= \frac{Z_2}{\sqrt{1 - (Z_2/Z_1)}} = \frac{50}{\sqrt{1 - (50/75)}} = 86.6 \Omega \end{aligned}$$

The loss is

$$\begin{aligned} K &= \frac{2(75)}{50} - 1 + 2\sqrt{(75/50)(75/50 - 1)} \\ &= 3.73 \\ K_{\text{dB}} &= 10 \log(3.73) = 5.7 \text{ dB} \\ K &= P_1/P_2 = (V_1^2/Z_1)/(V_2^2/Z_2) \\ V_1/V_2 &= \sqrt{Z_1 K/Z_2} = \sqrt{75(3.73)/50} = 2.37 \end{aligned}$$

Transformers

Transformers can be used to match impedances for measurement purposes. A transformer consists of two separate coils that share the same core. The coupling of the magnetic fields of the coils causes a voltage on one coil to induce a voltage on another coil. Since the coupling mechanism depends on a changing magnetic field, a transformer works only for AC signals and not DC.

The ideal transformer is a two-port device with the following voltage and current relationships (Figure 12-19a).

$$V_2 = nV_1 \quad (12-29)$$

$$I_2 = I_1/n \quad (12-30)$$

where n is the turns ratio of the transformer.

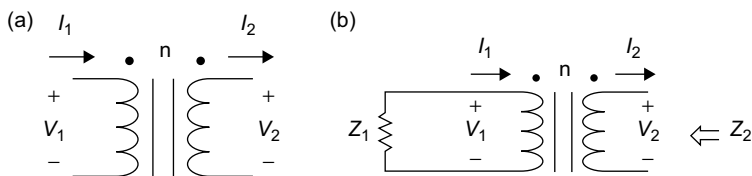


Figure 12-19 (a) The ideal transformer. (b) A transformer can be used to produce an impedance change.

The voltage is changed according to the turns ratio, while the current is transformed inversely proportional to the turns ratio. If an impedance, Z_1 , is connected to the V_1 , I_1 port of the transformer, the impedance looking into the V_2 , I_2 port will be given by (Figure 12-19b)

$$Z_2 = \frac{V_2}{-I_2} = \frac{nV_1}{-I_1/n} = n^2(-V_1/I_1) = n^2Z_1 \quad (12-31)$$

So the impedance is transformed by the square of the turns ratio.

A transformer is usually optimized for some particular frequency range, and the user should consider this when selecting one for measurement use.

Ideally, the transformer is lossless—the power in will equal the power out. But in reality there will be some loss in the transformer. Such a loss can be characterized and normalized out of the measurement.

Another use of transformers in electronic measurements is to provide DC isolation between the device under test and the measuring instrument. Since there is no direct connection between the two ports of the transformer, transformers are isolated for DC. In other words, a transformer can be used to convert a grounded input instrument into one with floating inputs.

12.13 Measurement Filters

It is sometimes desirable to condition a signal's frequency content before it enters the instrument input. For example, an undesirable out-of-band signal might be large enough to cause distortion in the analyzer. Often, a simple filter can remove the offending signal or signals. High-quality filters are commercially available for many different applications, but they can also be built by the instrument user. Electronic filter design has consumed the pages of many other books, and only a few low-pass and high-pass topologies will be discussed here.

High-Impedance Filters

Two filters suitable for use in situations with high-impedance instrument inputs are shown in Figure 12-20. Both of these are single-pole filters: one low-pass and one high-pass.

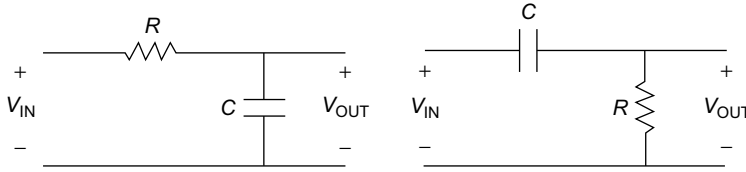


Figure 12-20 (a) A low-pass filter for high-impedance inputs. (b) A high-pass filter for high-impedance inputs.

The low-pass transfer function is

$$V_{\text{OUT}}/V_{\text{IN}} = \frac{1}{1 + j(f/f_{3\text{dB}})} \quad (12-32)$$

and the high-pass transfer function is

$$V_{\text{OUT}}/V_{\text{IN}} = \frac{j(f/f_{3\text{dB}})}{1 + j(f/f_{3\text{dB}})} \quad (12-33)$$

where

$f_{3\text{dB}}$ = frequency at which the response is reduced by 3 dB

For both filters

$$f_{3\text{dB}} = \frac{1}{2\pi RC} \quad (12-34)$$

Being a single-pole filter, the transfer function rolls off at the rate of 20 dB per decade above or below the 3 dB frequency (depending on whether it's a high-pass or a low-pass design).

Z_0 Filters

If the previous filters are used in a Z_0 system, the loading of the Z_0 impedance on the filter output would likely distort the response of the filter. A different approach is needed, one which takes into account the Z_0 loading of such a system. In fact, these filters are required to be loaded in Z_0 to obtain the desired response.

Two filter networks, a high-pass and a low-pass, are shown in Figure 12-21. The values for the low-pass version are computed from⁷

$$L = \frac{\sqrt{2}Z_0}{2\pi f_{3\text{dB}}} \quad C = \frac{\sqrt{2}}{2\pi f_{3\text{dB}} Z_0} \quad (12-35)$$

For the high-pass filter, the equations are

$$L = \frac{Z_0}{\sqrt{2}(2\pi f_{3\text{dB}})} \quad C = \frac{1}{\sqrt{2}(2\pi f_{3\text{dB}}) Z_0} \quad (12-36)$$

⁷ The Z_0 filters discussed here are second-order Butterworth filters.

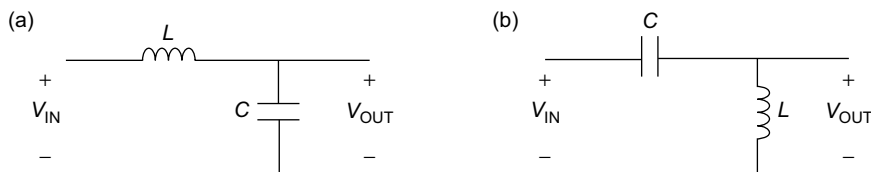


Figure 12-21 (a) A low-pass filter for Z_0 systems. (b) A high-pass filter for Z_0 systems.

The low-pass transfer function is

$$H(f) = \frac{f_{3\text{dB}}^2}{f_{3\text{dB}}^2 - f^2 + j\sqrt{2}f f_{3\text{dB}}} \quad (12-37)$$

and the high-pass transfer function is

$$H(f) = \frac{f^2}{f_{3\text{dB}}^2 - f^2 + j\sqrt{2}f f_{3\text{dB}}} \quad (12-38)$$

The filter response is 3 dB down at $f_{3\text{dB}}$. After that, it rolls off at 40 dB per decade because it has two poles. If the filter characteristics must be sharper, the reader is encouraged to consult one of the many books devoted to filter design.

Example 12.5

Determine the component values for a low-pass $50\ \Omega$ filter with a 3 dB frequency equal to 10 MHz. What is the approximate filter attenuation at 100 MHz?

$$\begin{aligned} f_{3\text{dB}} &= 10\ \text{MHz} \\ L &= \frac{\sqrt{2}Z_0}{2\pi f_{3\text{dB}}} = \frac{\sqrt{2} 50}{2\pi(10 \times 10^6)} \\ &= 1.125\ \mu\text{H} \\ C &= \frac{\sqrt{2}}{2\pi f_{3\text{dB}} Z_0} = \frac{\sqrt{2}}{2\pi(10 \times 10^6) 50} \\ &= 450\ \text{pF} \end{aligned}$$

At 100 MHz, which is one decade above the 3 dB frequency, the response will be attenuated by 40 dB. (The filter response rolls off at 40 dB/decade.)

Bibliography

Agilent Technologies. “Agilent Active Differential Probes U1818A U1818B,” Publication Number 5990-4148EN, Santa Clara, CA, January 2010.

Agilent Technologies. “Agilent RF and Microwave Test Accessories Catalog 2012/13,” Publication number 5990-8661EN, September 2011.

Agilent Technologies. "Amplifier Measurements Using the Scalar Network Analyzer," Application Note 345-1, Publication Number 5954-1599, May 2001.

Agilent Technologies. "Differences in Application between Power Dividers and Power Splitters," Publication Number 5989-6699EN, Santa Clara, CA, August 2007.

Daniels, Jerry W., and Robert L. Atchley, "Attenuating the Classical Attenuator Problem," *Hewlett-Packard Journal*, May 1975.

Johnson, Russell A. "Understanding Microwave Power Splitters," *Microwave Journal*, December 1975.

Williams, Arthur, and Taylor, Fred. *Electronic Filter Design Handbook*, 4th ed. New York: McGraw-Hill Book Company, 2006.

Witte, Robert A. *Electronic Test Instruments: Analog and Digital Measurements*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, Inc., 2002.

Two-Port Networks

Two-port network theory provides the theoretical basis for network measurements. Two-port network theory can be expanded to N -port theory for networks having more than two ports, while one-port measurements are a subset of two-port measurements. The simplest of two-port measurements is the gain or transfer function of the device. This assumes a fairly simple model of the device under test (DUT). More complete two-port models such as impedance parameters provide a more complete view of device behavior, while scattering parameters present a two-port model that is consistent with transmission line theory and measurements.

13.1 Sinusoidal Signals

The standard forcing function for network analysis is the sinusoid, either the sine or cosine function. This stimulus is especially appropriate if we can make the assumption that the network being measured is a linear, time-invariant (LTI) system. Applying a sinusoid to the network's input and measuring the amplitude and phase of the network's output (both as a function of frequency) adequately characterizes the network. Selecting the cosine representation, the input forcing function (or stimulus) is

$$v(t) = V_M \cos(2\pi ft + \theta) \quad (13-1)$$

which is equal to the real part of an exponential function

$$v(t) = \text{Re}[V_M e^{j(2\pi ft + \theta)}] \quad (13-2)$$

This can be verified easily using Euler's identity

$$e^{jx} = \cos(x) + j \sin(x) \quad (13-3)$$

Splitting the exponential gives

$$v(t) = \text{Re}[V_M e^{j2\pi ft} e^{j\theta}] \quad (13-4)$$

In a linear system, the output signal will always be the same frequency as the input signal (with no other frequencies present). The $\text{Re}[\]$ and the $e^{j2\pi ft}$ term are dropped to produce the *vector* or *phasor form* of equation (13-4)

$$\bar{V} = V_M e^{j\theta} \quad (13-5)$$

This is the *polar form* of the vector. Expanding the exponential using Euler's identity gives the *rectangular form*

$$\begin{aligned}\bar{V} &= V_M[\cos \theta + j \sin \theta] \\ &= V_M \cos \theta + j V_M \sin \theta\end{aligned}\quad (13-6)$$

$$V_{RE} = \text{Re}[\bar{V}] = V_M \cos \theta \quad (13-7)$$

$$V_{IM} = \text{Im}[\bar{V}] = V_M \sin \theta \quad (13-8)$$

Other useful conversion formulas are

$$\theta = \tan^{-1}(V_{IM}/V_{RE}) \quad \text{for } V_{RE} \geq 0 \quad (13-9)$$

$$\theta = \tan^{-1}(V_{IM}/V_{RE}) + 180 \quad \text{for } V_{RE} < 0, V_{IM} > 0 \quad (13-10)$$

$$\theta = \tan^{-1}(V_{IM}/V_{RE}) + 180 \quad \text{for } V_{RE} < 0, V_{IM} < 0 \quad (13-11)$$

$$V_M = \sqrt{V_{RE}^2 + V_{IM}^2} \quad (13-12)$$

A further change in notation produces a yet more concise vector form:

$$\bar{V} = V_M \angle \theta \quad (13-13)$$

To summarize, the following three expressions all represent the same signal:

$$v(t) = V_M \cos(2\pi ft + \theta) \quad (13-14)$$

$$\bar{V} = V_M e^{j\theta} \quad (13-15)$$

$$\bar{V} = V_M \angle \theta \quad (13-16)$$

The vector form is a first step toward a true frequency domain representation of the signal. Note that the vector forms do not explicitly state the frequency of the input and output signals. So far, we have considered the vector form to be valid only at one specific frequency. Later we will expand on this concept and consider vectors that are a function of frequency.

Example 13.1

Express the following signal in polar and rectangular vector forms: $v(t) = 5 \cos(20t + 30^\circ)$.

The amplitude of the waveform is 5, and the phase is 30° . Therefore, the polar vector form is $5e^{j(30^\circ)}$ or $5\pi 30^\circ$.

The rectangular form is found by

$$V_R = 5 \cos(30^\circ) = 4.33$$

$$V_I = 5 \sin(30^\circ) = 2.5$$

$$\bar{V} = 4.33 + j2.5$$

Example 13.2

Express the following 20 MHz vector signal as a time domain function:

$$\bar{V} = 8 + j4$$

First, convert to polar form.

$$V_M = \sqrt{8^2 + 4^2} = 8.94$$

$$\theta = \tan^{-1}(8/4) = 63.43^\circ$$

$$\bar{V} = 8.94 \angle 63.43^\circ$$

$$v(t) = 8.94 \cos[2\pi(20 \times 10^6)t + 63.43^\circ]$$

13.2 The Transfer Function

Engineering circuit theory textbooks introduce the concept of the *transfer function* (or *system function*), which is the output voltage of a network divided by its input voltage, both in vector form and both a function of frequency (Figure 13-1).

$$H(f) = \frac{V_2(f)}{V_1(f)} \quad (13-17)$$

The transfer function may be shown as a function of scalar frequency, ω or f , or complex frequency, s , where $s = \sigma + j\omega$. The former is compatible with Fourier theory, and the latter is oriented toward Laplace transforms. The complex frequency approach supports transient analysis, which is not used for classic network measurements. Thus, the transfer function is shown here as a function of $f(\text{Hz})$.

In some cases, the transfer function is simplified by ignoring the phase information contained in $H(f)$. Taking the magnitude of the voltages results in voltage gain as a function of frequency.

$$G_V(f) = |H(f)| = \frac{|V_2(f)|}{|V_1(f)|} \quad (13-18)$$

As shown in Chapter 2, the voltage gain can be expressed in decibel form as

$$G_{V(\text{dB})}(f) = 20 \log[G_V(f)] \quad (13-19)$$

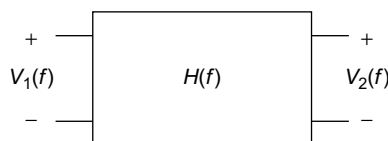


Figure 13-1 The transfer function of a network, $H(f)$, relates the input voltage to the output voltage.

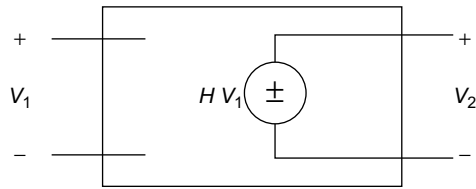


Figure 13-2 This circuit models the behavior of a network, ignoring loading conditions.

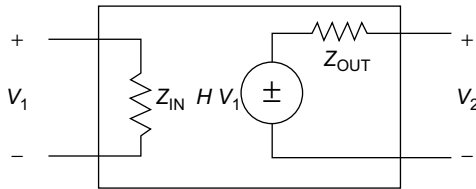


Figure 13-3 This improved circuit model shows the input and output impedances of the network.

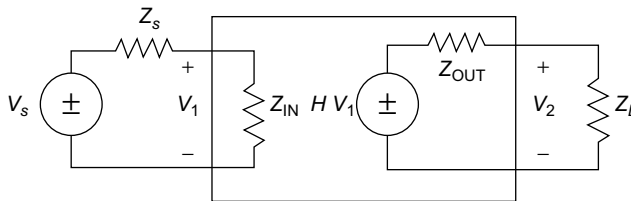


Figure 13-4 When driven by a nonzero source impedance or loaded with a finite load impedance, loading effects will occur.

A circuit model for the transfer function is shown in Figure 13-2. Since V_1 , V_2 , and H are all functions of frequency, we will drop the explicit designation at this point. This simple model ignores any loading considerations and is an appropriate model for studying network theory but needs to be improved on for practical use.

13.3 Improved Two-Port Model

The previous circuit model implies that the impedance looking into the input terminals is infinite and the impedance looking into the output terminals is zero. This model is inaccurate for circuits having finite input and output impedances. An improved model is shown in Figure 13-3. The circuit providing the input voltage to the two-port network is loaded by the input impedance of the network. Similarly, the output of the network is loaded by the impedance connected across its output (Figure 13-4). The voltage source driving the network is shown as having finite output impedance, Z_s . The relationship of the source and output voltages is now

$$\frac{V_2}{V_s} = \frac{Z_{IN} Z_L H}{(Z_s + Z_{IN})(Z_L + Z_{OUT})} \quad (13-20)$$

The ratio of the output voltage to source voltage is clearly affected by the source impedance and load impedance. If the source impedance is low relative to the network input impedance, the effect may be negligible. Similarly, if the load impedance is large relative to the output impedance of the network, that effect may be ignored. In many practical measurement situations, the source impedance and load impedance are specified. That is, the measurement procedure calls for a source with a particular output impedance and for a particular load impedance to be installed at the output. If only the forward transfer characteristics, V_2/V_S , of the network are required, specifying the terminating impedances may be sufficient to ensure an accurate characterization of the network.

13.4 Impedance Parameters

So far in this chapter, we have discussed two-port network models that are incomplete in that they do not provide for both forward and reverse transfer characteristics. What is needed is a model which takes into account the influence of the output port on the input port. There is an entire class of linear two-port models that completely characterizes a two-port network. The first of these we will discuss is the use of *impedance parameters*.

The impedance parameter model is shown in Figure 13-5. Note that in the circuit are two impedances and two dependent voltage sources. Alternatively, the model can be expressed as two linear equations:

$$V_1 = Z_{11}I_1 + Z_{12}I_2 \tag{13-21}$$

$$V_2 = Z_{21}I_1 + Z_{22}I_2 \tag{13-22}$$

The input and output voltages are expressed as linear functions of the input and output currents. Since the coefficients in the equations are the ratio of a voltage and a current, and have the units of ohms, they are called impedance (or Z) parameters. The first subscript of the particular Z parameter indicates the port at which the effect occurs, while the second subscript indicates the source of the effect. For example, Z_{21} represents the effect on the voltage at port 2 due to the current at port 1. In general, Z parameters are complex numbers that vary as a function of frequency.

To solve for the particular impedance parameter, the currents I_1 and I_2 are selectively set to zero. In an actual measurement situation this means that the appropriate port is left open.

$$Z_{11} = \left. \frac{V_1}{I_1} \right|_{I_2=0} \tag{13-23}$$

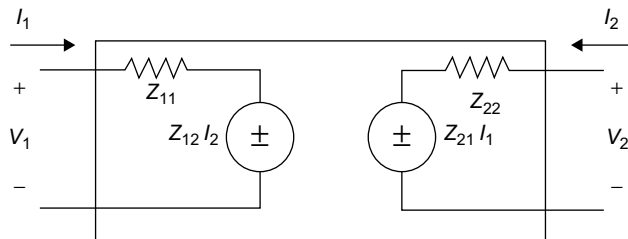


Figure 13-5 The circuit model associated with two-port impedance parameters.

So Z_{11} is the input impedance of the network, with the stipulation that the output is left open. Therefore, Z_{11} is called the *open-circuit input impedance*. The other parameters can be determined in a similar manner and have similar names.

$$Z_{12} = \left. \frac{V_1}{I_2} \right|_{I_1=0} \quad \text{open-circuit reverse transfer impedance} \quad (13-24)$$

$$Z_{21} = \left. \frac{V_2}{I_1} \right|_{I_2=0} \quad \text{open-circuit forward transfer impedance} \quad (13-25)$$

$$Z_{22} = \left. \frac{V_2}{I_2} \right|_{I_1=0} \quad \text{open-circuit output impedance} \quad (13-26)$$

13.5 Admittance Parameters

Another set of two-port parameters is the *admittance parameters*. The linear equations that relate the admittance parameters to the terminal voltages and currents are

$$I_1 = Y_{11}V_1 + Y_{12}V_2 \quad (13-27)$$

$$I_2 = Y_{21}V_1 + Y_{22}V_2 \quad (13-28)$$

The admittance coefficients can be solved for by setting one of the port voltages to zero. This corresponds to placing a short on the appropriate port.

$$Y_{11} = \left. \frac{I_1}{V_1} \right|_{V_2=0} \quad \text{short-circuit input admittance} \quad (13-29)$$

$$Y_{12} = \left. \frac{I_1}{V_2} \right|_{V_1=0} \quad \text{short-circuit reverse transfer admittance} \quad (13-30)$$

$$Y_{21} = \left. \frac{I_2}{V_1} \right|_{V_2=0} \quad \text{short-circuit forward transfer admittance} \quad (13-31)$$

$$Y_{22} = \left. \frac{I_2}{V_2} \right|_{V_1=0} \quad \text{short-circuit output admittance} \quad (13-32)$$

13.6 Hybrid Parameters

The *hybrid parameters* (or *h* parameters) are often used to describe transistor characteristics. The coefficients of the linear equations are not consistently impedances or admittances—hence the name hybrid parameters.

The parameters are defined by

$$V_1 = h_{11}I_1 + h_{12}V_2 \quad (13-33)$$

$$I_2 = h_{21}I_1 + h_{22}V_2 \quad (13-34)$$

Solving for the h parameters

$$h_{11} = \left. \frac{V_1}{I_1} \right|_{V_2=0} \quad \text{short-circuit input impedance} \quad (13-35)$$

$$h_{12} = \left. \frac{V_1}{V_2} \right|_{I_1=0} \quad \text{open-circuit reverse voltage gain} \quad (13-36)$$

$$h_{21} = \left. \frac{I_2}{I_1} \right|_{V_2=0} \quad \text{short-circuit forward current gain} \quad (13-37)$$

$$h_{22} = \left. \frac{I_2}{V_2} \right|_{I_1=0} \quad \text{open-circuit output admittance} \quad (13-38)$$

13.7 Transmission Parameters

Yet another variation on the basic concept of two-port parameters is the *transmission parameter* (also called *ABCD parameter*). These parameters are defined by

$$V_1 = AV_2 - BI_2 \quad (13-39)$$

$$I_1 = CV_2 - DI_2 \quad (13-40)$$

The parameters are given by

$$A = \left. \frac{V_1}{V_2} \right|_{I_2=0} \quad \text{open-circuit voltage ratio} \quad (13-41)$$

$$B = \left. \frac{V_1}{-I_2} \right|_{V_2=0} \quad \text{negative short-circuit transfer impedance} \quad (13-42)$$

$$C = \left. \frac{I_1}{V_2} \right|_{I_2=0} \quad \text{open-circuit transfer admittance} \quad (13-43)$$

$$D = \left. \frac{I_1}{-I_2} \right|_{V_2=0} \quad \text{negative short-circuit current ratio} \quad (13-44)$$

13.8 Scattering Parameters

Scattering parameters (or S-parameters) are most commonly used at high frequencies and are the most important set of two-port parameters relating to network measurements. Unlike the previous sets of two-port parameters, S-parameters use a traveling wave approach to describe the activity at each port.

A lightwave analogy is often used to describe this behavior, and the term scattering parameter originates from the optical world. When a lightwave encounters a clear lens, part of the incident wave is reflected back while the majority of the wave travels through the lens is transmitted out the other side (Figure 13-6). Alternatively, if the lens is highly reflective, a large portion of the incident wave is reflected and only a small portion is transmitted through the lens.

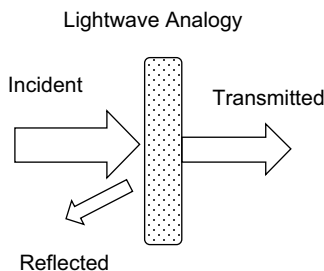


Figure 13-6 Using a lightwave analogy, when a lightwave encounters a lens part of the incident wave is reflected while the rest of it is transmitted through the lens.

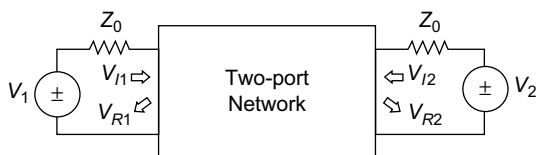


Figure 13-7 A two-port network can be characterized using S-parameters, which describe the behavior of the incident and reflected voltages at each port.

This lightwave analogy can be applied to a two-port electrical network. A traveling wave is incident on the port: a portion of it is reflected, and the remainder is transmitted to the network (Figure 13-7). A signal source, incident wave and reflected wave are shown for both ports of the network.

S-parameter measurements are referenced to a characteristic impedance (Z_0), which is usually the nominal input and output impedance of the network. The approach is consistent with transmission line theory; hence, its common usage in high-frequency measurements.

The defining equations for S-parameters are

$$V_{R1} = S_{11}V_{I1} + S_{12}V_{I2} \tag{13-45}$$

$$V_{R2} = S_{21}V_{I1} + S_{22}V_{I2} \tag{13-46}$$

Notice that the equations are made up only of incident and reflected voltages (and not currents). Again, we will solve for the individual coefficients, the S-parameters.

To solve for S_{11} it is necessary to set V_{I2} to zero. One might be tempted to attach mentally a short circuit to port 2 to accomplish this. But remembering that the S-parameter model of the network deals with incident and reflected voltages, we will attach a Z_0 load to port 2 (Figure 13-8). This sets V_{I2} to zero while maintaining a nonreflecting load at port 2.

$$S_{11} = \left. \frac{V_{R1}}{V_{I1}} \right|_{Z_0 \text{ load on port 2}} \tag{13-47}$$

So S_{11} is the same as the reflection coefficient at port 1 when port 2 is terminated with a Z_0 load. S_{11} is called the *input reflection coefficient*. This is the same concept as the complex reflection coefficient (Γ) discussed in Section 11.8.

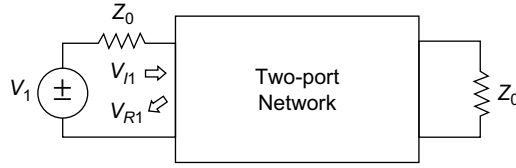


Figure 13-8 S_{11} is found by determining the amount of reflection at the input port while applying a Z_0 load to the output port.

Solving for S_{21} with a Z_0 load still connected to port 2:

$$S_{21} = \left. \frac{V_{R2}}{V_{I1}} \right|_{Z_0 \text{ load on port 2}} \quad (13-48)$$

S_{21} is called the *forward transmission coefficient*, which loosely corresponds to the transfer function of the device.¹

Continuing on

$$S_{22} = \left. \frac{V_{R2}}{V_{I2}} \right|_{Z_0 \text{ load on port 1}} \quad (13-49)$$

So S_{22} is the reflection coefficient at port 2 (with port 1 terminated in Z_0) and is called the *output reflection coefficient*. This also corresponds to the complex reflection coefficient shown in Section 11.8, but now at the output port.

$$S_{12} = \left. \frac{V_{R1}}{V_{I2}} \right|_{Z_0 \text{ load on port 1}} \quad (13-50)$$

S_{12} is called the *reverse transmission coefficient*, which represents the effect on the input port due to the incident voltage on the output port.

The S-parameter equations are often shown with the incident and reflected voltages normalized by the square root of Z_0 . The defining equations are then

$$b_1 = S_{11}a_1 + S_{12}a_2 \quad (13-51)$$

$$b_2 = S_{21}a_1 + S_{22}a_2 \quad (13-52)$$

where

$$a_1 = \frac{V_{I1}}{\sqrt{Z_0}}, \quad a_2 = \frac{V_{I2}}{\sqrt{Z_0}}, \quad b_1 = \frac{V_{R1}}{\sqrt{Z_0}}, \quad b_2 = \frac{V_{R2}}{\sqrt{Z_0}}$$

This notation is introduced to provide continuity with other literature that the reader may encounter. This notation is summarized and shown graphically in Figure 13-9.

It is worth mentioning again that, in general, S-parameters are a function of frequency. We could emphasize this point by always writing the S-parameters in the form $S_{11}(f)$, $S_{21}(f)$, and so on. This is not done in practice, but instead it is understood that the parameters vary with frequency.

¹ We will examine this statement more closely later on in the chapter.

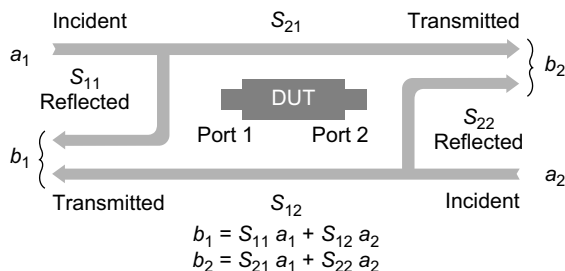


Figure 13-9 Summary of S-parameter notation for two-port networks.

13.9 Transfer Function and S_{21}

As previously stated, S_{21} , the forward transmission coefficient, is similar to the conventional notion of transfer function, but there are some differences. Normally, the concept of a V_2/V_1 type of transfer function means that the voltages at the input and output ports of the network are measured directly, perhaps with some specified source and load impedances. Given that the source and load impedance constraints are observed, the transfer function measurement degenerates to a vector voltage measurement (preserving the phase information). This transfer function model inherently assumes that the voltage at the output port depends only on the voltage at the input port (Figure 13-2).

Now consider S_{21} , which is equal to the reflected voltage at the output port, V_{R2} , divided by the incident voltage at the input port, V_{I1} (with a Z_0 load on the output port). The reflected voltage at the output port is somewhat of a misnomer, since it is really a traveling wave leaving the output port, due to activity on the input port. Since there is no incident wave on the output port and since V_{R2} will be totally absorbed by the Z_0 load, V_{R2} is the same as the V_{OUT} of the transfer function measurement.

Things are not quite the same on the input side. The incident wave, V_{I1} is the voltage that would be delivered by the Z_0 voltage source to a Z_0 load. If the input to the network is a perfect Z_0 , then none of the incident voltage will be reflected from the input port. In that case, V_{I1} is the same as V_1 of the transfer function measurement. However, if the input impedance of the network is not Z_0 , some of the incident voltage will be reflected, making V_1 different from V_{I1} . Another way of saying this is that S_{21} expresses the output voltage relative to the voltage available from the source when it is driving Z_0 . For devices that have input impedances near Z_0 , S_{21} is basically the same as the notion of a transfer function. When the input impedance is not close to Z_0 , the two measurements will be different.

13.10 Why S-Parameters?

So why use S-parameters at all? The answer is multifaceted. The main reason for using S-parameters is that voltages and currents are difficult to measure directly at high frequencies. However, measurement techniques have been developed to measure the traveling voltage waves required by S-parameters. In fact, S-parameters are closely related to and are an extension of transmission line theory, in that the input and output voltages are treated as

incident and reflected traveling waves. For low-frequency design and measurement, this may not be a factor, but at higher frequencies transmission line concepts are unavoidable due to the shorter wavelengths.

S-parameters are measured with the network ports terminated in Z_0 impedances. Other two-port parameters require open or short circuits to be connected at the input and output ports. At higher frequencies, open and short circuits can be difficult to implement. Stray capacitance and inductance as well as transmission line effects get in the way. More importantly, many circuits will not behave well when presented with an open or short termination—distortion or oscillations may occur.

Directional devices (discussed in Chapter 15) provide a means of separating incident and reflected waves while maintaining a Z_0 match in the system. Thus, the individual traveling waves can be measured without significantly disturbing the device under test.

A wide body of design methodology has been developed that relates directly to and relies on S-parameters. Reflection coefficients, S_{11} and S_{22} , are often plotted on the Smith chart to design impedance matching and other networks. Flow graphs can be used to analyze systems that are characterized by S-parameters. Entire textbooks have been written on these design techniques and are beyond the scope of this book.

Bibliography

Agilent Technologies Inc. “S-Parameter Design,” Application Note 154, Publication Number 5952-10 87, June 2006.

Agilent Technologies, Inc. “Understanding the Fundamental Principles of Vector Network Analysis,” Application Note, Publication Number 5965-7707E, September 2012.

Carson, Ralph S. *High Frequency Amplifiers*, 2d ed. New York: John Wiley & Sons, Inc., 1982.

Gonzalez, Guillermo. *Microwave Transistor Amplifiers*, 2d ed. Englewood Cliffs, NJ: Prentice Hall, Inc., 1996.

Hayt, William H., Kemmerly, Jack E., and Durbin, Steven M. *Engineering Circuit Analysis*, 8th ed. New York: McGraw-Hill Book Company, 2011.

Hewlett-Packard Company. “S-Parameter Techniques for Faster, More Accurate Network Design,” Application Note 95-1, Palo Alto, CA, February 1967.

Irwin, J. David, and Nelms, Robert M. *Basic Engineering Circuit Analysis*, 10th ed. New York: Macmillan Publishing Company, 2010.

Network Analyzers

Network measurements can be divided into two types: transmission through the network and reflection at the network's input or output port. Full two-port network analysis normally requires the use of a multichannel network analyzer and a scattering parameter (S-parameter) test set. Simpler measurements, such as transmission-only measurements, can be performed with less sophisticated equipment.

14.1 Basic Network Measurements

The transmission through a two-port network is measured by applying a signal to one port and measure the response at the other port (Figure 14-1). The forward transmission characteristics of a network are measured by connecting the signal source to the input port and measuring the response at the output port. The reverse transmission characteristics of the network can be measured by driving the network at the output port and measuring the response at the input.

The reflection at the input port of a network is measured by applying a signal to the port and measuring the traveling wave reflected by the input. The reflection at the output port can be measured in a similar manner while driving the output port.

14.2 Oscilloscope and Sweep Generator

An oscilloscope and a sweep generator can be used to implement a simple scalar network analyzer (Figure 14-2). This measurement setup is useful for making transmission measurements when only the magnitude (and not the phase) of the transfer function is required. The oscilloscope is operated in X - Y (channel versus channel) mode: the X (horizontal) input is the sweep voltage from the generator, and the Y (vertical) input is the output of the device under test. The sweep voltage of the generator is proportional to the generator frequency. As the generator sweeps, the output of the device under test is plotted across the oscilloscope display.

Marker outputs from the sweep generator may be used to identify accurately frequencies on the oscilloscope display. These marker outputs are used to drive the intensity (Z -axis) input of the oscilloscope. When the marker frequency is swept through, the intensity on the scope changes. Some sweep generators can increase the output level slightly as the generator passes through the marker frequency, causing a blip on the oscilloscope display.

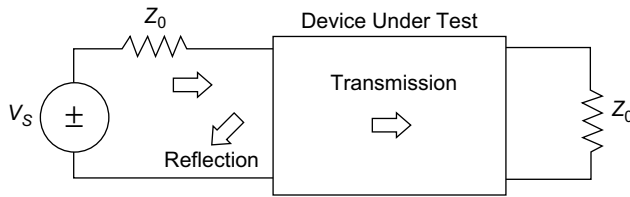


Figure 14-1 Transmission and reflection in a two-port network.

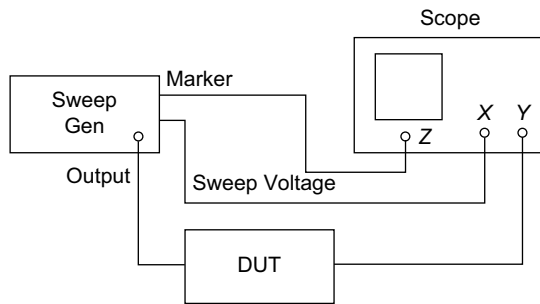


Figure 14-2 A sweep generator and an oscilloscope can be configured to perform basic transmission measurements.

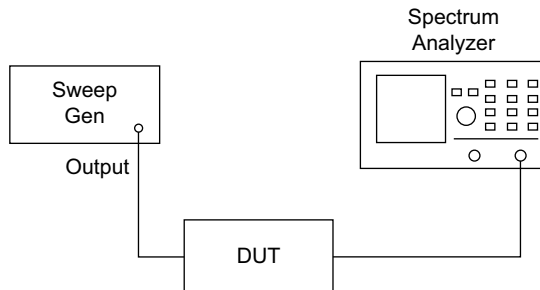


Figure 14-3 A sweep generator and spectrum analyzer can be used to make transmission measurements.

This method relies upon the sweep generator output to be constant with changes in frequency. Any imperfection in the flatness of the sweep generator will show up as an error in the measurement. Another major disadvantage of this network measurement technique is that the oscilloscope display is linear and has a limited dynamic range.

14.3 Network Measurements Using a Spectrum Analyzer

A spectrum analyzer can also be used to make basic network measurements. Figure 14-3 shows a sweep generator being used with a spectrum analyzer to perform scalar network measurements. Most spectrum analyzers provide a “max hold” feature, which causes the

display to retain the largest measured value at each frequency. Both the sweep generator and spectrum analyzer are set to sweep the frequency range of interest, with the spectrum analyzer set to max hold. As the sweep generator excites the device under test, the spectrum analyzer measures the device's output. Gradually, the spectrum analyzer accumulates the entire response of the device.

The sweep times of the generator and analyzer may interact, slowing down the measurement. The worst case is when the analyzer and generator are sweeping at approximately the same speed, but offset in frequency so that they tend not to be at the same frequency simultaneously. To alleviate this problem, one of the instruments is set to sweep very fast (usually the analyzer) while the other is set to sweep slowly. In this case, just a few sweeps of the slower instrument is required to produce a useful plot on the analyzer display.

The accuracy of this technique is limited by the amplitude flatness of the generator. However, the flatness can be measured and calibrated out of the measurement if the spectrum analyzer has storage and subtraction capability. First, the device under test (DUT) is removed and the generator is connected directly to the analyzer so that the generator's amplitude response is measured and stored. The DUT is reinstalled and the spectrum analyzer is set to subtract the generator response from the measured DUT response.¹

If the sweep generator and spectrum analyzer sweeps are synchronized, the measurement can be completed in one sweep and max hold is unnecessary. The spectrum analyzer may have an external trigger that can be driven by a trigger signal from the generator. Or perhaps both instruments can be triggered simultaneously by an external signal. Synchronizing the two instruments is usually more difficult than it first appears due to latency between the trigger signal and the start of sweep and the difference in sweep rates between the two instruments.

Tracking Generator

The problem of synchronizing the generator and analyzer sweeps can be circumvented by having the generator integrated into the spectrum analyzer. Since the generator tracks the analyzer's receiver frequency, it is called a *tracking generator*. (See Chapter 5 for more information on tracking generators.) This combination is essentially a simple network analyzer, capable of making basic transmission measurements (Figure 14-4).

The flatness of the generator is still a source of error in the measurement, but again it can be calibrated out by measuring it and subtracting it from the DUT response.

14.4 Vector Network Analyzer

A *vector network analyzer* (VNA) has a built-in source and multiple receiver channels that are closely matched in amplitude and phase accuracy. The term *vector* means that the analyzer's receivers can measure both amplitude and phase. The VNA has largely displaced the *scalar network analyzer*, which can perform only amplitude measurements.

¹ The desired mathematical operation is actually division, not subtraction, but the spectrum analyzer normally displays the response in logarithmic (decibel) form. Since the subtraction is done after the log, it is equivalent to division before the log function: $\log(a/b) = \log a - \log b$.

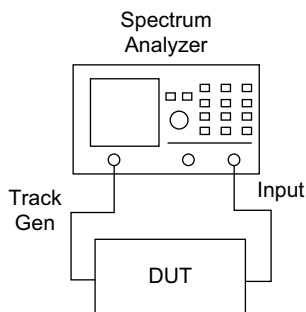


Figure 14-4 Spectrum analyzers that include a tracking generator are capable of making basic transmission measurements.

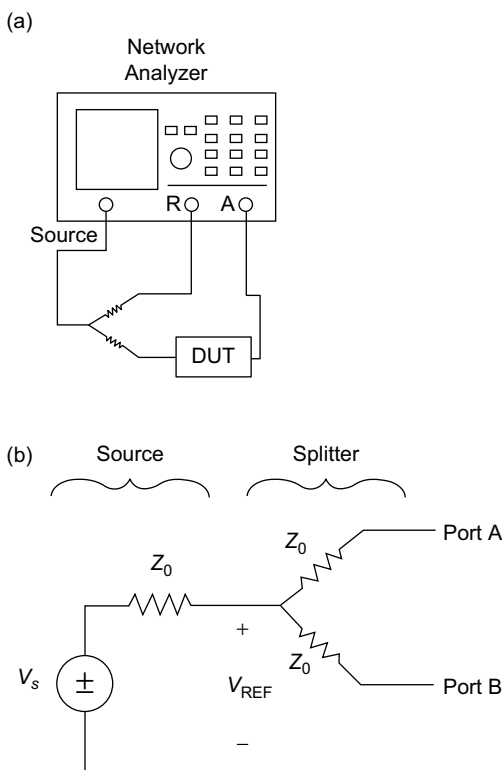


Figure 14-5 A multichannel network analyzer and a power splitter allow a ratio measurement to be performed.

Figure 14-5a shows how ratio techniques are used to measure the transmission characteristics of a DUT. The VNA source drives a two-way power splitter, with one splitter output going to the DUT and the other output connected directly to the reference (R) channel of the analyzer. The output of the DUT is connected to the analyzer (A) channel. The network analyzer uses the split-off source signal to correct for frequency response imperfections in the source. The network analyzer is configured to display the DUT response at channel A divided

by the signal present at the reference channel (resulting in an A/R measurement). To the extent that the two channels are matched, the source flatness is removed from the measurement.

When making ratio measurements, the power splitter should always be a two-resistor type splitter. (See Chapter 12 for a discussion of other types of power splitters and dividers.) Figure 14-5b shows a Z_0 source connected to a two-resistor power splitter. V_{REF} is considered a virtual voltage source, with both port A and port B seeing V_{REF} through a Z_0 impedance. This implies that both ports always receive the same incident voltage. V_{REF} and the incident voltage may change with frequency, but this effect will be removed by measuring the ratio of the two channels.

14.5 Directional Bridges and Couplers

A *directional bridge* or *directional coupler* has the ability to sense the energy traveling in one direction along a transmission line. It is used to separate the incident voltage from the reflected voltage on the line when performing network measurements.

As configured in Figure 14-6, the directional coupler senses the incident wave at the input port of the device under test by diverting a small amount of the incident power to the auxiliary port. The diverted incident wave is connected to the reference channel of the network analyzer and is used to obtain a ratio measurement. Reversing the directional coupler allows the reflected wave at the DUT input port to be measured. Figure 14-7 shows the use of a directional coupler, along with a power splitter, to produce a reflection measurement. The power splitter is used to produce the reference channel signal, while the directional coupler measures the reflected wave from the DUT. Here the DUT is shown as a one-port device since this is a one-port measurement.

Directional bridges and couplers are discussed further in Chapter 15.

14.6 S-Parameter Test Set

A power splitter and one or more directional couplers may be combined to create a *test set* for a specific set of network measurements. Figure 14-8 shows a *transmission/reflection test set* for measuring the transmission through the DUT and the reflection at the input. The transmission through the device, S_{21} , and input match, S_{11} , can be displayed simultaneously

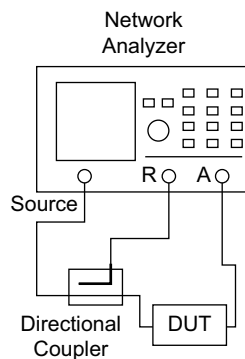


Figure 14-6 A directional coupler can also be used to produce a ratio measurement.

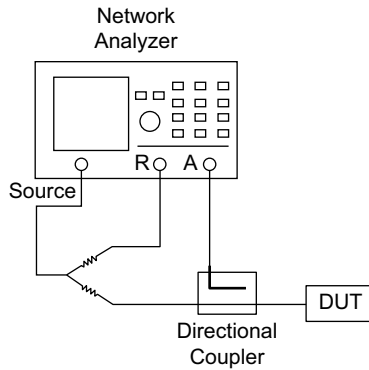


Figure 14-7 A directional coupler and a power splitter are shown here configured to measure the reflection from the device under test.

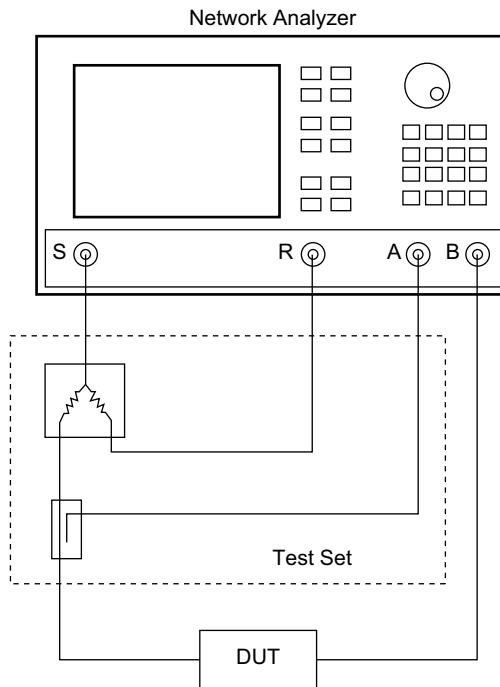


Figure 14-8 A transmission/reflection test set enables the measurement of transmission characteristics through the DUT and reflection measurements at one port.

using a three-channel network analyzer. Both measurements are ratio measurements using the R channel as the reference. The transmission measurement uses the (B) channel divided by the (R) channel, producing B/R , while the reflection measurement is A/R .

For complete characterization of a linear two-port network, all four of the two-port parameters must be measured. The required directional couplers and power splitter are assembled

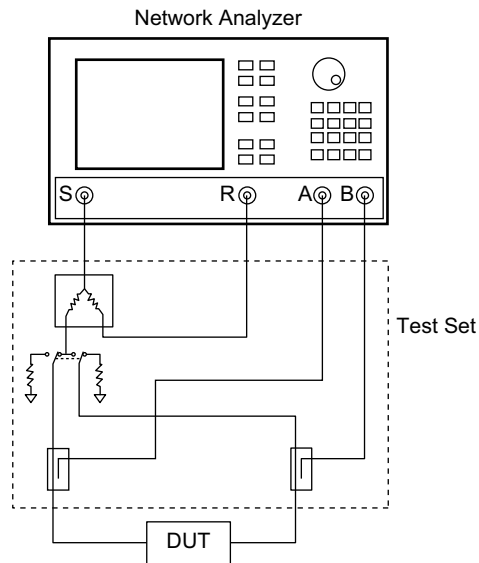


Figure 14-9 An S-parameter test set allows all four S-parameters to be measured without reconnecting the device under test.

into one device called an *S-parameter test set*. A test set is commonly configured with a power splitter, two directional devices, and switching relays as shown in Figure 14-9. This configuration allows a three-channel network analyzer to measure either S_{11} and S_{21} or S_{12} and S_{22} , depending on the position of the relay. Thus, the transmission through the device and the reflection at the driven port can be displayed simultaneously. All four S-parameters can be measured without reconnecting the device under test (although the relay must change position).

The reader may encounter older network analyzers with separate test sets configured as shown in Figures 14-8 and 14-9. As network analyzers have evolved, the test set hardware has been integrated into the analyzer, allowing for a more compact system and improved usability.

14.7 Modern Vector Network Analyzer Configurations

The VNA configuration shown in Figure 14-9 is a classic configuration that provides for basic two-port S-parameter measurements. VNA block diagrams have continued to evolve and incorporate more measurement ports and more types of measurements. The modern VNA is really a measurement system in a box, able to completely characterize the critical parameters of the DUT (Figure 14-10). Rather than cover a wide range of configurations, we will look at the key features that are commonly provided:

- *Additional ports*: Some VNAs provide additional measurement ports for handling more complex DUTs. The basic concepts of two-port measurements are extended to cover more ports.
- *Multiple sources*: Additional signal sources (usually with a combiner) enable test flexibility including two-tone intermodulation measurements.

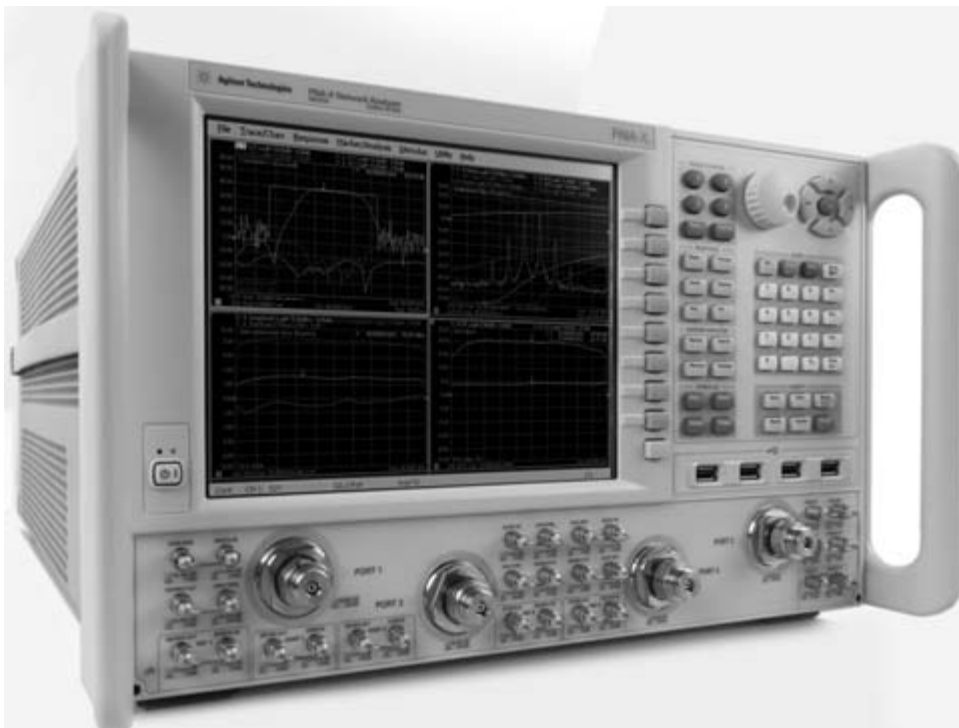


Figure 14-10 The modern VNA is able to completely characterize the critical parameters of the DUT. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

- *Bias tee*: Allows a DC bias to be inserted into the measurement port to test devices that require or need to be characterized based on DC voltages.
- *Switches*: With the increased complexity of the VNA, internal switches are provided to change the configuration programmatically.
- *Front/rear panel access loops*: These jumpers allow access to the internal points of the VNA, providing more measurement flexibility.
- *Low-noise amplifier*: Allows noise figure measurements to be made on the DUT.

Signal sources for network analyzers used to be separate, stand-alone sources that were synchronized with the network analyzer receiver. Modern network analyzers have one or more sources integrated into the analyzer, consistent with the overall trend toward a highly integrated measurement system.

14.8 Sweep Limitations

A network analyzer has sweep rate (Hz/sec) limitations just as the spectrum analyzer does. Neither analyzer can be swept arbitrarily fast. The sweep rate limitation of a spectrum analyzer is proportional to the square of the resolution bandwidth, as discussed in Chapter 5. The sweep

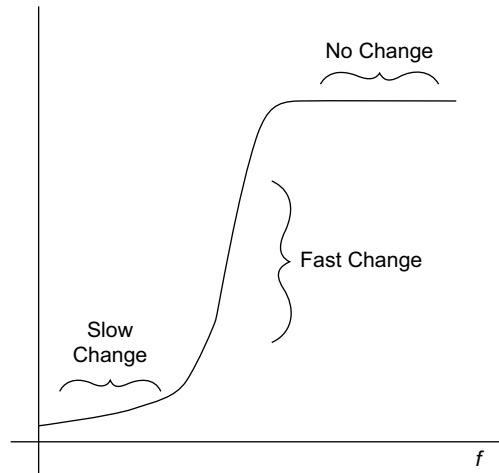


Figure 14-11 The rate of amplitude change seen by the network analyzer depends on the shape of the DUT's transfer function and the sweep speed of the analyzer.

rate limitation in a network measurement is less obvious since it depends on both the intermediate frequency (IF) bandwidth of the analyzer and the response of the device under test.

Since the source and receiver of a network analyzer are usually tuned to the same frequency, the signal out of the source lands in the center of the receiver passband.² Compare this with the spectrum analyzer case, where the signal is generated external to the analyzer, usually at some fixed frequency. The analyzer's receiver is swept past the stationary signal such that the signal is seen passing across the passband of the receiver. In the spectrum analyzer case, the signal is assumed to be a pure spectral line and the characteristics of the resolution bandwidth filter determines the maximum sweep rate.

In the network analyzer case, the signal is not sweeping past the receiver since the source and receiver are moving together. However, the signal amplitude at the output of the DUT will change as the sweep progresses. Consider a device under test that has the filter shape shown in Figure 14-11. As the sweep starts, the receiver will see a small signal due to the finite stop band of the filter. For the first part of the sweep, the signal amplitude will change slowly. When the sweep progresses to the filter skirt, the signal amplitude will begin to increase. The rate of amplitude increase will depend on the sweep rate and also the steepness of the filter response. The steeper the filter, the faster the amplitude will change. During this time, the receiver must respond quickly enough to track the increasing signal amplitude; otherwise, the measured response will be smeared as the receiver is unable to keep up. The wider the receiver bandwidth, the quicker it will respond.

To provide a little more perspective, consider a device that has a very flat amplitude response—a cable. If the cable were perfectly flat, the analyzer could sweep extremely fast because there are no amplitude variations for the analyzer to track.

² One exception is when the device under test has a large amount of delay relative to the sweep rate such that the receiver moves in frequency by a significant amount before the source's signal propagates through the device. When this happens, the sweep rate must be decreased to minimize the effect.

Because sweep rate is essentially determined by the device under test, network analyzers usually place the entire burden for determining the sweep rate on the user. The sweep rate is rarely automatically chosen (as in the case of a spectrum analyzer).

So how does one determine the optimum (fastest) sweep rate? Unfortunately, the most common way of setting the sweep rate is by trial and error. First, a starting sweep rate is chosen (probably by past experience). The frequency response of the device is noted, and the sweep rate is decreased. If the response does not change, the device is not being swept too fast. If the response does change, the device is being swept too fast and the sweep rate must be decreased. This process is repeated until the response is stabilized, implying that the sweep rate is adequate.

14.9 Power Sweep

Some network analyzers can automatically vary the amplitude of their source, producing a *power sweep* or *amplitude sweep*. This produces a measurement with the source power on the x -axis and some other parameter on the y -axis. For instance, the gain of an amplifier might be measured as a function of its input power to determine the gain compression point (Figure 14-12).

Unlike most other network measurements, power sweep measurements imply that the network is nonlinear. Normal S-parameter measurements assume that the network is either linear or nearly linear. The purpose of an amplitude sweep is to uncover and measure network parameters that change with varying amplitude, and such changes are inherently nonlinear.

14.10 Flexible Source Frequency

Most network measurements are made with the analyzer source and receiver set to the same frequency. However, many network analyzers now allow the source and receiver to be offset in frequency, enabling some additional measurements.

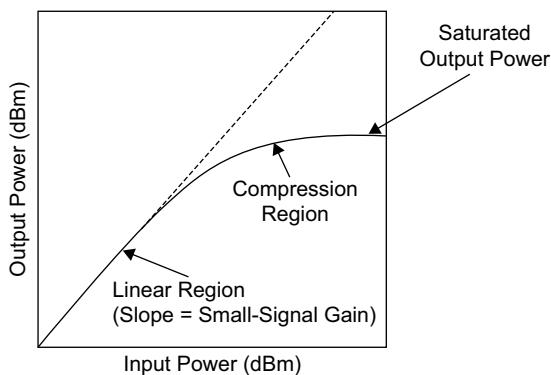


Figure 14-12 The linearity and compression characteristics of an amplifier can be measured using power sweep.

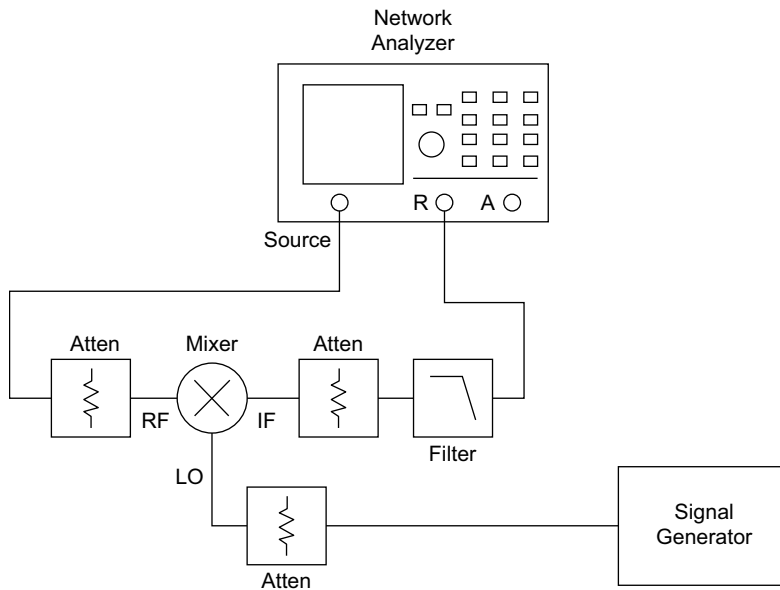


Figure 14-13 A network analyzer with frequency offset capability can be configured to characterize the conversion loss of a mixer.

Mixers and other frequency converters translate the input frequency to a different output frequency. To make network measurements on such a device, we'll need to sweep the source frequency over a frequency range that is different from the receiver frequency. The test configuration shown in Figure 14-13 is used to characterize the conversion loss of a mixer. The network analyzer source provides a swept signal to the RF port of the mixer, while a signal generator provides the local oscillator (LO) signal. The IF signal out of the mixer is measured by the network analyzer receiver. The network analyzer sweeps its source and receiver offset in frequency by the LO frequency, producing a swept response of the mixer's conversion loss. As shown, the measurement instruments are connected to the mixer ports with attenuators to minimize errors due to impedance mismatch at the mixer ports. Also, a low-pass filter is included at the mixer output to remove the unwanted mixer products.

Swept harmonic measurements characterize the harmonic distortion performance of a device over a range of frequencies. For swept harmonic measurements, the source is set to the fundamental frequency, f , and the receiver is tuned to nf , where n is the number of the harmonic to be measured. As the network analyzer sweeps in frequency, the receiver automatically tracks and measures the chosen harmonic. Since the entire frequency range can be measured in one sweep, it represents a large measurement speed improvement over other methods where the user must individually measure the harmonic level at each fundamental frequency with a spectrum analyzer.

Similarly, some network analyzers can perform a swept intermodulation distortion (IMD) measurement. The network analyzer must have two sources with their outputs combined to produce the two-tone stimulus required for an IMD measurement.

In Figure 14-14, the VNA is configured to measure simultaneously a robust set of DUT characteristics: S-parameters, intermodulation distortion, and gain compression.

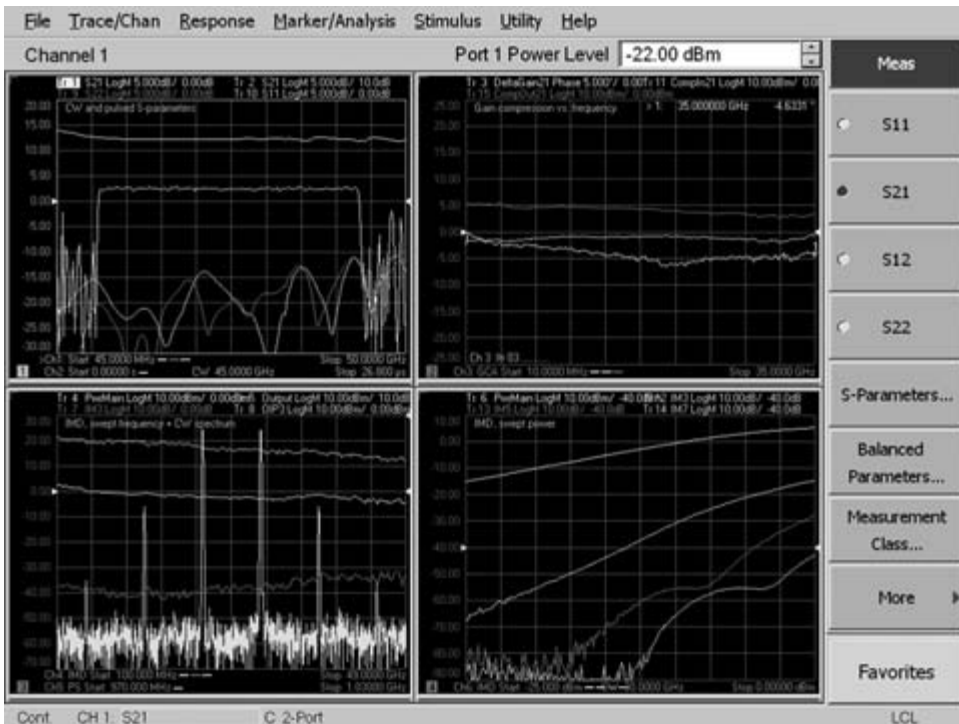


Figure 14-14 This VNA display is showing four different measurements simultaneously. (upper left) S-parameters. (lower left) Intermodulation distortion versus frequency. (upper right) Gain compression versus frequency. (lower right) Intermodulation versus power. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

14.11 VNA Time Domain Measurements

The inverse fast Fourier transform (IFFT) can be used to change a vector frequency domain network measurement into the time domain response of the network. This time domain response is usually displayed as either the impulse response or the step response of the network. This feature can be of great use if the time domain characteristics of the network are of interest. This measurement is roughly equivalent to *time domain reflectometry* (TDR) measurement, which is done using pulsed waveforms in the time domain.

Even if the time domain characteristics of the network are not desired, the time domain response can be used to remove certain imperfections in the measurement. For example, the reflections due to a poor connector may not be easily visible in the frequency domain, but when transformed to the time domain the connector's reflection will be apparent. This reflection can be removed by use of a gating function in the time domain. Then the gated time domain response can be transformed back into the frequency domain using a fast Fourier transform (FFT). The result is a frequency domain measurement that no longer contains the errors due to the reflections of an imperfect connector.

14.12 Nonlinear VNA Measurements

Classic network measurements are based on linear network and system theory, using sinusoidal stimulus and response. S-parameters are based on linear network analysis. Measurements using power sweep are not strictly linear measurements since the transfer function of the DUT changes depending on signal amplitude. Similarly, measurements of harmonic and intermodulation distortion are not strictly linear measurements. These nonlinear measurements are often made by stretching the measurement capability of a VNA.

To provide a more comprehensive approach to nonlinear measurements, Agilent Technologies developed the nonlinear vector network analyzer (NVNA) and an extension to S-parameters called X-parameters.³ X-parameters provide a more robust network model that accurately represents the nonlinear behavior of electronic circuits. For more information on NVNA measurements, see Agilent Technologies (2011) and Root et al. (2013).

Bibliography

Agilent Technologies. “Agilent Nonlinear Vector Network Analyzer (NVNA),” Publication Number 5989-8575EN, September 2011.

Agilent Technologies. “Agilent PNA Family Microwave Network Analyzers Configuration Guide,” Publication Number 5990-7745EN, October 2012.

Agilent Technologies. “Agilent PNA-X Series Microwave Network Analyzers,” Publication Number 5990-4592EN, September 2010.

Agilent Technologies. “E5061B Network Analyzer Data Sheet,” Publication Number 5990-4392EN, July 2011.

Agilent Technologies. “Making Stimulus/Response Measurements,” Application Note, Publication Number 5991-1877EN, March 2013.

Agilent Technologies. “Time Domain Analysis Using a Network Analyzer,” Application Note 1287-12, Publication Number 5989-5723EN, May 2012.

Curran, Jim. “Simplify Your Amplifier and Mixer Testing,” *RF Design*, April 1988.

Hewlett-Packard Company. “High Frequency Swept Measurements,” Application Note 183, Publication Number 5952-9200, December 1978.

Hewlett-Packard Company. “Vector Measurements of High Frequency Networks,” Application Note, Publication Number 5954-8355, March 1987.

Root, David E., Verspecht, Jan, Horn, Jason, and Marcu, Mihai. *X-Parameters: Characterization, Modeling, and Design of Nonlinear RF and Microwave Components*. Cambridge, UK: Cambridge University Press, 2013.

³ X-parameters is a trademark of Keysight Technologies, Inc.

Vector Network Measurements

Network measurements characterize the transmission through the device under test (DUT) and the reflections at each port. The device under test can have any number of input and output ports, but we'll focus on classic two-port measurements. For distortionless transmission through a device, the output signal must be identical to the input signal, perhaps delayed in time and scaled in amplitude. This implies the device that must have a flat amplitude response and a linear phase response. Group delay is the derivative of phase, which provides a useful way to view time delay through a device.

The use of scalar network measurements has decreased over time, and most measurements are now vector, including magnitude and phase. *Vector error correction* is a powerful technique for reducing measurement error, especially at high frequencies.

15.1 Distortionless Transmission

A system or network is called *distortionless* if its output is an exact replica of its input, except for amplitude scaling and time delay. Put mathematically,

$$y(t) = kx(t - t_0) \quad (15-1)$$

where

$y(t)$ = output signal

$x(t)$ = input signal

k = amplitude scale factor

t_0 = time delay in the system

Note that k and t_0 are constants and are not allowed to be a function of frequency. Figure 15-1 shows an example of input and output signals of a linear system. The input pulse has an amplitude of 1 and a pulse width of T . The output is the same shape as the input but is delayed by t_0 , and the amplitude of the output has been changed by the amplitude scale factor, k .

Now let us see how the criterion of distortionless transmission relates to a frequency domain measurement. Traditional network measurements are performed by exciting the network with a known sinusoid and measuring the amplitude and phase of the output relative to the input.

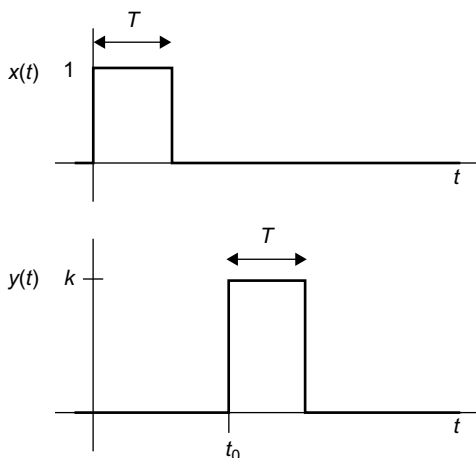


Figure 15-1 For distortionless transmission, the output of a system, $y(t)$, must be the same as the input, $x(t)$, except for time delay and amplitude scaling.

Let

$$x(t) = A \cos \omega t \tag{15-2}$$

For distortionless transmission,

$$y(t) = kA \cos[\omega(t - t_0)] \tag{15-3}$$

$$y(t) = kA \cos[\omega t - \omega t_0] \tag{15-4}$$

$$y(t) = kA \cos[\omega t - \theta(\omega)] \tag{15-5}$$

where

$\theta(\omega)$ = phase response of the system

$$\theta(\omega) = \omega t_0 = 2\pi f t_0$$

Thus, for distortionless transmission, the amplitude response of the system is a constant (flat with frequency) and the phase response is a linear function of frequency (Figure 15-2). Phase measurements are usually limited to $\pm 180^\circ$, and the phase plot in Figure 15-2 is shown wrapping around to stay within this range.¹ Note that the output sinusoid has the same frequency as the input sinusoid and that no other frequencies are present.

Having introduced a strict definition of distortionless transmission, we will note that only devices with infinite bandwidth, a completely flat magnitude response, and a linear phase response will meet this definition of distortionless. Often, this definition is too strict and can be relaxed based on knowledge of the intended application of a device. That is, a device may be considered distortionless over a particular range of frequencies or amplitudes.

¹ It certainly can be argued that the phase response continues on in a straight line, but the usual measurement techniques limit it to a principal angle covering one 360° range.

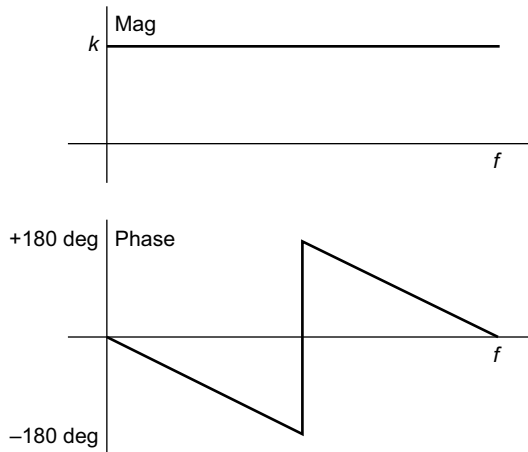


Figure 15-2 In the frequency domain, distortionless transmission implies a constant amplitude response and a linear phase response.

15.2 Nonlinearity

Many practical networks have nonlinearities that produce distortion in signals. Mathematically, these nonlinearities can be modeled using

$$y(t) = k_0 + k_1x(t) + k_2x^2(t) + k_3x^3(t) + k_4x^4(t) + \dots \quad (15-6)$$

As shown in Chapter 7, a sinusoidal input into this type of network produces output frequencies at the harmonics of the input. This violates the distortionless transmission criterion since the output does not have the same shape as the input.

Many networks that are considered linear will exhibit this type of response under some operating conditions. For instance, a typical “linear” amplifier will have some finite level of harmonics at its output due to distortion introduced in the amplifier. As long as these harmonics are small enough (as determined by the system requirements), we may choose to ignore them and consider the amplifier to be linear. However, if the amplifier is overdriven, the harmonic distortion may become severe, in which case we may need to treat it as a nonlinear system.

Nonlinearities are not limited to solid-state circuits or active circuits since passive circuits can exhibit nonlinear behavior. For instance, many iron core inductors will saturate at high current levels. This causes a nonlinear inductance in the circuit, which can cause the usual distortion products. Again, we may choose to consider such a network as distortionless as long as the distortion products remain below a certain level.

15.3 Linear Distortion

Many networks that do not produce frequencies other than the input frequency still do not meet the strict definition of distortionless transmission. They distort the signal by introducing

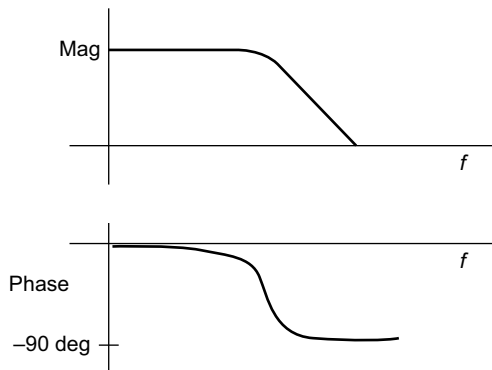


Figure 15-3 The amplitude and phase response of a single-pole low-pass filter.

amplitude characteristics that are not flat or phase characteristics that are not a linear function of frequency. These networks are sometimes said to have *linear distortion*.

One simple example of linear distortion occurs with a single-pole low-pass filter (Figure 15-3). At low frequencies, both the amplitude and the phase are constant and introduce no distortion. As the frequency increases, the amplitude rolls off and the phase changes. The amplitude rolling off is a clear violation of distortionless transmission. The phase is allowed to change, but it must change in a linear manner over the entire frequency range. In a single-pole filter, this is not the case, so the phase characteristic also introduces distortion.

But now consider the purpose of such a network: to remove undesirable high frequencies while retaining the low frequencies. At low frequencies, little or no amplitude or phase distortion is introduced, so the frequencies that appear at the output are not distorted. At higher frequencies, distortion is introduced, but these frequencies tend to be removed from the system anyway. So for frequencies of interest, we may choose to think of this as a distortionless network even though it does not meet the rigorous definition.

Many other examples exist. Band-pass networks are common in radio frequency (RF) receivers, and their amplitude response may be flat over some limited frequency range but not over all frequencies. Still, we may choose to consider them distortionless over that range. Even broadband amplifiers, which are usually considered flat in amplitude response, introduce linear distortion because they do not have infinite bandwidth. Amplifiers that are AC coupled do not pass arbitrarily low frequencies (and certainly not DC), and the amplifier's response rolls off on the high-frequency side. Over its frequency range, we may wish to consider an amplifier as linear.

In summary, *linearity* and *distortionless transmission* are assumed with many networks that do not meet the strict definition. In practice, this is not a problem as long as it is understood what is meant by *distortionless* for a particular application.

15.4 Importance of Linear Phase

Distortion is often discussed in terms of amplitude distortion of the waveform, particularly in the form of harmonics. But a device could have a perfectly flat amplitude response and still severely distort the waveform if the device's phase is not linear with frequency.

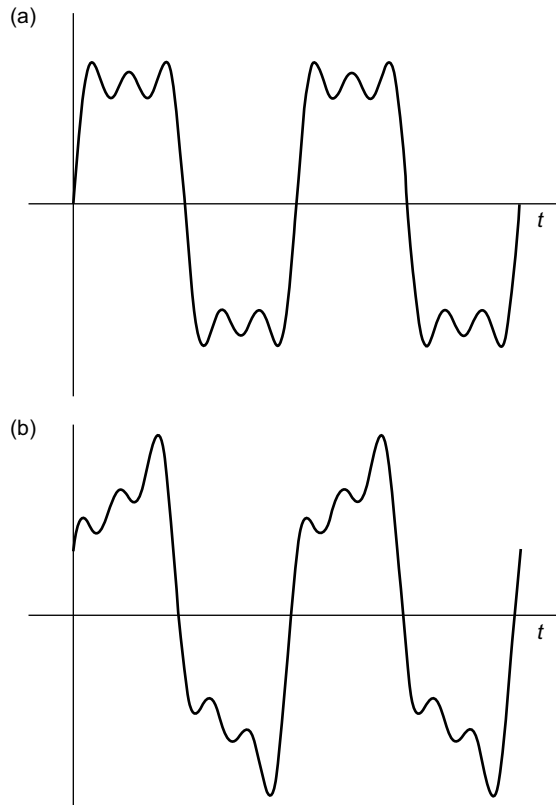


Figure 15-4 (a) The waveform that results from the first fundamental and third harmonic of a square wave. (b) The same waveform after passing through a system with nonlinear phase.

Consider the square wave as an example. Figure 15-4a shows the square waveform that results from including only the fundamental and third harmonic of a square wave. Recall that it would take an infinite number of harmonics to recreate the square wave exactly. However, the two sine waves combine to produce a waveform that is recognizable as a square wave. Notice how the fundamental is lined up in phase with the square wave and that the third harmonic adds to the fundamental in just the right places to make the combined waveform more square.

Suppose this signal is passed through a device that alters the phase relationship between the two sine waves (Figure 15-4b). Now the two sine waves add together in such a way that produces a new waveform—one that does not approximate the ideal square wave nearly as well. The amplitudes of the sine waves are still the same, but the phase relationship has been altered.

If both sine waves are delayed by an equal amount of time, the waveform appears at the output undistorted but delayed. A constant delay implies linear phase since a degree of phase at a low frequency is a much longer time than a degree of phase at high frequency. Thus, for the delay to be constant, the amount of phase shift required will increase with frequency.

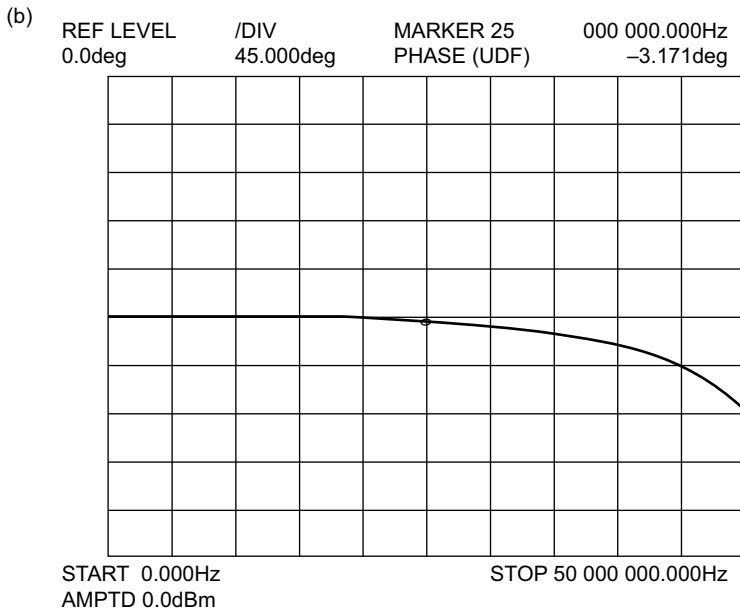
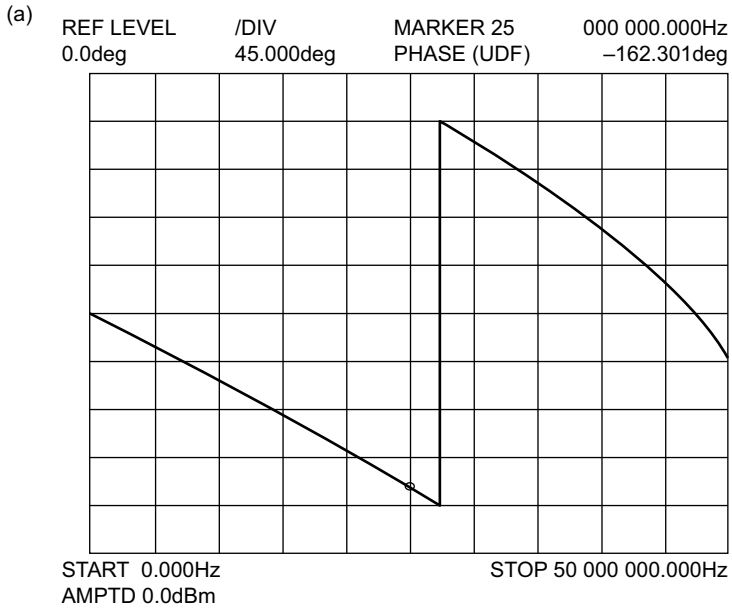


Figure 15-5 (a) A typical measurement of a high-order filter with a large amount of delay in the passband. (b) The same measurement after a large amount of linear phase is removed.

Linear phase response has become much more critical in electronic systems with the increased use of digital data and pulsed signals. If a single sine wave is passed through a system, it may not be important that the phase response is well controlled. However, when a pulsed or digital signal is transmitted, the phase response of the system must be linear so that the fidelity of the pulse is maintained.

Many devices such as high-order narrowband filters have a large amount of phase change over a small frequency range. It is difficult to determine how linear the phase response of a device is from a plot as shown in Figure 15-5a. Most network analyzers provide a feature to introduce or remove a user selectable amount of linear phase. The user can adjust the amount of linear phase introduced to flatten the measured phase response of the device under test. After the phase is made as flat as possible, the deviation from linear phase can be measured (Figure 15-5b).

Phase Error

So far the error analysis has concentrated on the magnitude error introduced by various mechanisms. Phase error is also important and can be derived from the magnitude error. Consider the vector diagram shown in Figure 15-6. In general, a magnitude error, ΔV , may have an arbitrary phase relationship with the signal, V , and introduces some corresponding phase error, $\Delta\theta$. The worst case for the phase error is when the new vector is perpendicular to the error vector. Under such conditions the phase error can be determined from

$$\sin(\Delta\theta) = \Delta V/V \quad (15-7)$$

$$\Delta\theta = \sin^{-1}(\Delta V/V) \quad (15-8)$$

Example 15.1

A particular measurement has a worst-case magnitude error of ± 1.0 dB. What is the corresponding phase error?

A magnitude error of ± 1.0 dB implies that

$$\begin{aligned} 1\text{ dB} &= 20 \log(1 + \Delta V/V) \\ \Delta V/V &= 10^{(1/20)} - 1 = 0.122 \end{aligned}$$

The worst-case phase error is

$$\begin{aligned} \Delta\theta &= \sin^{-1}(\Delta V/V) \\ &= \sin^{-1}(0.122) = \pm 7.01^\circ \end{aligned}$$

15.5 Group Delay

The group delay through a device is defined as the negative of the derivative of its phase response.

$$t_g = -\frac{d\phi}{d\omega} \quad (15-9)$$

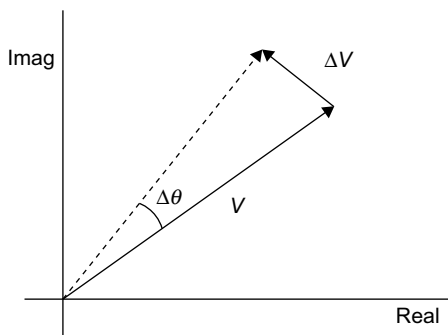


Figure 15-6 Phase error or uncertainty can be related to magnitude error using this vector diagram.

where

ϕ = phase response in radians

ω = frequency in radians/sec

If degrees and hertz are used,

$$t_g = -\frac{1}{360} \frac{d\phi}{df} \quad (15-10)$$

Because of the differentiation, a linear phase response produces a constant group delay. Deviations from linear phase show up as changes in the group delay as a function of frequency. Therefore, group delay flatness is used to specify and measure phase-related distortion. For instance, the group delay of a filter may be specified to be flat within some tolerance over its passband.

Delay Aperture

In modern network analyzers, group delay is usually derived from the phase response by calculating an approximation to the derivative.² The derivative of the phase is approximated by taking a small Δf in frequency and determining the corresponding phase change, $\Delta\phi$, as shown in Figure 15-7. The group delay is computed as

$$t_g = -\frac{\Delta\phi}{360\Delta f} \quad (15-11)$$

The term Δf is called the *delay aperture*, since it is the frequency aperture over which the delay measurement is computed. The delay aperture is usually selectable by the instrument user to optimize the measured results. Ignoring noise considerations, a small delay aperture seems appropriate since it more closely approximates the true derivative operation. However, the derivative operation tends to exaggerate any noise that happens to be present in the measurement. Since Δf appears in the denominator of the group delay calculation,

² In some network analyzers, group delay is measured using a modulated source. However, the concept of delay aperture is still valid.

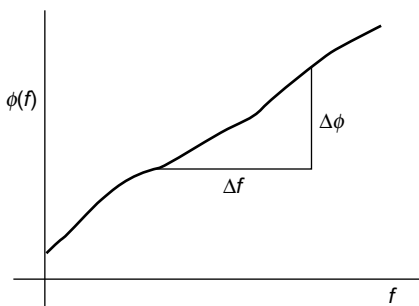


Figure 15-7 Group delay can be calculated by approximating the slope of the phase response.

making Δf smaller makes the noise in the group delay larger. Making the group delay larger tends to minimize the noise effects, but at the expense of frequency resolution (Figures 15-8a and 15-8b). Phase perturbations that are narrower in frequency than the delay aperture tend to be smeared over and will not be measurable. For fine frequency resolution or rapidly changing phase response, a narrow aperture is needed.

Delay aperture is often set on the network analyzer as a percent of the frequency span. For example, a 10 MHz span and a delay aperture set to 1% of the span produces a 100 kHz delay aperture.

15.6 Normalization

Normalization is a basic first-order error correction method that removes the frequency response of a test setup. Figure 15-9 shows a network analyzer with test set used to measure the transmission characteristics of a device. First, a *through connection* is used to connect the two ports of the network analyzer system. The response of the network analyzer, test set, and cabling is measured and stored in digital memory inside the network analyzer. Figure 15-10 shows the typical amplitude response of the measurement system. Besides the amplitude unflatness there is often significant linear phase response due to delays in the system. After the test system responses are stored, the measurement is made relative to them.³ After normalization is performed but with the through still connected, the analyzer display will show a flat 0 dB magnitude response and a flat 0° phase response. Next, the device under test is inserted into the measurement path and its characteristics are measured. The normalization data may be valid only with the particular analyzer setting, so if the analyzer setup is changed the test system may need to be renormalized.

Normalization is a simple but effective technique for removing error from the measurement. It requires that the network analyzer be stable with time, so if the analyzer response drifts significantly it will need to be normalized often. Assuming that the network analyzer is stable from one measurement to another, any absolute measurement errors are removed by the normalization.

³ The measurement system response is usually stored in digital memory as an array of complex (vector) numbers. The normalized measurement is computed by dividing the device response by the system response.

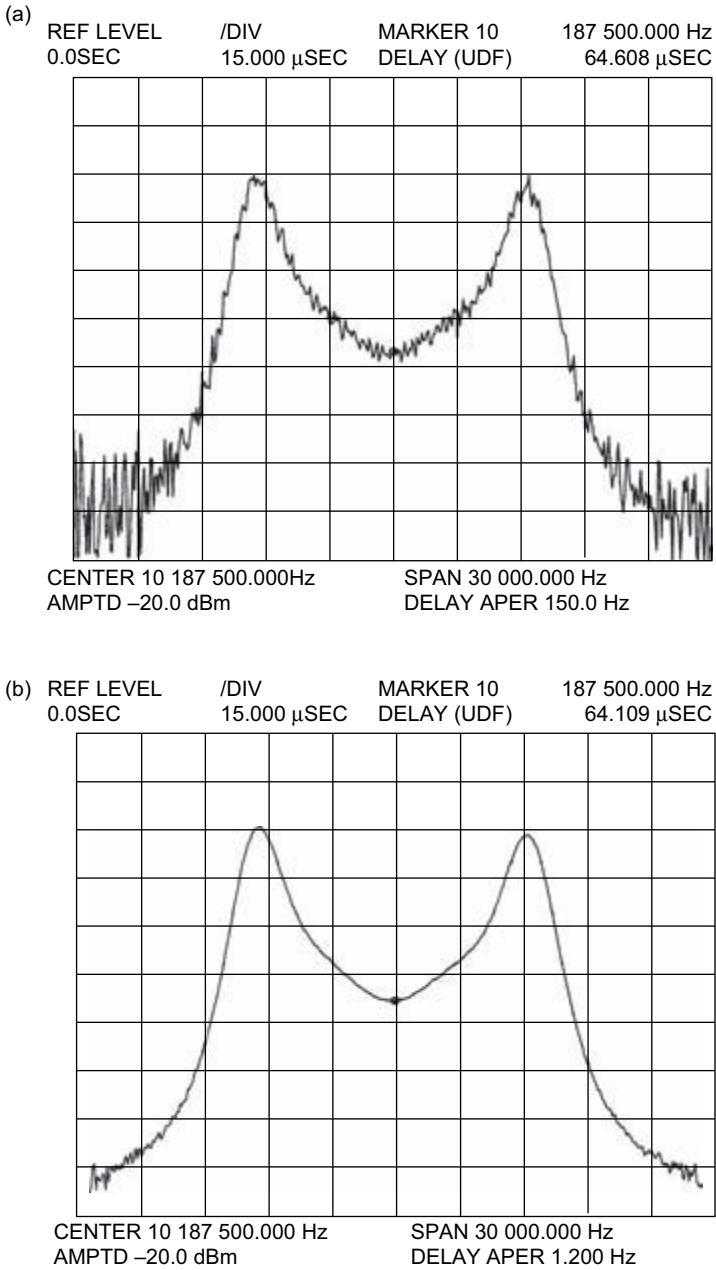


Figure 15-8 (a) A group delay measurement of a crystal filter using a narrow delay aperture.
 (b) The same group delay measurement with a wider delay aperture.

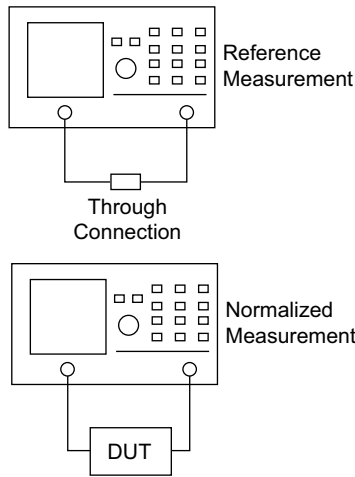


Figure 15-9 During normalization, a *through connection* is substituted for the device under test and the system response is measured and stored. Then the device under test is reinserted and the measurement is made relative to the stored normalization data.

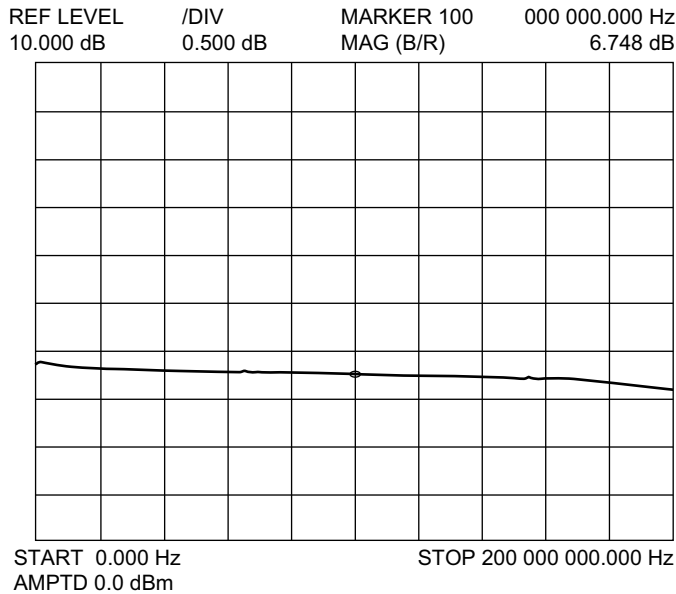


Figure 15-10 A typical system response measured during normalization. After normalization, the measured response will be perfectly flat.

Take the example of a perfect 0 dB insertion loss device. During normalization, the output power of the source is measured by the receiver. This measurement may have some absolute error in it, but when the perfect 0 dB device is inserted the network analyzer will read the exact same value. At 0 dB, which is where the normalization occurred, we can

achieve extremely good accuracy. Now suppose that we insert a 10 dB attenuator as the DUT. The power level at the network analyzer's receiver drops by 10 dB. It is not critical that the analyzer can measure the exact power level at the DUT output, but we do need the analyzer to measure accurately the power output *relative to the power level during normalization*. The network analyzer absolute accuracy is not important, but its *relative* accuracy is critical. In this example, we are interested in how accurately the analyzer can measure this 10 dB change, which is normally specified as *dynamic accuracy*.

Example 15.2

A network analyzer is used to make a normalized measurement of the insertion loss through a filter with a nominal loss of 3 dB. The network analyzer specifications include the following:

Source level accuracy: ± 1 dB

Source flatness: 1.5 dB peak to peak over the frequency range of the analyzer

Receiver absolute accuracy: ± 0.15 dB

Receiver dynamic accuracy: ± 0.02 dB

How much error may be introduced into the measurement due to these instrument errors?

Since normalization is used in this measurement, most of these instrument errors are removed. The effects of source level accuracy, source flatness, and receiver absolute accuracy are all removed during normalization. This leaves only the receiver dynamic accuracy contributing to error in the measurement. The total error introduced by these effects is ± 0.02 dB. Note that this is much better than the other specifications would imply.

15.7 Measurement Plane

The transmission line that connects the network analyzer to the device under test will introduce a delay in the signal. At low frequencies and with short cables, this delay may be negligible. As the frequency increases, the phase angle corresponding to a fixed amount of delay also increases. For example, suppose that a transmission line is 1 m long with a velocity factor, k_v , of 1. A 1 MHz sinusoid will have a wavelength of 300 m, so the 1 m cable represents $(1/300) \times 360 = 1.2^\circ$. At 100 MHz, the wavelength changes to 3 m, and the cable causes 120° of phase shift. The effect of the cable will obviously be noticed at 100 MHz when measuring a device connected to it.

With even a reasonably short length of transmission line changing the phase response that is measured by the network analyzer, it is necessary to define exactly where the measurement is taking place. This point is called the *measurement plane*.

One way to remove the effects of the transmission line delay is to use normalization. The through connection is placed at the ends of the measurement cables while any adapters and test fixturing should be left connected to the cable to remove their effect. In other words, the measurement connections should be the same as when the actual measurement is performed, except that the device under test is removed. This cannot be achieved exactly since a through connection must be inserted, which may introduce a small delay. The system response, including cable delay, is stored in the analyzer and the measurement is made relative to the

system response. Using this method the measurement plane exists at the ends of the transmission line.

Electrical Delay

Another method of removing transmission delay from the measurement is the use of *electrical delay* math function. Electrical delay (also called *line stretch* or *electrical length compensation*) is a network analyzer feature that adds or subtracts the effect of an ideal transmission line. This compensation is done mathematically by adding linear phase into the measured response, with the user specifying the physical length of the line or the amount of delay in seconds. The network analyzer will normally allow the user to enter the propagation velocity of the transmission line. The user may choose to enter a specific line length or delay or may simply adjust the amount of line strength until the phase response flattens.

15.8 Reflection Measurements

Reflection measurements characterize the impedance match of a two-terminal port of a device, which may have one or more ports. The fundamental reflection measurement is the complex reflection coefficient, which is the ratio of the reflected voltage to the incident voltage. (The reflection coefficient is based on transmission line theory introduced in Chapter 11.)

$$\Gamma = V_R/V_I \quad (15-12)$$

The magnitude of the reflection coefficient may be displayed on a linear scale but is more commonly shown in decibel form as return loss.

$$\text{RL} = -20 \log(|\Gamma|) = -20 \log(\rho) \quad (15-13)$$

The return loss calculation includes a minus sign that causes the return loss values to be positive. When measured on a network analyzer, the minus sign is often omitted, producing measured values that are negative. For example, a network analyzer may read -40 dB, corresponding to a return loss of 40 dB. Figure 15-11 shows a typical return loss measurement of a band-pass filter.

SWR and Impedance

Other parameters can be derived from the complex reflection coefficient and displayed on a rectangular graticule. Equations for standing wave ratio (SWR) and complex impedance were covered in Chapter 11. An example of an SWR plot is shown in Figure 15-12.

Polar Display Formats

The reflection coefficient is a complex number and is often displayed in polar format, on a complex plane with a horizontal real axis and a vertical imaginary axis (Figure 15-13). In a polar display, the frequency axis is lost, but the network analyzer marker or cursor will provide the user with the frequency of any particular data point. Any particular frequency point is plotted according to the magnitude, ρ , and phase, θ , of the reflection coefficient.

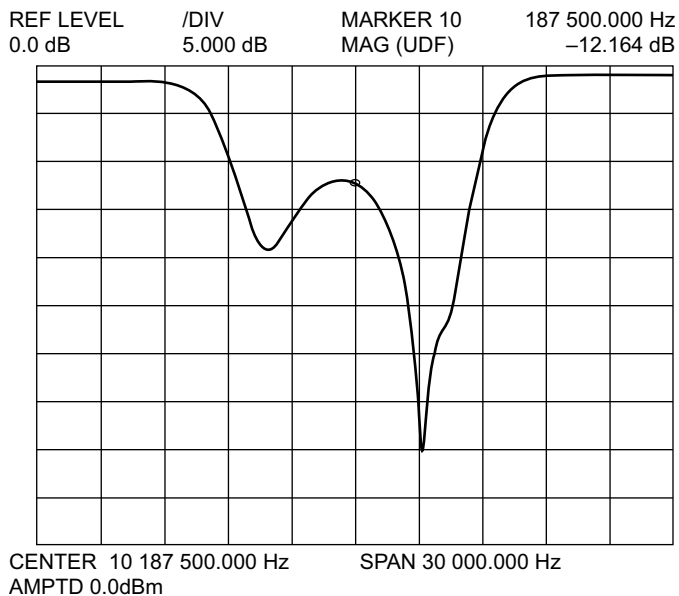


Figure 15-11 A return loss measurement of the input port of a band-pass filter. The return loss exceeds 10 dB in the center of the passband (the lower the trace, the better the match.)

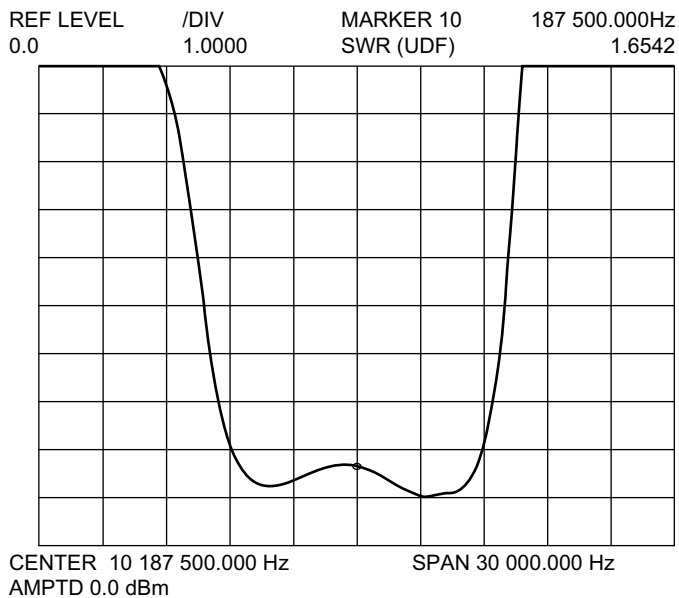


Figure 15-12 An SWR measurement of the same band-pass filter measured in Figure 15-11.

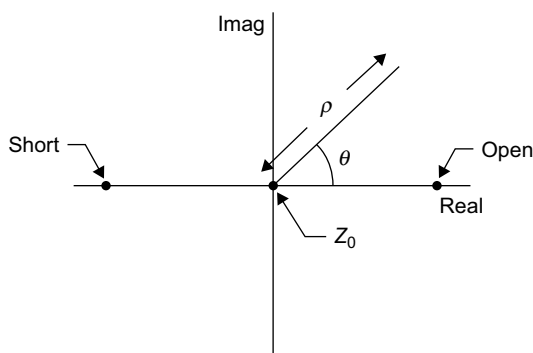


Figure 15-13 The complex reflection coefficient is often plotted in polar format.

The point is plotted at the end of a vector that starts from the center of the polar plot and extends outward a distance equal to ρ and at an angle of θ . The angle is determined relative to the right hand real axis, which is defined as 0° .

If a perfect Z_0 load is connected to the test port, there will be no reflection and the reflection coefficient will be $0 \angle 0^\circ$, or $0 + j0$, which is the center of the polar plot. An open circuit produces complete reflection of the incident wave with a reflection coefficient of $1 \angle 0^\circ$, which is plotted on the right-hand side of the polar display. A short on the test port causes complete reflection of the incident wave and the reflection coefficient is -1 or in polar format, $1 \angle 180^\circ$. This point is plotted on the left-hand side of the polar plot. All three of these points are shown plotted on the complex plane in Figure 15-13.

The Smith Chart

A variation on the polar plot of the complex reflection coefficient is the *Smith chart*. The reflection coefficient is still plotted in polar form, but with a different graticule called the Smith chart (Figure 15-14). The Smith chart is a familiar tool to radio frequency engineers and is used extensively in design work. As a network analyzer graticule, the Smith chart converts the complex reflection coefficient to normalized impedance. (Many other conversions are possible with the Smith chart.)

The normalized impedance, z , is given by

$$z = Z/Z_0 \quad (15-14)$$

where

Z = unnormalized impedance

Z_0 = characteristic impedance of the system

The normalized and unnormalized impedances can also be expressed in terms of their resistive and reactive components.

$$Z = R + jX \quad (15-15)$$

$$z = r + jx \quad (15-16)$$

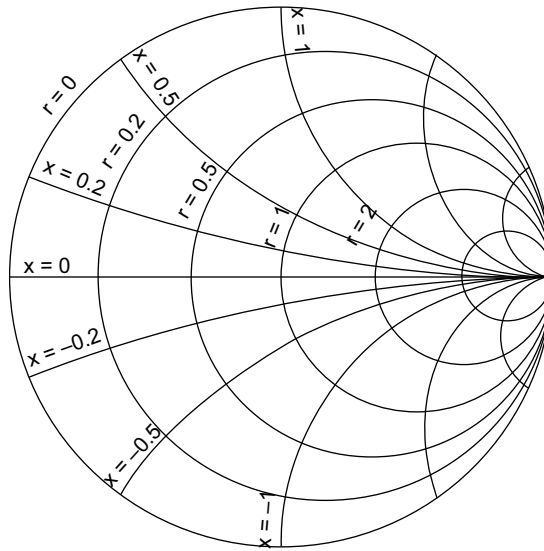


Figure 15-14 The Smith chart is a graphical mapping of the complex reflection coefficient into normalized complex impedance.

and

$$r = R/Z_0 \quad (15-17)$$

$$x = X/Z_0 \quad (15-18)$$

Example 15.3

In a 50Ω system, a particular value of complex reflection coefficient is plotted on the Smith chart and the normalized impedance is $0.3 - j2$. What is the impedance (unnormalized) for this value of reflection coefficient?

$$z = 0.3 - j2$$

$$z = Z/Z_0$$

$$Z = Z_0 z = 50(0.3 - j2) = 15 - j100 \Omega$$

Evaluating the complex reflection coefficient and plotting the locus of points that have the same normalized resistance produces circles of constant resistance as shown in Figure 15-15a. Similarly, the locus of points having the same reactance can be plotted on the complex reflection coefficient plane, producing arcs of constant normalized reactance shown in Figure 15-15b. Note that the normalized reactance can be either positive or negative, corresponding to inductive and capacitive impedances, respectively. Combining these two loci of points produces the complete Smith chart.

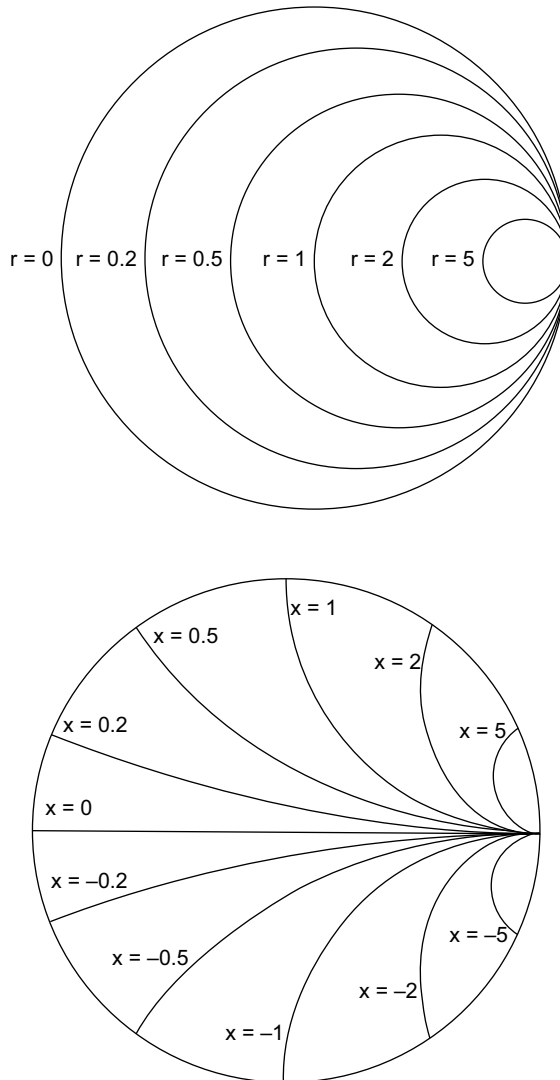


Figure 15-15 (a) Circles of constant resistance on the Smith chart. (b) Lines of constant reactance on the Smith chart.

Figure 15-16 shows the reflection measurement from Figure 15-11 in polar form with a Smith chart graticule.

The Smith chart's circles of constant resistance and lines of constant reactance provide a graphical conversion from reflection coefficient to normalized impedance.⁴ But the Smith chart is much more than just a graphical conversion technique. A wide variety of analysis

⁴ See Hayt (2000) for a more complete derivation of the Smith chart.

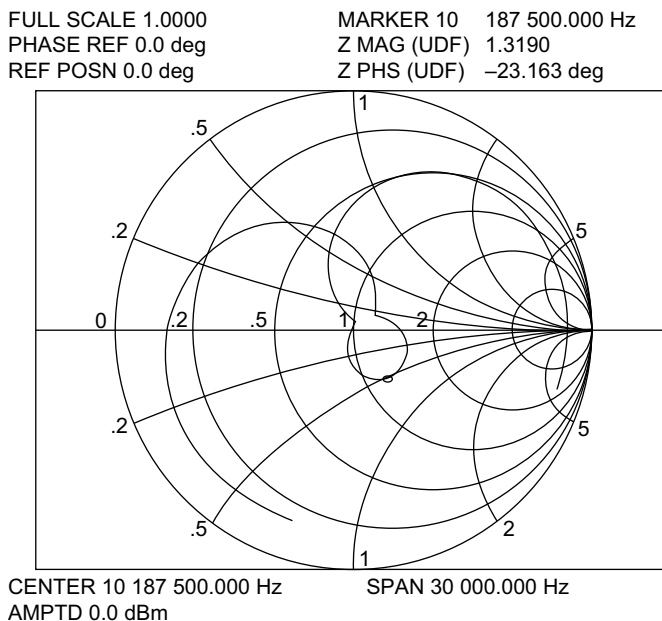


Figure 15-16 Polar plot of input reflection of band-pass filter from Figure 15-11, with Smith chart graticule.

and design methods, using the Smith chart, have been developed, making it a standard design tool for engineers working in the radio frequency and microwave areas. Entire textbooks have been written on these design techniques, which is beyond the scope of this book.⁵

15.9 Directional Bridges and Couplers

A directional bridge or a directional coupler can be used to extract the incident or reflected voltage along a transmission line or at a port of a device under test. Ideally, the bridge or coupler measures only the wave traveling in the desired direction and ignores any traveling waves going the other way. However, bridges and couplers have practical limitations that will be discussed later in the chapter.

The directional bridge and directional coupler perform the same basic function, but with different techniques. First, let's examine the operation of the directional bridge. A simplified circuit diagram of a directional bridge is shown in Figure 15-17. When the test port is terminated in a perfect Z_0 , the bridge is balanced and the detector will measure 0 V, indicating that no reflected wave is present. Now suppose the test port is left open. The detector port would receive half of the source voltage, indicating a large reflection. On the other hand, if the test port were shorted the detector port would again receive half of the source voltage, but with opposite polarity. Thus, this directional bridge produces a detector voltage that is

⁵ See Gonzalez (1996) for more information on design techniques using the Smith chart.

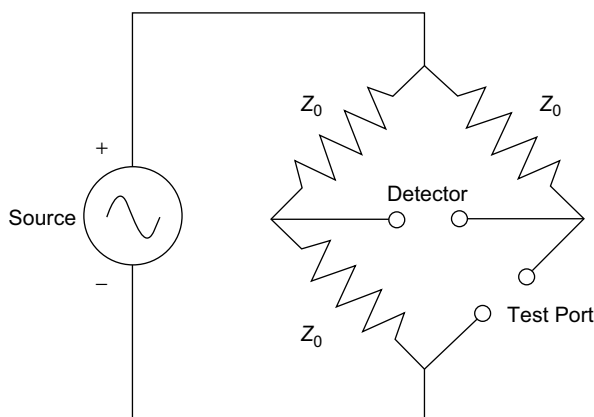


Figure 15-17 A simplified circuit of a directional bridge.

proportional to the amount of reflection. Just as important, the detector voltage phase or polarity indicates the phase of the reflection.

Notice that the detector port has a voltage on it that is a replica of the source voltage. More specifically, this means that the detector port signal is at the source frequency and must be detected. When used with a network analyzer, the analyzer receiver will perform the detection and measurement of the detector signal. A directional bridge is also used in meters designed to display SWR, called a *reflectometer* or simply *SWR meter*. In that case, a diode detector circuit converts the detector voltage to a DC level, which drives a conventional voltage meter.

One practical issue that must be handled is that the detector port is balanced precariously on the directional bridge and is not referenced to ground. Network analyzers usually have one side of their receiver inputs connected to ground. The network analyzer's source is also usually grounded, so driving such a network analyzer directly with the bridge circuit shown in Figure 15-17 would cause the bridge circuit to be unbalanced. To sidestep this problem, the ground connection to the source or detector port must be broken. Directional bridges usually have a transformer (or balun) that either creates a balanced source (floating with respect to ground) or converts the balanced detector output to a single-ended output with one side connected to ground. While this transformer solves the problem, it does not operate at or near DC so the frequency response of the bridge is forced to roll off at some low frequency, typically 10 kHz to 100 kHz.⁶

At microwave frequencies, a directional coupler is most often used to separate traveling waves. A directional coupler provides the same basic function as the directional bridge, but using waveguide techniques to separate the traveling waves. For our purposes, directional bridges and directional couplers perform the same basic function. Therefore, the term *directional device* will be used to loosely refer to both bridges and couplers.⁷

⁶ For a more detailed discussion of a practical directional bridge, see Spaulding (1984).

⁷ See Dunsmore (2012), Laverghetta (1984), and Adam (1969) for more information on directional couplers.

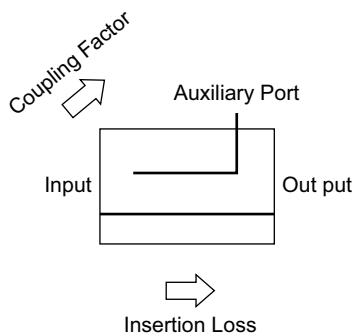


Figure 15-18 The representation of a three-port directional device.

Insertion Loss

Figure 15-18 shows a representation of a directional device, with the three ports labeled. (Sometimes a fourth port is shown, but since it is almost always terminated with a $50\ \Omega$ load it is not shown here.) The signal applied to the input will appear at the output with some amount of loss. The insertion loss of a directional device is the ratio of the input power to the output power, expressed in decibels, typically in the range of 0.5 to 3 dB. For a given source power level, large insertion loss means less power delivered to the test port.

Coupling Factor

The coupling factor of a directional device is a measure of how much signal appears at the auxiliary port for a given signal level at the input port (Figure 15-18). The coupling factor is defined as the ratio of the input power to the auxiliary port power, expressed in decibels.

A typical coupling factor is in the range of -10 to -20 dB. Note that the coupling factor is shown as a negative number in decibels, indicating that the signal at the auxiliary port is smaller than the input signal. However, similar to other parameters expressed in decibels, it is common to refer to the coupling factor as a positive number (e.g., 20 dB instead of -20 dB).

The power present at the auxiliary port of a directional device is proportional to the directional traveling wave that is being measured. The terminology used here and shown in Figure 15-18 assumes that the forward traveling wave is being measured (i.e., a wave traveling from left to right in the figure).

Directivity

The most important figure of merit for a directional device is *directivity*, which indicates how well a directional device can separate opposite traveling waves. Ideally, the directional device can completely separate the forward and reverse waves, but some of the forward wave is present in a measurement of the reverse wave (and vice versa). Directivity is defined as the ratio of the power present at the auxiliary port when the signal is traveling in the forward direction to the power present at the auxiliary port when the same signal is traveling in the reverse direction. This ratio is expressed in decibel form, is ideally infinite but typically is

30 to 40 dB. Finite directivity can be thought of as being a leakage path that lets the undesired traveling wave couple into the auxiliary port.

Directivity is important because it limits the maximum return loss that can be measured using a particular directional device.

Example 15.4

A directional bridge has an insertion loss of 6 dB, a coupling factor of 6 dB and a directivity of 40 dB. Configured as shown in Figure 15-16 and with a source power of 0 dBm, determine the power levels at the auxiliary port and the output port with the output terminated in Z_0 .

The insertion loss of the bridge is 6 dB; therefore, the output level is 0 dBm -6 dB = -6 dBm. The coupling factor is also 6 dB, so the power at the auxiliary port is 0 dBm -6 dB = -6 dBm. Since there is no reflected signal from the output port (due to the Z_0 termination), the directivity does not affect the signal level at the auxiliary port. Had there been a reflection, a portion of that signal would have also appeared at the auxiliary port.

15.10 Reflection Configuration

Now let's reverse the orientation of the directional device and rename the ports to be more appropriate for a reflection measurement (Figure 15-19). A signal will be applied to the input port (formerly the output port) that will excite the DUT at the test port. If the DUT is a perfect Z_0 match, no reflection will occur and no signal will be present at the auxiliary port. (In fact, the auxiliary port will have a small signal present due to the finite directivity of the directional device.) If the test port has a non- Z_0 load connected to it, a reflection will occur and the signal level at the auxiliary port will be proportional to the size of the reflected wave. A directional bridge or coupler is used in this manner to perform a reflection measurement.

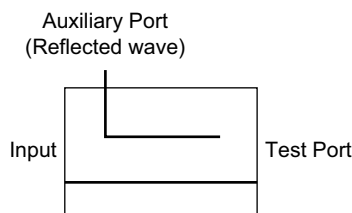


Figure 15-19 The representation of a directional device labeled with reflection measurement terminology.

Example 15.5

The bridge described in Example 15.3 is configured for a reflection measurement as shown in Figure 15-19. The input signal level is 0 dBm. Determine the signal level at the auxiliary port for the following loads at the test port: (a) short and (b) Z_0 load.

- (a) Load = short. The wave incident on the test port is $0 \text{ dBm} - 6 \text{ dB}$ (insertion loss) = -6 dBm . Since the load is a short circuit, all the incident wave is reflected and appears at the auxiliary port, but reduced by the coupling factor. Thus, the power at the auxiliary port is $-6 \text{ dBm} - 6 \text{ dB} = -12 \text{ dBm}$. The directivity of the device will also contribute to the power at the auxiliary port. The auxiliary port signal due to the directivity is the input signal level reduced by the insertion loss, the coupling factor, and the directivity. This signal level is $0 \text{ dBm} - 6 \text{ dB} - 6 \text{ dB} - 40 \text{ dB} = -52 \text{ dBm}$. This signal will introduce an error at the auxiliary port, but since -52 dBm is much smaller than -12 dBm , its effect will be slight.
- (b) Load = Z_0 . The incident wave (which is fully absorbed at the test port) is still -6 dBm . Since none of the wave is reflected, the auxiliary port would have (ideally) no signal. However, the finite directivity of the device will cause a signal to be present. As previously calculated the power in this signal will be -52 dBm . Therefore, even with a perfect Z_0 load this bridge will indicate a reflected power of $-52 \text{ dBm} + 6 \text{ dB} = -46 \text{ dBm}$, referred to the test port. Since -6 dB was incident on the test port, the reflected power of -46 dBm corresponds to a device with a return loss of 40 dB . Therefore, a directional device cannot directly measure a return loss value greater than the device's directivity.
-

15.11 Reflection Normalization

Normalization of transmission measurements was discussed earlier in the chapter. This same concept can be applied to reflection measurements.

The coupling factor of the directional coupler is a nuisance when determining the return loss measured with a directional device. Since return loss is defined to be the ratio (in dB) of the reflected power to the incident power, the coupling factor and the insertion loss of the directional device must be accounted for. Also, the coupling factor and insertion loss will not be constant with frequency and will introduce a frequency response error. Both of these problems can be eliminated by the use of normalization.

During normalization of a reflection measurement, a short or open is placed on the test port, causing the whole incident wave to be reflected. This will cause a signal to appear at the auxiliary arm that is equal to the input signal reduced by the insertion loss and coupling factor of the directional device. Since this signal represents complete reflection at the test port, it corresponds to 0 dB return loss. More significantly, when swept over the frequency range of interest this signal will exhibit the frequency response of the directional coupler. Saving such a frequency sweep in the analyzer's memory and measuring relative to it results in a reflection measurement with the frequency response error removed.

We briefly mentioned that a short or open can be used to provide a totally reflected signal. At frequencies below a few hundred megahertz, either one will suffice so the choice is not critical. At higher frequencies, the short circuit is often preferred since it provides a more reliable and repeatable termination than the open circuit, which suffers from stray capacitance effects. The short circuit, having a reflection coefficient of -1 , introduces a 180° phase change at the test port. (The magnitude will be unaffected.) Some network analyzers provide a special normalization feature that removes the effect of this phase change. If not, the user must remember to invert the phase of the measured data.

15.12 Error in Reflection Measurements

The standard error model for a directional device is described by⁸

$$\Gamma_M = D + \frac{(1 + T_R)}{(1 - M_S \Gamma_A)} \Gamma_A \quad (15-19)$$

where

Γ_A = actual reflection coefficient

Γ_M = measured reflection coefficient

D = directivity error

T_R = frequency response error

M_S = source match error

(All these variables are complex and are a function of frequency.)

Notice that the directivity error term appears as a constant in the equation. This implies that the absolute error introduced by the finite directivity is independent of the load. (Recall that directivity can be thought of as a leakage from the input port to the auxiliary port.) The $(1 + T_R)$ term in the equation represents the frequency response of the system. Any source, receiver, or directional device unflatness will be accounted for here.

The last term in the equation is $1/(1 - M_S \Gamma_A)$, which itself depends on the reflection coefficient. This term accounts for source match errors, that is, error introduced due to the lack of a perfect Z_0 impedance looking back into the directional device. This type of mismatch will rereflect a portion of the signal that is reflected from the test port. The double reflection will show up at the test port and introduce an error.

The uncertainty in the reflection coefficient measurement is defined by

$$\Delta\Gamma = \Gamma_M - \Gamma_A \quad (15-20)$$

$$\Delta\Gamma = D + \frac{(1 + T_R)}{(1 - M_S \Gamma_A)} \Gamma_A - \Gamma_A \quad (15-21)$$

$$\Delta\Gamma = D + \frac{T_R \Gamma_A + M_S \Gamma_A^2}{(1 - M_S \Gamma_A)} \quad (15-22)$$

For most measurement situations, the product of the source match coefficient, M_S , and the actual reflection coefficient, Γ_A , is much smaller than 1. The equation can be simplified to

$$\Delta\Gamma = D + T_R \Gamma_A + M_S \Gamma_A^2 \quad (15-23)$$

which is the classical result for predicting errors in reflection measurements. Since the phase of the complex reflection coefficient is not usually known, Γ is replaced by ρ as we consider the worst case.

$$\Delta\rho = D + T_R \rho_A + M_S \rho_A^2 \quad (15-24)$$

⁸ This error model was originally described in Ely (1967).

Example 15.6

Determine the measurement uncertainty in the following reflection measurement. The actual return loss is 10 dB, the directivity of the directional bridge is 40 dB, the effective source match is 20 dB, and the frequency response error is ± 1 dB.

First, convert each of the measurement parameters from decibels into linear values.

$$\rho_A = 10^{-(20/20)} = 0.316$$

$$D = 10^{-(40/20)} = 0.01$$

$$M_S = 10^{-(20/20)} = 0.1$$

$$\text{freq. response (db)} = 20 \log(1 + T_R)$$

$$T_R = 10^{(1/20)} - 1 = 0.122$$

$$\begin{aligned} \Delta\rho &= D + T_R\rho_A + M_S\rho_A^2 \\ &= 0.01 + (0.122)(0.316) + 0.1(0.316)^2 \end{aligned}$$

$$\Delta\rho = \pm 0.585$$

ρ could actually be anywhere between $0.316 + 0.0585 = 0.375$ and $0.316 - 0.0585 = 0.258$ or, expressed as return loss, 8.52 to 11.77 dB.

15.13 Vector Error Correction

The three main error mechanisms just discussed can be characterized and eliminated from the measurement through *vector error correction*, also known as *accuracy enhancement*. It has already been mentioned that normalization provides a means of removing frequency response errors from a reflection measurement, but it does not improve the directivity of the directional device. Recall that finite directivity means that the traveling wave going in the opposite (undesired) direction will contribute to the measured level of the desired traveling wave. These two signals will add vectorally and may add constructively, destructively, or anywhere in between, showing up as a ripple in the response of the reflection measurement. When the two signals add destructively (totally canceling), the error approaches infinity. When the two signals add together in phase, the measured value may be off by as much as a factor of 2 (or 6 dB).

Error correction involves the measurement of the characteristics of the directional device over the frequency range of interest, storing them in digital form and correcting the measured values to produce a more accurate measurement. In the past, the complexity of the error correction computation and the amount of digital storage needed required the use of an external computer. With large memories and powerful microprocessors, the error correction calculations can be performed internal to the instrument. Typically, the user is prompted to attach the appropriate terminations (Z_0 , open and short) during error correction calibration, and the instrument does the rest. In some cases, the error correction may be good only for the selected frequency range and changing the frequency range may require a new error correction procedure to be initiated by the user.⁹

⁹ This calibration procedure should not be confused with the internal calibration of the instrument, periodically done by a metrology lab.

15.14 Normalization Revisited

Let us reexamine the concept of normalization, given our error model for directional devices. Normalization of a reflection measurement required the placement of an open or short at the test port of the directional device. This causes the entire signal to be reflected back to the auxiliary port. For a device with good directivity (30 or 40 dB), this reflected signal will be much larger than the error signal caused by finite directivity. Also, for a directional device and a source producing a good Z_0 impedance, the source match error can be ignored. Thus, reflection normalization measures the frequency response error term of the error model (also called the *reflection tracking error*) while setting $D = 0$ and $M_S = 0$.

$$\Gamma_M = (1 + T_R)\Gamma_A \quad (15-25)$$

The error correction or normalization term, E_N , is

$$E_N = 1 + T_R \quad (15-26)$$

which is a complex function of frequency stored away in digital memory. To produce the corrected measurement, the error model equation is reversed.

$$\Gamma_A = \Gamma_M / E_N \quad (15-27)$$

15.15 Two-Term Error Correction

A two-term error correction can be performed by ignoring the source match error and correcting the directivity and frequency response errors. The advantage of this technique is shorter calibration time (compared with three-term error correction, which will be described shortly), since only two terminations (open and Z_0 load) must be measured. Ignoring the source match error is not too much of a compromise in many cases, particularly if the impedance looking back into the directional device is very close to Z_0 or the input of the device under test is close to Z_0 . The equation for the two-term error model is

$$\Gamma_M = D + (1 + T_R)\Gamma_A \quad (15-28)$$

Two calibration measurements are required, the first one with a Z_0 load. This sets the reflection coefficient equal to zero so that the directivity term can be measured. On the second calibration measurement, an open (or sometimes a short) is placed on the test port, causing the reflection coefficient to be equal to one. Thus, the frequency response term can be measured. The directivity term is still present in the second measurement but may be subtracted off by the analyzer error correction algorithm.

15.16 Three-Term Error Correction

Using the full three-term error correction model necessitates the use of three calibration measurements and three unique terminations: a Z_0 load, an open, and a short. With the Z_0 load attached, the reflection coefficient is zero and only the directivity appears in the

measurement. When a short is attached, the reflection coefficient is -1 , and the error model reduces to

$$\Gamma_M = D + \frac{(-1)(1 + T_R)}{(1 + M_S)} \tag{15-29}$$

Then an open circuit is attached, making the reflection coefficient equal to one. This gives

$$\Gamma_M = D + \frac{(1 + T_R)}{(1 - M_S)} \tag{15-30}$$

With two equations and two unknowns, T_R and M_S , the two unknowns can be calculated, resulting in a complete three-term error correction. Again, the correction factors are stored in memory and the error model is used to enhance the accuracy of the measurement.

Figure 15-20 shows the return loss measurement of a Z_0 load, with and without three-term error correction. The difference between the two traces represents the improvement in directivity that the error correction provides. The directivity errors have been removed from the corrected measurement, effectively leaving no detected auxiliary port voltage. Thus, the analyzer measures its own input noise, and the measurement appears noisier than the uncorrected trace. Even though the trace is noisier with a Z_0 load connected, the substantial directivity improvement provides a much more accurate measurement.

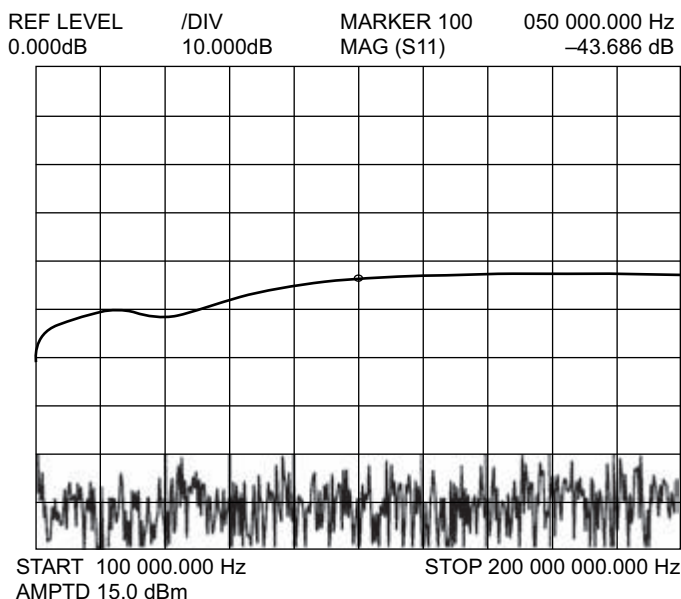


Figure 15-20 A return loss measurement of a Z_0 load without error correction (upper trace) and with three-term error correction (lower trace). The directivity is improved from approximately 43 dB to over 80 dB.

15.17 Two-Port Error Correction

Error correction is not limited to reflection measurements with a single directional device. The same concepts can be expanded to apply to a two-port measurement using a scattering parameter (S-parameter) test set. In addition to the directivity, source match, and reflection frequency response errors, two-port measurements incur errors due to load match, transmission frequency response, and crosstalk. Therefore, correction for these errors requires a more complex model, with more correction terms and more calibration measurements, resulting in a more accurate measurement. The error correction procedure for an S-parameter test set is more involved since both ports must be calibrated for reflection measurements as well as transmission measurements. In addition to the short-, open-, and load-type calibration discussed under one-port error correction, a through connection between the two ports is measured.

Figure 15-21 shows the forward error model for the two-port error correction. For simplicity, only half of the error terms for the two-port model are shown: the ones relevant to *forward* measurements. The model shown is sufficient for error-corrected S_{11} and S_{21} measurements. There is a corresponding error model for the reverse measurements, S_{22} and S_{12} .

The forward error model includes six error terms, listed in the left column of Table 15-1. The forward directivity error (EDF) is caused by the imperfect directivity of the directional device in Figure 15-21. Similarly, error terms for the *forward source match* (ESF), *forward reflection tracking* (ERF), *forward load match* (ELF), *forward transmission tracking* (ETF), and *forward crosstalk* (EXF) are associated with the imperfections of the test setup in Figure 15-21. Again, these are only the error terms for the forward measurement case (as shown in Figure 15-21.) There is a complementary set of error mechanisms working in the reverse direction captured in the right column of Table 15-1.

This book covers the basic concepts associated with vector error correction. Appendix A shows more detail concerning the mathematical model for the two-port case. The nomenclature for the two-port error terms is consistent with Dunsmore (2012), which is an excellent reference book for diving deeper into error correction.

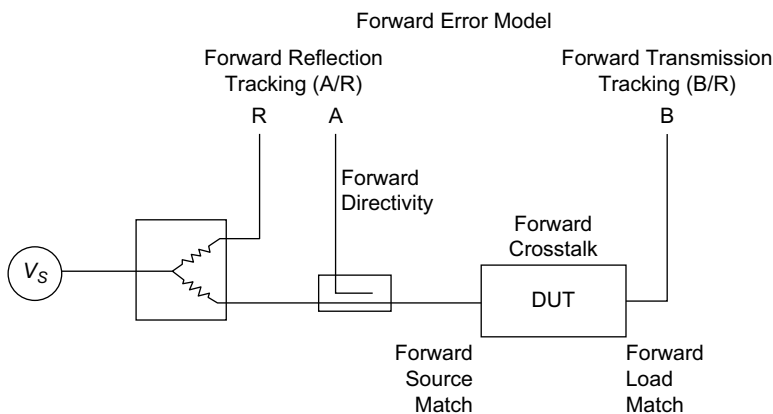


Figure 15-21 The forward error model is used to show the error terms for a two-port S-parameter measurement.

Table 15-1 Error Mechanisms Associated with Two-Port Vector Network Analysis^a

Forward Error Terms	Reverse Error Terms
EDF Forward Directivity	EDR Reverse Directivity
ESF Forward Source Match	ESR Reverse Source Match
ERF Forward Reflection Tracking	ERR Reverse Reflection Tracking
ELF Forward Load Match	ELR Reverse Load Match
ETF Forward Transmission Tracking	ETR Reverse Transmission Tracking
EXF Forward Crosstalk	EXR Reverse Crosstalk

^aThe nomenclature is consistent with Dunsmore (2012).

Opens, Shorts, and Z_0 Loads

Vector error correction depends on the use of high-quality terminations, most commonly open circuits, short circuits, and Z_0 loads. A calibrated measurement uses terminations as a reference standard, which means that the quality of the measurement is ultimately limited by these terminations.

First, consider a short circuit. Ideally, a short circuit has $Z = 0 \Omega$ and $\Gamma = -1$ for all frequencies of interest. At low frequencies, a good-quality short circuit can be obtained. This might be as simple as a BNC connector with a wire shorted across it. However, as the frequency increases the inductance of the short will become significant, so a higher-quality connector with a low-inductance short is required. At microwave frequencies, some finite inductance is unavoidable. Therefore, the impedance of the short is characterized and controlled so that it can be included in the vector error correction calculations.

A common model for the inductance of a short circuit standard is

$$L(f) = L_0 + L_1(f) + L_2(f^2) + L_3(f^3) \quad (15-31)$$

Note that the inductance is modeled as a polynomial function of frequency.

An ideal open circuit has $Z = \infty$ and $\Gamma = 1$ for all frequencies of interest. Below a few hundred megahertz, we can just leave the test connector open to create a useful open circuit for error correction purposes. At microwave frequencies, an open connector has significant stray capacitance, so we must use a special termination specifically designed to be an open.¹⁰ This termination is designed to control the stray capacitance between the inner conductor and the outer shield, such that the error correction algorithm can account for it.

A common model for the capacitance of an open circuit standard is

$$C(f) = C_0 + C_1(f) + C_2(f^2) + C_3(f^3) \quad (15-32)$$

An ideal Z_0 load has an impedance of Z_0 and $\Gamma = 0$ for all frequencies of interest. A good quality termination may have a return loss of 30–50 dB, depending on frequency range. Since this is the reference load for the measurement, the corrected measurement cannot reliably exceed this return loss. The device under test is simply being compared with the

¹⁰ This only seems like a contradiction in terms.

reference load used during the error correction calibration. So an alternative view is that a corrected measurement indicating that the device under test is a perfect Z_0 load really means that the device under test has the same impedance as the reference load.

SOLT Calibration

The most common calibration technique is known as short-open-load-through (SOLT) calibration. This calibration procedure requires the vector network analyzer (VNA) user to measure the error mechanisms of the VNA system by installing short, open, load, and through reference standards. In modern VNAs, this is done by invoking a calibration procedure in the instrument that instructs the user to install the proper standard while automatically accumulating the required measurements. This makes the calibration procedure easy to do, but it does take some time to change the standards and perform the measurements.

TRL Calibration

Another common calibration method is the through-reflect-line (TRL) method. This method is often used in systems where standard connectors are not suitable for making the measurement connections. For example, on-wafer probing of integrated circuits and some printed circuit board applications may not allow for conventional Z_0 connectors.

The through connection is achieved by making the shortest possible connection between the two ports of the DUT. The reflect standard needs to provide a significant reflection but does not need to be well characterized. It does need to provide the same reflection to both test ports. The line standard is a transmission line that is significantly longer in electrical length than the through connection. On-wafer calibration standards often consist of precision thin-film resistors, short-circuit connections, and Z_0 transmission lines fabricated on the DUT wafer or a separate calibration substrate.

Electronic Calibration Standards

To improve the consistency, convenience, and speed of performing a VNA calibration, the instrument manufacturers have created electronic standards that can be switched by the network analyzer. The user just has to connect the standard to the VNA and initiate the cal—the instrument will take care of switching in different terminations. These calibration standards are automatically switched using solid state devices, resulting in an easy, fast, accurate calibration.

These electronic calibration standards do not have shorts, opens, and Z_0 loads inside but use a different set of well-characterized impedances. The characteristics of these standards are measured at the factory and stored in nonvolatile memory in the standard.

Bibliography

Adam, Stephen F. *Microwave Theory and Applications*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1969.

Agilent Technologies. "Agilent RF Bridges," Publication Number 5990-5352EN, June 2012.

Agilent Technologies. "Network Analyzer Basics," Publication Number 5965-7917E, August 2004.

Agilent Technologies. “Agilent Electronic vs. Mechanical Calibration Kits: Calibration Methods and Accuracy,” Publication Number 5988-9477EN, June 2003.

Agilent Technologies. “Applying Error Correction to Network Analyzer Measurements,” Application Note 1287-3, Publication Number 5965-7709E, March 2002.

Agilent Technologies. “Understanding the Fundamental Principles of Vector Network Analysis,” Publication Number 5965-7707E, September 2012.

Agilent Technologies. “Agilent RF and Microwave Test Accessories Catalog,” Publication Number 5990-8661EN, September 2011.

Cascade Microtech. “On-Wafer Vector Network Analyzer Calibration and Measurements,” Cascade Microtech Application Note, n.d.

Dunsmore, Joel P. *Handbook of Microwave Component Measurements*. New York: Wiley, 2012.

Ely, Paul C., Jr. “Swept Frequency Techniques,” *Proceedings of the IEEE*, June 1967.

Gonzalez, Guillermo. *Microwave Transistor Amplifiers*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, Inc., 1996.

Haefner, Sylvester J. “Amplifier-Gain Formulas and Measurements,” *Proceedings of the IRE*, July 1946.

Hayt, William H., Jr. *Engineering Electromagnetics*, 6th ed. New York: McGraw-Hill Book Company, 2000.

Hewlett-Packard Company. “High Frequency Swept Measurements,” Application Note 183, Publication Number 5952-9200, Palo Alto, CA, December 1978.

Hewlett-Packard Company. “Vector Measurements of High Frequency Networks,” Application Note, Publication Number 5954-8355, Palo Alto, CA, March 1987.

Laverghetta, Thomas S. *Practical Microwaves*. Indianapolis, IN: Howard W. Sams & Company, Inc., 1984.

Oliver, Bernard M., and Cage, John M. *Electronic Measurements and Instrumentation*. New York: McGraw-Hill Book Company, 1971.

Spaulding, William M. “A Broadband Two-Port S-Parameter Test Set.” *Hewlett-Packard Journal*, November 1984.

Ziemer, R. E., and Tranter, W. H. *Principles of Communications*, 6th ed. Boston: Houghton Mifflin Company, 2008.

EMC Measurements

In most countries, electronic products are required to meet established standards for electromagnetic compatibility (EMC). These regulations attempt to allow the wide array of electronic devices to live together in electromagnetic peace and harmony. In this chapter, we'll take a look at using spectrum analyzers to measure two important aspects of EMC: *radiated emissions* (signals radiated from the device under test [DUT]) and *conducted emissions* (signals conducted via the power cable from the device under test).

16.1 Electromagnetic Compatibility

EMC is the ability of an electronic device or system to function properly in the presence of an electromagnetic environment while not polluting the electromagnetic environment. *Electromagnetic interference* (EMI) refers to the situation when interference does occur between electronic systems. In an ideal world, electronic devices would be immune to all ambient electromagnetic (EM) fields and would not create any harmful EM emissions. In reality, electronic devices are sensitive to external fields and do emit electronic noise.

There are two main categories of interference: (1) radiated, which occurs via electromagnetic radiation between devices; and (2) conducted, which is transmitted via the power line from one device to another.

EMC should be an integral part of the design process from the very start. The designer of an electronic system should consider likely emitters and take steps to control their behavior. In particular, high-speed digital signals with fast rise times deserve careful attention. As with most design issues, it is more effective and less expensive to detect EMC issues early in the process and not wait until the final product units are available for testing. Consistent with this, some early EMC measurements should be made to verify the level of emissions from the product. Full compliance measurements usually require the use of a certified test lab. While some organizations have fully equipped EMC test labs, most companies must have their products tested in a third-party test lab. However, early in the design cycle it may be more effective to use less rigorous bench testing known as *precompliance testing*, which can be performed on the test bench using a spectrum analyzer and some specialized accessories. For radiated measurements, the use of a shielded anechoic chamber or screen room minimizes the effects of ambient electromagnetic signals. A comparison of EMC measurement techniques for precompliance and full compliance testing is shown in Table 16-1.

Table 16-1 Comparison of EMC Measurement Techniques

Test Type	Purpose	Equipment
Precompliance Testing	Early testing of critical frequencies	Spectrum analyzer, test antennas, magnetic and electric near-field probes.
	Troubleshooting known EMC issues	Limited conformity to EMC regulations
Full-Compliance Testing	Final certification to EMC regulatory requirements	Certified or approved in-house test site with full conformity to EMC regulations

Precompliance testing is also very useful for troubleshooting EMC problems found during full compliance testing. Attempting to resolve a problem at a certified test site is both expensive and inconvenient. Often, the preferred strategy is to identify which frequencies exceed the emission standards at the certified site and take the device under test back to the lab for experimentation. There will be more on this later.

Radiated and conducted emission limits and measurement techniques are defined by standards bodies. For commercial products, CISPR 11 and CISPR 22 are the most common international standards for emission limits while CISPR 16 defines the required test methods and apparatus. CISPR is the abbreviation for the International Special Committee on Radio Interference, under the International Electrotechnical Commission (IEC). In the United States, the Federal Communications Commission (FCC) sets the regulatory limits in Part 15 of its regulations. For U.S. military applications, MIL-STD-461F is the most common standard.

16.2 Radiated Emissions

Radiated emissions are specified in terms of field strength, with units of $\text{dB}\mu\text{V}/\text{m}$. Both CISPR and FCC emission limits recognize two classes of equipment. Class A limits apply to equipment intended for commercial or industrial use. The lower Class B limits apply to equipment used in domestic surroundings (i.e., homes).

As shown in Figure 16-1, the frequency range for the FCC and CISPR measurements are from 30 MHz to 1 GHz. The FCC requirements also include measuring up to the fifth harmonic of the highest frequency oscillator in the system (but no higher than 6 GHz).

Figure 16-2 shows the typical test configuration for radiated emission measurements. The equipment being tested is placed on a nonconductive table. The radiated emissions are picked up by a calibrated EMC antenna located a prescribed distance away (usually 3 m or 10 m). An EMI receiver, often a spectrum analyzer, measures the amplitude of all emissions across the frequency range of interest. Since this is a book about spectrum and network measurements, we'll use the term EMI receiver to refer to a spectrum analyzer that meets the special requirements of EMC testing (see Section 16.5).

An example radiated emissions measurement using an EMI receiver is shown in Figure 16-3. This radiated emissions test is shown failing due to a number of spectral lines exceeding the defined limits. Note that the vertical scale is electric field strength, in units of $\text{dB}\mu\text{V}/\text{m}$.

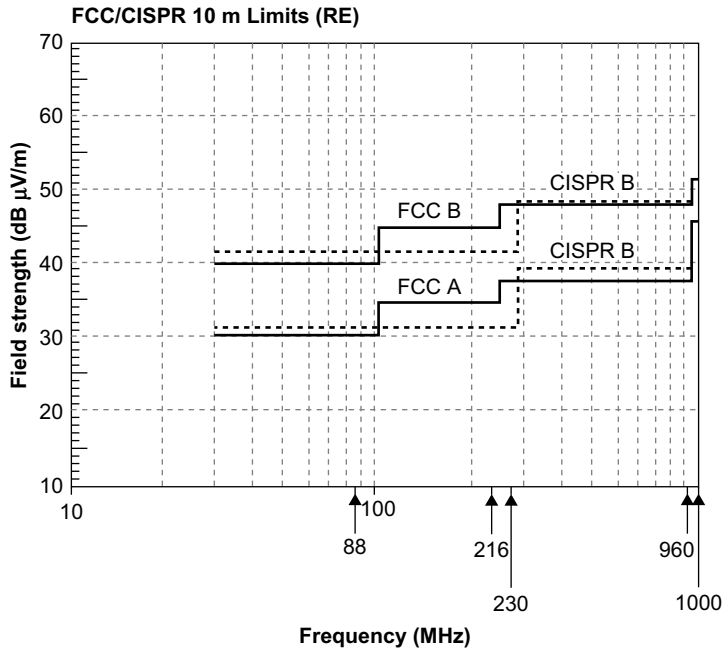


Figure 16-1 The field strength limits for radiated emissions at a distance of 10 m, as defined by the FCC and CISPR.

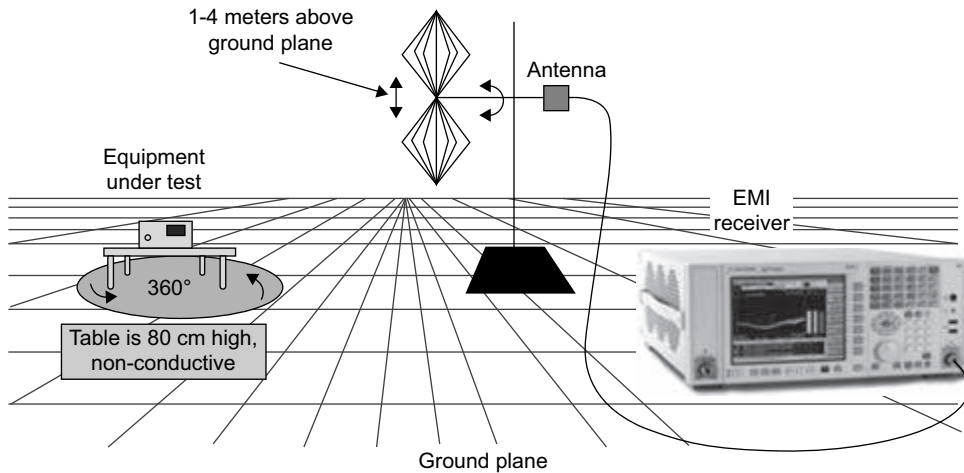


Figure 16-2 The test setup for measuring radiated emissions uses a calibrated antenna connected to a spectrum analyzer or EMI receiver. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

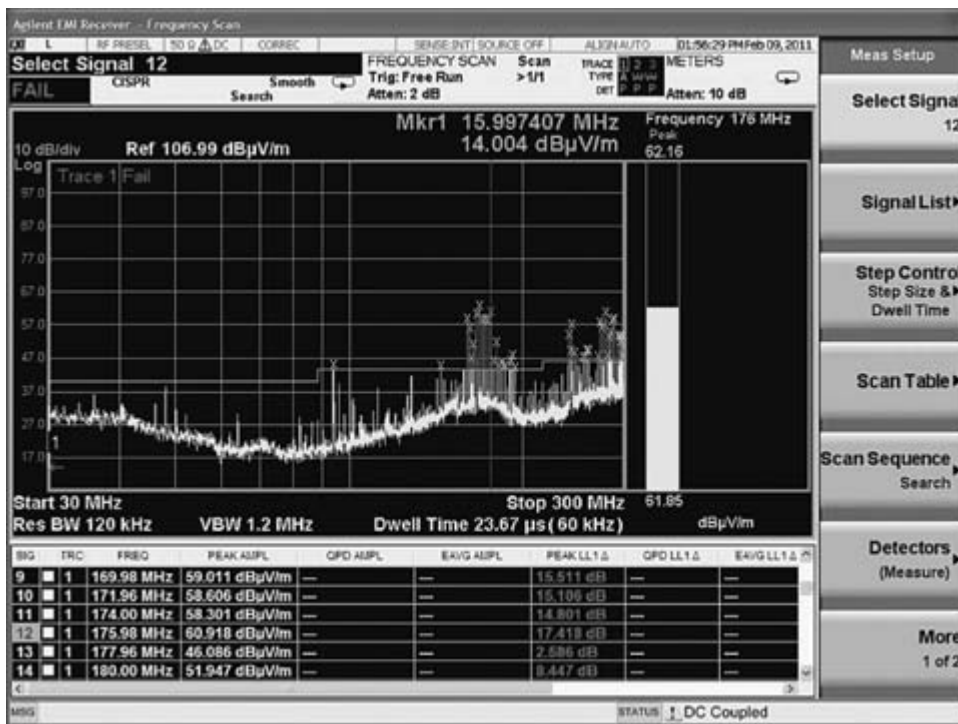


Figure 16-3 This radiated emissions measurement includes specific EMC features, such as frequency list and limit lines for relevant standards. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

16.3 Antennas

To measure radiated emissions, an antenna is used to capture the electromagnetic field and feed it to the input of the spectrum analyzer. A wide variety of EMC antennas are available with various trade-offs made between size, gain, and frequency coverage. Since EMC is inherently a broadband spectral measurement, having an antenna that covers the entire frequency of interest is desirable. One popular antenna design uses a hybrid approach that combines a biconical antenna with a log-periodic antenna to cover a wide range of frequencies (Figure 16-4).

The antenna is a transducer that converts the electric field strength to a known voltage at the antenna terminals. The conversion factor is known as *antenna factor* and is given by

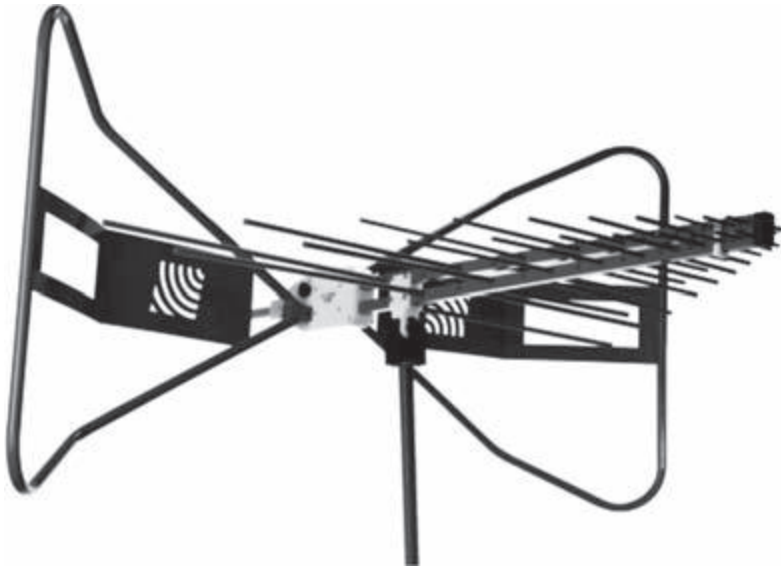
$$AF = \frac{E}{V_L} \tag{16-1}$$

where

AF = antenna factor, m^{-1}

E = electric field, V/m

V_L = voltage at the antenna terminals



ETS-Lindgren's Model 3143B BiConiLog Antenna

Figure 16-4 The BiConiLog is a hybrid antenna that combines a log periodic antenna and a biconical antenna to cover 30 MHz to 1 GHz. (Courtesy of ETS-Lindgren.)

Most EMC antennas will have a nominal impedance of 50Ω , and the antenna factor is specified assuming the antenna is connected to a 50Ω load (normally, the input of the EMI receiver). Antenna factor has the units of $1/\text{m}$, often shown as m^{-1} . When expressed in decibels, the AF has the units of dBm^{-1} or dB/m not to be confused with dBm . The electric field strength is usually shown in units of $\text{dB}\mu\text{V}/\text{m}$.

Modern spectrum analyzers can automatically apply the antenna factor correction and show the resulting measurement in terms of electric field strength.

$$E = AF \cdot V_L \quad (16-2)$$

Antenna factor is often expressed in decibel form, and the electric field strength is usually shown in units of $\text{dB}\mu\text{V}/\text{m}$. Other devices may be inserted in line that should be included in the measurement. If a preamp is used to boost the signal, the gain of the preamp should be included in the measured result. Sometimes a small attenuator is inserted to improve the 50Ω match of the antenna. Finally, the cable loss may be significant.

The E-field measurement, in dB, is given by

$$E_{(\text{dB}\mu\text{V}/\text{m})} = V_{SA(\text{dB}\mu\text{V})} - G_{pa(\text{dB})} + L_{c(\text{dB})} + L_{a(\text{dB})} + AF_{(\text{dB}/\text{m})} \quad (16-3)$$

where

$V_{SA(\text{dB}\mu\text{V})}$ = the spectrum analyzer reading in $\text{dB}\mu\text{V}$

$G_{pa(\text{dB})}$ = the gain of the preamp in dB

$L_{c(\text{dB})}$ = the cable loss in dB

$L_{a(\text{dB})}$ = the attenuator loss in dB

$AF_{(\text{dB}/\text{m})}$ = the antenna factor in dB

Example 16.1

An EMC antenna with antenna factor shown in Figure 16-5 is used to measure an E field at a frequency of 500 MHz. A 20 dB preamplifier is inserted inline and the cable loss at this frequency is 3 dB. If the spectrum analyzer reading is 47 dBμV, what is the field strength?

$$E_{(\text{dB}\mu\text{V}/\text{m})} = V_{SA(\text{dB}\mu\text{V})} - G_{pa(\text{dB})} + L_c(\text{dB}) + L_a(\text{dB}) + AF_{(\text{dB}/\text{m})}$$

From the figure, the antenna factor (AF) at 500 MHz is 19 dB/m

$$E_{(\text{dB}\mu\text{V}/\text{m})} = 47 - 20 + 3 + 0 + 19 = 49 \text{ dB}\mu\text{V}/\text{m}$$

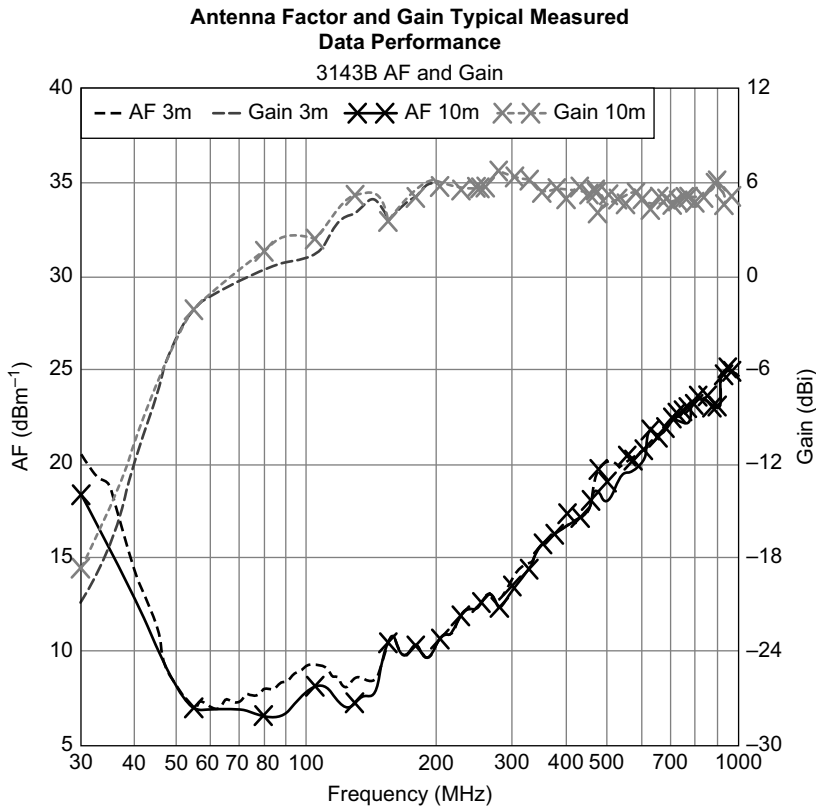


Figure 16-5 For EMC measurements, the AF is supplied by the manufacturer. (Courtesy of ETS-Lindgren.)

16.4 Near Field and Far Field

As the name implies, electromagnetic fields are a combination of an electric field (E) and a magnetic field (H). When a device produces an EM field that propagates some distance through free space, the wave is considered to be in the *far field*. In the far field, EM waves are relatively well behaved and propagate according to classic radio theory.

The impedance of free space is given by the ratio of the E field and the H field

$$Z_0 = \frac{E}{H} = \mu_0 c \approx 377 \Omega \quad (16-4)$$

where

μ_0 = the magnetic constant ($4\pi \cdot 10^{-7} \text{V} \cdot \text{s}/(\text{A} \cdot \text{m})$)

c = the speed of light in free space ($3 \cdot 10^8 \text{m/s}$)

When the EM field is close to the radiating source, it is said to be in the *near field*. In this region, the relative strengths of the E and H fields depend heavily on the nature of the radiating source and its immediate surroundings. If the source has high current and low voltage ($E/H < 377 \Omega$), the field is predominately magnetic. If the source has high voltage and low current ($E/H > 377 \Omega$), the field is predominately electric.

The free-space wavelength is determined by

$$\lambda = \frac{c}{f} \quad (16-5)$$

where

f = the frequency of the waveform

The dividing line between near and far field is somewhat arbitrary, with a transition region. A common rule of thumb is to define the far field as starting at a distance of $\lambda/2\pi$ from the radiation source. Radiated emission measurements are done in the far field, with the antenna some distance away from the DUT. Typically, the lowest frequency of interest is 30 MHz, which has a wavelength of 10 m. Thus, the far field starts at $10/2\pi$, or 1.6 m, away from the source.

Example 16.2

For a frequency of 250 MHz, calculate the distance from a radiation source where the far field begins.

The wavelength is

$$\lambda = \frac{c}{f} = (3 \cdot 10^8) / (250 \cdot 10^6) = 1.2 \text{ m}$$

The far field starts at this distance from the radiating source

$$\lambda/2\pi = 1.2/2\pi = 0.19 \text{ m or } 19 \text{ cm}$$

16.5 EMI Receiver Requirements

The classic definition of an EMI receiver is a fixed-tuned receiver with specific filters and detectors for EMI measurements. These receivers are designed with a robust front end that is highly immune to overload from strong out-of-band signals. A spectrum analyzer is a useful tool for measuring radio frequency (RF) signals, so it follows that it can be applied to EMC measurements. However, a spectrum analyzer may not have the required filters and detectors for EMI measurements and typically has a broadband front end that is more susceptible to overload. Some spectrum analyzers have been optimized for EMC use and meet the CISPR 16-1-1 requirements that are considered the standard definition for a true EMI receiver. (Even if a spectrum analyzer does not meet the EMI receiver requirements, it can still be a useful tool for identifying and correcting unwanted emissions.) As the product categories have matured, the distinction between EMI receivers and spectrum analyzers have blurred with each type of instrument adopting the others attributes. In this chapter, we'll use the term EMI receiver to mean spectrum analyzers that meet the CISPR requirements.

The CISPR requirements for an EMI receiver include the following:

- Specific resolution bandwidth filters
- Four types of detectors: peak, quasi-peak, EMI average, and root mean square (RMS) average
- Amplitude accuracy of ± 2 dB (9 kHz to 1 GHz) and ± 2.5 dB (1–18 GHz)
- Ability to pass radiated immunity in a 3 V/m field
- Ability to pass the CISPR pulse test (implies having a preselector below 1 GHz)

A few additional features tailor the instrument for EMC measurements: predefined frequency ranges, EMC limit lines, signal lists, and antenna factor correction.

Resolution Bandwidth

As discussed in Chapter 4, the resolution bandwidth (RBW) of a spectrum analyzer determines how finely the individual spectral components can be resolved in frequency. A narrower RBW allows the analyzer to measure each spectral component separately, while a wider RBW may allow multiple spectral lines to be measured by the detector. Clearly, the choice of RBW can change the measured amplitude for dense spectral lines. For EMC measurements, the CISPR 16 standard removes this variable by specifying the required RBW (see Table 16-2).

Table 16-2 Resolution bandwidths for EMI measurements specified by CISPR 16-1-1

Frequency Range	CISPR Band	Resolution Bandwidth (6 dB)
9–150 kHz	A	200 Hz
150 kHz to 30 MHz	B	9 kHz
30 MHz to 1 GHz	C/D	120 kHz
>1 GHz	E	1 MHz

Detectors

The CISPR specification also includes specific requirements for the characteristics of the receiver's detector. For continuous wave (CW) signals, the detector does not make a difference since the signal is not time varying. Often, EMI emissions are modulated, pulsed, intermittent, or otherwise time varying, so the detector characteristics come into play.

Consider the case of a pulsed RF signal as discussed in Chapter 9. For RF pulses with low repetition rates, the energy in the signal is relatively low. As the repetition rate is increased, the energy in the signal increases, eventually approaching the CW case. From an EMI perspective a low repetition rate signal produces much less interference than a high rep rate signal. The quasi-peak detector was defined to provide a consistent way of measuring pulsed signals.¹

16.6 Peak, Quasi-Peak, and Average Detectors

Spectrum analyzers intended to be used for EMC measurements will have several different detectors available:

Peak detector: Responds to the peak level of the signal present in at the detector for a given frequency bin.

Quasi-peak detector: Responds to the peak level and repetition rate of the signal present at the detector

EMI average detector: Responds to the average level of the signal present at the detector.

RMS average detector: Responds to the average of the RMS level of the signal present at the detector.

The peak detector is mostly useful for EMC troubleshooting since it quickly acquires signals (but may overstate their amplitude from an EMC point of view). The quasi-peak detector is designed specifically for EMC measurements. It responds to the peak level of the signal but then decays away with a prescribed time constant. Signals with a low repetition rate will have a lower reading, whereas high repetition rate signals will produce a larger measured response. This mimics the impact of the signal in an EMC environment. The quasi-peak detector slows down the measurement significantly, so a common approach is to scan with the peak detector until a more precise measurement needs to be made. The peak detector will always read greater than or equal to the quasi-peak measurement, so it is a conservative way to measure. Figure 16-6 shows how the peak, quasi-peak and average detectors respond to emissions with high and low repetition rates.

The quasi-peak, EMI average, and RMS average detectors have precise and somewhat complicated definitions in the CISPR standard. The RMS average detector was developed to respond to interference in a manner consistent with modern digital radio formats and was added to the standard in 2007. The RMS average detector is implemented as an RMS detector followed by a linear average detector and a peak reading meter algorithm.

¹ Sometimes the quasi-peak detector is said to detect the *annoyance factor* of an EM emission.

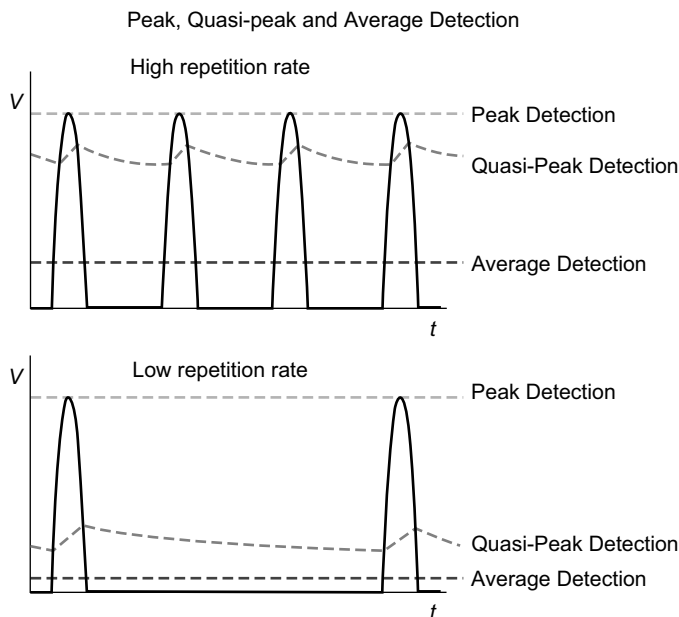


Figure 16-6 A comparison of detector behavior for high repetition rate and low repetition rate waveforms. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

In general, each type of detector may produce a different measurement reading than the other detectors. However, there is a consistent ordering to the amplitude of their readings. The peak detector always reads greater than or equal to the quasi-peak detector, which reads greater than or equal to the RMS average detector, which reads greater than or equal to the EMI average detector.

16.7 Conducted Emissions

Conducted emission regulations limit the spectral content that is coupled through the AC power cord of the DUT. For both CISPR and FCC regulations, the frequency range of interest is 150 kHz to 30 MHz as shown in Figure 16-7.

16.8 Line Impedance Stabilization Network

The typical test configuration for measuring conducted emissions uses a line impedance stabilization network (LISN) as shown in Figure 16-8. The AC power passes through the LISN to the DUT while also providing a match between the line impedance and the $50\ \Omega$ input of the EMI receiver. Often, a limiter is inserted in the measurement line to protect the receiver from large transients that can occur during power on. Keep in mind that the incoming AC power is delivering 100 to 240 V RMS, depending on the type of line voltage being tested, while the EMI receiver is set to measure microvolt levels.

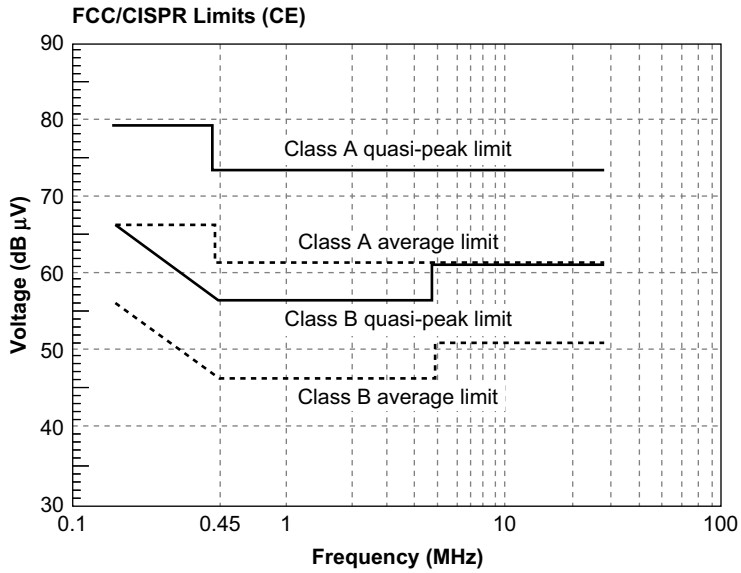


Figure 16-7 The CSPR and FCC test limits for conducted emissions.

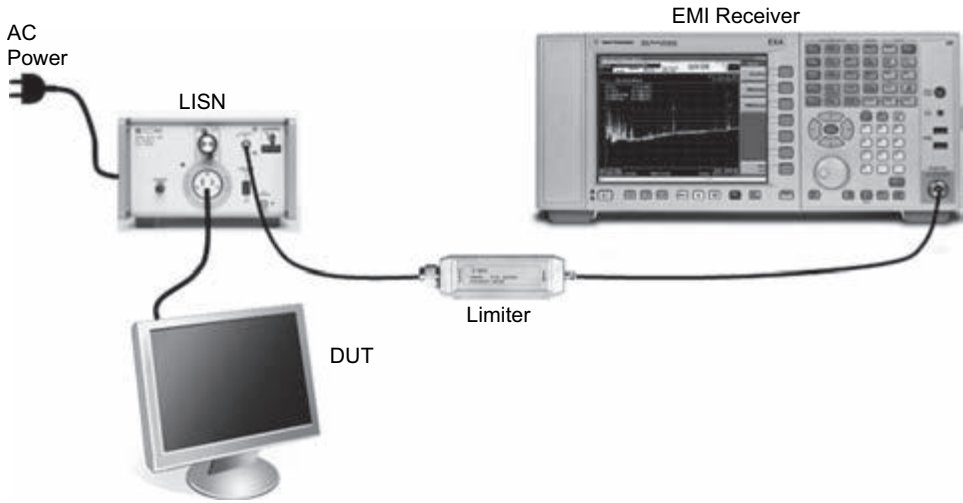


Figure 16-8 A typical test configuration for conducted emissions using an LISN, limiter, and EMI receiver.

A typical LISN is shown in Figure 16-9. The LISN provides an AC power outlet for the DUT to plug into along with a BNC output port that is connected to the EMI receiver. The schematic diagram in Figure 16-10 shows how the AC power is supplied to the DUT while providing an output port to measure the conducted emissions. The spectrum analyzer display of a conducted emissions measurement is shown in Figure 16-11. The emissions from the LISN are plotted as dBμV and compared to the regulatory test limits.



Figure 16-9 Photograph of a typical line impedance stabilization network (LISN). (Courtesy of ETS-Lindgren.)

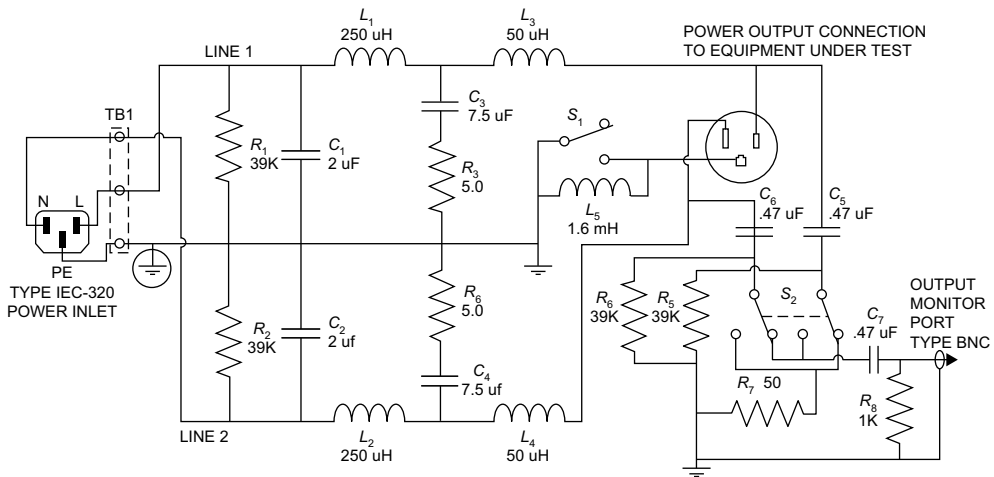


Figure 16-10 The circuit of typical LISN shows the AC power passing through left to right while also providing the measurement connection for the EMI receiver. (Courtesy of ETS-Lindgren.)

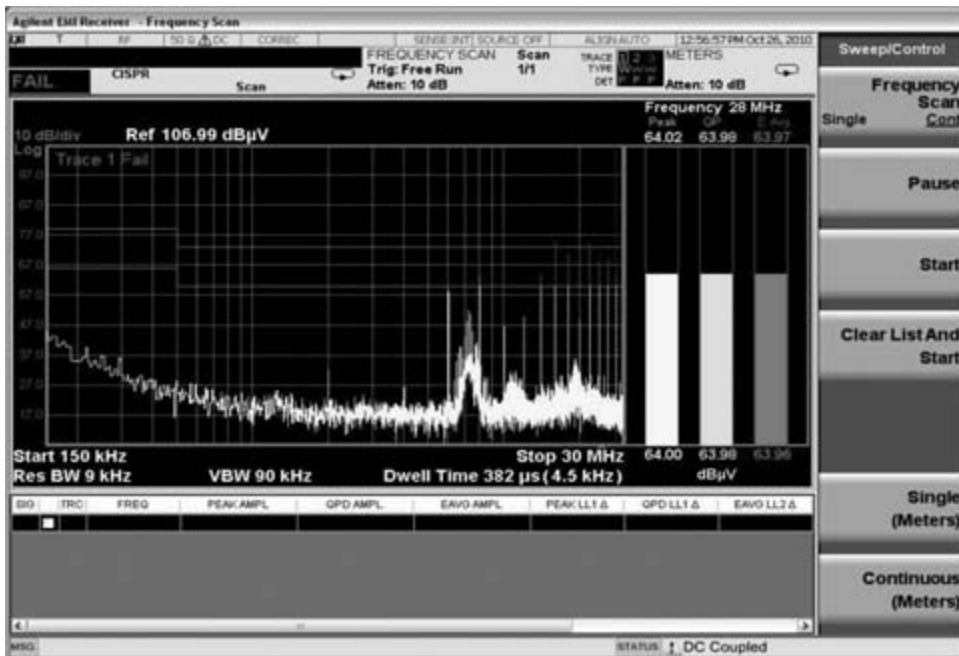


Figure 16-11 This conducted emissions measurement includes specific EMC features, such as frequency list and limit lines for relevant standards. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

16.9 EMC Troubleshooting

While the final compliance tests for both radiated and conducted emissions must be done on a certified test range, it is also important to be able to do less formal testing on the bench. This may be precompliance testing, to improve the odds of passing the full compliance test, or it may be for troubleshooting a particular test failure.²

Noncompliance measurements of radiated emissions can be done in a normal lab environment. A common configuration is to set the DUT on a table or workbench with a small antenna positioned about 1 m away. We are trying to emulate the *far-field* measurement at the official test. It is best to find a location that is electrically quiet, with very few ambient RF signals present. A basement location may offer some advantages in terms of avoiding radiation from broadcast stations and wireless base stations.

A calibrated EMC antenna can be used for troubleshooting, although it will generally be difficult to maintain highly accurate measurements on the bench. A lower cost and simpler antenna such as a basic TV antenna may be sufficient for troubleshooting purposes (Figure 16-12 and 16-13). A good strategy is to carefully measure the radiated emissions,

² For a more detailed discussion of EMC troubleshooting, see André and Wyatt (2014).

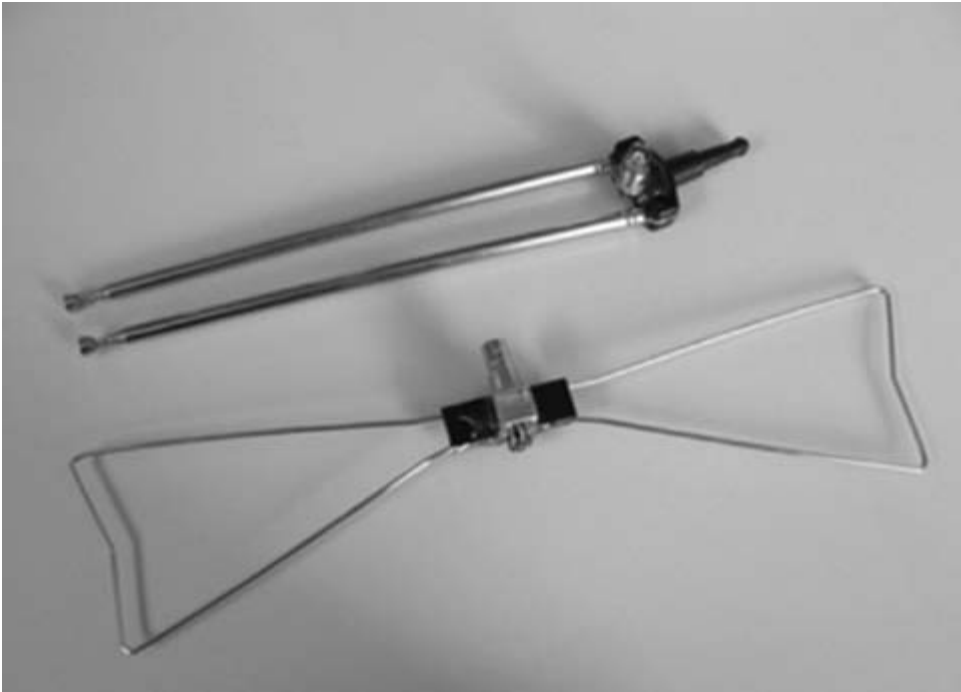


Figure 16-12 Simple television antennas can be used for troubleshooting radiated emissions. (Courtesy Kenneth Wyatt.)

focusing on the specific frequencies that are failing or are likely to fail. Then you can make design changes (e.g., installing filters, shielding) and measure the *relative* improvement in radiated emissions. Of course, for repeatability your measurement setup must be the same every time.

16.10 Near-Field Probes

To investigate the source of the emissions, whether radiated or conducted, we are going to have to poke around inside the DUT. Two types of near-field probes can be used: *magnetic field* and *electric field*. Recall that EM fields in the near field tend to be dominated by either the electric field or magnetic field, depending on the source.

A magnetic probe has a loop on the end of the probe for efficient coupling of magnetic fields. The larger the loop, the better the sensitivity. Smaller loops are better at locating the EMC hotspots since their spatial resolution is better. An electric field probe has a very small dipole at the tip of the probe, which is effective at picking up electric fields. Some typical magnetic and electric field probes are shown in Figure 16-14.

Figure 16-15 shows a magnetic-field probe being passed over a suspected circuit while monitoring its output on a spectrum analyzer. This is a powerful troubleshooting tool for finding the hot spots in a circuit assembly that are producing the harmonic currents at the particular frequencies that are failing the compliance test.



Figure 16-13 This 400–1000 MHz log periodic antenna is usable for troubleshooting purposes down to 100 MHz. The antenna is fashioned from PC board material (available from <http://wa5vjb.com>) and mounted on a simple tripod. (Courtesy Kenneth Wyatt.)

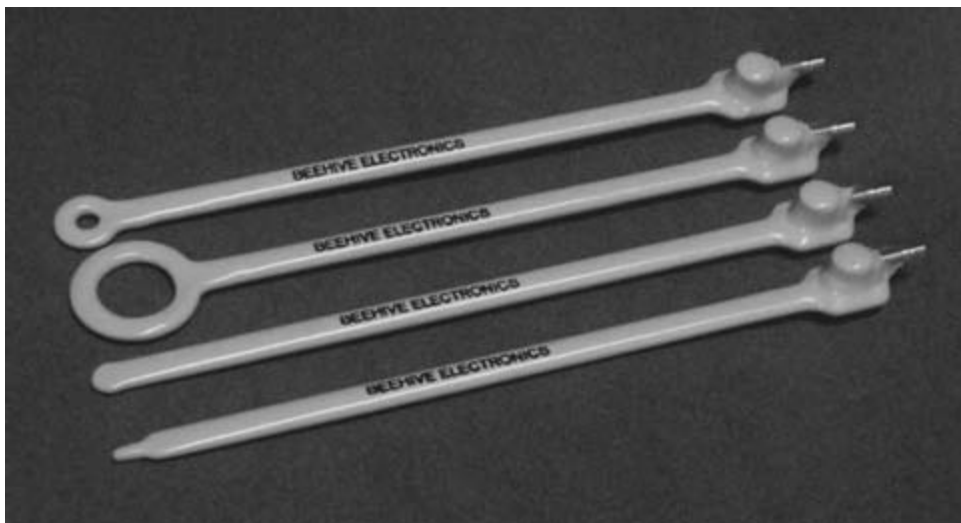


Figure 16-14 Three magnetic-field probes (with loops) and one electric-field probe (bottom). (Courtesy Beehive Electronics.)

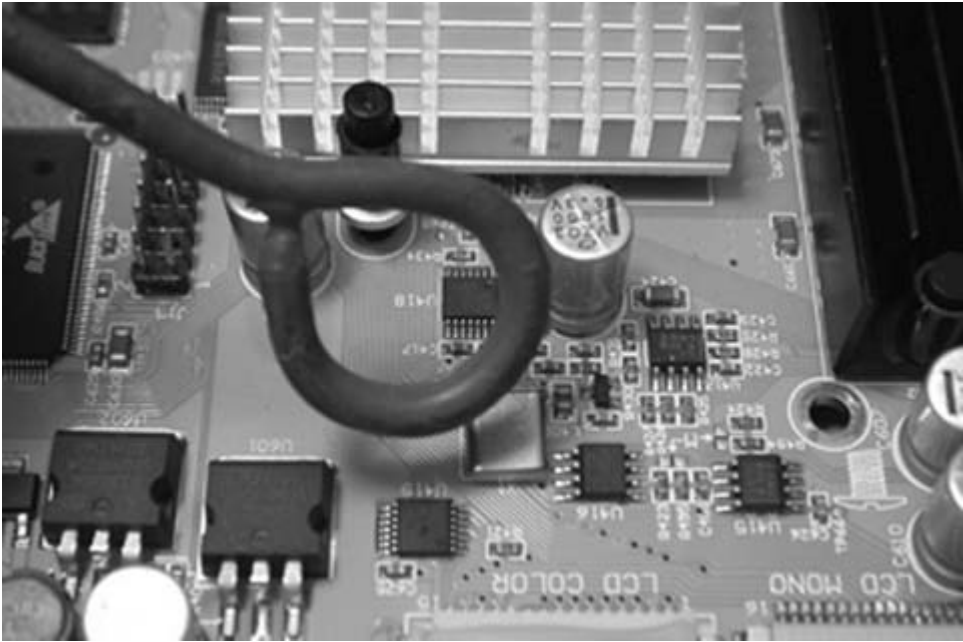


Figure 16-15 A magnetic probe is passed over a suspect circuit board, looking for evidence of radiated emissions. (Courtesy Kenneth Wyatt.)

The E-field and H-field probes are used for identifying sources of emissions, but it is important to use an antenna in the far field to see if it's really radiating. Not all hot spots are efficient radiating structures, so even though the near-field probes pick up some strong fields they may not be causing the radiated emissions.

16.11 Current Probe

A common source of EMI problems is related to currents flowing along unshielded cables or on the outside of shielded cables in a system. It is sometimes difficult to tell which cable is radiating and at what frequencies. The use of current probes can give us insight into the frequency content running through the cables. Figure 16-16 shows several current probes that clamp into a cable and connect to a spectrum analyzer.

This type of current probe is also called a *current transformer*, because that is the coupling mechanism used. This means the probe will pass only AC signals and the response will roll off as the frequency approaches DC.

The scale factor associated with the current probe is its *transfer impedance*.

$$Z_T = \frac{V_L}{I_P} \quad (16-6)$$

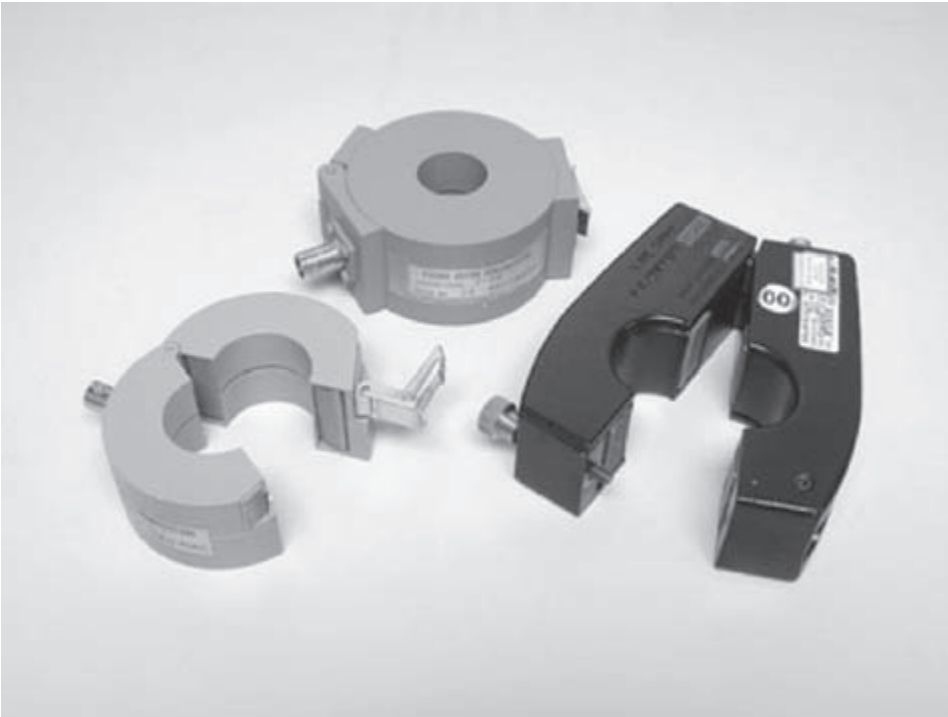


Figure 16-16 Examples of current probes that can be used to measure the frequencies flowing along unshielded cables or on the outside of a shielded cable. (Courtesy Kenneth Wyatt.)

where

Z_T = the transfer impedance(Ω)

V_L = the voltage produced by the probe(V)

I_P = the current flowing through the probe

The probes are configured with a suitable RF connector such as a BNC connector and are meant to drive a 50 Ω load.

16.12 Preamplifiers

Signals associated with EMC work are often very weak and may be near the noise floor of the EMI receiver. This is especially true in situations where the antenna factor or probe transfer impedance is low, further attenuating the signal. A low-noise preamplifier can be used to boost the signal ahead of the receiver input. Sometimes EMC antennas have a low-noise preamp installed at the factory.

A key point is that the preamp must have very low noise to improve overall measurement sensitivity. Commercially available preamps produce gains in the range of 20 to 40 dB with frequencies starting as low as 100 Hz and extending to tens of GHz. With this additional gain in the system, it is important not to overload the amplifier with large signals, producing distortion products.

Bibliography

Agilent Technologies. “EMC Measurement Application Measurement Guide—X-Series Signal Analyzer,” Publication Number N6141-90002, April 2011.

Agilent Technologies. “Essential Capabilities of EMI Receivers,” Application Note, Publication Number 5991-1756EN, October 2013.

Agilent Technologies. “Making EMI Compliance Measurements,” Publication Number 5990-7420EN, 2011.

André, Patrick G. and Kenneth Wyatt, *EMI Troubleshooting Cookbook for Product Designers: Concepts, Techniques, and Solutions*. SciTech Publishing, 2014.

Osburn, John D. M. “EMC Antenna Parameters and Their Relationships,” <http://www.ets-lindgren.com/pdf/antparameters.pdf>.

Ott, Henry W. *Electromagnetic Compatibility Engineering*. New York: John Wiley & Sons, Inc., 2009.

Wyatt, Kenneth. “The HF Current Probe: Theory and Application,” *Interference Technology, EMC Directory and Design Guide*, 2012.

Wyatt, Kenneth. <http://www.emc-seminars.com/>, n.d.

Analyzer Performance and Specifications

Spectrum and network analyzer specifications are the instrument manufacturer's way of communicating to the user the level of performance that should be expected from a particular instrument. Understanding and interpreting instrument specifications enable the instrument user to predict how the instrument will perform in a specific measurement situation, including the accuracy of the measurement.

The form and style of the specifications are usually related somewhat to the block diagram and measurement techniques internal to the instrument. These specifications may appear to be more complex than necessary. However, oversimplifying an instrument data sheet can force the manufacturer to understate the performance level of an instrument to cover all possible cases in a single specification. In general, the details present in analyzer data sheets provide a better understanding of instrument performance, so that the user can obtain the best measurement possible.

17.1 Source Specifications

Normal operation of a network analyzer signal source (or spectrum analyzer tracking generator) is to create a pure sine wave with the desired amplitude and frequency. There will always be some error in amplitude and frequency, which are described in the data sheet. Amplitude errors can be described as *level accuracy* across some frequency range or as *absolute accuracy* at one frequency combined with a *flatness specification*. Most modern analyzers use synthesized frequency references derived from a precision reference oscillator, so the frequency accuracy is normally very good.

Ideally, the source output produces only one frequency—the desired, fundamental frequency. In practice, other frequencies will be present, both harmonic and nonharmonic frequencies. A typical specification for a network analyzer source harmonic content might be -25 or -30 dBc, which may seem like rather poor distortion performance. However, if the source and receiver are tracking in frequency and the device under test is reasonably linear, the harmonic content will simply fall outside the passband of the receiver and not affect the measurement. Nonharmonic spurious signals are not necessarily so well behaved and may show up at what appears to be arbitrary frequencies. The analyzer manufacturer must make

sure that the level of these spurious signals does not introduce significant error in the measurement. If the spurious signals stay out of the receiver passband, they will not effect the network measurement.

- *Frequency resolution*: The frequency resolution specification indicates the smallest change possible when setting the source frequency. For example, a source with a 0.1 Hz frequency resolution may be set to 1000.1 Hz, 1000.2 Hz, and 1000.3 Hz but not 1000.05 Hz. This specification does *not* define how accurate the source frequency is, but only how finely it may be set.
- *Frequency stability*: A source's frequency varies over time due to thermal, aging, and other effects. The frequency stability specification describes this long-term frequency drift, usually in terms of parts per million per day, often with a temperature range specified. (Very short-term frequency fluctuation is specified in terms of phase noise.) A typical frequency stability specification might be $\pm 5 \times 10^{-8}$ /day. With this spec, a 100 MHz source frequency could vary as much as $\pm(5 \times 10^{-8})(100 \text{ MHz}) = \pm 5 \text{ Hz}$ per day.
- *Level accuracy*: This specification indicates how much error there can be in the output or power level of the source. This may be specified at only one frequency and power level, with a linearity specification to describe the accuracy at other output levels and a flatness specification to describe how the level varies with frequency. A typical specification is $\pm 0.5 \text{ dB}$ at 50 MHz and 0 dBm output power.
- *Level linearity*: Level linearity describes how the level accuracy changes with changing output level. It is often specified in table form with an accuracy specified for a particular range of output power. For example,

ERROR	OUTPUT LEVEL
$\pm 0.2 \text{ dB}$	-5 dBm to $+15 \text{ dBm}$
$\pm 0.5 \text{ dB}$	$+15 \text{ dBm}$ to $+20 \text{ dBm}$

- *Flatness*: The flatness specification represents the frequency response of the source power level. The flatness spec alone does not indicate anything about the absolute accuracy of the source power but instead indicates how much it varies over frequency. For example, a flatness specification of $\pm 1 \text{ dB}$ means that for a given amplitude setting, the actual source power level may vary over a 2 dB range when swept in frequency, usually measured relative to a low-frequency amplitude value.
- *Impedance*: The nominal output impedance of the source is important in that a Z_0 system should be driven by a source having an output impedance of Z_0 . The quality of this Z_0 source impedance will usually be specified in terms of return loss or standing wave ratio (SWR). This is important in predicting measurement error due to imperfect source match. Typical specification: $>20 \text{ dB}$ return loss.
- *Phase noise*: Very-short-term variations in frequency are specified in terms of phase noise. The phase noise is specified in dBc (dB relative to the carrier or source frequency) at some frequency offset away from the source frequency and normalized to a 1 Hz bandwidth. A typical specification is stated as $<-90 \text{ dBc}$ (1 Hz BW) at a 10 kHz offset.
- *Harmonics*: The harmonic content present in the output signal is specified in terms of dBc (decibels relative to the carrier, in this case the fundamental frequency). Typical specification: $<-30 \text{ dBc}$.

- *Nonharmonic spurious signals*: The source may produce other spurious signals that are not harmonically related to the source frequency. These spurious signals are also usually specified in terms of dBc. Typical specification: <-50 dBc.

17.2 Receiver Characteristics

The receiver characteristics indicate how accurately signals can be measured. Ideally, the receiver only responds to the intended signal present at the input but noise, distortion and other signals may be included in the measurement.

- *Input impedance*: Important if the input is required to properly terminate a port of the device under test. A Z_0 (50 Ω) input will usually have a return loss or SWR specification associated with it so that mismatch errors can be estimated. Typical specification: >25 dB return loss. Analyzers operating below 50 MHz may also supply a high impedance input, whose nominal resistive and capacitive components are specified (e.g., 1 M Ω and 30 pF).
- *Displayed average noise level (DANL)*: Describes the basic *sensitivity* and *noise performance* of the analyzer, usually measured in a 1 Hz resolution bandwidth (RBW). This specification represents the reading of the analyzer due to noise with no signal present. Signals obscured by this noise level cannot be detected, while signals right at this level will be measurable but with some error (see Chapter 8). Typical specification: -140 dBm (1 Hz BW).
- *Second harmonic distortion*: Can be specified as *second harmonic intercept* (SHI) or *second order intercept* (SOI) using the distortion model described in Chapter 7. Alternatively, the harmonic level may be specified in dBc at some specific input level. Typical specification: SHI = $+45$ dBm.
- *Third-order intercept (TOI)*: Describes the distortion performance of the receiver using the distortion model described in Chapter 7. Typical specification: $+15$ dBm.
- *Spurious responses*: Erroneous signals that appear on the analyzer display that are not harmonically related to the input signal. Typical specification: >100 dB below maximum input level.
- *Residual responses*: One type of spurious response—signals that appear on the analyzer display due to imperfections internal to the analyzer with no input signal connected.
- *Input-related spurious responses*: Input-related spurious responses are signals that appear on the analyzer display due to imperfections internal to the analyzer when an input signal is connected. When the input is disconnected, these responses disappear. These responses are artifacts of the analyzer's internal block diagram, and they differ from distortion products in that they occur at frequencies not directly related to the input frequency.
- *DC response/LO feedthrough*: Most analyzers that operate near 0 Hz generate a response at DC. In a swept analyzer, this is due mainly to the local oscillator feedthrough. In a fast Fourier transform (FFT) analyzer, this response is due to DC offsets in the signal path. The level of this response at 0 Hz is usually specified in decibels relative to a full-scale response. This specification may be omitted on analyzers whose low-frequency limit is significantly above 0 Hz (e.g., 100 kHz). Typical specification: >33 dB below full-scale input level.

Multichannel network analyzer amplitude characteristics are usually specified for the single-channel case as well as the dual-channel (or ratio) case. The accuracy of the ratio

case is usually better since the channels are designed and built to be matched in magnitude and phase characteristics. Normal network analyzer measurements use the ratio of two channels, taking advantage of the matching between channels.

- *Amplitude absolute accuracy*: Usually specified for a full-scale signal and may be restricted to center of screen. A dynamic accuracy specification is added to the absolute accuracy spec to determine the accuracy at lower amplitudes. Alternatively, amplitude accuracy may be composed of several different specifications such as *intermediate frequency (IF) gain uncertainty*, *radio frequency (RF) gain uncertainty*, and *amplitude temperature drift*.
- *Amplitude dynamic accuracy*: Also known as *incremental accuracy* or *log scale fidelity*, this specification describes how accurate the analyzer is in a relative sense. That is, if a signal changes by 1 dB at the input, what change is shown on the analyzer display? A typical spec is stated as ± 0.05 dB/dB, meaning that for a 1 dB change in signal level an error of ± 0.05 dB may be introduced. Alternatively, it may be specified in table form, with an error limit for each measurement range. This specification is important because it represents the main error remaining after a normalization is performed.
- *Amplitude resolution*: The smallest change in amplitude that can be detected by the analyzer, often related to the marker or cursor readout. The resolution should be significantly smaller than the typical amplitude accuracy, so that the resolution does not limit accuracy.
- *Amplitude frequency response or flatness*: This specification describes the variation in amplitude response due to changing frequency. In cases where the absolute accuracy of the receiver is specified at only one point, the amplitude flatness must be added in to determine the error at other frequencies. The amplitude flatness is also important in cases where a network measurement is performed without the use of normalization. Phase specifications usually only apply when two receiver channels are used together in a relative or ratio phase measurement.
- *Phase accuracy*: The absolute phase accuracy of the receiver is usually specified for a full-scale signal and may be restricted to center of screen. A dynamic accuracy specification is added to the absolute accuracy spec to determine the phase accuracy at lower amplitudes.
- *Phase dynamic accuracy*: This specification describes how accurate the phase response of the analyzer is with changes in signal amplitude. It is important because it represents the main error remaining after a normalization is performed.
- *Phase resolution*: The smallest change in phase that can be detected by the analyzer, often related to the marker or cursor. The resolution should be significantly smaller than the typical phase accuracy, so that the resolution does not limit accuracy.
- *Phase frequency response*: This specification describes the variation in phase response due to changing frequency. In cases where the absolute phase accuracy of the receiver is specified at only one point, the phase frequency response must be added in to determine the error at other frequencies. The phase frequency response is also important in cases where a network measurement is performed without the use of normalization.
- *Delay specifications*: Since most analyzers calculate the delay measurement from the phase measurement, delay specifications are obtained by translating the phase specs into delay specs. For example, delay accuracy might be given as

$$\text{delay accuracy} = (\text{phase accuracy}) / (360 \times \text{delay aperture})$$

with phase accuracy in degrees and delay aperture(Hz)

17.3 Spectrum Analyzer Dynamic Range

Spectrum analyzer dynamic range is an important receiver specification, or set of receiver specifications, that deserves to be treated separately. Dynamic range describes the range of signal levels that can be reliably measured simultaneously. In particular, it describes the analyzer's ability to measure small signals in the presence of large signals. This ability is critical to the function of an analyzer since its main function is to measure the individual frequency components of a signal or the frequency response of a network.

Dynamic range is defined as the maximum ratio of two signal levels simultaneously present at the input that can be measured to a specified accuracy (IEEE, 1979). We can imagine connecting two signals to the analyzer input: one that is the maximum allowable level for the analyzer's input range, and the other much smaller one (Figure 17-1). The smaller one is reduced in amplitude until it is no longer detectable by the analyzer. When the smaller signal is just measurable, the ratio of the two signal levels (in dB) defines the dynamic range of the analyzer.

What effects might make it undetectable? Such things as residual responses of the analyzer, harmonic distortion of the large signal (due to analyzer imperfections), and the internal noise of the analyzer could all be large enough to cover up the smaller signal as we decrease its amplitude. The smaller signal might not appear at the same frequency as a spurious response or the harmonic of the larger signal, but when considering the general case we must assume that it could. Another way to say this is we cannot tell the difference between the smaller signal and an imperfection of the analyzer such as a distortion product or residual response. Thus, the dynamic range of the instrument determines the amplitude range over which we can reliably make measurements.

Figure 17-2 shows the error mechanisms that limit the dynamic range of the analyzer. A single frequency is at the input of the analyzer along with its harmonics generated internal to the analyzer. For simplicity, we have shown only one input frequency. Had there been more than one frequency, we would also have intermodulation distortion products (in addition to the harmonic distortion products). Other sources of error shown are the residual and input-related spurious responses in the analyzer. The third factor in dynamic range limitations is the internal noise of the analyzer, which produces a noise floor below which a signal cannot be measured. The measured level of this noise depends on the resolution bandwidth used.

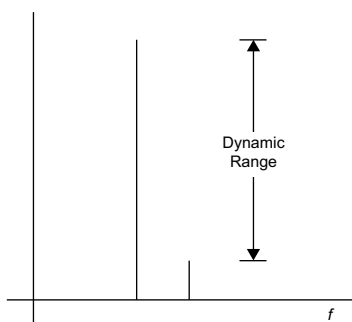


Figure 17-1 The dynamic range of a spectrum analyzer is the ratio (expressed in dB) of the largest and smallest signals that can be reliably measured at the same time.

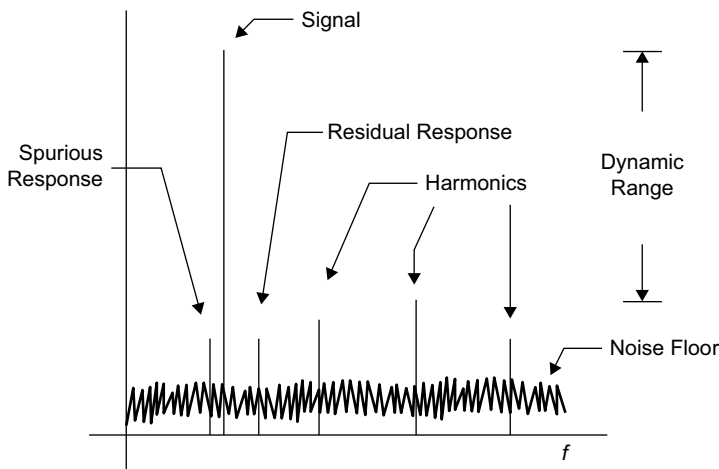


Figure 17-2 Dynamic range is limited by the analyzer’s harmonic distortion, internal noise, residual responses, and input-related spurious responses.

A narrower bandwidth will allow less noise into the measurement, thereby reducing the measured level of the noise. Any of these three mechanisms (distortion, residual/spurious responses, and noise) can limit the dynamic range of the instrument.

The instrument user can take some steps to optimize the dynamic range for the user’s particular measurement application. If noise is limiting the dynamic range, reducing the predetection bandwidth (by averaging or filtering) will reduce the noise level measured by the analyzer without affecting the measured signal level. If the distortion products are the limiting factor, they can be reduced by reducing the signal level. As discussed in Chapter 7, the distortion products will drop in amplitude by a larger amount than the signal level, causing an increase in dynamic range. An external attenuator supplied by the user or the analyzer’s internal attenuator can reduce the signal level. Of course, as the signal level is reduced, the dynamic range may be limited by the noise floor.

The effect of distortion and noise on dynamic range is sometimes shown on an instrument data sheet using a figure similar to Figure 17-3. The horizontal axis is the level of the signal at the input mixer. We can just think of this as corresponding to the input signal level, plus or minus any amplification or attenuation in the analyzer front end.

Looking at the plot of DANL, we see that at a mixer level of -25 dBm the DANL is 130 dB below the mixer level. In this case, the DANL would limit the dynamic range to 130 dB. (This implies that the DANL is -25 dBm $-$ 130 dB = -155 dBm.) As the mixer level is decreased (moving to the left on the plot), the *relative* DANL level increases until it is -75 dB when the mixer input is -80 dBm. In other words, having a lower signal level at the mixer causes DANL to limit the dynamic range of the measurement.

Now consider the plot of second-harmonic distortion, which has the opposite slope. At a mixer level of -25 dBm, the second harmonic distortion is 75 dB below the mixer level. As the mixer level is decreased, the *relative* level of the second harmonic distortion drops until it is -130 dB when the mixer level is -80 dBm. Thus, having a high signal level at the mixer causes the second harmonic distortion to limit the dynamic range but at lower signal levels the harmonic performance improves dramatically. The third-order intermodulation distortion

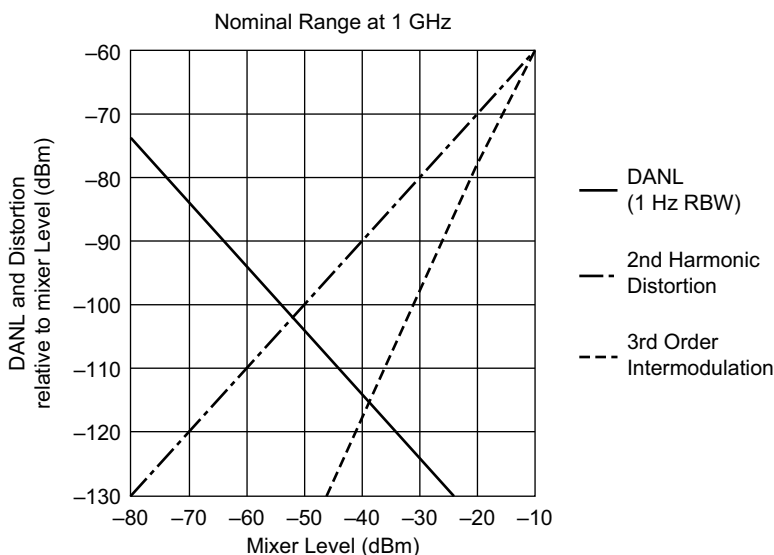


Figure 17-3 Dynamic range for a spectrum analyzer may be described using a plot of DANL, second harmonic, and third-order intermodulation performance. (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

plot behaves in a similar manner but has a larger slope, meaning that the third-order products drop more quickly as the signal level is decreased. This is consistent with the distortion model described in Chapter 7.

Figure 17-3 provides a quantitative way to assess spectrum analyzer dynamic range and set up the instrument for the best measurement. Keep in mind that the DANL specification is relative to a 1 Hz bandwidth, which may require the use of a narrow resolution bandwidth that may slow down the measurement. Using a wider RBW can speed up the measurement but with a higher noise level.

17.4 Network Analyzer Specifications

As a product category, network analyzers have evolved from having separate sources, receivers, and test sets into well-integrated measuring instruments. The modern network analyzer is a complete system in a box, able to accurately test high frequency components. Accordingly, the specifications have evolved to reflect the entire system performance. Ultimate network analyzer performance depends on the vector error correction as discussed in Chapter 15. Most of the analyzer specifications apply only with a measurement calibration and error correction applied.

Dynamic Range

Network analyzers have a different set of factors that affect the dynamic range. The performance of both the source and receiver must be considered, since either can limit the dynamic

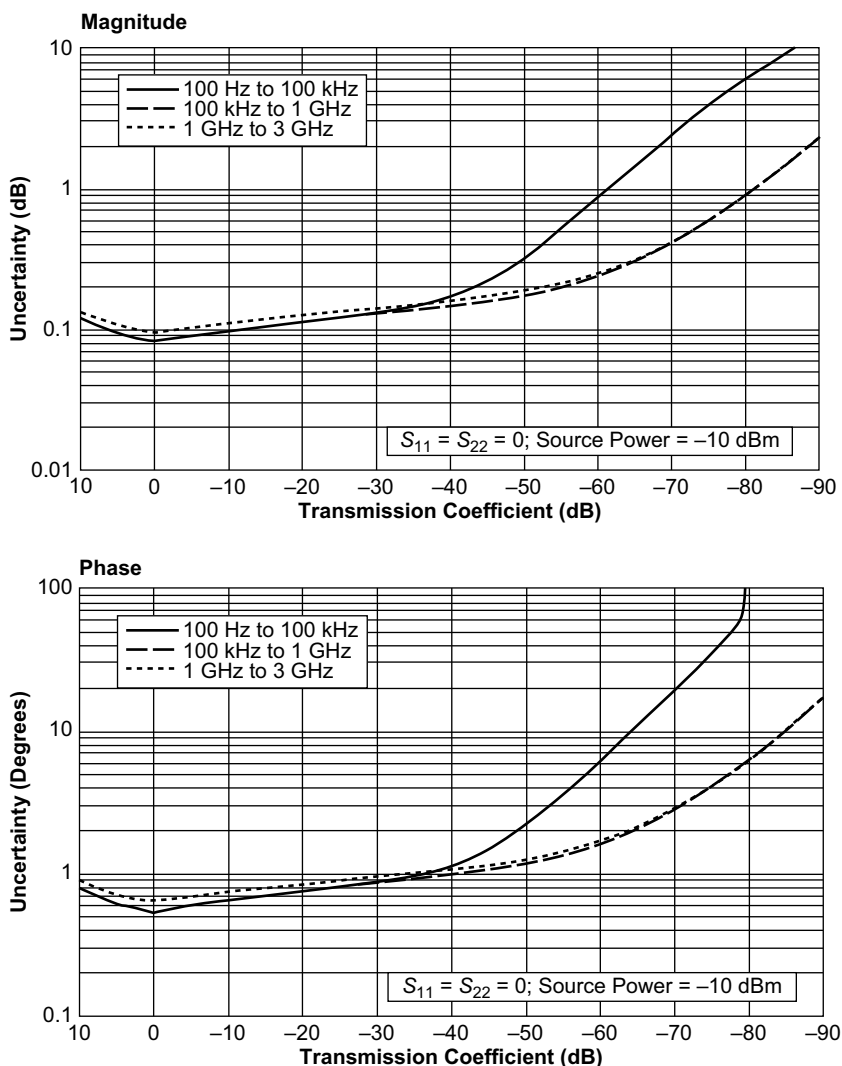


Figure 17-4 A typical example of a network analyzer specification of transmission coefficient uncertainty (magnitude and phase). (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

range. With the general case of a spectrum measurement, we have to assume the signal is largely unknown. In the case of a classic network measurement, the source and receiver are both tuned to the same frequency. Therefore, harmonic distortion (in either the source or receiver) is not usually a problem since the harmonics fall outside the receiver passband. Intermodulation distortion is usually negligible since classic network measurements are performed with only one frequency stimulating the device under test (DUT).

Source spurious responses can potentially cause measurement error but do not usually limit dynamic range. Since the source is always at the measurement frequency, its amplitude

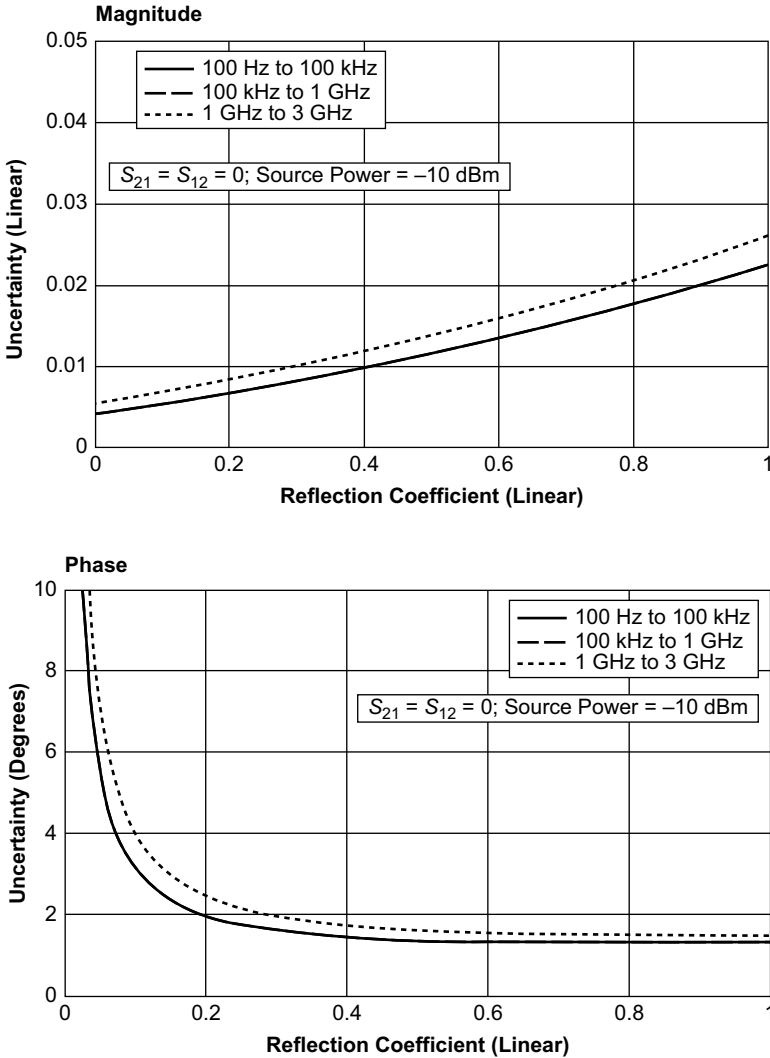


Figure 17-5 A typical example of a network analyzer specification of reflection coefficient uncertainty (magnitude and phase). (© Keysight Technologies, Inc. Reproduced with Permission, Courtesy of Keysight Technologies, Inc.)

tends to dominate over small spurious frequencies. The device under test attenuates these spurious responses along with the desired source frequency.

The remaining analyzer imperfections that normally limit the dynamic range of a network measurement are receiver residual responses and the receiver noise floor. As in the spectrum analyzer case, the noise floor can be reduced by narrowing the predetection bandwidth, perhaps at the expense of increased measurement time.

Network analyzer dynamic range is usually specified in dB, for a particular frequency range or other configuration.

Transmission and Reflection Measurements

The most common network analyzer measurement is the classic two-port measurement of S_{11} , S_{12} , S_{21} , and S_{22} . The measurement uncertainty for the transmission measurements, S_{12} and S_{21} , and the reflection measurements, S_{11} and S_{22} , is specified in terms of magnitude and phase (Figures 17-4 and 17-5). These uncertainty specifications are for the corrected system performance, after vector error correction has been applied, using the specified procedure and calibration kit.

Bibliography

Agilent Technologies. “Agilent 2-Port and 4-Port PNA-X Network Analyzer Data Sheet and Technical Specifications,” Publication Number N5247-90002, October 2011.

Agilent Technologies. “CXA X-Series Signal Analyzer N9000A Data Sheet,” Publication Number 5990-4327EN, July 2013.

Agilent Technologies. “E5061B Network Analyzer Data Sheet,” Publication Number 5990-4392EN, July 2011.

Agilent Technologies. “N9020A MXA X-Series Signal Analyzer Data Sheet,” Publication Number 5989-4942EN, October 2013.

Agilent Technologies. “Optimizing RF and Microwave Spectrum Analyzer Dynamic Range,” AN-1315, Publication Number 5968-4545E, 2000.

Institute of Electrical and Electronics Engineers (IEEE). “I.E.E.E. Standard for Spectrum Analyzers,” I.E.E.E. Standard 748-1979. New York: IEEE, September 1979.

Two-Port Vector Error Correction

For the most accurate network measurements, vector error correction is employed, as discussed in Chapter 15. In this Appendix, we will examine the two-port model in more detail.

Figure A-1 shows the forward error model for the two-port error correction. For simplicity, only half of the error terms for the two-port model are shown: the ones relevant to forward measurements. The model shown is sufficient for error-corrected S_{11} and S_{21} measurements. There is a corresponding error model for the reverse measurements, S_{22} and S_{12} .

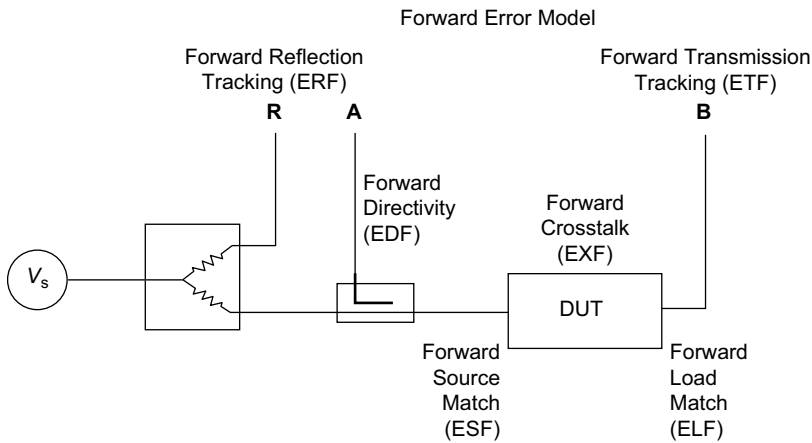


Figure A-1 The forward error model is used to show the error terms for a two-port S-parameter measurement. (Adapted from Joel P. Dunsmore, *Handbook of Microwave Component Measurements*, Wiley, 2012, Chapter 3.)

S_{11M} is the measured version of S_{11A} , which is the actual S_{11} value for the device under test (DUT). The error terms included in the equation are all from the left column of Table A-1, which means the forward error model is sufficient to describe the measured result for S_{11} . Note that S_{11M} depends on all four actual S parameters: S_{11A} , S_{21A} , S_{22A} , and S_{12A} .

$$S_{11M} = EDF + \frac{ERF \left(S_{11A} + \frac{S_{21A} \cdot ELF \cdot S_{12A}}{(1 - S_{22A} \cdot ELF)} \right)}{\left[1 - ESF \cdot \left(S_{11A} + \frac{S_{21A} \cdot ELF \cdot S_{12A}}{(1 - S_{22A} \cdot ELF)} \right) \right]}$$

Table A-1 Error Mechanisms Associated with Two-Port Vector Network Analysis

Forward Error Terms	Reverse Error Terms
EDF Forward Directivity	EDR Reverse Directivity
ESF Forward Source Match	ESR Reverse Source Match
ERF Forward Reflection Tracking	ERR Reverse Reflection Tracking
ELF Forward Load Match	ELR Reverse Load Match
ETF Forward Transmission Tracking	ETR Reverse Transmission Tracking
EXF Forward Crosstalk	EXR Reverse Crosstalk

Similarly, S_{21M} is the measured version of S_{21A} , which is the actual S_{21} value for the DUT. The error terms included in the equation are all from the left column of Table A-1, which means the forward error model is sufficient to describe the measured result for S_{21} . Note that S_{21M} depends on all four actual S parameters: S_{11A} , S_{21A} , S_{22A} , and S_{12A} .

$$S_{21M} = \frac{S_{21A} \cdot ETF}{(1 - S_{11A} \cdot ESF)(1 - S_{22A} \cdot ELF) - (ESF \cdot S_{21A} \cdot S_{12A} \cdot ELF)} + EXF$$

Not surprisingly, the reverse error model is roughly the mirror image of the forward model (Figure A-2).

The equations for S_{12M} and S_{21A} use the reverse error model, and the equations are the mirror image of the forward case. In this case, the error terms are all from the right column of Table A-1. Note that S_{22M} and S_{12M} both depend on all four actual S parameters: S_{11A} , S_{21A} , S_{22A} , and S_{12A} .

$$S_{12M} = \frac{S_{12A} \cdot ETR}{(1 - S_{22A} \cdot ESR)(1 - S_{11A} \cdot ELR) - (ESF \cdot S_{12A} \cdot S_{21A} \cdot ELR)} + EXR$$

$$S_{22M} = EDR + \frac{ERR \left(S_{22A} + \frac{S_{12A} \cdot ELR \cdot S_{21A}}{(1 - S_{11A} \cdot ELR)} \right)}{\left[1 - ESR \cdot \left(S_{22A} + \frac{S_{12A} \cdot ELR \cdot S_{21A}}{(1 - S_{11A} \cdot ELR)} \right) \right]}$$

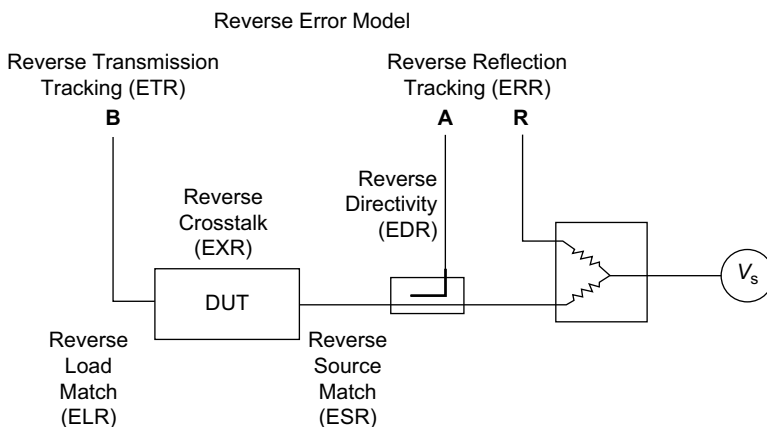


Figure A-2 The reverse error model is used to show the reverse error terms for a two-port S -parameter measurement.

One-Port Case

This degenerates into the one-port case by setting $S_{21A} = 0$:

$$S_{11M} = EDF + \frac{ERF(S_{11A} + 0)}{[1 - ESF \cdot (S_{11A} + 0)]}$$

$$S_{11M} = EDF + \frac{ERF \cdot S_{11A}}{(1 - ESF \cdot S_{11A})}$$

Chapter 15 uses the error model described by Paul Ely (1967) to explain vector error correction:

$$\Gamma_M = D + \frac{(1 + T_R)}{(1 - M_S \Gamma_A)} \Gamma_A$$

where

- Γ_A = actual reflection coefficient
- Γ_M = measured reflection coefficient
- D = directivity error
- T_R = frequency response error
- M_S = source match error

We can equate this to the nomenclature used in the two-port model by setting

$$\Gamma_M = S_{11M}$$

$$\Gamma_A = S_{11A}$$

For these two equations to be equal, the following equations must be valid:

$$EDF = D$$

$$ERF = 1 + T_R$$

$$ESF = M_S$$

Bibliography

Paul C. Ely, Jr. "Swept Frequency Techniques." *Proceedings of the IEEE*, June 1967.

Index

- 1/f noise 154, 155
- ABCD parameters 247
- absolute decibel values 14–17
- accuracy enhancement 290
- adjacent channel power (ACP) 133, 134
- adjacent channel power ratio (ACPR) 134
- adjacent-point averaging 187, 191
- admittance parameters 246
- aliasing 49, 50
- amplitude absolute accuracy 318
- amplitude dynamic accuracy 318
- amplitude errors 315
- amplitude frequency response/flatness 318
- amplitude modulation (AM) 107, 108–11
 - measurement 112–14
 - sinusoidal modulation 110–11, 112
 - time domain 111
- amplitude resolution 318
- amplitude sweep 262
- analyzer performance and specifications 315
 - network analyzer specifications 321
 - dynamic range 321–3
 - transmission and reflection measurements 324
 - receiver characteristics 317–18
 - source specifications 315–17
 - spectrum analyzer dynamic range 319–21
- angle modulation 107
- antennas, EMC 300–2
- anti-alias filter 49, 50, 53, 65
- attenuating coupler 224
- attenuating probes 222–3
- attenuators
 - classical attenuator problem 232–4
 - high-impedance 228–9
 - Z_0 attenuators 229–30
- audio oscillator, harmonic distortion of 63
- autocorrelation function 73–5
- automatic noise level measurement 159
- averaging 177, 183
 - exponential weighting 185–7
 - versus filtering 191–2
 - general averaging 184–5
 - linear weighting 185
 - root mean square (RMS) average 188
 - smoothing 191
 - in spectrum and network analyzers 187
 - log detector problem 187–8
 - variance ratio 183–4
 - vector averaging 188–90
- band-pass networks 270
- band selectable analysis 53–4
- bank-of-filters analyzer 43–4, 87
- bayonet Neill Concelman (BNC) connector 224–5, 294
- BiConiLog 301
- binary phase-shift keying (BPSK) 127, 129
- bins, defined 51
- bit rate 127
- cardinal values, for decibels 13
- carrier amplitude, measuring 112
- carrier null method 122–3
- Carson’s rule 121
- channel power 133
- characteristic impedance 196, 216, 248
- coaxial cables 195
- coaxial lines 216–17
- coherence measurements 70–2
- combined spectrum/network analyzers 9
- complex reflection coefficient 203, 281, 282

- conducted emissions 297, 298, 306
- correlation 72–3
- cross-correlation 75–6
- cross-power spectrum 68–70
- cumulative density function (CDF) 76
- current probe 312–13
- current transformer 312

- DC bin 51
- DC response/LO feedthrough 317
- decibels (dB) 11
 - cardinal values 13
 - definition of 11–12
 - error expressed in 20–1
 - gain and loss calculations 17
 - multiple blocks 18–19
 - voltage gain 17–18
 - and percent 19–20
 - values 14
 - dB μ V 15
 - dBm 14
 - dBm/dBV conversions 15–16
 - dBmV 15
 - dBV 15
 - dBW 15
 - high-impedance measurements 16–17
- decimating low-pass filters 99, 100
- delay aperture 274–5
- delay specifications 318
- desensitization 175
- device under test (DUT) 7, 65, 151, 212–13, 255, 267
- digital modulation 125–6, 131, 133
 - formats 127–30
- directional bridges 257, 284–7
- directional coupler 257, 258, 284, 285, 288
- directional devices 251, 285, 287, 291, 293
 - coupling factor of 286
 - insertion loss of 286
 - standard error model for 289
 - three-port 286
- directivity 286–7, 291, 292
- direct spectrum technique 164
- discrete Fourier transform (DFT) 23, 36–8
 - inverse DFT (IDFT) 38
 - limitations of 38
- displayed average noise level (DANL) 317, 320
- distortionless system/network 267
- distortion measurements 137
 - distortion internal to the analyzer 148–9
 - distortion model 137
 - harmonic distortion measurements 145–6
 - higher-order models 142
 - intercept concept 142–5
 - intermodulation distortion measurements 147–8
 - single-tone input 138–9
 - two-tone input 139–42
 - use of low-pass filter on source 146–7
- double-sideband (DSB) modulation 115
- double-sided frequency domain representations 25
- duty cycle 168, 175
- dynamic range
 - network analyzer 321–3
 - spectrum analyzer 319–21
- dynamic signal analyzer: *see* fast Fourier transform (FFT) analyzer

- EDR Reverse Directivity Error 294
- ESR Reverse Source Match Error 294
- ERR Reverse Reflection Tracking Error 294
- ELR Reverse Load Match Error 294
- ETR Reverse Transmission Tracking Error 294
- EXR Reverse Crosstalk Error 294
- effective impulse bandwidth (IBW) 176
- effective pulse width 169–70
- E-field measurement, in dB 301
- electrical delay 279
- electrical length compensation: *see* electrical delay
- electric field probe 310, 311, 312
- electromagnetic compatibility (EMC)
 - measurements 97, 297–8
 - antennas 300–2
 - conducted emissions 306
 - current probe 312–13
 - full-compliance testing 297, 298
 - line impedance stabilization network (LISN) 306–9
 - near field and far field 303
 - near-field probes 310–12
 - peak, quasi-peak, and average detectors 305–6
 - preamplifiers 313–14
 - precompliance testing 297–8
 - radiated emissions 298–300
 - troubleshooting 309–10

- electromagnetic interference (EMI) 297
 - average detector 305
 - receiver requirements 304
 - detectors 305
 - resolution bandwidth 304
- electronic calibration standards 295
- electronic filter characterization 67–8
- equivalent noise bandwidth (NBW) 155–7
- error correction, vector 290
 - three-term 291–2
 - two-port 293–5
 - two-term 291
- error expressed in decibels 20–1
- error vector magnitude (EVM) 131–2
- exponential weighting 185–7
- exponential window 60–2

- far field 303
- fast Fourier transform (FFT) analyzer 6, 39, 43, 45–6, 103, 167, 177, 180, 181
 - autocorrelation 73–5
 - averaging in 183
 - band selectable analysis 53–4
 - bank-of-filters technique 43–4
 - coherence 70–2
 - controlling the frequency span 52–3
 - correlation 72–3
 - cross-correlation 75–6
 - cross-power spectrum 68–70
 - electronic filter characterization 67–8
 - exponential window 60–2
 - flattop window 58–9
 - frequency resolution 44–5
 - Hanning window 55–8
 - histogram 76–8
 - leakage 55
 - network measurements 65–6
 - octave measurements 84
 - oscillator characterization 62–3
 - overlap processing 81–3
 - phase 66–7
 - properties 51–2
 - real-time bandwidth 78–9
 - and RMS averaging 79–80
 - and transients 80–1
 - sampled waveform 46–7
 - sampling theorem 47–50
 - spectral maps 63–5
 - swept sine 83
 - versus swept spectrum analyzers 98–9
 - time average in 190
 - time domain display 65
 - time domain waveform 46–7
 - uniform window 59–60
 - window function, selecting 62
- feedthrough termination 225
- filtering 177
 - averaging *versus* 191–2
 - postdetection filtering 180–1, 182
 - postdetection filters 181, 183
 - predetection filtering 177–9
 - predetection filters 179–80
- finite measurement time 40–1
- flat line 204
- flatness specification 316
- flattop window 58–9, 62, 63
- flicker noise 154
- folding frequency 48, 50
- forward crosstalk (EXF) error 293
- forward directivity error (EDF) error 293
- forward load match (ELF) error 293
- forward reflection tracking (ERF) error 293
- forward source match (ESF) error 293
- forward transmission coefficient 249, 253
- forward transmission tracking (ETF) error 293
- Fourier theory 2, 23
 - discrete Fourier transform (DFT) 36–8
 - limitations of 38
 - fast Fourier transform (FFT) 39
 - finite measurement time 40–1
 - Fourier series 23, 24–5
 - of a square wave 25–9
 - of other waveforms 30–1
 - periodic function, representation of 24
 - Fourier transform 31–2
 - properties of 36
 - of a pulse 32–3
 - relationships 33
 - inverse Fourier transform 33, 34–5
 - periodicity 23
 - relating theory to measurements 39–40
- Fourier transform 23, 31–2, 153
 - properties of 36
 - of a pulse 32–3
 - relationships 33
- free-space wavelength 303
- frequency division multiplexing (FDM)
 - systems 5

- frequency domain 23
 - measurements, advantages of 4–5
 - representation 2, 23, 32
 - double-sided 25
 - single-sided 25
 - of square wave 29
- frequency mask triggers 101, 103, 104
- frequency-modulated carrier 116
- frequency modulation (FM) 107, 115, 116
 - combined AM and FM 123–5
 - measurements 122–3
 - carrier null method 122–3
- frequency resolution specification 316
- full-compliance measurements, EMC 297, 298
- group delay 267, 273–5
- Hanning window 55–8
- harmonic distortion 62–3, 138
 - measurement of 4–5, 145–6
- harmonics 316
- high-impedance attenuators 228–9
- high-impedance filters 236–7
- histogram 76–8
- hybrid parameters 246–7
- image filter 89
- impedance matching attenuator 234
- impedance matching pad 234
- impedance parameters 245
- incremental accuracy: *see* amplitude dynamic accuracy
- in-phase modulation 126
- input impedance 317
 - open-circuit 246
 - of transmission line 207–9
- input reflection coefficient 248
- input-related spurious responses 317
- insertion gain/loss 212–16
- intermediate frequency (IF) 88
 - amplifier stage 142
 - filter 88, 92
- intermodulation distortion (IMD) 140
- intermodulation distortion measurements 147–8
- inverse discrete Fourier transform (IDFT) 38
- inverse fast Fourier transform (IFFT) 264
- leakage 38, 55
- level accuracy 315, 316
- level linearity 316
- lightwave analogy 202, 248
- linear distortion 269–70
- linear phase, importance of 270–3
- linear time invariant (LTI) system 4, 137, 241
- linear weighting 185
- line impedance stabilization network (LISN) 306–9
- line losses 216
- line spectra 29, 170–1
- line stretch: *see* electrical delay
- local oscillator (LO)
 - frequency 89
 - signal 88
- log scale fidelity: *see* amplitude dynamic accuracy
- lower-sideband (LSB) modulation 115
- low-pass filter 146–7
 - transmission characteristics of 8
- magnetic-field probe 310, 311
- manual sweep 95
- maximum voltage and power transfer 220
- mean value 152
- measurement connections 219
 - active high-impedance probes 223–4
 - attenuators 228–30
 - classical attenuator problem 232–4
 - high-impedance inputs 220
 - attenuating probes 222–3
 - high-impedance probes 220–2
 - impedance matching devices 234
 - minimum loss pads 234–5
 - transformers 235–6
 - input connectors 224–5
 - loading effect 219
 - maximum voltage and power transfer 220
 - measurement filters 236
 - high-impedance filters 236–7
 - Z_0 filters 237–8
 - power dividers and splitters 225–7
 - return loss improvement 230–2
 - Z_0 terminations 225
- measurement filters 236
 - high-impedance filters 236–7
 - Z_0 filters 237–8
- measurement instrumentation 1
- measurement plane 278–9
- minimum loss pads 234–5

- mismatch errors 213–14
- mismatch loss 210
- mismatch uncertainty 211, 214
- modular instruments 10
- modulating frequency, measuring 112
- modulation error ratio (MER) 131
- modulation index, measuring 112–13
- modulation measurements 107
 - amplitude modulation (AM) 108–11
 - measurement 112–14
 - sinusoidal modulation 110–11, 112
 - time domain 111
 - angle modulation 115–18
 - carrier 107–8
 - channel measurements 133–4
 - combined AM and FM 123–5
 - common digital modulation formats 127–30
 - digital modulation 125–6
 - error vector magnitude 131–2
 - frequency modulation (FM) measurements 122–3
 - narrowband angle modulation 118–19
 - quadrature modulation 126–7
 - varieties of 115
 - wideband angle modulation 119–22
 - zero-span operation 114–15
- multichannel network analyzer amplitude characteristics 317
- multiple phase detector techniques 165

- narrowband angle modulation 118–19
- narrowband frequency domain measurements 4
- near-field probes 310–12
- network analyzer 7, 253, 285
 - basic network measurements 253
 - directional bridges and couplers 257
 - flexible source frequency 262–4
 - network measurements using spectrum analyzer 254–5
 - oscilloscope and sweep generator 253–4
 - power sweep 262
 - S-parameter test set 257–9
 - specifications 321
 - dynamic range 321–3
 - transmission and reflection measurements 324
 - sweep limitations 260–2
 - vector network analyzer (VNA) 255–7
 - configurations 259–60
 - nonlinear VNA measurements 265
 - time domain measurements 264
- network measurements 3, 7–8, 253
 - FFT analyzer 65–6
 - using spectrum analyzer 254–5
 - see also* vector network measurements
- noise and noise measurements 151, 158–9, 177–9
 - automatic noise level measurement 159
 - equivalent noise bandwidth 155–7
 - frequency distribution 153–5
 - mean, variance, and standard deviation 152
 - noise floor 159–60
 - correction for 160–1
 - noise units and decibel relationships 157–8
 - phase noise 161–5
 - power spectral density (PSD) 153
 - random noise, statistical nature of 151–2
- noise equivalent bandwidth: *see* equivalent noise bandwidth (NBW)
- noise floor extension 160
- noise marker 159
- nondeterministic noise 151
- nonharmonic spurious signals 317
- nonlinearities 269
- nonlinear vector network analyzer (NVNA) measurements 265
- normalization 275–8, 291
 - of reflection measurement 288
- Nyquist rate 47

- occupied bandwidth (OBW) of signal 133
- octave measurements 84
- open-circuit input impedance 246
- open circuits 294
- oscilloscope 253–4
 - probes 220, 221
- output reflection coefficient 249
- overlap processing 81–3

- pads: *see* attenuators
- peak detector 305
- periodic function, Fourier series representation of 24
- periodic signal 23, 24, 31, 40
- periodic waveforms, autocorrelation of 75
- phase accuracy 318
- phase dynamic accuracy 318

- phase error 273, 274
- phase frequency response 318
- phase lock loop 2
- phase-modulated carrier 116
- phase modulation (PM) 107, 115
- phase noise 161–5, 316
- phase resolution 318
- polar display formats 279
- postdetection filtering 180–1, 182
- postdetection filters 181, 183
- power attenuator 224
- power average 187–8
- power combiner 225
- power dividers 225–7
- power gain 17
- power loss 17
- power spectral density (PSD) 153, 157
- power splitters 225–7
- power sweep 262
- preamplifiers 313–14
- precompliance testing 297–8, 309
- predetection filtering 177–9
 - noise 177–9
- predetection filters 179–80
- probability density function (PDF) 76, 151–2
- probability of intercept (POI) 101
- propagation velocity 197
- pseudorandom noise (PRN) 59
- pulse measurements 167
 - effective pulse width 169–70
 - line spectrum 170–1
 - pulse desensitization 175–6
 - pulsed RF 174
 - pulse spectrum 171–4
 - spectrum of pulsed waveform 167–9
 - sweep time 172
- pulse repetition frequency (PRF) 167–9

- quadrature amplitude modulation (QAM)
 - 127–8, 130
- quadrature modulation 126–7
- quadrature phase-shift keying (QPSK) 127, 129, 131, 132
- quasi-peak detector 97, 305

- radiated emissions measurement 297, 298–300
- random noise, statistical nature of 151–2
- real-time bandwidth (RTBW) 78–9, 101
 - and RMS averaging 79–80
 - and transients 80–1
- real-time spectrum analyzer (RTSA) 101–3, 104
- receiver characteristics 317–18
- reflection coefficient 203
- reflection configuration 287
- reflection measurements 7, 279–84
- reflection measurements, error in 289–90
- reflection normalization 288
- reflection tracking error 291
- reflectometer 285
- relative constellation error (RCE): *see* error vector magnitude (EVM)
- residual responses 317
- resolution bandwidth (RBW) 44, 92, 158, 159–60
 - of EMI receiver 304
- return loss 203–4, 288
 - improvement 230–2
- reverse error model 326
- reverse transmission coefficient 249
- RF analyzer 9
- RF signal 88, 175
- RMS average detector 305
- root mean square (RMS) average 188
- root mean square (RMS) value
 - of a pulsed RF signal 175
- Rosenfell detector 97

- sampling theorem 47–50
- scalar network analyzer (SNA) 8
- scalar reflection coefficient 203, 211
- scattering parameters 247–50, 325–6
 - test set 257–9
 - two-port networks 250–1
- second harmonic intercept (SHI) 317
- second-order intercept point 144
- selective level meter 87
- self-windowing 59, 61
- short circuits 288, 294
- short-open-load-through (SOLT) calibration 295
- signal analyzer 104
- signals and systems 1–2
- single-conversion receiver 91
- single pulse, spectrum of 33
- single-sideband (SSB) modulation 115
- single-sided frequencies 25
- single-tone input 138–9
- sinusoidal modulation 110–11, 112, 116–17
- sinusoidal signals 241–3

- sinusoidal voltages 202–3
- Smith chart 281, 282, 283
- smoothing 191
- source specifications 315–17
- S-parameters: *see* scattering parameters
- specifications 315
 - delay 318
 - network analyzer 321
 - dynamic range 321–3
 - transmission and reflection measurements 324
 - receiver characteristics 317–18
 - source 315–17
- spectral lines, defined 29
- spectral maps 63–5
- spectrum analyzer 3, 5, 7, 112
 - dynamic range 148–9, 319–21
 - fast Fourier transform (FFT): *see* fast Fourier transform (FFT) analyzer
 - frequency resolution of 39
 - types 103–4
 - zero-span mode of 114
- spectrum measurements 5–6
- spurious responses 317
- square wave
 - Fourier series of 25–9
 - frequency domain representation of 29
- standard deviation 152, 183, 184
- standing wave 204–7
- standing wave ratio (SWR) 204–7, 279
 - SWR meter 285
- stationary signal 40
- sweep generator 253–4, 255
- swept-sine analysis 83
- swept spectrum analyzers 6, 87, 89–91
 - detectors, types of 97
 - digital IF section 96–7
 - discrete sweep 95
 - FFT versus 98–9
 - heterodyne block diagram 88–9
 - IF detector section of 96
 - IF response of 94
 - input section 91
 - local oscillator (LO) feedthrough 95
 - manual sweep 95
 - modern spectrum analyzer block diagrams 99–101
 - power sweep 98
 - practical considerations 91
 - program sweep/list sweep 95
 - quadrature detector 100
 - real-time spectrum analyzer (RTSA) 101–3
 - resolution bandwidth 92
 - specialized sweep modes 95
 - sweep limitations 92–5
 - tracking generator 98
 - types of 103–4
 - wave analyzer 87
 - zero-span operation 95
- synchrotune operation: *see* zero-span operation
- system function: *see* transfer function
- system transfer function 3–4, 7
- third-order intercept (TOI) 317
- third-order intercept point 144
- three-term error correction model 291–2
- through-reflect-line (TRL) method 295
- time average in FFT analyzers 190
- time domain 23
 - and frequency domain relationships 2–3
- time domain display 65
- time domain pulse 32
- time domain reflectometry (TDR) 264
- total harmonic distortion (THD) 146
- trace-to-trace averaging 187, 192
- tracking generator 98, 255
- transfer function 243–4
 - and forward transmission coefficient 250
 - of a system 3–4, 7
- transformers 235–6
- transform pairs, defined 33
- transient events, real-time bandwidth and 80–1
- transmission and reflection measurements 324
- transmission lines 195
 - characteristic impedance 196
 - coaxial lines 216–17
 - complex reflection coefficient 203
 - distributed model 195–6
 - generator, line, and load 197
 - non- Z_0 load 198–200
 - open load 200
 - short load 201
 - Z_0 load 197–8
 - impedance changes 201–2
 - input impedance of 207–9
 - insertion gain and loss 212–16
 - line losses 216

- measurement error due to impedance
 - mismatch 209
 - imperfect source, imperfect load (case) 210–12
 - perfect source, imperfect load (case) 209–10
- need for 195
- propagation velocity 197
- return loss 203–4
- sinusoidal voltages 202–3
- standing waves 204–7
- transmission parameters 247
- transmission/reflection test set 257
- two-channel cross-correlation technique 165
- two-port error correction model 293–5
- two-port networks 241
 - admittance parameters 246
 - hybrid parameters 246–7
 - impedance parameters 245–6
 - improved two-port model 244–5
 - scattering parameters 247–50
 - sinusoidal signals 241–3
 - S-parameters 250–1
 - transfer function 243–4
 - and forward transmission coefficient 250
- two-port vector error correction 325–7
- two-term error correction model 291
- two-tone signal 139–42

- uniform window 59–60
- upper-sideband (USB) modulation 115

- variance 183
 - of waveform 152
- variance ratio (VR) 183–4
 - for linear averaging 185
- vector averaging 188–90
- vector error correction 267, 290
 - two-port 325–7
- vector network analyzer (VNA) 8, 255–7, 295
 - configurations 259–60
 - nonlinear VNA (NVNA) measurements 265
 - time domain measurements 264
- vector network measurements 267
 - directional bridges and couplers 284–7
 - distortionless transmission 267–9
 - group delay 273–5
 - linear distortion 269–70
 - linear phase, importance of 270–3
 - measurement plane 278–9
 - nonlinearity 269
 - normalization 275–8, 291
 - reflection configuration 287
 - reflection measurements 279–84
 - error in 289–90
 - reflection normalization 288
 - three-term error correction model 291–2
 - two-port error correction model 293–5
 - two-term error correction model 291
 - vector error correction 290
- vector signal analyzer (VSA) 104
- video filter 90, 180
 - see also* postdetection filters
- voltage-controlled oscillator (VCO) 90
- voltage divider relationship 219
- voltage gain 17–18
- voltage standing wave ratio (VSWR):
 - see* standing wave ratio (SWR)

- waterfall display: *see* spectral maps
- wave analyzer 87, 89
- wave meter: *see* wave analyzer
- weighting
 - exponential 185–7
 - linear 185
- white noise 153–4
- wideband angle modulation 119–22
 - Carson's Rule 121

- X-parameters 265

- Z_0 attenuators 229–30
- Z_0 filters 237–8
- Z_0 loads 294–5
- zero-span operation 114–15
- zoom operation: *see* band selectable analysis