Signals and Communication Technology

Sean A. Fulop

Speech Spectrum Analysis

Springer

Signals and Communication Technology

For further volumes:
http://www.springer.com/series/4748

Sean A. Fulop

Speech Spectrum Analysis

Springer

Dr. Sean A. Fulop
Department of Linguistics
California State University Fresno
N. Backer Ave. 5245
Fresno CA 93740-8001
USA
e-mail: sfulop@csufresno.edu

ISSN 1860-4862

ISBN 978-3-642-17477-3 e-ISBN 978-3-642-17478-0

DOI 10.1007/978-3-642-17478-0

Springer Heidelberg Dordrecht London New York

Cover design: eStudio Calamar, Berlin/Figueres

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

For Billy, who helped me write this.

Preface

The analysis and measurement of the spectrum of a speech signal is one of the most important areas of sound signal processing for a number of fields, yet it is not an area to which a book has been specifically devoted. The accurate determination of the speech spectrum is commonly pursued in diverse areas including speech processing, recognition, and acoustic phonetics. With this book I hope to make the subject of spectrum analysis understandable to a wide audience, which I imagine could include those with a solid background in general signal processing (but not necessarily in speech), and also speech scientists and students with some acoustic phonetics experience who have limited knowledge of signal processing. In keeping with these goals, this is not a book that replaces or attempts to cover the material found in a general signal processing textbook. Some essential signal processing concepts are presented in Chap. 2, but even there the concepts are presented in a generally understandable fashion as far as is possible. Throughout the book, the focus will be on applications to speech analysis and the measurement of important descriptive speech parameters. No attention is paid to parametrizing speech purely for coding or decorrelation for further processing. Mathematical theory will be provided for completeness, but many of these developments are set off in boxes for the benefit of those readers with sufficient background. Other readers may proceed through the main text, where the key results and applications will be presented in plain language as far as possible, and illustrated with software routines and practical "show-and-tell" discussions of the results.

At some points, the book refers to and uses the implementations in the Praat speech analysis software package, which has the advantages that it is used by many scientists around the world, and it is free and open source software, obtainable on the internet from the Praat homepage. At other points, special software routines have been developed and made available to complement the book, and these are provided in the Matlab programming language. If the reader has the basic Matlab package, he/she will be able to immediately implement most of the programs in that platform—only Chap. 7 requires the extra Signal Processing toolbox. A few other freely available toolboxes are also needed, and all the Matlab code is made available for download at the Springer website for additional materials.

And finally, as was written by Lord Kelvin and Professor Tait in their *Treatise on Natural Philosophy* (1912), "I confidently hope that few erratums of serious note will now be found in the work."

Fresno, October 2010 Sean Fulop

Acknowledgments

I should definitely thank Kelly Fitz for collaborating with me on developing reassignment algorithms; Doug Nelson for sharing some of his knowledge (and code) on that subject; Paul Boersma for always answering Praat-related questions; François Auger for telling me a few things about the Time–Frequency Toolbox; and Sandy Disner for helping to perfect my reassigned spectrograms. I should also thank Steven Lulich and Stefanie Shattuck–Hufnagel for trying to figure out what a voice bar is. The writing of this book was carried out over a two-year period which included summer visits to the Department of Linguistics, University of Calgary in 2009 and 2010; thanks to John Archibald, Betsy Ritter, and the other members of the department who tolerate my perennial intrusion (and get me a library card). The book was finished while I had a sabbatical at Fresno State, but was never a funded project. Thanks to Christoph Baumann at Springer for facilitating its publication.

Contents

1 Introduction 1
References 4

2 Phonetics and Signal Processing 5
2.1 Essentials of Phonetics 5
2.1.1 Speech Production Fundamentals 5
2.1.2 Syllables and Speech Sounds 7
2.1.3 Vowels and Consonants 8
2.1.4 Uses of Vocal Pitch 12
2.2 Essentials of Digital Signal Processing 12
2.2.1 Periodic and Aperiodic Signals 13
2.2.2 Sampling of Analog Signals 15
2.2.3 Autocorrelation 15
2.2.4 Fourier's Series and Transform Spectra 18
2.2.5 Practical Computing of Fourier Spectra 27
2.2.6 Filters 34
2.2.7 Analytic Signals 35
2.2.8 Concepts of Frequency 37
References 39

3 History of Speech Spectrum Analysis 41
3.1 Fourier Analysis of Speech 41
3.1.1 Early History 41
3.1.2 The Physical Reality of Fourier Components 43
3.1.3 Recording Sound Signals 45
3.1.4 Early Methods of Fourier Analysis 49
3.2 History of Speech Spectra 52
3.2.1 Vowels and Formants: Early Years 52
3.2.2 Vowel Spectra: 1915–1960 57
3.2.3 Spectrographic Analysis 63

3.3 Parametric Spectral Analysis . . . 65
References . . . 66

4 The Fourier Power Spectrum and Spectrogram . . . 69
4.1 The Power Spectrum in Speech Analysis . . . 69
4.1.1 Vowel Spectra . . . 70
4.1.2 Obstruent Spectra and Averaging Techniques . . . 75
4.1.3 Phonation Types . . . 78
4.2 Principles of the Spectrogram . . . 80
4.2.1 Definitions of the Spectrogram . . . 80
4.2.2 Development of Spectrogram Theory . . . 87
4.2.3 Uncertainty Principle . . . 87
4.3 Spectrographic Analysis of Speech . . . 88
4.3.1 General Guidelines . . . 89
4.3.2 Short Window (Wideband) Analysis . . . 92
4.3.3 Long Window (Narrowband) Analysis . . . 99
4.4 Appendix: Praat and Matlab Techniques . . . 102
4.4.1 Praat Functions . . . 102
4.4.2 Matlab Code . . . 104
References . . . 105

5 Alternative Time–Frequency Representations . . . 107
5.1 Wigner–Ville Distribution . . . 108
5.1.1 Definition and Theory . . . 108
5.1.2 Discrete Implementation . . . 109
5.1.3 Features of the Wigner–Ville Distribution . . . 111
5.2 Zhao-Atlas-Marks Distribution . . . 113
5.2.1 Quadratic Distributions . . . 114
5.2.2 Discrete Implementation . . . 115
5.2.3 Speech Analysis with ZAM . . . 118
5.3 Appendix: Matlab Routines . . . 122
References . . . 124

6 The Reassigned Spectrogram . . . 127
6.1 Reassignment: History and Definitions . . . 128
6.2 Reassigning the Spectrogram . . . 131
6.2.1 Nelson's Algorithm . . . 131
6.2.2 Reassigned Power Spectrum . . . 135
6.3 Pruning the Reassigned Spectrogram . . . 136
6.3.1 General Definitions . . . 136
6.3.2 Cross-Spectral Method . . . 138
6.3.3 Justifying the Interpretation of the Phase Derivative . . . 139
6.3.4 Separation of Formants from Glottal Impulses . . . 139
6.4 Analyzing Phonation . . . 140

6.4.1 Beyond Source-Filter Theory 140
6.4.2 Observations on Phonation Types 145
6.4.3 Phonation as a Biometric 149
6.5 Dynamics of Formants 149
6.6 Nasals and Nasalization 156
6.7 Stops and Fricatives 158
6.8 Pitch Tracking 160
6.9 Appendix: Matlab Techniques 162
References 164

7 Linear Prediction and ARMA Spectrum Estimation 167
7.1 Preliminaries 168
7.1.1 Linear Filters and the z-Transform 169
7.1.2 Source-Filter Model of Speech 171
7.2 Speech Spectra from Linear Prediction 173
7.2.1 Computing the LP Coefficients 173
7.2.2 Computing the Gain and Power Spectrum 174
7.3 Interpretation as Filter Spectrum 175
7.3.1 Poles and Resonances 176
7.3.2 Picking Peaks Versus Solving Roots 179
7.4 Practical Spectrum Analysis and Formant Extraction 181
7.4.1 Linear Prediction Accuracy Studies 181
7.4.2 Analysis Windows 182
7.4.3 Filter Order and Pre-Emphasis 184
7.4.4 Pitch-Asynchronous Versus Closed Phase 186
7.5 Application to Real Speech 188
7.6 Autoregressive Moving Average Modeling 192
7.6.1 A Little ARMA Theory 192
7.6.2 ARMA Computation 193
7.6.3 Applications to Speech 194
7.7 Appendix: Praat and Matlab Techniques 197
7.7.1 Praat Functions 197
7.7.2 Matlab Functions 198
References 200

Index 203

Chapter 1
Introduction

The vibration of sound itself contains melody, harmony, and rhythm

Vangelis

The quote from the composer and musician Vangelis sounds at first like an artist's mystical musing, but really it couldn't be more true, particularly of speech sound. When the vocal cords vibrate (as they do during "voiced" speech sounds), they contact each other and produce air pressure pulses in a repeating rhythm between 70–250 times each second. This rhythm is so rapid that it yields a sound in the surrounding air having the vibration rate as its fundamental frequency, which is heard as a pitched tone (melody). Moreover, the complicated mechanical nature of the vocal cord vibration gives rise to a series of harmonic frequencies in the sound, which are integer multiples of the lowest frequency.

Spectrum analysis is essentially a tool for separating the melody, harmony and rhythm of a complicated sound. Its generalization to time–frequency analysis provides a view of how the sound spectrum changes through time, and such analyses are sometimes described as akin to a "musical score." The spectrum of frequencies within a speech sound is the primary means by which the human auditory system can distinguish different sounds, since it is the spectrum which characterizes the distinct sound timbres or "qualities of tone." The features of the spectrum also permit us to infer properties of the speech mechanism, so accurate spectrum analysis will always be an invaluable tool in speech science and linguistics. Indeed, numerous current problems in speech science can be addressed through improvements in spectrum analysis. The measurement of speech formant frequencies is the most obvious example, for which no adequate solution has yet been developed.

A spectrum, in the narrowest sense, is a pure frequency representation of a sound that ignores any changes that may occur over time. Since speech is a dynamically changing signal, the analysis of the frequency spectrum as such has immediate limitations. The development of methods to show how the spectrum of a sound varies through time is commonly known as *time–frequency analysis*, and new techniques in this area have significant potential to improve upon the older methods which most practitioners use. Time–frequency representations with "high resolution" compared to the standard can be used to investigate less understood aspects of speech sound, such as the finely detailed and rapid

S. A. Fulop, *Speech Spectrum Analysis*, Signals and Communication Technology,
DOI: 10.1007/978-3-642-17478-0_1, © Springer-Verlag Berlin Heidelberg 2011

variations of the formants and vocal cord impulses within the confines of individual glottal periods.

The term *spectrum* has an interesting history; it was first used by scientists in the 17th century to refer to the range of colors observed from light passed through a prism. By the 19th century, light was understood as a wave phenomenon, and the spectrum of colors was known to result from a spectrum of frequencies of the waves. In the early 19th century, J. Fourier showed how a complicated function, such as a sound signal, could be represented as a number of superposed functions each with a simple frequency. During the 20th century, the range of such functions (represented by their frequencies) present within a signal, showing their relative amplitudes, came to be called the "spectrum" of frequencies, although I have not determined precisely when this term was first applied to sound. Early practitioners in speech science used other terminology to refer to what was essentially a frequency spectrum of a speech sound, like "harmonic analysis" [4] and "composite frequency analysis" [2], but then Lewis [3] can be found using "spectrum" in a matter-of-fact way in his 1936 study of vocal resonance.

The present book treats speech spectrum analysis like any other complete subject, with its own foundation, history, established practice, and new frontiers. The next two chapters deal with the fundamentals and history of the subject, respectively. These chapters go beyond the usual nutshell treatments, and I hope that even highly experienced readers will find something of interest in them. Chapter 2 is specifically designed to permit readers with less mathematical experience to learn the theory behind signal processing, Fourier analysis, and digital implementation. A quick overview of phonetics is also provided, chiefly for readers from an engineering background.

Chapter 4 discusses mostly time-worn techniques in spectrum analysis involving the Fourier power spectrum (a pure frequency analysis) and short-time Fourier spectrogram (a time–frequency analysis). I have found that, in spite of the age and prevalence of these methods, there is no literature which coherently and systematically outlines the many parameters which govern the analyses generated. While I expect that the treatment in Chap. 4 will be of greatest value to beginning researchers and students, I again believe that even highly experienced readers will find some tidbits of useful information about these well-known methods.

Chapter 5 is the first which deals with "non-standard" methods of time–frequency analysis. The so-called *quadratic* or *bilinear* time–frequency representations are similar to a spectrogram but use different computational methods to obtain considerably different images; they have been introduced and studied in the signal processing literature since the 1980s. Applied researchers would benefit from understanding something about this area, but the literature tends to be inscrutable, offering little beyond theoretical equations and proofs of mathematical properties. There is no "how-to" literature. I digested some of the theoretical literature and scoured around for implementations by experts in the field; the fruits of my hunting are distilled and provided here tailored to the speech researcher. Some of these techniques have the potential to outperform

conventional spectrograms, or perhaps provide a complementary view of speech sounds, and I have chosen one such bilinear representation to demonstrate.

Chapter 6 introduces the method of *reassigning* a spectrogram, which yields a new sort of time–frequency image that has unparalleled precision in its representation of components and impulses. It achieves this in part by discarding information about the bandwidth of components, but much of this information in a spectrogram is tainted by the short-time Fourier procedure in any case. A discussion of the concept of the *instantaneous frequency* is included in Chap. 2, and it is especially relevant to reassigned spectrograms. Reassignment acts upon the information found in a spectrogram, changing the location of points in the time–frequency plane so that they line up more closely to the locations of the instantaneous frequencies of components, as well as to the precise time instants at which events have been recorded. It will be shown how the increased precision of the resulting time–frequency analysis of speech reveals a considerable amount of previously obscured information, leading to a variety of avenues for the investigation of hitherto unobservable properties of speech spectra.

Both the theory and practice of so-called *parametric* spectrum estimation is described in the final chapter, which focuses largely on the method of *linear prediction*. This class of techniques is a double-edged sword; it yields simplified spectral analyses which are easy to mine for measurements automatically, but there is no guarantee of accurate results. Once again, my goal is to enable readers to understand how the various inputs to a linear prediction analysis affect the spectrum estimate which results, and to therefore select the best possible procedure for their purposes. The chapter concludes by introducing a promising extension of the linear prediction framework known as *autoregressive moving average* modeling.

While I have tried to discuss all methods of speech spectrum analysis which I feel deserve detailed understanding by applied researchers in many fields, this book was never intended to be completely comprehensive like the typical book on "speech signal processing." At least two major topics related to our subject have not been treated at all here, namely cepstral processing and automated tracking of either formants or pitch. The former of these, while widely used in speech engineering, has little to offer the practicing speech scientist in my opinion. Cepstral smoothing of a speech spectrum cannot show any advantages over the other methods presented, so it seemed that including a chapter on it would chiefly serve tradition rather than expedience, and would risk bloating a fairly concise book. On the other hand, a discussion of tracking algorithms has been left out because these are essentially heuristic ad hoc methods for extracting maximally consistent and reasonable results from spectral analyses. Tracking is a post hoc procedure that has no theory governing it, and which relies for its effectiveness on a high-quality spectrum analysis in the first place.

In writing this book I have tried to straddle a deadly fence, which has tripped up many who have gone before. The presentation is intended to be especially useful for applied speech scientists and linguistic phoneticians, but it is also more mathematical than probably any other book with this audience in view. On the

other hand, while it is essential to include a few of the most important equations in the discussion, most things are explained in English, and I have also tried to segregate the "heavier" mathematics that can be safely ignored into areas with a gray background which I refer to as "math boxes." In this way, I have made an attempt at providing "everything a speech scientist wanted to know about signal processing, but was afraid to ask." At the same time, I have also tried to make the book useful for engineers and scientists from a richer mathematical background. I believe it can serve this second purpose because all the relevant mathematics is presented in some way, each chapter has an extensive bibliography of primary sources, and there are many discussions of methodology specific to speech analysis which are not found in other literature.

In order to make the book as practical and inviting as possible, Matlab code has been written for it which implements nearly all procedures discussed, and particularly all that cannot be performed using Praat analysis software [1]. Chapters 4, 5, 6, 7 which deal with practical methodology each conclude with an appendix detailing how Praat can be used and manipulated to perform the procedures when possible, and also detailing how the accompanying Matlab code should be used.

References

1. P. Boersma, D. Weenink, Praat: doing phonetics by computer. Computer software (2009)
2. I.B. Crandall, C.F. Sacia, A dynamical study of the vowel sounds. Bell Syst. Tech. J. **III**, 232–237 (1924)
3. D. Lewis, Vocal resonance. J. Acoust. Soc. Am. **8**, 91–99 (1936)
4. E.W. Scripture, *The Elements of Experimental Phonetics* (Charles Scribner's Sons, New York, 1902)

Chapter 2
Phonetics and Signal Processing

2.1 Essentials of Phonetics

As this book is only about speech spectrum analysis, it is not intended to give a serious overview of phonetics or speech science. Most readers are probably acquainted with this subject, and so the purpose of this section is merely to present the essential concepts of phonetics as I interpret them. This will allow the definition of all the phonetic terminology to be employed throughout the book, rather than trying to rely on other authors' definitions found throughout the literature. After all, even well-informed readers may not have read all the same books that I have, or believed all the same controversial phonetic theories, so the goal of this section is to put everyone on the same page, as it were.

When the need arises, speech examples throughout the book may be presented in the standard transcription of the International Phonetic Alphabet, although knowledge of this system will rarely be critical for understanding the text. Readers who desire a reference on the IPA are invited to either visit the internet address of the International Phonetic Association, or consult a recent phonetics textbook (e.g. [26]).

2.1.1 Speech Production Fundamentals

2.1.1.1 Phonation

The process of human speech production relies foremost on breathing out. The lungs expel air during speech at a carefully controlled rate (often called "speech breathing" [47]). The air passes through the larynx, which contains the *vocal folds* (often called "vocal cords," with the term *glottis* technically referring to the space between them), whose positioning can be finely tuned by a panoply of laryngeal muscles [30]. When the vocal folds are adducted (closed) somewhat gently, then a

S. A. Fulop, *Speech Spectrum Analysis*, Signals and Communication Technology,
DOI: 10.1007/978-3-642-17478-0_2,

certain amount of air pressure from the lungs below (subglottal pressure) can set the folds into a self-sustaining oscillation. This condition is called *phonation* or *voicing*, and it involves a delicate balance of tissue coupling forces and aerodynamic forces whose overall description is still a current research topic; a relatively recent state of our understanding has been called the "myomucoviscoelastic-aerodynamic theory" of phonation [30]. The length of time taken by one complete phonation cycle we may call the fundamental period of phonation; the reciprocal of this period generally determines the perceived pitch of the voice, and is called the *fundamental frequency* of phonation.

The laryngeal muscles are able to position the vocal folds in such a wide variety of postures that a number of distinct *phonation types* are important in speech, and are indeed more important in some languages than in others. The most common type of phonation generally involves a vocal fold oscillation from a completely closed position (called the *closed phase* of phonation) to a more open state (called the *open phase*) and back again, and is usually termed *modal* phonation. It is also possible to maintain phonation with a larger adductive force acting to close the glottis. At its most extreme, such phonation is called *creaky*, and is marked by a very low fundamental frequency and a very small amount of air flow, both of which result from an extremely long closed phase within the phonation period. Phonation with considerably higher frequency and airflow than in creaky voice, but still with greater adductive force than modal phonation, has often been termed *stiff* phonation [25].

At the other end of this "scale" of phonation types, one may point to many instances of phonation which do not in fact exhibit any closed phase. Such phonation is termed *breathy*, and is caused by there being only a very small adductive force acting to close the vocal folds. In such phonation, the vocal folds vibrate more in the fashion of the double reed in a woodwind instrument such as the oboe. As one might imagine, such phonation is also characterized by a very large amount of air flow, which often creates additional aeroacoustic noise. One may also point to instances of phonation which involve considerably less adductive force than is typical of modal phonation, but which do not yet allow the vocal folds to remain open throughout the phonation period. Such phonation has often been termed *slack* phonation [25], and is usually characterized by a long open phase within the phonation period, together with a relatively lower fundamental frequency. In the study of languages which involve an important opposition between stiff and slack or breathy phonation, these states of the larynx have very often been called *registers*.

It is also possible to position the vocal folds so that they are closed through approximately 60% of their length anteriorly, while being abducted at the posterior points where they attach to the arytenoid cartilages, thus producing a small triangular opening. It is this posture which is naturally used in whispering, and so it is called *whisper* when there is no phonation. It is also possible, however, for the anterior portion of the vocal folds to undergo a more or less modal phonation in the posture of whisper, and the superposition of this phonation upon the aeroacoustic noise of the whisper is usually called *whispery* phonation [8, 27].

2.1.1.2 Sources and Filters

In modern phonetics and speech science, speech production is usually conceived in the "source-filter" paradigm that grew out of early twentieth century acoustic speech studies. The paradigm is commonly credited to the work of Fant [11], but the idea was essentially introduced to the phonetics community by Joos [22], and had earlier precursors [28]. A speech sound can generally be described as created from a sound source whose output is modified by the vocal tract, viewed as a resonating chamber, or resonant filter (see the section below on filters). Vowels and many other voiced sounds chiefly rely on the vocal cords as a source, and the phonation output is then filtered through the prominent vocal tract resonances called *formants*. Many consonant sounds have their main source at some place of articulation in the vocal tract, where the release or passage of air past that point generates an aeroacoustic noise that may subsequently be filtered by the vocal tract. The filtering will be neglible in the case of consonants whose primary sound source involves articulation with the lips, since there is no need for the sound to pass through the vocal tract before it emerges.

The source-filter model of speech has been very fruitful, and facilitated a greater understanding of how vocal tract resonance shapes speech sounds such as vowels. On the other hand, speech in fact involves complex and dynamic aeroacoustic sound production in which the flowing air is very important. The source-filter theory effectively ignores any specific effects from the airflow, and so it is not a perfect model of speech production by any means. In later chapters, we will have several opportunities to test the limits of the source-filter model, and to present findings that can lead to a more accurate understanding of speech production.

2.1.2 Syllables and Speech Sounds

Although fluent speech clearly presents a continuous stream of sound, it has often been said that the smallest "units" of speech are *syllables*. That is to say, syllables are postulated as the smallest pieces of speech which are relevant for the human prodution and perception mechanisms [47, 23]. To quote a simple truism from Ladefoged, "nobody, not even a baby, can utter anything less than a syllable; he certainly cannot make a [p] or a [b] by itself" [23]. I present this view here because I currently subscribe to it, but it is by no means universally held. Fortunately, for the purposes of this book, the absolute truth of this idea is not particularly important, and we can safely adopt it as our methodological viewpoint.

Whatever one thinks about the syllable, it is clear that the notion has never been clearly and completely defined—a fact that will not be rectified here. A syllable can be related to speech production by noting that speech inevitably involves moving the jaw, tongue, and lips from more open to more closed postures. A syllable is then approximately a single speech gesture of this kind, a movement of or within the mouth from a more closed to a more open posture (and often includes a subsequent return to a more closed posture).

Of course, even upon beginning to define a syllable, one must immediately note that syllables can potentially be further subdivided. This is apparent from a variety of simple facts, such as the English speaker's ability to distinguish the syllables constituting the words *bee* and *pea*, which differ only in their intial speech gestures. A speech gesture, i.e. a component of a syllable, which can be contrasted or swapped in this way with other speech gestures at particular linear positions in the speech output is what is commonly called a *segment*. Other roughly equivalent terms are "speech sound" and "phone." I will at no point (except in this very sentence) make use of the theoretically loaded term "phoneme," since there is no need to engage that debate in this book.

Linguists commonly consider the "structure" of syllables, viewing them as containers for segments. The central position of a syllable, containing the main segment that gives a speaker or listener the sensation that there is a syllable, is known as the *nucleus*. Any segments preceding the nucleus are together called the *onset*, while segments following the nucleus are called the *coda*. For example, the English word *pea* includes an onset as well as the nucleus (which is obligatory for a syllable), while the word *please* includes a (somewhat larger) onset as well as a coda in its single syllable.

Unfortunately, a proper definition of a segment is as elusive as that of a syllable. It is common to assert that a segment is some kind of subunit of syllables, but how big is it, and which gestures of speech production are included in it? For example, standard treatments of English phonetics would transcribe the one-syllable word *plea* as either [p^hli] or [p̥li] using three segments. The notation [p^h] indicates a single segment whose final phase involves a sort of "h"-sound called *aspiration* in the context of a stop consonant. In this sense the segment is "complex." The alternative includes a symbol [l̥] for a "devoiced" 'l'-sound. Is the voiceless portion part of the 'p', or part of the 'l'? For a further example, in the Southeast Asian language Hmong, a very similar syllable would be transcribed by linguists as [p^{lh}i] using just two segments, where the first "laterally released and aspirated" 'p' is more complex than any one of the English segments. There is, however, no way to demonstrate a purely phonetic difference between a complex segment and a sequence of simpler segments; the reason for prefering one analysis over another in a given case is a methodological one only.

2.1.3 Vowels and Consonants

Phonetics has traditionally classified the segments of speech into two basic varieties which are called *vowels* and *consonants*. Once again, there has never been a straightforward definition of these terms. Early linguists in India also grappled with the concepts of vowel, consonant, and syllable around 800 BC, and they recognized that the three notions are hopelessly intertwined [1]. The definitions used here will be similar to those of the ancient Sanskrit scholars, and in fact, the development of modern phonetics in the West owes much to the transmission of knowledge in translation from the Sanskrit sources.

A *vowel* is defined as a "vowel-like segment" (what Pike [32] termed a *vocoid*) that occupies the nucleus of a syllable. A segment is considered to be a vocoid when its articulation permits the relatively free passage of air through the center of the mouth. This definition is also rather loose, but in roughly familiar terms, most segments that are at least as open as an English *w* or *y*-sound (the latter is transcribed [j] in IPA) are vocoids, all others being non-vocoids. A *consonant* is then defined simply as a non-vocoid, no matter what syllable position it occupies. This imperfect dichotomy leaves room for a middle category, that of the *semivowel*, which is defined as a vocoid located outside the nucleus of a syllable. Semivowels, in spite of being vocoids, are usually regarded as a special sort of consonant (often called a "glide") in the interests of preserving the consonant–vowel dichotomy. The interplay of consonants, vowels, and syllables in the speech stream is given a slightly different (more acoustic) view by Orlikoff and Kahane:

> Consonants differ from vowels primarily by the amount of vocal tract constriction employed in their production … Speech can be considered to be an overlay of consonants on the vocal signal. The dispersion of consonants results in an amplitude modulation of the acoustic energy that, for the most part, gives rise to our perception of syllables. [30] p. 158

As so many of the speech examples shown in this book are drawn from American or Canadian English, it will be illustrative to give an inventory of the segments found in most major "standard" English dialects of North America, to provide a specific example while the classification of speech sounds is described. Table 2.1 lists the vowels and shows a one-syllable word using each. The one vowel not found in the list is the "schwa" transcribed [ə], as this vowel is found only in unstressed syllables of English (e.g. *about*; [əbaʊt]) and so is not normally present in isolated one-syllable words.

Vowels are traditionally classified using a number of phonetic features which have more recently been determined to have a largely auditory basis. The feature of *height* (see Table 2.2) is chiefly measured by the frequency of the lowest

Table 2.1 The vowels of North American English

Example word	IPA symbol
Beat	[i]
Bit	[ɪ]
Bait	[eɪ]
Bet	[ɛ]
Bat	[æ]
Bought	[ɔə] or [ɑ]
Bot	[a] or [ɑ]
Boat	[oʊ]
Book	[ʊ]
Boot	[u]
But	[ʌ]
Bird	[ɝ]
Bite	[aɪ]
Bout	[aʊ]
Boy	[ɔɪ]

Table 2.2 Vowels of English in the traditional chart

i				u	*high*
	ɪ		ʊ		
	eɪ	ə		oʊ	
		ɛ	ʌ	ɔə	
		æ	a	ɑ	*low*
front				*back*	

Table 2.3 The stops of American English

Example word	IPA symbol
Purr	[p^h]
Rubber	[b]
Upper	[p]
Too	[t^h]
Stew	[t]
Patty/paddy	[ɾ]
Redo	[d]
Cool	[k^h]
School	[k]
Lagoon	[g]

characteristic resonance of the vowel, known as F_1, the first formant frequency. The feature of *front-back* or "backness" is chiefly measured from the frequency of F_2, the second formant. At this point, the discussion gets tricky, since F_2 is also the chief auditory determinant of the degree of *lip rounding* in a vowel, which is the last of the three major vowel features. Many authors also suggest that the most important backness metric is in fact the difference $F_2 - F_1$, but in any case the second formant is crucial. The role of F_3 is not fully understood in this connection, but it is known to play at most a secondary role except in determining the *rhoticity* or "r"-ness of a vowel as in *bird*. Because the measurement of formants is so important for characterizing vowels, we will find in later chapters that formant measurement is, and indeed has always been, one of the chief reasons for undertaking speech spectrum analysis.

The consonants of a language are traditionally classified into a number of different varieties using articulatory criteria—these are the *manners of articulation*. Table 2.3 shows the *stops* found in English, which are defined as those consonants whose oral airflow is completely occluded for a brief period (and which are not nasals). A number of these segments are restricted to occur in certain environments, as is reflected in the choice of exemplifying words; for example, the voiced stops [b, d, g] normally do not initiate a word (although they may). The uniquely brief *tap* sound [ɾ] is included with the other stops because of its overall similarity to them.

The *fricatives* are defined as those consonants whose oral airflow passes through a constriction which is sufficiently narrow to yield an aeroacoustic noise

Table 2.4 The fricatives of American English

Example word	IPA symbol
Fat	[f]
Vat	[v]
Thin	[θ]
This	[ð]
Sip	[s]
Zip	[z]
Pressure	[ʃ]
Pleasure	[ʒ]
hot	[h]

source. Table 2.4 gives the nine fricatives of English together with exemplifying words. It is useful to note that most treatments of English phonetics also recognize two special complex single segments [ʧ, ʤ], found at the beginning and end of the respective words *church, judge*. These combinations of a stop and a fricative are known as "affricates," but in truth there is no special phonetic evidence to determine that these sound sequences are indeed single segments. It should also be pointed out that the example words in Table 2.4 illustrating the fricatives [ʃ, ʒ] show them in the word-medial position.

Having come this far in the description of consonants, it must be pointed out that, in addition to the place and manner of articulation, consonants are usually identified as either *voiced* or *voiceless*. A voiceless consonant is one during which the vocal cords do not phonate, while a voiced consonant is one during which phonation does take place. We often encounter pairs of consonants in a language which are distinguished only by this means, such as [f, v].

The other manners of articulation have smaller inventories in English, and so do not require tabulation. The *nasals* [m, n, ŋ] are defined as consonants whose airflow is completely diverted through the nasal sinus. The words *meat, neat, king* are sufficient to exemplify. It is noteworthy that the third one of these cannot initiate an English word. An *approximant* is defined as a consonant whose articulatory constriction is slightly more open than that of a fricative, to the point that aeroacoustic noise is not produced. The only examples of such a segment in English are provided by the voiced [l], as in *less*, or its voiceless counterpart [l̥] exemplified above in *please*. The remaining segments of English are the semivowels [j, w, ɹ], as in *yet, wet, right*.[1]

[1] It might be noted that most treatments (including the IPA itself) classify English [ɹ] of *right* as an approximant rather than a semivowel, but I believe this is not consistent with the definitions of the terms, and is mostly done because of tradition. I have to thank my colleague Chris Golston for convincing me that [ɹ] is a semivowel.

2.1.4 Uses of Vocal Pitch

As nearly all words in the languages of the world include a voiced vowel or approximant (though there are noteworthy exceptions [20]), a word must be spoken with a particular choice of vocal cord vibrational frequency. The fundamental frequency of phonation is usually named by the perceptual term *pitch* in phonetics, since the scale of perceived pitch is very nearly equal to the physical frequency scale in the range of normal speech (50–350 Hz). There are, however, sometimes reasons for saying that the perceived voice pitch is not equal to the fundamental frequency, usually because there is some kind of "abnormal" phonation process that makes the pitch ambiguous. In any case, it is clear that the voice pitch is very important to the linguistic phonetic pronunciation of language, acting to manifest many *prosodic* "above the segment" aspects of the speech.

The overall pitch melody of an utterance is called its *intonation*. Different intonations are frequently used for the purpose of questions, emphasis, and many forms of speaker attitude expression. On a different note, many languages require words to have specified pitch melodies applied to each syllable; this phenomenon is called *tone*, with such languages being called tone languages. Finally, pitch is of considerable importance in other prosodic aspects of language besides tone; English syllable stress provides but one example, where we normally find that stressed syllables differ substantially in pitch from neighboring syllables.

2.2 Essentials of Digital Signal Processing

The term *signal* is used to mean any quantity y that varies over time. In general, a signal could be multidimensional, but in a book about speech sound our concern is entirely with the acoustic output of the voice. This is usually measured as the varying voltage output from a microphone responding to the acoustic pressure wave of the voice, a single scalar quantity which varies as a function of time.

Signals can be classified in a number of fundamental ways. A natural signal is generally *continuous-time* (or simply *continuous*), meaning that $y(t)$ has values for every real number time value in the time interval through which the signal exists. This is true of the microphone voltage signal, for example. It is possible to use an analog device to perform analog signal processing on such a signal, but these approaches will only be considered in the historical review chapter. Normally the first thing that happens in modern signal processing is the *sampling* of a continuous signal at regular time intervals to yield a *discrete* or *digital* version of the signal that is also discrete in the y-axis showing the signal (voltage) values. In this case, the signal is represented as a sequence $s(n)$ of the sample values for each whole number time sample point n, which in Matlab software is a one-dimensional *matrix* or *vector*.

Another important dichotomy is the distinction between *deterministic* and *stochastic* or *random* signals. A random signal is formally a single realization of a

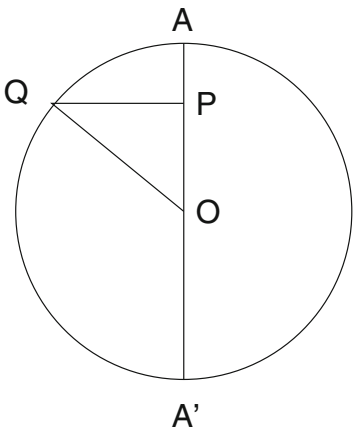
stochastic process that generates it; each time the signal is realized, it is somewhat different owing to the randomness of the process. A deterministic signal can then simply be defined as resulting from a process that is not at all random. As the reader can already imagine, there is often sufficient randomness in human speech production to allow the treatment of speech signals as random. On the other hand, speech is never totally random. In practice, speech signal analysis treats speech as random or deterministic, whichever is most convenient for a given approach. This chapter will focus almost entirely on defining concepts for deterministic signals; spectrum analysis procedures which treat speech as a random signal will be dealt with in other chapters.

2.2.1 Periodic and Aperiodic Signals

A *periodic* signal is one that repeats a pattern at regular time intervals. The time taken for one repetition or cycle of the pattern is called the *period*, and the reciprocal of the period is called the *fundamental frequency* of the signal. With time measured in seconds, the unit of frequency then becomes the "cycle per second" or Hertz, written Hz. A signal that does not have a recurrent period is an *aperiodic* signal. We will see that even this sort of signal can be represented as a sum or integral of periodic functions, however.

Periodic signals are of great importance, not least because the most fundamental kind of signal is periodic, and is described by either of the fundamental functions of trigonometry.

> When a point Q moves uniformly in a circle, the perpendicular QP drawn from its position at any instant to a fixed diameter AA' of the circle, intersects the diameter in a point P, whose position changes by a simple harmonic motion. ([36] p. 38)



Then, thanks to trigonometry, the displacement s of such a point P performing a simple harmonic motion back and forth across a zero position is a trigonometric function of time:

$$s(t) = A\cos(2\pi ft + \phi), \tag{2.1}$$

in which A is the maximum displacement (amplitude), f is the fundamental frequency, and ϕ is a time offset called the *phase* which is included for generality (so it can be specified where we start the motion at $t = 0$). The angle value inside parentheses is always expressed in *radians* rather than degrees; recall that a complete circle subtends an angle of 2π radians. Those familiar with signal processing will be well aware of its many conventions, including the convention of using the *angular frequency* ω (radians/s) to stand for $2\pi f$.

Let us remind ourselves of Euler's relations between sinusoidal and complex exponential functions involving exponents of the number e, also written exp[·]:

$$e^{i\omega t} = \cos(2\pi ft) + i\sin(2\pi ft) \tag{2.2}$$

$$A\cos(2\pi ft) = \frac{A}{2}e^{i\omega t} + \frac{A}{2}e^{-i\omega t} \tag{2.3}$$

$$A\sin(2\pi ft) = \frac{A}{2i}e^{i\omega t} - \frac{A}{2i}e^{-i\omega t}. \tag{2.4}$$

It will usually prove convenient to use the complex exponential functions instead of sin and cos, since the mathematics of spectrum analysis is much easier to deal with by involving complex numbers.

In the discrete-time regime, a sinusoid is expressed as a function of a set of integer time indices marking points at which the signal samples occur; the variable ranging over the time indices (or *points*) is normally symbolized as n.

$$s(n) = A\cos(2\pi fn + \phi), \tag{2.5}$$

where now f is a frequency in *cycles per sample* rather than Hz. In fact, the highest frequency that can exist in the discrete-time regime is $1/2$ cycle per sample. Let us note that a periodic discrete-time signal with period N meets the condition that $s(n + N) = s(n)$. The total energy in a discrete-time signal is defined as the sum of the squared absolute signal values:

$$E \stackrel{\text{def}}{=} \sum_{n=-\infty}^{\infty} |\, s(n)\,|^2 \tag{2.6}$$

Any signal whose total energy is finite is usually called an *energy signal*. A periodic signal represented for the entire time domain $(-\infty, \infty)$ then has to have infinite total energy.

2.2.2 Sampling of Analog Signals

In order to perform digital signal processing on a real signal it has to be sampled and quantized. This is often called *analog-to-digital conversion*. Sampling is the process by which values of the signal (what used to be called ordinates) are recorded at equally spaced time intervals separated by the sampling period. The reciprocal of the sampling period is the *sampling frequency* or *sampling rate*.

If the signal values could be recorded with arbitrary precision when sampling was performed, then one would have a discrete-time signal with continuous ordinates. This type of signal is theoretically of importance, but in practice the signal values are also assigned to a discrete set of values. This was true even when sampling was (a long time ago) performed by hand from some kind of analog record like an oscillogram. When a digital computer is used, the signal values are assigned to a value within a discrete set by a process known as *quantizing*. Such a fully discrete signal is also called a *digital* signal.

When a digital signal results from sampling an analog signal at a frequency f_s (e.g. 44.1 kHz in the case of standard digital audio), the highest frequency that can be represented in the digital signal is the *Nyquist frequency* $f_{\max} = f_s/2$, or 22.05 kHz in this case. This is because the highest frequency for a digital signal must be $1/2$ cycle per sample. For speech, a very useful sampling rate available on most soundcards is 22,050 Hz, since this preserves the sound frequencies up to 11,025 Hz, and very little important energy is present in speech at higher frequencies than this.

As a result of the above "Nyquist theorem" governing sampling rates and the frequencies of digital signals, it is critical to somehow filter or cut off the frequency range of an analog signal when it is sampled, so that it is *bandlimited* below $f_{\max}$. In modern practice, the hardware A-to-D converters found in computer sound cards and digital recorders do all the hard work of bandlimiting, sampling and quantizing, so all the user generally needs to worry about is the sampling rate. In modern consumer-grade equipment, the quantizing is not user-adjustable, but it is fixed to such a high fidelity that almost no distortion will be introduced that way. As a result, no one besides a hardware engineer needs to be concerned with the particulars of amplitude quantizing; the typical computer sound card digitizes a sound using 65,536 (2^{16}) discrete amplitude steps, and professional hardware often provides more than this.

2.2.3 Autocorrelation

A useful quantity that is computable from a signal is its *autocorrelation*, which is, roughly speaking, a function obtained by multiplying the signal $s(n)$ by a time-shifted copy of itself $s(n - \ell)$; the time shift ℓ is standardly called the *lag*, and the autocorrelation is really a function of the lag. The subject of autocorrelation brings

us to a terminology problem (highlighted in [3]), because it is one of the subjects where the signal processing literature intersects with literature on statistics of sampled signals (usually called *time series*), and these two disciplines have distinct terminological fashions and traditions. In the time series literature, what signal engineers call the autocorrelation is instead called the *autocovariance*, and then the autocorrelation is usually defined as the autocovariance normalized by (i.e. divided by) the particular autocovariance value for zero lag—which in turn is just the statistical variance of the signal.

In this book I will feel free to follow signal processing authors in using the term "autocorrelation" to refer to what is strictly a form of the autocovariance, since as a practical matter it is not too important which of the various defining equations is used (see box for details). This is because we do not usually care about the actual value of the autocorrelation function, we chiefly care about the lag times where it has large peaks. Any of these quantities allows us to probe the same properties of a signal; in particular these quantities are highly sensitive to periodicity which may be present in an otherwise adulterated or noisy signal. A lag time showing a large peak in the autocorrelation function is indicative of a period in the signal equal to that lag. Intuitively, the autocorrelation takes a large value whenever the time-shifted copy of the signal matches the original well—the signal is highly correlated with itself at those lags.

In literature on time series statistics such as [33], there are two common estimates (from a finite sample) of the theoretical autocovariance of a random process, expressed as a function of lag time ℓ. The first is:

$$R_{ss}(\ell) \stackrel{\text{def}}{=} \frac{1}{N-\ell} \sum_{n=\ell}^{N-1} (s(n) - \bar{s})(s(n-\ell) - \bar{s}) \tag{2.7}$$

in which $\bar{s}$ is the sample mean of the signal, used as an estimate of the overall process mean. The above "sample autocovariance" is asymptotically unbiased.

An alternative definition is given by:

$$R_{ss}(\ell) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=\ell}^{N-1} (s(n) - \bar{s})(s(n-\ell) - \bar{s}). \tag{2.8}$$

This sample autocovariance is often referred to as the "periodogram-based estimate" (the meaning of this will be made clear in a later section). It is biased, but has smaller mean square error than the first alternative. It is this quantity that is computed by Matlab time series tools, as shown in Fig. 2.1. A related sample autocorrelation estimate can be obtained from either of the above, by simply dividing by the signal variance, which is the same thing as the autocovariance at $\ell = 0$.

In the signal processing literature on the other hand, we must be aware of a number of differences in the definitions [39]. First of all, what is above called the autocovariance is generally called the autocorrelation. Moreover, the signal mean is usually not subtracted, and further there is usually no factor of N or $N-\ell$ put into the denominator. Finally, the resulting quantity is applied willy–nilly to random signals or deterministic ones. The "autocorrelation" of an energy signal $s(n)$ is nowadays standardly defined in the signal processing literature as [34]:

$$r_{ss}(\ell) \stackrel{\text{def}}{=} \sum_{n=-\infty}^{\infty} s(n)s(n-\ell) \tag{2.9}$$

For a finite-duration signal consisting of N samples, the definition becomes:

$$r_{ss}(\ell) \stackrel{\text{def}}{=} \sum_{n=j}^{N-|k|-1} s(n)s(n-\ell), \tag{2.10}$$

in which for non-negative lags one sets $j=\ell, k=0$, and for negative lags one sets $j=0, k=\ell$. In turns out that this function is symmetric about zero, so it is sufficient to compute it only for non-negative lags.

Example 2.1 Figure 2.1 shows a 45 ms snippet of the English vowel in the word "how'd." The obvious periodicity results from the vocal cord vibration. Also shown in the figure is a plot of the autocorrelation, also known as the *correlogram*, produced in Matlab. This example shows how the correlogram can be used to find the main period in a periodic signal—it is equal to the first positive lag where there is a high autocorrelation value. The highest positive peak occurs at a lag of 417 samples, which

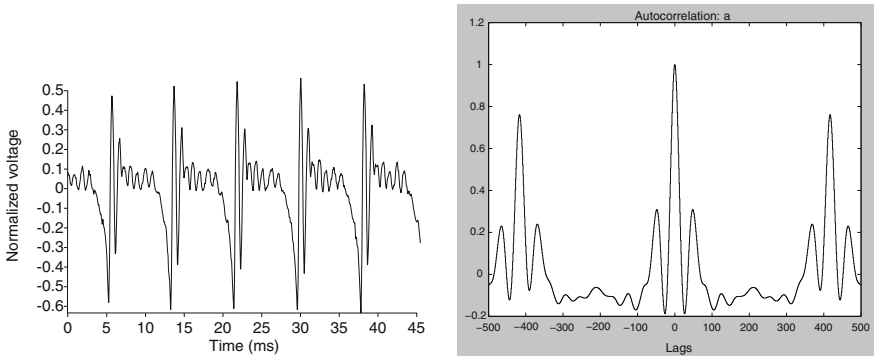


Fig. 2.1 *Left panel* a brief portion of a natural vowel, the first part of the English diphthong [aʊ] in "how'd." *Right panel* a plot of the autocorrelation of the signal at left, produced by Matlab time series tools. The lag time is shown in units of samples. Note how the function is symmetric about zero

equals 8.14 ms for the signal's sampling rate (51.2 kHz). This is the fundamental period of the vocal cord vibration, on average, during the 45 ms displayed.

2.2.4 Fourier's Series and Transform Spectra

Drawing on the work of predecessors Leonhard Euler and Daniel Bernoulli the elder, Jean–Baptiste Joseph Fourier devised the first proof [12] that a periodic function can be represented as a sum of elementary periodic functions (i.e. sinusoidal or their equivalent exponential functions). In this section I will present the fundamentals of Fourier's results, which are at the heart of spectrum analysis. The treatment is drawn chiefly from [33] and [34].

2.2.4.1 Fourier's Series Spectrum

Consider a periodic signal $s(t)$ in continuous time, having period T, over the entire time domain $-\infty < t < \infty$. Fourier's series representation of the signal is then given by the following formula:

$$s(t) \stackrel{\text{def}}{=} \sum_{k=-\infty}^{\infty} c_k \exp[2\pi \mathrm{i} k f_0 t] \tag{2.11}$$

in which $f_0 = 1/T$ is the fundamental frequency of the periodic signal $s(t)$. The content of this historic equation is that a periodic function (one which also meets certain broad conditions) can be expressed as a sum of exponential (i.e. sine and cosine) "component" functions, and each of these has a frequency which is an integer multiple of the fundamental frequency. Components with such frequencies are called "harmonics" of the function.

This formula is only valid for representing a signal with infinite support in the time domain, and thus Fourier's series representation is strictly a theoretical object only. The values $\{c_k\}$ for all integers k, called the *Fourier series coefficients*, represent both the amplitudes and phases of the harmonics; as such they are in general complex numbers, and are specified by the following integral formula:

$$c_k = f_0 \int_T s(t) \exp[-2\pi \mathrm{i} k f_0 t] \mathrm{d}t \tag{2.12}$$

where the interval of integration covers the signal over one period T exactly. One useful fact is that when the signal is real, then the pair of coefficients c_k and c_{-k} are complex conjugates[2] for all k; as a result, for real signals it is sufficient to compute

[2] Recall that for complex number $a + b\mathrm{i}$, its conjugate is $a - b\mathrm{i}$. For complex number $A\mathrm{e}^{\mathrm{i}\theta}$ expressed in polar form, its conjugate is $A\mathrm{e}^{-\mathrm{i}\theta}$.

the Fourier series coefficients for non-negative integers k, meaning we do not have to bother with the negative frequencies for physical signals.

The total energy in an infinite-time periodic signal $s(t)$ is also infinite; the energy over a span of one period $(-T/2, T/2)$, however, is given by the integral expression:

$$\int_{-T/2}^{T/2} s^2(t)\mathrm{d}t,$$

and a consequence of "Parseval's relation" (another historic result from Fourier's era) is that

$$\int_{-T/2}^{T/2} s^2(t)\mathrm{d}t = T\sum_{k=0}^{\infty} |\ c_k\ |^2\ . \tag{2.13}$$

The above is an extremely important equation at the foundation of spectrum analysis; it shows that *all the information* about energy in a signal is present in the (squared) magnitudes of the Fourier series coefficients.

The energy in a certain number of periods is usually thought of in terms of energy per unit time, or *power*. From the above Eq. 2.13, dividing by the time span T yields the total power in the signal as $\sum_{k=0}^{\infty} |c_k|^2$. Then each quantity $|c_k|^2$ (which is now real) measures the contribution to the total power from the particular term in Fourier's series for $s(t)$ with frequency k/T Hz. If we plot the y-axis values $|c_k|^2$ against the x-axis values k/T, we obtain a *discrete power spectrum* describing how the power is distributed over the harmonic components of $s(t)$. This kind of spectrum is sometimes called a power density spectrum, but this terminology is inaccurate because the spectrum is discrete, and so is formally not a density function.

In graphing the spectra of acoustic signals, it is typical to show the energy/power on a logarithmic axis that is defined using the *decibel*, a relative unit of power that is formally dimensionless. Representing the squared signal amplitude by $|c_k|^2$ for a single Fourier series component, which would have units of squared normalized voltage in the case of a signal from a microphone, the decibel value of this amplitude is calculated relative to a reference level p_{ref}^2 using the following definition:

$$A_{\text{dB}} \stackrel{\text{def}}{=} 10\log_{10}\frac{|c_k|^2}{p_{\text{ref}}^2}. \tag{2.14}$$

When graphing a spectrum in practice, the reference level is somewhat arbitrary, and quite irrelevant when one considers that negative dB values can just as easily be shown. This means that the reference level in practice serves to determine the position of the 0 dB point on the magnitude axis, and does not at all affect the overall shape of the spectrum graph, or the relative dB amplitudes of the signal components.

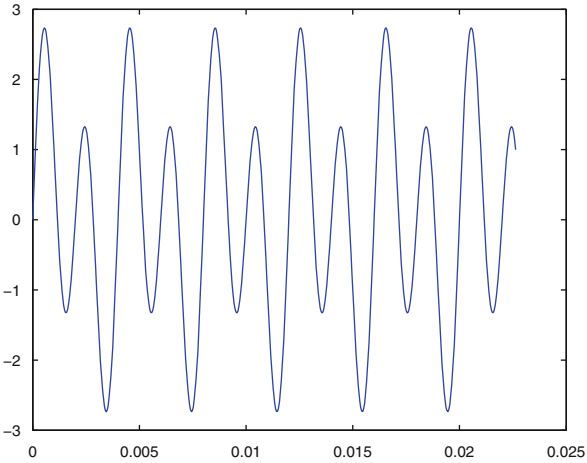


Fig. 2.2 Portion of a function (supposed to exist across the entire real line) composed of sine waves at 250 and 500 Hz frequencies

Once again, let me emphasize that an infinite-time signal and its discrete spectrum are purely theoretical objects. No physical signal has a precisely discrete Fourier series spectrum. It is therefore not possible to illustrate the foregoing considerations with any real physical example or numerical computation. Let us instead concoct a simple theoretical example, in the form of an infinite-time real signal composed of two sinusoids, one at the fundamental frequency $f = 250$ Hz and another of twice the amplitude at the next harmonic frequency $2f = 500$ Hz. A snippet of this function is shown in Fig. 2.2.

Euler's relation tells us that such a signal comprises four exponential components in two complex-conjugate pairs:

$$\sin(2\pi ft) + 2\sin(4\pi ft) = \frac{1}{2\mathrm{i}}\mathrm{e}^{2\pi \mathrm{i} ft} + \frac{1}{2\mathrm{i}}\mathrm{e}^{-2\pi \mathrm{i} ft} + \mathrm{i}\mathrm{e}^{4\pi \mathrm{i} ft} + \mathrm{i}\mathrm{e}^{-4\pi \mathrm{i} ft} \tag{2.15}$$

$$c_1 = c_{-1} = 1/2\mathrm{i} \tag{2.16}$$

$$|c_1|^2 = |c_{-1}|^2 = 1/4 \tag{2.17}$$

$$c_2 = c_{-2} = \mathrm{i} \tag{2.18}$$

$$|c_2|^2 = |c_{-2}|^2 = 1 \tag{2.19}$$

The end result of the math in the box is to find that the fundamental frequency component, which has half the amplitude of the double frequency component in the signal, ends up contributing $1/4$ as much to the total power. On a decibel scale, we find that a component with half the amplitude is 6 dB less (v. Fig. 2.3). One also finds that using the complex exponential form of the Fourier series introduces

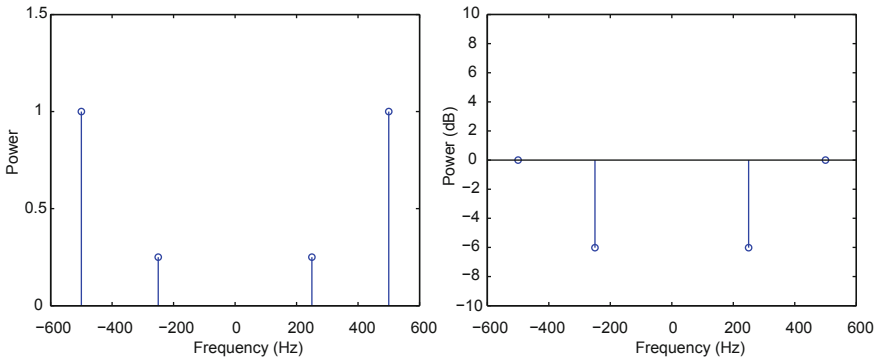


Fig. 2.3 Discrete power spectrum of the (complex) Fourier coefficients computed from the function shown in the previous figure. *Left panel* shows the raw power, *right panel* shows the components using a decibel scale, where the highest amplitude components have the reference level

a component with a negative frequency for every positive frequency component. Normally the negative components are ignored in applications, since they have no physical interpretation and are basically like mathematical "echoes" which result from working in the complex number plane.

2.2.4.2 Fourier's Series in Discrete Time

For a periodic signal in discrete-time, Fourier's series representation can have at most N terms, or frequency components, for a signal of period N (v. [34]). The first equation below represents a discrete-time signal as a Fourier series, while the second gives the formula for the Fourier coefficients in the series.

$$s(n) = \sum_{k=0}^{N-1} c_k \exp\left[\frac{2\pi \mathrm{i} kn}{N}\right] \tag{2.20}$$

$$c_k = \frac{1}{N} \sum_{n=0}^{N-1} s(n) \exp\left[\frac{-2\pi \mathrm{i} kn}{N}\right] \tag{2.21}$$

Owing to the complex conjugate pair $c_k^* = c_{-k}$ (asterisk indicates complex conjugation) for real signal $s(n)$, a further identity follows, namely $|c_k| = |c_{N-k}|$. The important result of this is that now the coefficients for $0 \leq k \leq N/2$ (for even period N) form a complete spectrum, since the "upper half" of the coefficient sequence simply repeats the same information. For odd period N, one uses the coefficients for $0 \leq k \leq \frac{N-1}{2}$. So, a Fourier spectrum of a discrete-time signal is invariably redundant, and only the lower half of the coefficients are needed to represent the complete information.

2.2.4.3 Fourier Transform and Spectrum

The Fourier integral representation of a signal is, in essence, the mathematical form taken by the Fourier series representation when the signal is aperiodic. Fourier showed this when he realized that an aperiodic function could be viewed as a periodic function taken to the limit of an infinite period (see box for a derivation).

Recalling Eqs. (2.11) and (2.12), one can represent a signal as its Fourier series in the following way:

$$s(t) = \frac{1}{T}\sum_{k=-\infty}^{\infty}\left(\int_{-T}^{T} s(x)\exp\left[\frac{-2\pi ikx}{T}\right]\mathrm{d}x\right)\exp\left[\frac{2\pi ikt}{T}\right] \tag{2.22}$$

Define $(2\pi k)/T$ as an angular harmonic frequency variable ω. In the limit as $T \to \infty$, $(2\pi)/T$ can be taken as an infinitesimal $d\omega$. This is about the point where mathematicians squirm because of the lack of rigor in making such a manoeuver, but this derivation is found in essentially this form in many textbooks, and pretty well represents Fourier's original derivation. It is possible to make it a more solid mathematical derivation. In any event, this "step" results in the following equation:

$$s(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} s(x)\mathrm{e}^{\mathrm{i}\omega x}\mathrm{d}x\right)\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}\omega \tag{2.23}$$

The above was derived by Fourier in 1811, and is now called the Fourier Integral Theorem. The inner integral defines what is now called the Fourier transform(ation) of the signal function, discussed below.

For a continuous function of time (i.e. a signal) $s(t)$, the Fourier transform (rewritten in the first equation below) defined by the Fourier Integral Theorem above is a function of frequency obtained from an infinite integral over the time domain, and is often taken to provide a mathematical definition of the pre-mathematical physical concept of "frequency."

$$S(f) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} s(t)\mathrm{e}^{-2\pi ift}\mathrm{d}t \tag{2.24}$$

$$s(t) = \int_{-\infty}^{\infty} S(f)\mathrm{e}^{2\pi ift}\mathrm{d}f \tag{2.25}$$

The second equation shows that the signal can in turn be expressed as a similar integral (over the frequency domain) of its own Fourier transform, which is then called the inverse Fourier transform. A function and its Fourier transform are

commonly called a *Fourier transform pair*, since they are in essence transforms of each other. The Fourier transform is often expressed as a function of angular frequency $\omega = 2\pi f$, but this induces slight changes in the form of the transform definitions (see box for details).

The Fourier transform and its inverse are commonly expressed using the angular frequency $\omega \stackrel{\text{def}}{=} 2\pi f$, for which there are two or three different conventions [44]. Here is one pair of equations of this sort:

$$S(\omega) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} s(t) \mathrm{e}^{-\mathrm{i}\omega t} \mathrm{d}t \tag{2.26}$$

$$s(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega) \mathrm{e}^{\mathrm{i}\omega t} \mathrm{d}\omega \tag{2.27}$$

In any such definition using ω in place of f for a frequency variable, one of the integrals (it makes no difference which) must be preceded by the factor $1/2\pi$ because of the substitution of angular frequency. A third way is to restore the symmetry of the transform pair by placing a factor of $1/\sqrt{2\pi}$ in front of both integrals.

An aperiodic signal has a well-defined Fourier transform only if it vanishes at infinity (unlike a periodic signal which is assumed to have infinite support), and therefore has a finite total energy. Just as with a periodic signal, the total energy in the continuous-time real signal $s(t)$ is $\int_{-\infty}^{\infty} s^2(t)\mathrm{d}t$, but now the analog of Parseval's relation leads to the following equation involving the integral transform in place of the Fourier series coefficients (cf. Eq. 2.13):

$$\int_{-\infty}^{\infty} s^2(t)\mathrm{d}t = \int_{-\infty}^{\infty} |S(f)|^2 \mathrm{d}f. \tag{2.28}$$

Although often credited to Parseval or Plancherel for the originating concepts, the above equation was derived in this particular form by Wiener [38]. The content of the equation is to say that, for aperiodic signals, the analog to the discrete spectrum of Fourier series coefficient magnitudes is now a continuous-frequency function, commonly called an *energy density spectrum*, which is determined by the squared magnitude of the Fourier transform. It is this equivalence which leads to the commonly expressed notion that the Fourier transform of a signal gives us its spectrum. As before, it is common to graph the spectrum as a function of frequency using a decibel-scaled axis for the magnitude.

The energy density spectrum is often called the *periodogram* in the context of random signals and other time series data. It can also be connected directly to the

autocovariance, since it is in fact the Fourier transform of the latter (e.g. [33]); the intrepid reader can try to verfy this by writing out the discrete Fourier transform (see below) of Eq. 2.8 (the periodogram-based estimate of the sample autocovariance).

Once again, let me emphasize that the preceding treatment relates a continuous-time signal $s(t)$ to its continuous-frequency Fourier transform $S(f)$, and so none of the above facts can be computed or illustrated numerically with a computer. In order to progress to the kind of Fourier spectrum which can be, and commonly is, computed, we must move from functions which are continuous in time and frequency to functions which are discrete in time and frequency.

2.2.4.4 Discrete-Time Fourier Transform

For a discrete-time function $s(n)$, the *discrete-time Fourier transform* is now defined as an infinite sum over the domain of time integer indices, instead of an integral over continuous time:

$$S(\omega) \stackrel{\text{def}}{=} \sum_{n=-\infty}^{\infty} s(n)\mathrm{e}^{-\mathrm{i}\omega n} \tag{2.29}$$

where now ω is an angular frequency value in radians/sample. If $s(n)$ is sampled from a continuous signal $s(t)$, the discrete-time Fourier transform does approximate the Fourier transform of $s(t)$.

The sampling causes the discrete-time Fourier transform to be periodic in the frequency domain, with period equal to the sampling frequency, which is 2π radians/sample [42]. The necessity of this periodicity in angular frequency derives from the periodicity of the exponential function, which is easily expressed [33]:

$$\exp[-\mathrm{i}(\omega + 2\pi k)n] = \exp[\mathrm{i}\omega n] \quad \text{for all integers } k. \tag{2.30}$$

The inverse discrete-time Fourier transform is now the following:

$$s(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\omega)\mathrm{e}^{\mathrm{i}\omega n}\mathrm{d}\omega \tag{2.31}$$

This is the same as in the continuous case, except that now the perodicity of the transform function in angular frequency allows the interval of integration to be limited to $(-\pi, \pi)$. Now that we have taken care of the time sampling, we can take the final step to define a Fourier transform which is discrete in the frequency domain as well.

2.2.4.5 Discrete Fourier Transform

In applications, one cannot perform a discrete-time Fourier transform literally because it would require calculating over the entire time domain. In practice, one

is limited to a sampled sequence having a finite length L, which is often called the window length. Then Eq. 2.29 for the discrete-time Fourier transform becomes:

$$S(\omega) = \sum_{n=0}^{L-1} s(n) \mathrm{e}^{-\mathrm{i}\omega n} \tag{2.32}$$

The above transform function is still impossible in digital computer applications because it is specified over a continuous range of frequencies. To represent the function digitally, we make it discrete-frequency by evaluating $S(\omega)$ at a finite set $\{\omega_k\}$ of N equally spaced frequencies across one interval of length 2π:

$$\{\omega_k\} \stackrel{\text{def}}{=} \frac{2\pi k}{N}, \quad k = 0, \ldots, N-1. \tag{2.33}$$

This makes the transform into the *discrete Fourier transform* (DFT), also called the *finite Fourier transform* [5], which is discrete in both time and frequency [42]:

$$S(\omega_k) \stackrel{\text{def}}{=} \sum_{n=0}^{L-1} s(n) \exp\left[\frac{-2\pi \mathrm{i} kn}{N}\right]. \tag{2.34}$$

In applications it is typical to make certain that the number of discrete frequency points N in the transform (also called the frame size) is greater than the number of signal samples L in the window, and when that is the case the transform takes the following standard form:

$$S(\omega_k) \stackrel{\text{def}}{=} \sum_{n=0}^{N-1} s(n) \exp\left[\frac{-2\pi \mathrm{i} kn}{N}\right]. \tag{2.35}$$

This form of the DFT no longer refers to the window length, but only the frame size, and so in order to ensure that the signal is not used beyond the intended window length, the signal value $s(n)$ is set equal to zero for all sample points $n \geq L$. This common procedure is called "zero-padding" beyond the analysis window to yield an analysis frame which is longer than the window but which does not include any extra information. By doing this, the frequency sampling resolution need not be small when we analyze a short window of the signal. The downside of this standard definition is that the frequency sampling resolution has to be large when we analyze a long window, which can be inconvenient because of computation time.

The DFT evaluates enough frequency components to allow its inverse transform to reconstruct the segment of the signal that was analyzed. It is sometimes interpreted as reconstructing one period of a periodic signal, assuming that exactly one period was analyzed in the first place. The following gives the inverse DFT formula:

$$s(n) = \frac{1}{N} \sum_{k=0}^{N-1} S(\omega_k) \exp\left[\frac{2\pi \mathrm{i} kn}{N}\right]. \tag{2.36}$$

The magnitude spectrum induced by the N-point DFT is just given by $|S(\omega_k)|^2$ analogous to the continuous Fourier transform, and is commonly graphed as a function of frequency with a decibel-scaled magnitude axis. Now, however, the spectrum is once again in a discrete form, which is usually taken to approximate the "ideal" continuous form. Each discrete frequency ω_k or f_k is referred to as a *frequency bin*. The reader should at this point note the similarity between the above Eq. 2.35 for the discrete Fourier transform and the earlier Eq. 2.21 for the discrete Fourier *series* coefficients. Modulo a multiplying factor, they are in fact the same. This shows that, for a digital approximation to a signal (resulting from time-sampling), the digital approximation to the Fourier transform (the DFT, which is also discrete in frequency) is essentially the digital version of the Fourier series [5]. Fourier's series and transform are thus unified in the digital realm.

Owing to the periodicity of the transform that was induced by the passage to discrete-time, the highest frequency that can be found in the DFT is equal to the signal's sampling frequency f_s. In addition, for a real signal the DFT exhibits a symmetry, so that $S(f_k) = S^*(f_{N-k}) \bmod N$ [41]. This means that the DFT spectrum is half-redundant, so that the upper half of the frequency bins have the same magnitude spectrum (in mirror image) as the lower half. We already obtained this result earlier, from deriving the Fourier series of a discrete-time (sampled) signal. In fact, then, the highest frequency that can be non-redundantly represented in a DFT spectrum is equal to the signal's Nyquist frequency $f_s/2$. It is a consequence of this redundancy that a 256-point DFT, for example, would yield a spectrum plot showing magnitudes for just the first 128 frequency bins. The other 128 simply repeat the information in reverse order, and so are not standardly shown.

2.2.4.6 Fast Fourier Transform

The above formula for computing the discrete Fourier transform has been known, in essence, since the 1870s [21], and it was also recognized at an early point that it is not a tractable formula—i.e. its computational complexity is too high. In particular, for a signal frame of N samples, the simple DFT formula requires N^2 multiplications of complex numbers and $N(N-1)$ complex addition operations [43].

Improving on the complexity of the basic DFT formula, a class of algorithms exists known as Fast Fourier transform algorithms. After extensive research and reinvention of these algorithms over a period of 100 years, it is now firmly established that the upper bound on the computational complexity of these methods is always on the order of $N \log_2 N$ when N is a power of 2. There is, interestingly, no extant proof that this upper bound is the lowest achievable upper bound on the complexity of an exact DFT computation.

Most of the algorithms are fastest when N is a power of 2, and the algorithms known as "radix 2" (e.g. [9]) reduce the number of complex multiplications to the order of $(N/2) \log_2 N$, although these methods require that N be a power of 2. Because of the popularity of the radix 2 FFT algorithms, it has often been

erroneously written (e.g. [24, 17]) that a fast Fourier transform can only be performed on a data frame consisting of some power of 2 samples, but this is not at all true of the other varieties of FFT algorithms [10].

2.2.5 *Practical Computing of Fourier Spectra*

2.2.5.1 DFT and Discrete Fourier Series

The earliest computations of the discrete Fourier transform of speech signals (e.g. [21, 35]) would invariably analyze a periodic signal one period at a time. The computation over a signal window of exactly one period makes the DFT analysis provide a discrete Fourier series of a (fictitious) signal that repeats the one period over all time. Moreover, these early analysts saw no reason to alter the recorded sound signal for analysis. They just simply "clipped out" a signal window equal to one period from an analog record, manually measured the signal values at equally spaced sample points, and computed the discrete Fourier series.

In later years, it was realized that the discrete Fourier series was constrained to find only those frequencies which are integer multiples of the fundamental, being the reciprocal of the period analyzed. The DFT can also be understood from a different perspective, since it also provides a digital approximation to the Fourier transform of a possibly aperiodic function. The DFT itself construes every function provided to it as periodic, having a period equal to the frame length. The only frequencies that can be correctly found by the DFT to have a significant amplitude are the reciprocal of the analysis frame length (which equals the frequency of the first DFT bin), and integer multiples of that—the DFT just *is* a discrete Fourier series, after all. This entails that a DFT energy spectrum, plotted as a graph, will look smoother for longer analysis frames—the discrete Fourier series frequencies will better "sample" the continuous frequency dimension.

Example 2.2 This example is computed using Praat speech analysis software. Consider the following signal that was examined theoretically before:

$$s(t) = \frac{1}{4}\sin(250 \times 2\pi t) + \frac{1}{2}\sin(500 \times 2\pi t), \tag{2.37}$$

being a superposition of two sine functions of frequencies 250 and 500 Hz, in which the higher frequency is double the amplitude of the lower. These two are harmonics, so this signal would have a very simple exact Fourier series if it were extended for all time—only two frequencies would have amplitudes in the spectrum. The squared magnitude of the DFT (its lower-frequency half, expressed in decibels) is plotted against frequency, providing a graph of the power spectrum—it is technically power and not energy, since the DFT inherently presumes the function extends for all time although we provide only one period, and so it is assumed to have infinite total energy.

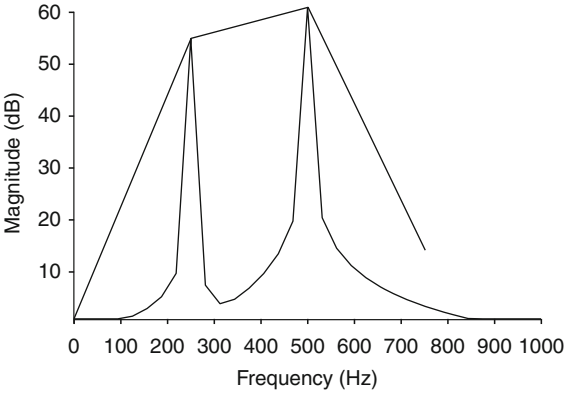


Fig. 2.4 The straight line segments connect the frequencies of just three components discovered from a DFT of one period of the sines signal. The smoother curve results from a DFT of eight periods of the same signal

Figure 2.4 shows two graphs; the one made up of straight line segments is the result of computing the DFT the "classical" way, using exactly one period of the signal as the analysis frame. The smoother line results from computing the DFT from an analysis frame encompassing eight periods. The first graph demonstrates what the Fourier series demands; the lowest frequency is the fundamental of the signal, and the next frequency is its harmonic, and above that there is one more frequency with negligible amplitude resulting from sampling error. The second graph is the result of computing a Fourier series using eight real signal periods as a new fundamental period. From this, the "fundamental" frequency will be 31.25 Hz (one eighth of the original), and all integer multiples of that will possibly have non-zero amplitude. It can be seen in the graph that, because there are nonetheless only two sinusoids, only those two "harmonics" of 31.25 Hz actually have large amplitudes; the small magnitudes in some other frequencies are largely a result of sampling error.

What all this means is that, whenever one computes a DFT, one is in fact computing a Fourier series of a fictitious infinite-time signal that is infinitely repeating the content of the analysis frame. The analysis frame length will then automatically determine the fundamental frequency of the Fourier series (regardless of whether this is the fundamental of the signal), and this will be the frequency of the first bin in the DFT. Simply "cutting out" the analysis frame as I have done so far is generally not going to give pleasing results, however. It works well here because the analysis frames were equal to signal periods to within sampling error. But what if we used an analysis frame of some other length, which might cut off the signal after, say, 2.5 periods, or 8.893 periods?

2.2.5.2 Window Functions

Example 2.3 Figure 2.5 shows the power spectrum that results from computing the DFT of the same signal as before, using a 34 ms analysis frame—2 ms longer than eight periods. The two component sinusoids are still the prominent peaks in

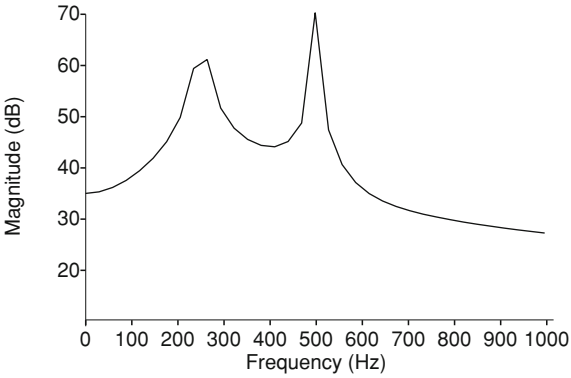


Fig. 2.5 The power spectrum computed by a DFT of the sines signal, using a window 2 ms longer than 8 periods

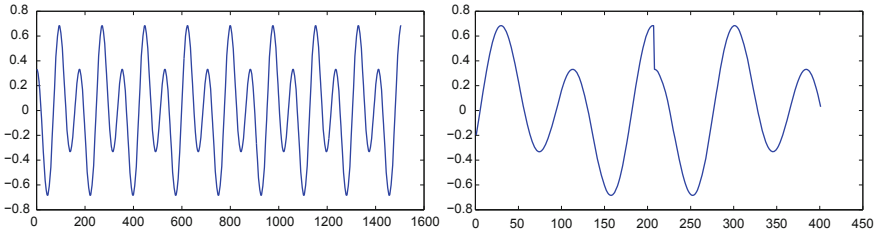


Fig. 2.6 The *left panel* shows the 34 ms frame clipped out of our manufactured signal consisting of two sinusoids (this contains 1,506 samples). The *right panel* is a close-up of the region where one instance of the frame is appended to another, as is implicitly done by the DFT computation. Note the obvious jump discontinuity, which will induce spectral leakage

the spectrum, but now they are obscured by an unwanted and uninformative broadening of the peaks at their bases. This phenomenon is often called "spectral leakage," because it appears that the power at the true component frequencies has "leaked out" into adjacent frequencies.

The cause of spectral leakage is not sampling error, it is the abrupt truncation and juxtaposition of signal frames in the fictitious periodic signal whose Fourier series we are computing, which introduces jump discontinuities at the boundaries between frames that are akin to impulses which have broad frequency content (v. Fig. 2.6). Leakage will come about whenever the frame length (34 ms, in this case) is not commensurate with the natural period of the signal. Another way to see the problem in the above example is to notice that the actual frequencies of the two signal components now lie between center frequencies of DFT bins, whereas in the preceding example the harmonic frequencies equalled (modulo sampling error) two of the bin frequencies. These two facts (one is a time-domain fact, the other a frequency-domain fact) go hand in hand, inducing leakage effects.

The way to mitigate leakage is to apply a *window function* to the analysis window on the signal which is tapered at the ends, in a more-or-less Gaussian bell

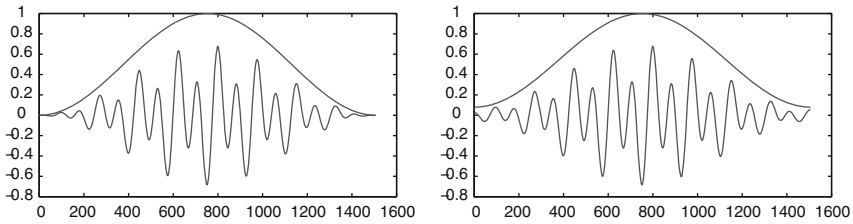


Fig. 2.7 The *left panel* shows a 1,506-point Hann window function, and also the result of multiplying this window by the 1,506-point frame of the sines signal. The *right panel* shows a similar figure using a Hamming window

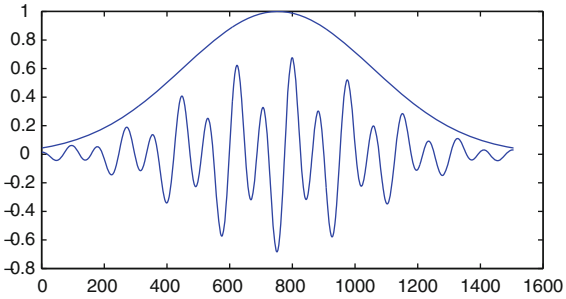


Fig. 2.8 A 1,506-point Gaussian window, $\alpha = 2.5$, and the result of multiplying by the sines signal

shape. Normally, a window function the same length as the analysis window is created using a formula, and then the signal window and window function are multiplied to yield a tapered or "windowed" signal window [46] (v. Fig. 2.7). Then, the fictitious signal which endlessly repeats the "windowed" window has much smaller jump discontinuities at the boundaries. There are many different window functions and families of functions which have been proposed over the years; we can look at just a few examples.

Two popular and effective windows use versions of a formula which derives the window from a cosine function. The Hann and Hamming windows were so named after their respective developers, although for reasons unknown to me, the Hann window came to be called by the made-up moniker "Hanning" as some sort of jokey allusion to "Hamming," and now the funny made-up name is more widely known. The Hann and Hamming windows are demonstrated in Fig. 2.7; their formulas are given in the math box below.

Another important class of windows is derived directly from the Gaussian function (see box for details). Figure 2.8 demonstrates. The reason there are so many different window functions proposed in the literature is chiefly due to the trade-off between a window's effect on leakage reduction and its deterioration of the component precision in the power spectrum, but there are a number of other minor factors which should be included in an assessment of a window's effectiveness in a DFT spectrum application. The reader is referred to Harris [18] for the most complete treatment of window functions I am aware of. By Harris'

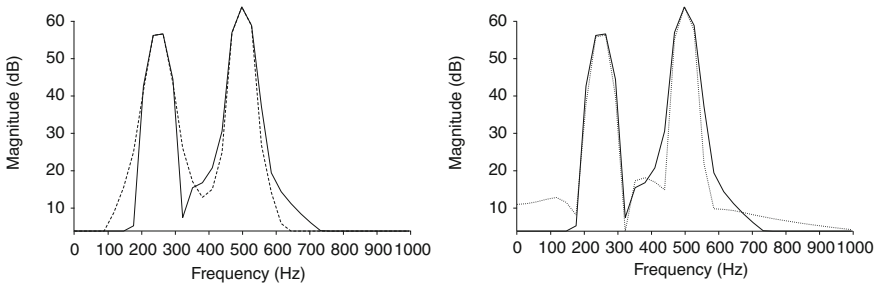


Fig. 2.9 The *left panel* compares the DFT power spectra of the 1,506-point sines signal frame as computed by Praat software, windowed using the Hann function (*dotted*) and a Gaussian. The *right panel* shows a similar comparison between the Hamming (*dotted*) and Gaussian windowed power spectra

criteria, the Hann and Hamming windows are quite effective, but the Gaussian windows (with e.g. $\alpha = 2.5$) are often considerably better at leakage reduction (also known as "sidelobe rejection") while maintaining comparable component precision. Figure 2.9 demonstrates the differences between the Hann, Hamming, and Gaussian window effects on a DFT spectrum, where it can be seen that the Hann and Hamming spectra generally have broader peaks (which is not what is wanted in this instance).

As yet another alternative, Kaiser derived a very effective window family from the zero-order modified Bessel function I_0 well-known in physics [2, 40]; frequently called the Kaiser–Bessel window in homage to its source, it has been judged the best-performing window function overall [18]. Figure 2.10

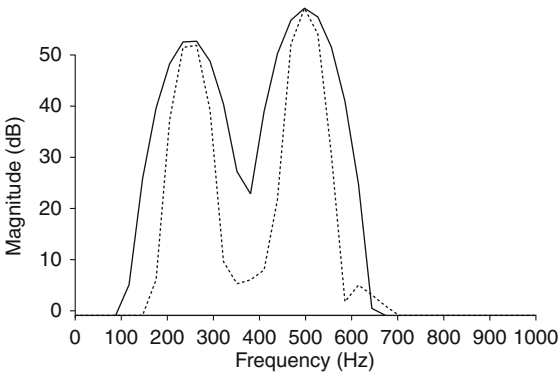


Fig. 2.10 Compares the power spectra of the 1506-point sines signal frame resulting from applying each of two Kaiser windows currently available in Praat. Observe how one of these (with a high value of the parameter β in the formula) fails to precisely locate the signal components. The other more moderate choice appears to be the best-performing window we have tried on this example function

demonstrates the performance of a couple of different Kaiser window functions. Detailed formulae are given in the math box below.

The Hann or "Hanning" window is commonly defined by the first equation below (as a function of discrete sample sequence $\{n\}$, containing N samples), while the Hamming window is given by the second equation [18].

$$w(n) \stackrel{\text{def}}{=} 0.5 - 0.5\cos\left(\frac{2\pi n}{N+1}\right) \tag{2.38}$$

$$w(n) \stackrel{\text{def}}{=} 0.54 - 0.46\cos\left(\frac{2\pi n}{N+1}\right) \tag{2.39}$$

The family of Gaussian windows is defined using the following equation which includes a parameter α, the reciprocal standard deviation of the normal density function defined by the Gaussian.

$$w(n) \stackrel{\text{def}}{=} \exp\left[-\frac{1}{2}\left(\alpha\frac{n}{N/2}\right)^2\right] \tag{2.40}$$

The following equation defines the Kaiser family of window functions, parameterized by β:

$$w(n) \stackrel{\text{def}}{=} \frac{I_0\left(\pi\beta\sqrt{1-\left(\frac{2n}{N}-1\right)^2}\right)}{I_0(\pi\beta)}, \tag{2.41}$$

in which I_0 denotes the standard zero-order modified Bessel function [2].

Because the tapering action of any window function effectively decreases the DFT spectrum resolution (in relation to the original unwindowed signal), it is better to use a rectangular window (in other words, don't use a window function, just clip the signal) in the event the analysis frame length can be made commensurate with the natural fundamental period of the signal. This fact is demonstrated in Fig. 2.11.

2.2.5.3 Zero-Padding the Analysis Frame

In our earlier description of the discrete Fourier transform, the procedure of *zero-padding* the analysis frame was briefly mentioned, but it will be useful to expand on those remarks now. When a discrete-time signal is submitted to a DFT, the number of samples in the DFT frame determines two important things: the frequency sampling resolution of the resulting spectrum, and (in most

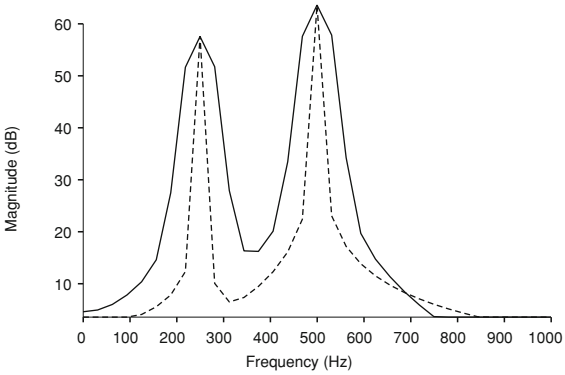


Fig. 2.11 A signal frame exactly eight periods long was cut from the sines signal, and the figure compares resulting DFT power spectra computed in Praat using a Gaussian and a rectangular window (*dashes*). In this case, the rectangular window is best because the signal frame length is commensurate with the natural signal period, obviating any need for a tapering window

implementations) the speed of the FFT algorithm. There can be some trade-offs involved, since oftentimes a very short analysis window on the signal is desired, but clearly the best sampling resolution in frequency will be achieved by a long DFT frame.

For example, suppose I have a speech signal which is sampled at only 10 kHz, which old-timers will remember as once being quite commonly done to save on storage space. Suppose next that, for reasons which will become clear in the upcoming chapters, I wish to compute a spectrum for an 8 ms slice of the signal. These 8 ms will be represented by just 80 sample points. Suppose I simply run an FFT using this 80-point frame. Because the DFT is redundant for its upper half, only 40 frequency bins will be available for sampling the 0–5,000 Hz frequency range that is represented in the discrete signal. This is, you may easily imagine, a pretty rough frequency sampling.

Next considering the FFT speed issue, which can still be quite important when spectrograms are computed (as will become apparent later), 80 is not a power of 2, so the FFT algorithm's speed will not be optimized. It is therefore desired to be able to compute an FFT of the 80-point window using, say, a 1,024-point frame. There is an easy way to do this; simply place the (windowed) discrete signal values into the first 80 points of a 1,024-point frame, and put zeros in for the remaining 944 points. Because the zeros amount to nothing, they have a zero spectrum, so their presence does not adulterate the spectrum of the signal window. But now, we are able to compute an FFT using a frame length which is a power of 2, and moreover, we will now have 512 frequency bins sampling the 0–5,000 Hz range, which will yield a much nicer-looking discrete spectrum than 40 bins.

This procedure seems like a cheap trick which is too easy to be free of defects, but in practice a zero-padded long frame can actually yield an improved look at the spectrum versus a shorter frame which exactly matches the signal window (v. Fig. 2.12), because the discontinuities that cause spectral leakage are even

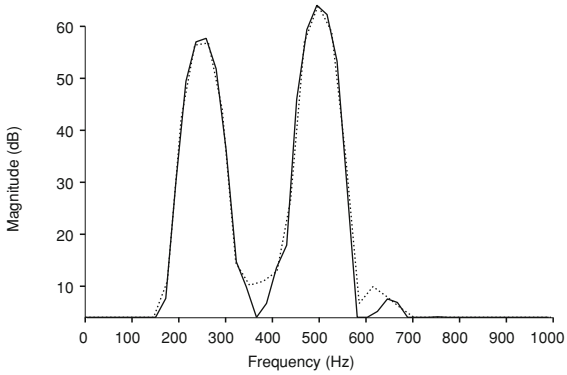


Fig. 2.12 Compares the Kaiser-windowed power spectra of the 1,506-point sines signal frame (*dotted*) and the same windowed signal zero-padded to 2,048 points. Note that zero-padding the frame provides almost the same spectral peak shapes, but the leakage between peaks is even further ameliorated. An added advantage here is that the FFT computation using 2,048 points should be faster than that using 1,506 points

further reduced by this method.[3] There are no disadvantages that I am aware of, except perhaps in the artificial cases when a window can be precisely tailored to the major periods in the signal. In spite of this, writers such as Hamming [16] can still be found who speak out against zero-padding. But Hamming's objections seem to be concerned purely with the precise coefficients of the discrete Fourier transform, and these will of course be different with different frame lengths, zero-padded or not. Here, our concern is solely with the power spectrum which these coefficients determine.

2.2.6 Filters

In signal processing, a *filter* is any type of system that acts on an input signal to yield an output signal, whose action changes the frequency content of the input. The action of a filter in the frequency domain is usually put in the following way [19]. Given an input signal $x(t)$ and an output signal $y(t)$, let us write their respective Fourier transforms as $X(\omega)$, $Y(\omega)$. The filter action, called the *frequency response*, can be written $H(\omega)$, and then the input and output spectra are related by the equation:

$$Y(\omega) = H(\omega)X(\omega). \tag{2.42}$$

[3] The reader has probably noticed that my intentional choice of terminology in discussing these matters uses the term *window* to refer to the possibly tapered (windowed) slice of the signal being analyzed, while the term *frame* is used to refer to the discrete Fourier transform analysis length. This terminology will be used throughout the book.

Filters are often classified according to broad aspects of their frequency response. For example, a *low-pass* filter is one which attenuates frequencies above a certain value. A *band-pass* filter is one which attenuates frequencies below and above a certain range. Filters can also be found to have more complicated responses, such as attenuating a number of different frequency ranges for the effect of a multiple band-pass.

An especially important type of filter from a speech perspective is often called a *resonant* filter or a tuned system. Such a filter has a frequency response that has a peak at some frequency ω_0, with a pass band surrounding the peak so that the transmitted power rolls off at increasingly distant frequencies. The filter bandwidth $2\Delta\omega$ is specified by the half-power points [19], so that $|H(\omega_0 + \Delta\omega|^2 = |H(\omega_0)|^2/2$ and $|H(\omega_0 - \Delta\omega|^2 = |H(\omega_0)|^2/2$. The bandwidth is often cited as a *quality factor* Q which is frequency-dependent, so that $Q \stackrel{\text{def}}{=} \omega_0/(2\Delta\omega)$. In speech science, a resonance frequency of the vocal tract may be characterized as a resonant filter which acts on the input sound provided by the vocal cords.

2.2.7 Analytic Signals

In speech signal processing, it is only natural that we wish to deal with signals that are real-valued only, since these are the only signals that have an obvious physical meaning. In signal processing, however, it is common to deal with complex-valued signals, and we will find some use for these for the reasons to be explained now. For any real signal $s(t)$, a particular complex signal $z(t)$ associated to it was originally defined by Gabor [14]:

$$z(t) \stackrel{\text{def}}{=} s(t) + \mathrm{i}z_i(t). \tag{2.43}$$

Observe that the real part of $z(t)$ is just $s(t)$, so what is the imaginary part for and how should it be specified? It was a matter of some importance at that time to figure out the best way of mathematically defining the phase of a signal, so that the derivative of the signal phase would correspond to the *instantaneous frequency* (see the next section). Gabor also noted that the spectrum of a real signal, as given by its Fourier transform, formally contains negative frequencies whose power spectrum mirrors the positive frequencies (recall Fig. 2.3). It is possible to define a complex signal which has no negative frequencies in its spectrum, and whose complex phase derivative is indeed the average frequency at each instant, thereby making it a good value to call the instantaneous frequency. Gabor defined $z(t)$ so that the Fourier transform $Z(\omega) = 0$ for all negative frequencies, while $Z(\omega) = 2S(\omega)$ for all positive frequencies, $S(\omega)$ being the Fourier transform of the original real signal. This associated complex signal effectively takes the power from the negative frequencies and "moves it" to the positive side where the values are doubled (thus preserving the energy).

It was Ville [37] who introduced the term *analytic signal* to refer to Gabor's complex $z(t)$ constructed from a real $s(t)$, because $z(t)$ meets the mathematical definition of being an analytic complex-valued function. Gabor had also shown that the analytic signal's imaginary part is guaranteed to be the *Hilbert transform* of the real part, and vice versa (see box for details). A very important raison d'être for the analytic signal, which was highlighted by Ville, is that this is the particular kind of signal for which the concept of instantaneous frequency can be provided a sensible mathematical definition as the phase derivative (see the next section).

The two main properties defining the analytic signal—viz. that it has no negative frequencies, and its real and imaginary parts form a Hilbert transform pair—provide us with two means of actually defining it, which ideally come to the same end, but in the discrete-time realm they are always slightly different. This leaves us with a choice of two definitions for the "discrete analytic" signal (which is no longer formally analytic) that is associated to a real discrete signal $s(n)$. We could compute a second real signal $z_i(n)$ by a discrete Hilbert transform of $s(n)$, and then simply set $z(n) \stackrel{\text{def}}{=} s(n) + \mathrm{i}z_i(n)$. This has been advocated as the best technique [4], but Marple [29] demonstrated that the frequency-domain method is superior. For this second method, I have implemented some recent improvements [31]; we first compute the N-point DFT of our real signal $s(n)$ (of length N samples), to yield the discrete complex sequence $X[m]$. Then we form the DFT $Z[m]$ of the associated "analytic" signal by setting it equal to $2X[m]$ in the range $1 \leq m \leq N/2$, and zero elsewhere (with a couple of adjusted points; see the paper [31] or the Matlab file `newhilbert.m` for the specific algorithm). The complex "analytic" signal $z[n]$ is then computed from $Z[m]$ by an inverse DFT. We will have several occasions in this book to consider analytic signals instead of real ones.

In the continuous-time regime, the Hilbert transform of a signal $s(t)$ (which may in general be complex-valued) has the following definition [19]:

$$\mathscr{H}[s(t)] \stackrel{\text{def}}{=} \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{s(\tau)}{t-\tau} \mathrm{d}\tau \tag{2.44}$$

$$\mathscr{H}[s(t)] = \frac{1}{\pi} \lim_{\varepsilon \to 0} \left[\int_{-\infty}^{t-\varepsilon} \frac{s(\tau)}{t-\tau} \mathrm{d}\tau + \int_{t+\varepsilon}^{\infty} \frac{s(\tau)}{t-\tau} \mathrm{d}\tau \right] \tag{2.45}$$

Because the integral in the defining expression is improper (the denominator has a zero), it is evaluated using its Cauchy principal value, which is a sum of two limit integrals that excludes the zero point.

The Hilbert transform produces a new function that is "in quadrature" with the original; formally, the two are orthogonal in their function space, having an inner product of zero. Resulting from this, the Hilbert transform of

a Hilbert transform gets us back to the original function with a minus sign: $-\mathscr{H}[\mathscr{H}[s(t)]] = s(t)$. For example, the Hilbert transform of $\sin(\omega t + \phi)$ is $-\cos(\omega t + \phi)$, and the Hilbert transform of $\cos(\omega t + \phi)$ is $\sin(\omega t + \phi)$ [19].

Another important mathematical fact concerns the relation between the Fourier transform (spectrum) of a signal and the Fourier transform of its Hilbert transform. Let us denote the Hilbert transform of $s(t)$ as $\hat{s}(t)$, and then the Fourier transform of it into a frequency function can be written $\hat{S}(f)$. The following facts are presented from [13].

$$\hat{S}(f) = -\mathrm{i}\,\mathrm{sgn}(f)S(f) \tag{2.46}$$

$$S(f) = \mathrm{i}\,\mathrm{sgn}(f)\hat{S}(f) \tag{2.47}$$

in which $\mathrm{sgn}(f)$ denotes the *signum* function that simply returns a value 1, -1 or 0 depending upon the sign of f. This relationship explains why the analytic signal created from a real signal ends up with a one-sided Fourier transform (zero for negative frequencies).

Originally, the Hilbert transform was conceived by David Hilbert in 1905, during work on a problem concerning analytic functions that is now called the Riemann–Hilbert problem [45].

2.2.8 Concepts of Frequency

In the end, we perform a spectrum analysis of a signal in order to find out about the frequencies that it contains, presumably because this information can in turn lead to useful inferences about the source of the signal. The trouble is, the notion of "frequencies in a signal" is not rigorously defined. Indeed, the notion of "frequency" is somewhat colloquial, referring as it does in simple language to the rate at which some process of interest repeats or recurs. The mathematical treatment of signals, however, demands a mathematical definition of "frequency," and here there are two distinct approaches that accord with the colloquial notion in differing ways.

The first definition emerges directly from the Fourier series or transform representation of a signal. The "frequencies" in a signal are represented therein by the trigonometric (or complex exponential) component signals, each of which exists only at one frequency. This definition has served well in many applications, and there is no doubt about its mathematical correctness. The question is, does this definition suitably model the colloquial notion of frequency? A problem arises when we consider the related colloquial notion of a changing frequency. Gabor [14] reminds us that "if the term *frequency* is used in the strict mathematical sense which applies only to infinite wave-trains, a 'changing frequency' becomes a

contradiction in terms." Therein lies the rub: the Fourier components of a signal are *infinite in time*, meaning that strictly speaking, only an infinitely long sine wave can have just a single frequency component. But surely, we all realize that a pure sine wave of just one frequency can be started and stopped, or its frequency can be changed (modulated) in time. Do these actions somehow give it more frequencies, or can we agree there is still only one frequency in a sine wave of a changing nature? This idea provides the intuitive concept of an "instantaneous frequency" of a modulated sine wave.

An arbitrary real signal $s(t)$ can be written in terms of an instantaneous amplitude and frequency modulation:

$$s(t) = A(t)\cos\phi(t), \tag{2.48}$$

and one might imagine that by doing so, the derivative of the frequency function $\phi(t)$ could provide us with an instantaneous frequency for the signal. However, it turns out that the above signal decomposition is not unique [5], so that no coherent definition may be derived from it. This is a moment in the sun for the associated analytic signal, because for any (continuous-time) signal $s(t)$ its analytic associate $z(t)$ (see Eq. 2.43) is unique, and when expressed in complex polar form:

$$z(t) = A_s(t)\,\mathrm{e}^{\mathrm{i}\phi_s(t)} \tag{2.49}$$

the pair of functions $A_s(t), \phi_s(t)$ giving respectively the amplitude and complex phase modulations is the unique *canonical pair* associated to $s(t)$ [5]. As we already saw, the real part of $z(t)$ is just $s(t)$, which may now be decomposed as

$$s(t) = A_s(t)\cos\phi_s(t), \tag{2.50}$$

which defines the *canonical representation* of $s(t)$ [5]. So now, the *instantaneous frequency* of $s(t)$ can be properly defined as the derivative of the canonical frequency modulation $\phi_s(t)$, a definition that dates back to Carson and Fry [7] based upon Carson's [6] earlier conception of the "generalized frequency."

It is important to note that while we require an analytic signal in order to define the instantaneous frequency, this is shown above to be the instantaneous frequency of the associated real signal also. Ville [37] showed, moreover, that the average frequency at each instant of a changing sinusoidal signal equals the complex phase derivative of its analytic associate. In plain language, this means that the mathematical instantaneous frequency just defined accords precisely with another intuitively defined "instantaneous" frequency, viz. the average sinusoidal frequency at an instant. In the end we have two useful coexisting models of the notion "frequency": the Fourier model, in which frequencies are formally a property only of infinite unchanging signals; and the model owing to Gabor, Ville, and Carson and Fry, in which instantaneous frequencies are defined by means of analytic signals.

References

1. W.S. Allen, *Phonetics in Ancient India* (Oxford University Press, Oxford, 1953)
2. G. Arfken, *Mathematical Methods for Physicists*, 3rd edn. (Academic Press, New York, 1985)
3. R.B. Blackman, J.W. Tukey, *The Measurement of Power Spectra from the Point of View of Communications Engineering* (Dover Publications, New York, 1958)
4. B. Boashash, A. Reilly, Algorithms for time–frequency signal analysis, in *Time–frequency Signal Analysis: Methods and Applications*, chap. 7, ed. by B. Boashash (Halsted Press, New York, 1992)
5. R. Carmona, W. L. Hwang, B. Torrésani, *Practical Time–Frequency Analysis: Gabor and Wavelet Transforms, with an Implementation in S.* (Academic Press, San Diego, 1998)
6. J.R. Carson, Notes on the theory of modulation. Proc. Instit. Radio Eng. **10**(1), 57–64 (1922)
7. J.R. Carson, T.C. Fry, Variable frequency electric circuit theory with application to the theory of frequency-modulation. Bell Syst. Tech. J. **16**(4), 513–540 (1937)
8. J.C. Catford, *Fundamental Problems in Phonetics* (Edinburgh University Press, Edinburgh, 1977)
9. J.W. Cooley, J.W. Tukey, An algorithm for the machine calculation of complex Fourier series. Math. Comput. **19**, 297–301 (1965)
10. D.F. Elliott, K.R. Rao, *Fast Transforms: Algorithms, Analyses, Applications* (Academic Press, New York, 1982)
11. G. Fant, *Acoustic Theory of Speech Production* (Mouton, The Hague, 1960, Reissued 1970)
12. J.B.J. Fourier, Théorie de la propagation de la chaleur dans les solides. Manuscript first published in [15] (1807)
13. L.E. Franks, *Signal Theory* (Prentice-Hall, Englewood Cliffs, 1969)
14. D. Gabor, Theory of communication. J. IEE Part III **93**(26), 429–457 (1946)
15. I. Grattan-Guinness, *Joseph Fourier 1768–1830*. (The MIT Press, Cambridge, 1972)
16. R.W. Hamming, *Digital Filters*, 3rd edn. (Prentice-Hall, Englewood Cliffs, 1989)
17. J. Harrington, S. Cassidy, *Techniques in Speech Acoustics*. (Kluwer, Dordrecht, 1999)
18. F.J. Harris, On the use of windows for harmonic analysis with the discrete Fourier transform. Proc. IEEE **66**(1), 51–83 (1978)
19. W.M. Hartmann, Acoustic signal processing. in *Springer Handbook of Acoustics*, ed. by T.D. Rossing (Springer, New York, 2007) pp. 503–530
20. J.E. Hoard, Syllabication in Northwest Indian languages, with remarks on the nature of syllabic stops and affricates. in *Syllables and Segments*, ed. by A. Bell, J.B. Hooper (North-Holland, Amsterdam, 1978)
21. F. Jenkin, J.A. Ewing, On the harmonic analysis of certain vowel sounds. Trans. R. Soc. Edinb. **28**, 745–777 (1878)
22. M. Joos, *Acoustic Phonetics* (No. 23 in Language Monographs, Linguistic Society of America, Baltimore, 1948)
23. P. Ladefoged, *Three Areas of Experimental Phonetics* (Oxford University Press, Oxford, 1967)
24. P. Ladefoged, *Elements of Acoustic Phonetics*, 2nd edn. (University of Chicago Press, Chicago, 1996)
25. P. Ladefoged, Linguistic phonetic features. in *The Handbook of Phonetic Sciences*, ed. by W.J. Hardcastle, J. Laver (Blackwell, London, 1997)
26. P. Ladefoged, *A Course in Phonetics*, 5th edn. (Thomson, Boston, 2006)
27. J. Laver, *The Phonetic Description of Voice Quality* (Cambridge University Press, Cambridge, 1980)
28. D. Lewis, Vocal resonance. J. Acoust. Soc. Am. **8**, 91–99 (1936)
29. S.L. Marple, Computing the discrete-time "analytic" signal via FFT. IEEE Trans. Sig. Proc. **47**(9), 2600–2603 (1999)

30. R.F. Orlikoff, J.C. Kahane, Structure and function of the larynx. in *Principles of Experimental Phonetics*, ed. N.J. Lass (Mosby, St. Louis, 1996) pp. 112–181
31. J.M. O'Toole, M. Mesbah, B. Boashash, A new discrete analytic signal for reducing aliasing in the discrete Wigner–Ville distribution. IEEE Trans. Sig. Proc. **56**(11), 5427–5434 (2008)
32. K.L. Pike, *Phonetics* (The University of Michigan Press, Ann Arbor, 1943)
33. M.B. Priestley, *Spectral Analysis and Time Series*, vol. 1. (Academic Press, London, 1981)
34. J.G. Proakis, D.G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, 2nd edn. (Macmillan, New York, 1992)
35. E.W. Scripture, *The Elements of Experimental Phonetics* (Charles Scribner's Sons, New York, 1902)
36. W. Thomson (Lord Kelvin), P.G. Tait, *Treatise on Natural Philosophy*. vol. 1, 2nd edn. (Cambridge University Press, Cambridge, 1912)
37. J. Ville, Thèorie et applications de la notion de signal analytique. Cables et Transmission **2A**(1), 61–77 (1948)
38. N. Wiener, Generalized harmonic analysis. Acta Math. **55**(1), 117–258 (1930)
39. Wikipedia: Autocorrelation. http://www.wikipedia.org (2009)
40. Wikipedia: Bessel function. http://www.wikipedia.org (2009)
41. Wikipedia: Discrete Fourier transform. http://www.wikipedia.org (2009)
42. Wikipedia: Discrete-time Fourier transform. http://www.wikipedia.org (2009)
43. Wikipedia: Fast Fourier transform. http://www.wikipedia.org (2009)
44. Wikipedia: Fourier transform. http://www.wikipedia.org (2009)
45. Wikipedia: Hilbert transform. http://www.wikipedia.org (2009)
46. Wikipedia: Window function. http://www.wikipedia.org (2009)
47. N.I. Žinkin, *Mechanisms of Speech*. (Mouton, The Hague, 1968)

Chapter 3
History of Speech Spectrum Analysis

This chapter traces the history of sound (and in particular, speech) spectrum analysis from its very beginnings in the theory developed by Fourier in the early 1800s. A particular goal of this historical outline is to describe not just the events and developments through the years, but also the beliefs and attitudes of scientists as these changed with the development of a better understanding. Some of the scientists whose work is discussed here are still widely known and cited, while others' contributions have been unjustly forgotten. With this chapter, I also hope to straighten out the historical record in this respect, giving all due credit to those pioneers who uncovered many facts about speech spectra that are now taken for granted.

3.1 Fourier Analysis of Speech

3.1.1 Early History

The concept of spectrum analysis is rooted in the idea of expressing one function (the signal) as a combination of other functions, each of which can be interpreted physically as some kind of frequency. The constituent functions are the basic functions of trigonometry, the sine and cosine, as shown in the previous chapter. The first mathematical work investigating this sort of decomposition of a function was probably that of Leonhard Euler, which is hardly a surprise when we remember that Euler's work is so often important to the history of modern mathematics. In a number of publications beginning around 1748, Euler explored how some functions could be rewritten as infinite series, such as infinite series of sine or cosine functions. Euler did not seem to have frequency spectrum analysis in mind as an application, however he did seek to solve differential equations with the aid of such series. It was also Euler who related the sine and cosine, which are real functions of a real variable, to the more general complex exponential function e^{ix} which underlies our fullest understanding of spectrum analysis.

S. A. Fulop, *Speech Spectrum Analysis*, Signals and Communication Technology,
DOI: 10.1007/978-3-642-17478-0_3, © Springer-Verlag Berlin Heidelberg 2011

Perhaps the first application of Euler's ideas to the physics of vibration was due to Daniel Bernoulli (the elder), who published a paper in 1753 proposing a novel solution to the vibrating string equation. He founded his solution on the physical theory, known from Pythagorean times, wherein a vibrating string of length l is modeled as a superposition of all its characteristic modes of vibration, which is naturally described as a sum of simple harmonic (trigonometric) functions:

$$f(x) = \alpha \sin \frac{\pi x}{l} + \beta \sin \frac{2\pi x}{l} + \cdots$$

This suggested form of the solution does indeed satisfy the differential equation for the vibrating string, the so-called *wave equation* which is foundational to linear acoustics.

In a manuscript dated 1807, Joseph Fourier showed how to derive both a sine and a cosine series representation of an "arbitrary" function, which he pursued as a means of solving the differential equation governing heat transfer [23]. Fourier was also aware that the vibrating string was governed by a similar differential equation, and he remarked in the manuscript that his methods would also work to represent functional solutions to that problem, thereby providing solutions to the wave equation. The following quote is my translation of Fourier.

> These objections make clear how much it was necessary to show that an unspecified function can always be developed in a series of sines or cosines of multiple arcs, and of all the proofs of this proposition, the most complete is that which consists of solving effectively an arbitrary function in such a series, by assigning the values of the coefficients. The preceding theorems satisfy this condition, and I am convinced, indeed, that the movement of the vibrating string is also exactly represented in all possible cases by trigonometric developments using the integral which contains arbitrary functions [23] art. 75.

Fourier's key contribution in this realm was not exactly the invention of the trigonometric series representation, but rather the work to establish the generality of such representation, and also his work to extend the discrete series form to an integral form capable of representing aperiodic functions. Indeed, most of the details of Fourier's series and integral were invented independently by another giant of history, Carl Gauss. Gauss's manuscript on what are now called Fourier's series and integral was not published until his collected works appeared in 1866, but historians have agreed that this particular manuscript was most likely written in 1805 [28].

In spite of contemporaneous work by Gauss, and despite Fourier's drawing upon previous publications for inspiration, his claims of such complete generality for his functional representations were greeted with skepticism. Fourier may have been convinced that his series representation was general enough to represent a very wide class of reasonable functions, but most of his contemporaries were not. Lagrange, although Fourier's doctoral advisor, famously remained opposed to Fourier's series until his death. The initial poor reception afforded Fourier's work seems in part to be due to academic politics, but was also due to the low level of mathematical rigor in many of Fourier's proofs. The wide applicability of

Fourier's series representation finally became a firmly proven result thanks to the labors of Johann P. G. L. Dirichlet, whose results on the issue were largely presented in two papers published in 1829 and 1837. In the end, the results have passed down to us as the "Dirichlet conditions" for the convergence of Fourier's series.

Further work on solving differential equations led Fourier from his series to what is now called the Fourier transform, by means of the Fourier Integral Theorem [26], a result he obtained in 1811 but which was not widely circulated until the publication of his book in 1822. He derived his integral representation of a function by considering the behavior of his earlier series expansions in the limit as the period approaches infinity, as was derived in the previous chapter. The Fourier transform relation and its reciprocal nature was discovered independently by yet another giant of the era, Augustin Cauchy (v. Note 19 in [9]).

3.1.2 The Physical Reality of Fourier Components

The application of Fourier's series to physical problems in which a natural signal was to be analyzed into its spectrum of harmonic sinusoids began to be explored in the late 1800s, after the initial skepticism surrounding Fourier's results had softened. Such analyses were performed on a variety of signals, including sunspot records, tidal height records, optical (light) signals,[1] and sound recordings. The results of these studies were not always universally accepted, and the "physical reality" of Fourier spectra was the subject of considerable debate for some time. A representative quote here comes from a paper by Godfrey published in 1900.

> [T]he equations of optics find their simplest solution in circular [trigonometric] functions. It is desirable to inquire how far we may resolve a natural luminous motion with a sum of simple wave-trains by means of Fourier's "Theorem of Double Integrals." This procedure was first suggested by Gouy [1886]. Doubts have often been entertained as to the permissibility of this process. Writers have been sceptical as to the physical meaning and independence of the simple waves thus introduced. ([24], p. 331)

Godfrey discussed Gouy's [25] description of analyzing "any disturbance" using Fourier's series. Gouy had considered a function defined on some interval of time; such may be analyzed into a sum of trigonometric functions of time, the periods of the terms being the interval itself and all integer "submultiples" (fractions $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$) of it, as we have already seen. To this point, Gouy was talking of the Fourier series analysis of a periodic signal. By 1886, this procedure was already being carried out both in the analog continuous regime (using analog calculating devices) and the discrete-time regime (using hand calculation).

[1] Until the early 1900s, light was not recognized as an electromagnetic particle/wave phenomenon, but rather was erroneously thought to involve waves in an otherwise undetectable substrate known as the *ether*.

Gouy went on to suggest that it was unnecessary to restrict the analysis to periodic functions. This suggestion invokes the passage to the limit (shown in the previous chapter), by which the Fourier integral transform is understood as the limit of Fourier's series when the fundamental period of oscillation goes toward infinity. Godfrey considered the question, prominent at the time, "have the simple elements meaning in the limit, when their number is infinite, and the sum becomes an integral?" [24]. He went on to mention Poincaré's objection to Gouy's idea, which highlighted an apparent paradox arising from Fourier's integral representation of a signal:

> Each of the component vibrations exists unchanged through all time. This is true whatever be the nature of the disturbance we are analysing. But this disturbance may, for instance, be zero, except within a certain definite interval of time [24] p. 333.

In other words, there was thought to be a problem with general Fourier analysis because the aperiodic signal at issue will generally have finite support, while the sinusoidal components have to have infinite support. Godfrey attempted in his paper to explain away this paradox by some incorrect reasoning, but he was trying his best to argue that Fourier's integral transform *can* provide a spectrum associated with a signal that has some physical meaning, even if "the different simple elements of the Fourier integral cannot *in general* be said to have any independent physical existence" [24].

Concerns about the physical meaning of a Fourier spectrum were aired somewhat later by Miller:

> If a curve representing some physical phenomenon is periodic, then each separate term of the Fourier [series] equation of the curve may be presumed to correspond to something which has a physical existence; it is the belief in this statement, amply supported by investigation, which leads one to analyze sound waves by this method; … each term is presumed to correspond to a simple partial tone which actually exists. If the curve representing the physical phenomenon is non-periodic, any portion of the curve may be analyzed, and it will be completely represented as to *form* by the Fourier equation, within the limits analyzed, *but not beyond these limits*. In this case, the separate terms of the Fourier series may not correspond to anything having a separate physical existence; … There is no general method for analyzing non-periodic curves, that is, for curves containing [inharmonic] or variable components; such a method is very much desired for the study of … all speech sounds except the simple vowels. ([40] pp. 140–141)

Note that Miller, who was a physicist investigating musical and speech acoustics, sees every possible spectrum as being the discrete kind computed from the Fourier series coefficients. While Godfrey was trying to argue in favor of physically interpreting the continuous spectrum implicit in the Fourier transform representation of an aperiodic signal, Miller seems unsympathetic to the prospect.

The desire to attribute an independent physical existence to the components found by Fourier analysis has often been a powerful influence on the interpretation of spectral representations. Nevertheless, it is now recognized that the Fourier analysis of a function provides us with a different representation which is mathematically correct overall, but whose individual components may not correspond to anything physical in an obvious way. For example:

> If we consider a quantity which can be represented in a Fourier manner, i.e., by a superposition of monochromatic components, then it is the superposition which has a physical meaning, but not the isolated Fourier components. If we deal, for example, with a swinging chord whose motion can be described by a series of harmonics, a movie of this motion would reveal that the chord has a very complicated form at each moment, and that it varies incessantly according to a complex rule. Nothing in this motion allows us to distinguish the various monochromatic components: these components exist only in the minds of theorists who endeavor an abstract analysis of this motion ([5] as translated in [19]).

Finally, Bouasse tartly reminds us that "… unless one has lost the most elementary common sense, it is impossible to attribute an *objective* existence to the harmonic oscillations which emerge in the Fourier series" ([4], as translated in [19]).

3.1.3 Recording Sound Signals

Before applying Fourier's theorem to periodic sound signals, it was first necessary to somehow display a sound signal as a function of time. A number of

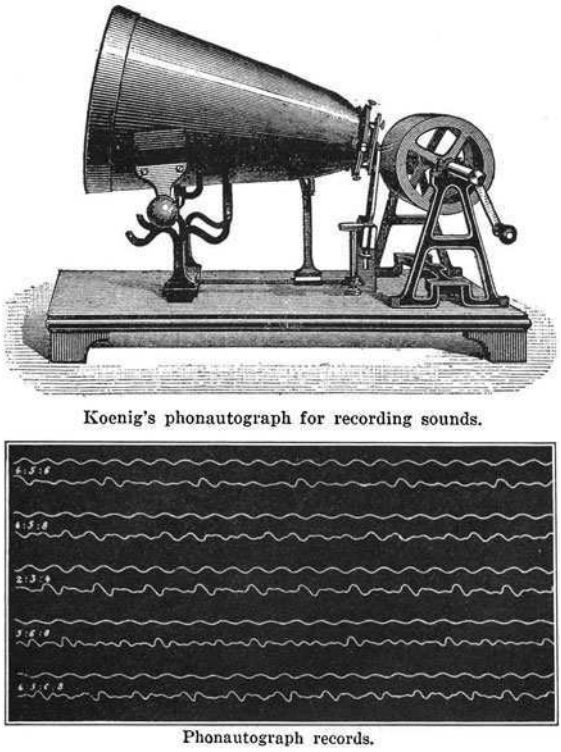


Fig. 3.1 A diagram of the phonautograph, together with waveform tracings obtained using it [40]

mechanisms for graphing a "wave-form" (a term perhaps due to Jenkin and Ewing, see below) visualizing sound as a vertical undulation against a horizontal time axis were developed beginning in the late 1800s. The first of these was the phonautograph of Leon Scott, which was a precursor of the earliest Edison phonograph. It did not make a sound recording, but used a rudimentary transducer to drive a small lever which would trace back and forth against a turning smoked cylinder (v. Fig. 3.1).

The phonograph of Thomas Edison, in its original configuration, was apparently inspired by the phonautograph since it, too, used a simple transducer to drive a lever whose stylus traced an impression on tinfoil wrapped around a turning cylinder. The phonograph record had the advantage that it could be replayed by a reading stylus that drove a simple loudspeaker, thus providing the earliest means of sound recording. The "groove" traced on the phonograph foil was very small, far too small to be useful as a graphical device. As a result, techniques for reproducing and magnifying the groove were devised, the first of which seems to be that of Jenkin and Ewing [30] (v. Fig. 3.2). The tin foil record on the cylinder at the upper right was read by a stylus which drove a system of levers ending in "one of Sir W. Thompson's electrical squirting recorder tubes, which magnified the depth of the indentations 400 times," [17] and thereby transmitted the waveform with ink droplets squirted onto a strip of telegraph paper wound around the large wheel *W*. The speed of the wheel was such as to magnify the length of the waveform seven times.

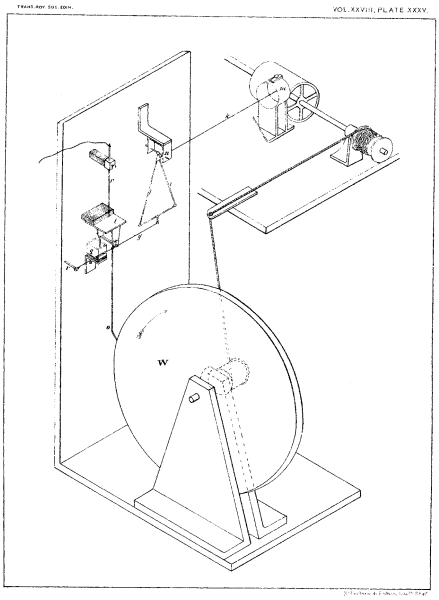
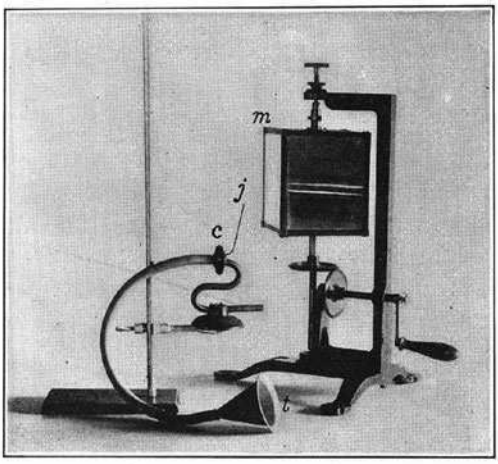


Fig. 3.2 An original figure from Jenkin and Ewing [30] diagramming their mechanism for visualizing the sound waveform recorded on a phonograph foil

A particularly interesting method of sound visualization known as the *manometric flame* method was first described in 1872 by Koenig [32]. It does not provide a waveform, but is worth discussing in our historical chapter if only because it is so peculiar and impractical, and thus became an historical curiosity after a few decades. In spite of this, it was quite an important tool during the last



Koenig's manometric capsule with revolving mirror.

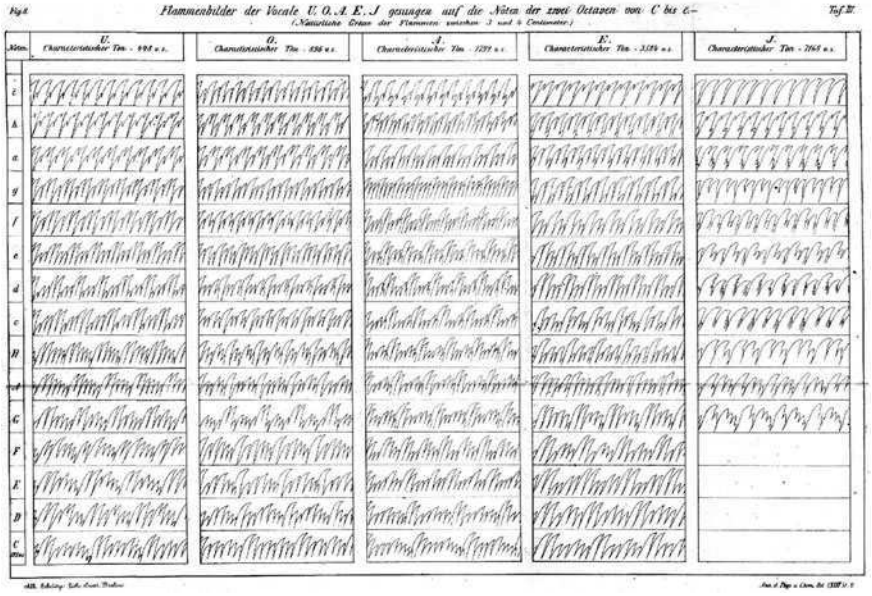


Fig. 3.3 At *top*, a photograph of the equipment set-up for creating manometric flames [40]. At bottom are Koenig's meticulous drawings of the flames generated by five vowel sounds sung on each of 15 musical notes

decades of the nineteenth century. The best concise description comes from Edward Scripture:

> The vowel is sung or spoken into a trumpet leading to a small box known as the 'manometric capsule.' This box is divided in two parts by a thin rubber membrane … One part is a tight chamber through which illuminating gas is flowing; the gas is lighted at the end of a small jet. As the sound waves descend they strike the rubber membrane, set it in vibration and thus produce movements of the gas analogous to those of the air in the sound waves [51] p. 26.

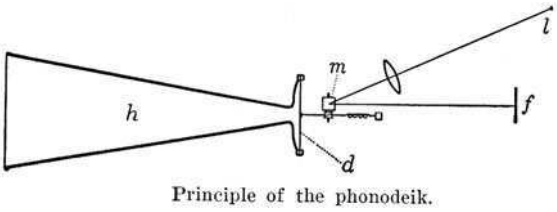
The rapidly changing flames were best viewed with the aid of revolving mirrors, and at first could only be recorded by hand-drawing. Figure 3.3 shows an old picture of the apparatus, together with flames produced from five French vowels sung on 15 musical notes.

An important development in the recording of sound was the invention of the microphone, which according to Miller [40] was due to Hughes in 1878. The original microphone transduced sound vibrations into electromagnetic waves, which could be received by the *oscillograph* developed by Blondel (published in 1893 by the Paris Academy of Science). The oscillograph responded to the electromagnetic waves as a galvanometer which vibrated a tiny mirror, which in turn was used together with a light source to produce a waveform of excellent quality recorded on photographic film.

Another means for recording a waveform photographically involved a device called the *phonodeik*, which was developed and described by Dayton Miller [40] (v. Figs. 3.4, 3.5).

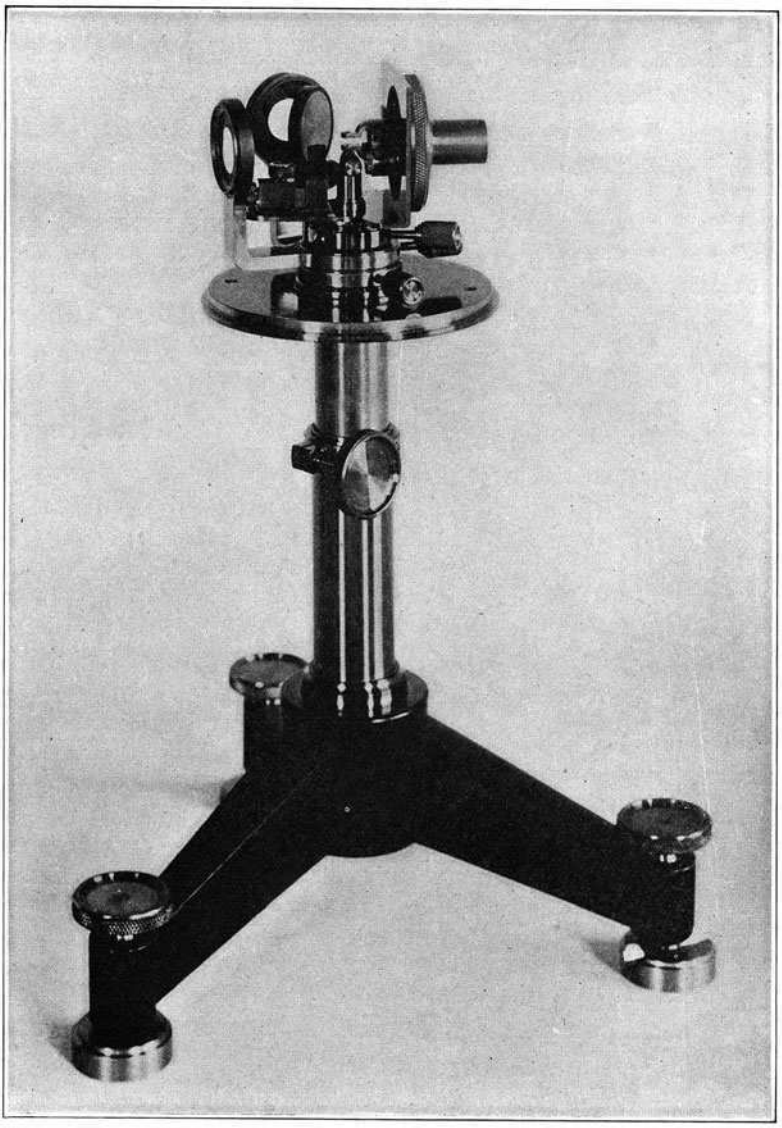
> The sensitive receiver of the phonodeik is a diaphragm, *d*, of thin glass placed at the end of a resonator horn *h*; behind the diaphragm is a minute steel spindle mounted in jeweled bearings, to which is attached a tiny mirror *m*; one part of the spindle is fashioned into a small pulley; a string of silk fibers, or a platinum wire 0.0005 in. in diameter, is attached to the center of the diaphragm and being wrapped once around the pulley is fastened to a spring tension piece; light from a pinhole *l* is focused by a lens and reflected by the mirror to a moving film *f* in a special camera. If the diaphragm moves under the action of a sound wave, the mirror is rotated by an amount proportional to the motion, and the spot of light traces the record of the sound wave on the film [40] p. 79.

The resulting 5-in. wide photographs of waveforms were of the highest fidelity obtainable at the time; the apparatus was sensitive to sound frequencies up to 10 kHz.



Principle of the phonodeik.

Fig. 3.4 Miller's [40] schematic of the phonodeik



The phonodeik used for photographing sounds.

Fig. 3.5 Photograph of a phonodeik [40]

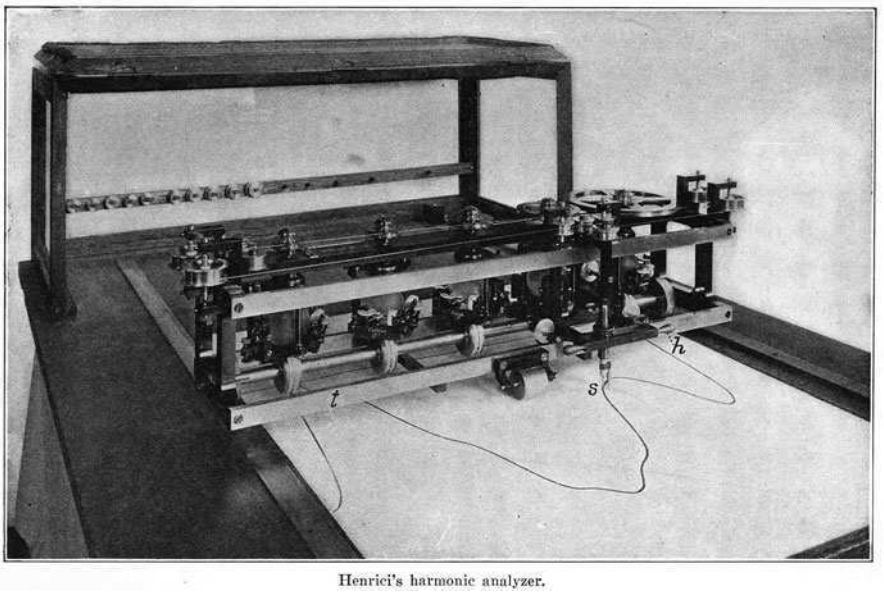
3.1.4 Early Methods of Fourier Analysis

In the earliest days of practical Fourier analysis, beginning in the late 1870s, investigators sought chiefly to compute the coefficients in a Fourier series that would serve to analyze a given naturally occurring periodic signal into a number of

harmonics. There quickly arose two types of methods for carrying out this task: using a machine of special design to perform what was in essence an "analog computation" of the harmonics and their amplitudes, or manually sampling a period of the signal and computing the discrete-time Fourier series coefficients by hand calculation. The special-purpose analog computers for performing the analysis were generally called *harmonic analyzers*. The first such instrument was designed by Thomson (Lord Kelvin) [55] for the computation of tidal frequencies, and was described in some detail in [56]. Harmonic analyzers worked by harnessing the natural relationship between trignometric simple harmonic motion and circular motion. The given natural periodic signal had to be presented as a waveform graph and carefully traced by a stylus, and the stylus in turn transferred components of its motion to a number of cylinders or spheres, each of which would output the amplitude of its particular harmonic. A harmonic analyzer sensitive enough to analyze a traced sound waveform was devised by Henrici in 1894, and thereafter perfected (v. Fig. 3.6) and frequently used for sound spectrum analysis. The mechanisms and operating principles behind Henrici's analyzer, as well as six other harmonic analyzers, were described in considerable detail by Carse and Urquhart [7].

Concerning the interpretation of the results provided by harmonic analyzers, which relates to the topic (discussed above) of the "reality" of Fourier transform components, Harry Hall [27] had this to say:

> Transient sounds represent continuous spectra whose amplitude and phase distributions are determined by the sound. These spectra are considered to exist throughout time, but



Henrici's harmonic analyzer.

Fig. 3.6 Photograph of a Henrici analyzer used in Miller's investigations [40]

> may be observed only during the lifetime of the original sound and are not fully indicated until the full sound has passed … If a Henrici analyzer is used to analyze selected waves taken from different parts of such a short sound, it will not give the true spectrum, but will give the line spectrum which would be produced if the particular wave in question were being repeated in a steady state.

These remarks represent an early attempt to define the notion of a "short-time" Fourier spectrum, meaning one which is valid only for a brief segment of a signal. We will later see how this notion was more firmly codified, so that it forms the foundation of the spectrogram.

The first investigators to compute the Fourier series coefficients (and thus a discrete spectrum) for a vowel sound were Schneebeli [50], who used a phonautograph to record speech waveforms, and Jenkin and Ewing [30], who used an Edison phonograph to trace the waveforms on tin foil (see above for some details). In the latter case the waveform was naturally bandlimited by the apparatus to frequencies below approximately 1,200 Hz, and was manually sampled 12 times per period, with the vertical (ordinate) value being measured to an accuracy of 0.005 in. The periodic motion was approximated as being composed of six harmonics. Jenkin and Ewing missed their opportunity to publish the first power spectrum graphs in a simple format, and instead presented their computed spectra as rather confusing tables of numbers.

In an appendix on "Practical Harmonic Analysis," Carslaw [8] summarizes the established methods of the early twentieth century for performing what is, in essence, a hand calculation of the discrete-time Fourier series coefficients using samples of a periodic signal. He then notes that "Runge gave a convenient scheme for evaluating these constants in the case of 12 equidistant points" in a single period, which was first published in 1903 and 1905. Standard forms which facilitated the implementation of Runge's method using either 12 or 24 sample points were made available by Whittaker and Robinson in the many editions of their book beginning in 1924 [57]. A different method of quick computation of the Fourier series coefficients, which improved upon the speed of Runge's techniques by using an approximation, was described by Thompson [54], and was thereafter used for speech spectrum analysis by Steinberg [52].

Although Runge's main method for computing the discrete-time Fourier series coefficients was quite efficient in exploiting trigonometric symmetries, cutting the number of required computations considerably, he had apparently given some hints which, when implemented, could further improve the efficiency of the scheme. Danielson and Lanczos [15] were seemingly the first to apply Runge's suggestions, implementing a matrix transformation which further improved the efficiency of Runge's more well-known scheme without resorting to approximation, effectively making the number of calculations proportional to $N \log N$ for N sample points or "ordinates". Their contribution was independently reinvented during the computer age, whereupon it was christened the Fast Fourier Transform [11]. So, the Fast Fourier Transform methods which render Fourier analysis relatively tractable for modern computers have surprisingly distant historical

origins—the general approach arose as a means of making "practical Fourier analysis" more tractable for calculation by hand!

3.2 History of Speech Spectra

3.2.1 Vowels and Formants Early Years

For many decades, speech science debated the question of which model of vowel production was more accurate, that of Robert Willis or that of Hermann Helmholtz.

> Willis [59] maintains two theses: 1. that a vowel consists of [at least] two tones, a cord tone and a mouth tone; 2. that the mouth tone is independent of the cord tone in regard to pitch. The theory of Willis was later taken up by Ludimar Hermann in the 1890s, but was criticised by Wheatstone, who supposed that the vowels arose from the vibrations of the vocal cords through the strengthening of certain overtones by the resonances of the mouth. Wheatstone's view was expounded as a general hypothesis by Grassmann and developed into a theory by Helmholtz [51] p. 407.

This debate, in the light of present understanding, has in essence been settled in favor of both positions, with each explaining a different aspect of a vowel sound. Indeed, the debate should have settled a long time earlier thanks to an authoritative treatment of the matter by Lord Rayleigh [53]; more will be said about that below. At any rate, the earlier view propounded by Willis reflects that the vocal cords indeed produce their own "tone," which is what is now called the voice source, by means of their periodic vibration. Then, as Willis had stated, each "puff" of air from the vocal cords excites the resonances of the mouth. We now know that the chief acoustic excitement comes from the pressure change that occurs when the cords close, rather than from the puff, which is more of an aerodynamic event than an acoustic one.

A vowel sound can also be profitably understood according to Helmholtz's view [29] that the periodic vocal cord vibration gives rise to a set of harmonics of the fundamental which are filtered by the vocal tract, thereby emphasizing certain harmonics over others. Helmholtz [29] measured what he called the "proper tones" of German vowels physically, by using tuning forks held up to the mouth configured for a particular vowel. He was able to find only one formant for the vowels [a, o, u], which was presumably the first formant, but he found that the front vowels [ɛ, e, i] "have each a higher and a deeper tone." It is reasonable to interpret his measurements as indicating he found the first and second formants of these vowels, which are much more widely separated than for the other three vowels.

Helmholtz credits Donders [16] with the first empirical finding that "the cavity of the mouth for different vowels is tuned to different pitches" [29]. Donders had measured by ear the main pitch of the noise produced by whispering vowels. Lloyd later refined an "analysis-by-synthesis" technique in which he drove a resonance bottle using a "hissing-tube" that produced a noise similar to a glottal whisper.

"On listening thus to the sounds issuing from the bottle it is soon found that they have often, and indeed generally, a more or less striking resemblance to whispered vowels …" [37]. Lloyd perfected the art of using such a device to determine formant frequencies, since the known dimensions of the adjustable resonators permitted these to be calculated, while the character of the artificial vowel could be altered to suit the experimenter's audition of a desired sound.

One of the first thorough investigations of vowels to include spectrum analysis was due to Jenkin and Ewing [30], who introduced the term *wave-form* for the tracing of the phonograph groove recording the sound. This term (without the hyphen) is used to this day in speech science to refer to the standard graph showing the output voltage of the microphone as a function of time. Their experiments focused on the English vowels [o] and [u] sung by six speakers. The vowels were sung on a wide range of notes of the musical scale, in order to see whether the relative amplitudes of the six partials characterizing a particular vowel would remain the same for different musical pitches.

Jenkin and Ewing found that the vowel spectra did not remain the same for different pitches at all. For the vowel [o], they noted a "specially strong reinforcement" of whatever upper partial was nearest to a particular frequency (representing the first formant of the vowel in this case), over a large range of the fundamental frequency. They struggled with strange data from the singing of [u], and from the nature of their lengthy discussion it can reasonably be inferred that they observed many instances in which the low first formant of the vowel was in close proximity to the fundamental. Additionally they found "that the experiments have given thorough confirmation of Helmholtz's discovery that vowel quality is not dependent on phase relation, so long as the constituent tones are unchanged." Their concluding remarks include the following remarkably prescient passage:

> In distinguishing vowels the ear is guided by two factors, one depending on the harmony of a group of relative partials, and the other on the absolute pitch of the reinforced constituents … We are forced to the conclusion … that the ear recognises *the kind of oral cavity* by which the reinforcement is produced. . . The vowel-producing resonance cavities are clearly distinguished in virtue of two properties—first, the absolute pitch at which they produce a maximum reinforcement; and, second, the area of pitch over which reinforcement acts. The latter property, when it is extensive, is very probably due to the existence of subordinate proper tones not far from each other in pitch [30] pp. 772–773.

Merritt [39] developed a method for photographing manometric flames, and was able to successfully measure the fundamental frequency of a sung vowel [a] across a wide range. The fundamental and harmonics appear to correspond to the individual small flames in the photographs (v. Fig. 3.7), with the harmonic series (up to around 1,000 Hz or so) visible numerous times as a repeating pattern from left to right. He measured the first formant of vowels by noting which harmonics were emphasized by more prominent flames in the photographs. He was thus able to measure the first formant frequency of [a] as averaging 736 Hz, a value in good agreement with modern measurements.

Scripture observed, in all recordings of vowels studied, at least one "tone of constant pitch—independent of the tone on which it is sung." According to

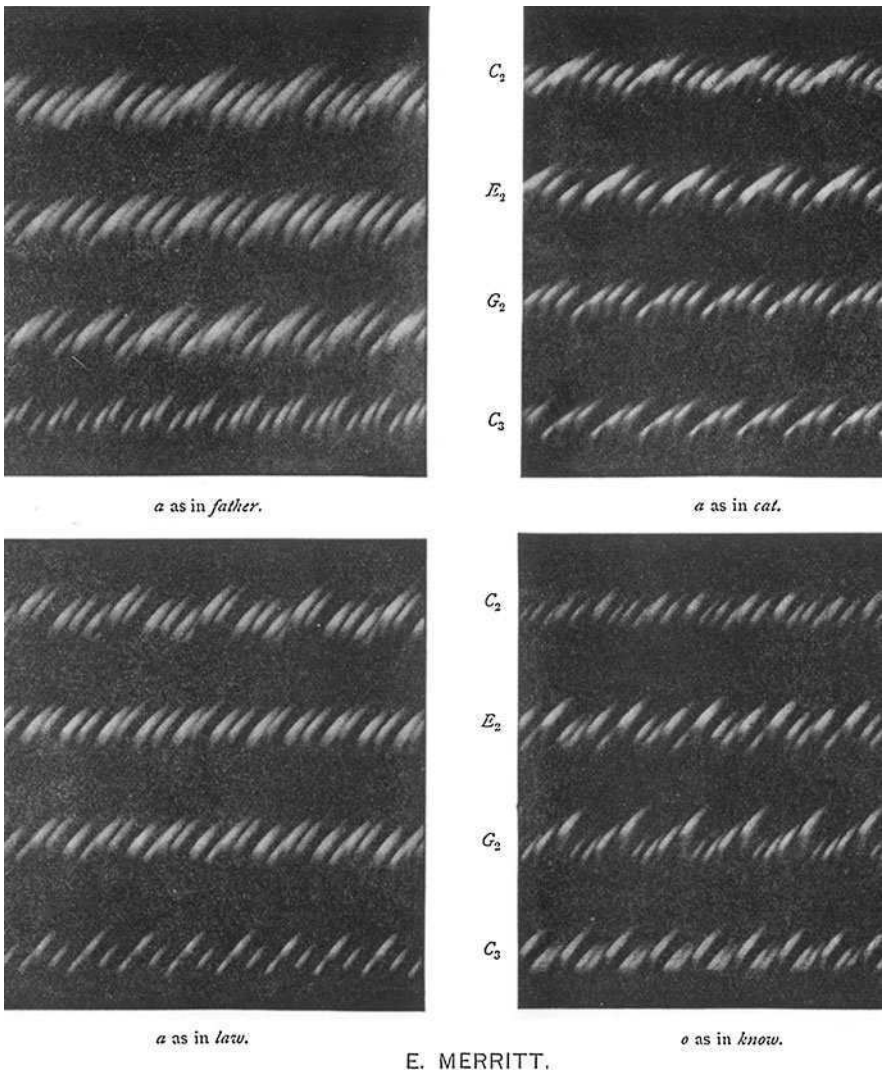


Fig. 3.7 Manometric flames photographed by Merritt [39], showing four vowels each sung on four notes

Scripture, this characteristic tone of a vowel was christened with the term *formant* by Ludimar Hermann.[2] The following quote is representative of Scripture's general understanding of speech production:

[2] I have been unable to ascertain which of Hermann's many papers first uses this term, but it was certainly published around 1890.

> It seems evident that in most cases the cords act by emitting a series of more or less sudden explosions that set the air in the resonance cavities in free oscillation. The periodical changes from strong to weak in these oscillations produce the cord tone as heard, just as a series of sharp noises from a card held against a toothed wheel or puffs from a siren will produce a note. The groups of similar vibrations indicate separate puffs from the cords... The cavity tones in a vocal sound probably always include more than two tones. These tones change with the constantly changing shapes of the cavities and probably never remain constant.

Scripture noted that Hermann had measured only one formant for most German vowels, but had found two formants for three of the vowels.

The question of how many formants were needed to characterize a vowel sound, and the manner in which they do so, was among the most important in acoustic phonetics during the late 1800s. It was Lloyd [36] who first championed the idea that more than one formant was essential for a correct account of vowel acoustics, noting that the different resonances which would be produced by speakers of differing sizes can then be understood to determine the same vowel by the prime importance of their relation, rather than their absolute frequency values. Scripture appeared to agree with Lloyd's general ideas about vowel formants, but it took some decades before the ideas were universally held. For example, a news article which appeared in Nature in 1901 [41] virtually ignored Lloyd's publications on vowels, while discussing at length the various theories of vowels and formants under the assumption that there is only one formant for a vowel.

Scripture made his own measurements of vowel formants, but did not seem to find the values worth publishing much of the time. His materials often consisted of waveforms traced from the grooves of grammophone records (i.e. discs rather than phonograph cylinders) of natural speech. He computed the harmonic series of naturally occurring vowels and other speech sounds using the above-mentioned manually calculated discrete-time Fourier series method initiated by Jenkin and Ewing, though he employed Hermann's specific methodology published in 1890. In his book [51], Scripture did report finding a lower formant for [a] at around 675 Hz, and also an upper formant around 1,150 Hz.

Scripture was also among the first to confront the conundrum of the lowest resonance found in voiced speech, or what is now called the *voice bar*. He suggested that he measured some low resonances which "may be trachea tones," and relied on suggestions made earlier by Pipping [45].

> Pipping considered as chest tones the low ones found in the neighborhood of the note 250 [Hz] in a series of Finnish vowels. The lower resonance tone of constant pitch found in a number of cases of [a] in [ai] ('I, eye,' etc.) may possibly arise from the chest instead of the mouth and pharynx... [51] p. 294.

We will see later that the voice bar remains something of an enduring mystery that was "swept under the rug" for a long time.

Scripture spilled some ink criticizing the Helmholtz vowel theory, and his concerns largely stemmed from a misunderstanding that seems to have been endemic in the era. Firstly, Helmholtz already recognized that if a vowel's

resonances act to reinforce harmonics emanating from the vocal cords, then the particular number of the harmonic which is emphasized must vary as the fundamental frequency varies. In his own (translated) words, "when I sing the vowel [a] on the note $e\flat^1$, the reinforced tone is... the 12th partial, and when I sing the same vowel on the note b^2 it is the second one." Secondly, Helmholtz mistakenly believed that the above fact necessitated that the mouth shape must accommodate itself to one harmonic, and when this changes to a certain extent due to changes in the fundamental, then the mouth must readjust itself to some other harmonic. This idea came to be called the "accommodation theory" of vowel production, and the need for such a strange aspect of speech gestures caused Scripture and others great concern. We now know that no such deliberate accommodation is necessary, because it does not matter if the formant exactly matches any particular harmonic. This is a key reason why both the Willis and Helmholtz theories of vowels are correct, and this fact was not grasped by most practitioners early on. The root of the misunderstanding was the mistaken view that a resonator had to be driven precisely at its resonance frequency in order to emphasize a harmonic component.

Insofar as Scripture favored the Willis–Hermann vowel theory completely, many of his proposed modifications to the theory are essentially correct. "I would amend the Willis–Hermann view by saying that the cords emit a series of puffs whose nature may vary from the sharpest of explosions to a perfectly smooth sinusoid. I would also add … that the character of the sound emitted by the cords depends essentially on the nature of the puff." Yet it is surprising that Scripture and others were still debating these matters (and would still for decades more), when any conflict between the Willis and Helmholtz vowel theories had already been shown to be illusory by Lord Rayleigh [53] in his typical correct and conclusive fashion:

> [Helmholtz's] view of the action of a resonator is of course perfectly correct; but at first sight it may appear essentially different from, or even inconsistent with, the account of the matter given by Willis. For example, according to the latter the mouth-tone may be, and generally will be, inharmonic as regards the larynx-tone. … Although the *natural* vibrations of the oral cavity may be inharmonic, the *forced* vibrations can include only harmonic partials of the larynx note … From these considerations it will be seen that both ways of regarding the subject are legitimate and not inconsistent with one another [53].

That the treatment of vowels in the second edition of Rayleigh's magnum opus *Theory of Sound*,[3] now probably the most currently read physics book from the nineteenth century, could have been so perfectly overlooked by an entire community of speech investigators from all backgrounds is surely one of the most nagging historical puzzles connected with speech analysis. Alas, the resolution of this puzzle today must be left for historians of science.

[3] The treatment of vowels was apparently not in the first edition of 1878, though I have not seen this version firsthand.

3.2.2 Vowel Spectra 1915–1960

Miller [43] obtained vowel waveforms using his phonodeik described above, and then obtained harmonic spectra from the waveforms with a specially constructed Henrici analyzer. His results were originally published in 1916. By drawing smooth curves over the harmonic spectra (v. Fig. 3.8) he was able to locate the first two formants of most of the English front vowels, and the values he obtained for F_1 at least are generally plausible. Owing to deficiencies of his equipment, Miller could find only the first formant of English back vowels, thereby propagating the earlier conclusions of Helmholtz and Scripture, to the effect that back vowels generally have one important formant and front vowels have two. Miller disagreed with Scripture's position on the vowel theories, and advocated the Helmholtz view as the most in accord with his experimental results. In contrast to many precursors, Miller apparently understood that the application of this theory required no special "accommodation" of the mouth to match the formant to a harmonic, but still the seeming ignorance of Lord Rayleigh's resolution of the matter is puzzling in the light of Miller's physics background.

Significant advances in speech spectrum analysis were made by Crandall and Sacia at Bell Labs, and published in a series of papers [12–14]. They used an improved oscillograph to obtain photographic records of speech waveforms, and then obtained Fourier spectra using a photomechanical harmonic analyzer designed by Sacia [48]. Although Crandall and Sacia began with the assumption (taken from Miller and also Fletcher [20]) that English back vowels were characterized by a single formant, they obtained more accurate results than any previous, and ultimately Crandall realized that all the vowels they recorded showed at least two important formants [13]. In fact, they also discovered the importance of the low F_3 in the English rhotic vowel [ɝ], and related its appearance to the retroflex tongue posture. They interpreted their results as favoring a "harmonic" theory of vowel production (i.e. the Helmholtz view), but also suggested that the inharmonic theory (i.e. Willis–Hermann) may also be needed as a complement to it, pointing to different decay properties of formants which were transiently excited by each vocal cord cycle (as Willis had originally asserted). The discussion by Crandall and Sacia is the first demonstrated understanding in the literature that, not only are the two models of vowel production equally correct, but that each model is needed to understand the disparate aspects of speech production. They also hold the distinction of being the first to publish power spectrum graphs of speech sounds in the format commonly used to the present day (v. Fig. 3.9).

Steinberg [52] analyzed vowels and other speech sounds by a discrete "fast" approximate Fourier transform computation owing to Thompson [54], and described the vowel spectra as consisting of harmonic partials which showed at least two regions of reinforced amplitude, one below 1,000 Hz and one between 1,000 and 2,000 Hz. Clearly he was able to locate the first and second formants by examining the spectra of harmonics. To record speech, Steinberg used a condenser microphone and an oscillograph which had a flat response up to 10 kHz to create an

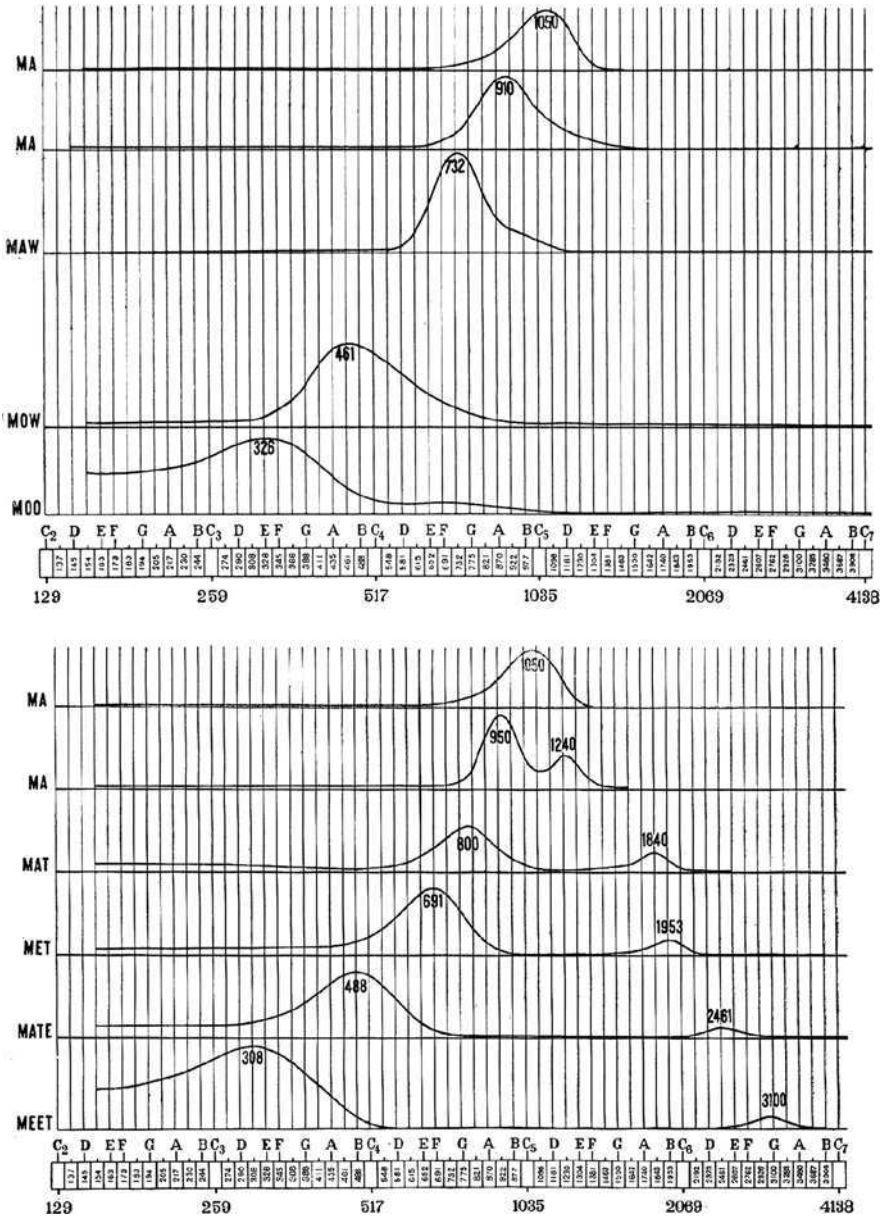


Fig. 3.8 Vowel spectra from Miller's book [40]. Characteristics curves for the distribution of the energy in vowels of Class I, having a single region of resonance. Characteristics curves for the distribution of the energy in vowels of Class II, having two regions of resonance

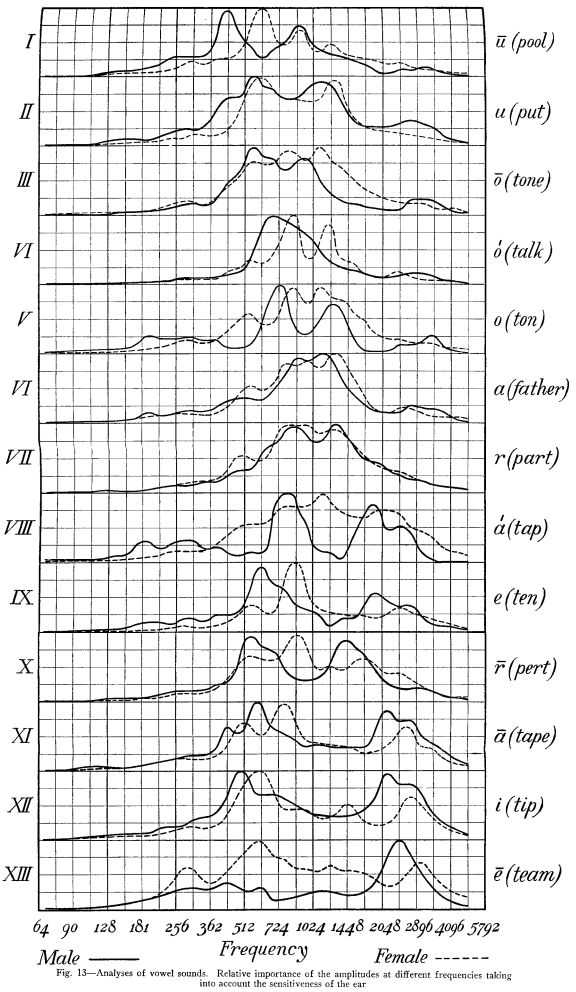
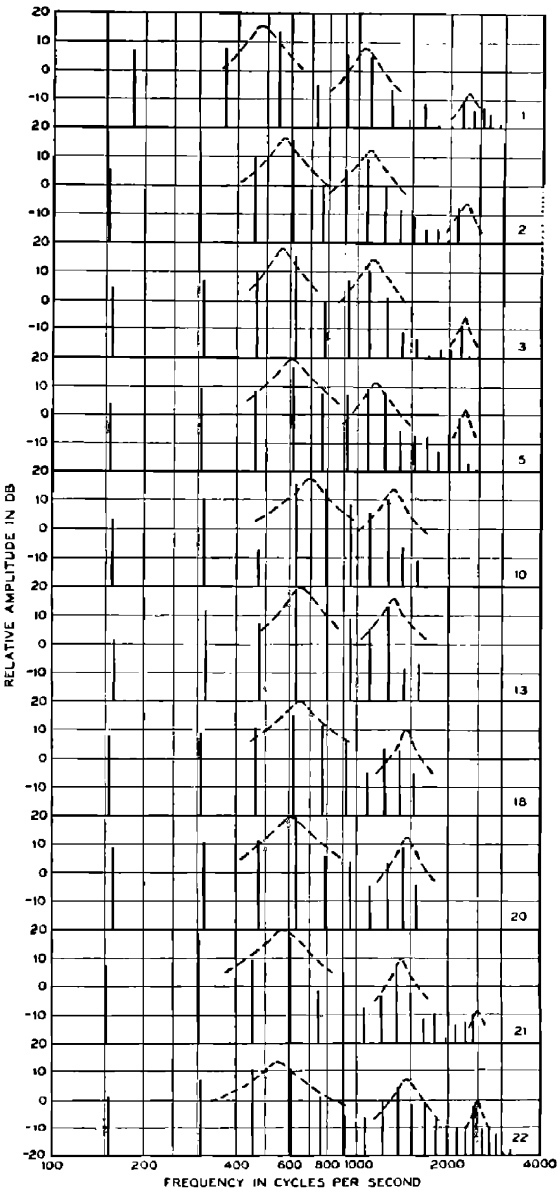


Fig. 3.9 Power spectrum graphs from Crandall [12] for 13 English vowels

oscillograph record of the waveform of the sentence "Joe took father's shoe bench out." Individual periods of the waveform were analyzed by a (hand-calculated) discrete-time Fourier series using up to 60 sample points over a single period. The power spectra thus calculated did reveal formants, and were laid out in a sequence to show the changes in the spectrum over time—Steinberg was the first to publish time-varying spectra, and his display technique is truly a forerunner of the spectrogram (v. Fig. 3.10). Steinberg presented his formant measurements for all the vowels in the utterance, and compared his values to those obtained previously by Crandall and Sacia, and also to those obtained by Paget [43] through analysis of synthetic vowels produced with a physical vocal tract model. His analysis method did not have a frequency resolution sufficient to resolve closely spaced formants of [u], [ʊ], which had been successfully resolved by Paget using synthetic speech.



Analysis of successive periods of *a* in "father's."

Fig. 3.10 A figure from Steinberg [52] showing a time sequence of power spectra. The formant peaks are hand-drawn over the computed discrete spectra

The next big advance in speech spectrum analysis was made by Lewis [35], who made oscillographic records of about a half-second long from a variety of vowels in sustained sung notes. He then traced the oscillographs over selected periods with a 40-component Henrici analyzer, yielding, approximately, the

Fourier series spectrum of the first 40 harmonics. The singer had been instructed to sing the vowels using his normal vibrato technique, with the result that different periods of the oscillograph had slightly different fundamental frequencies. For each vowel, a number of periods at different frequencies were selected for analysis, with the goal of checking whether the voice source and "filter" were independent. Indeed, Lewis published graphs showing the resulting vowel spectra which were created by overlaying the harmonic analyses from a number of different phonation periods at different fundamental frequencies (v. Fig. 3.11). No matter the

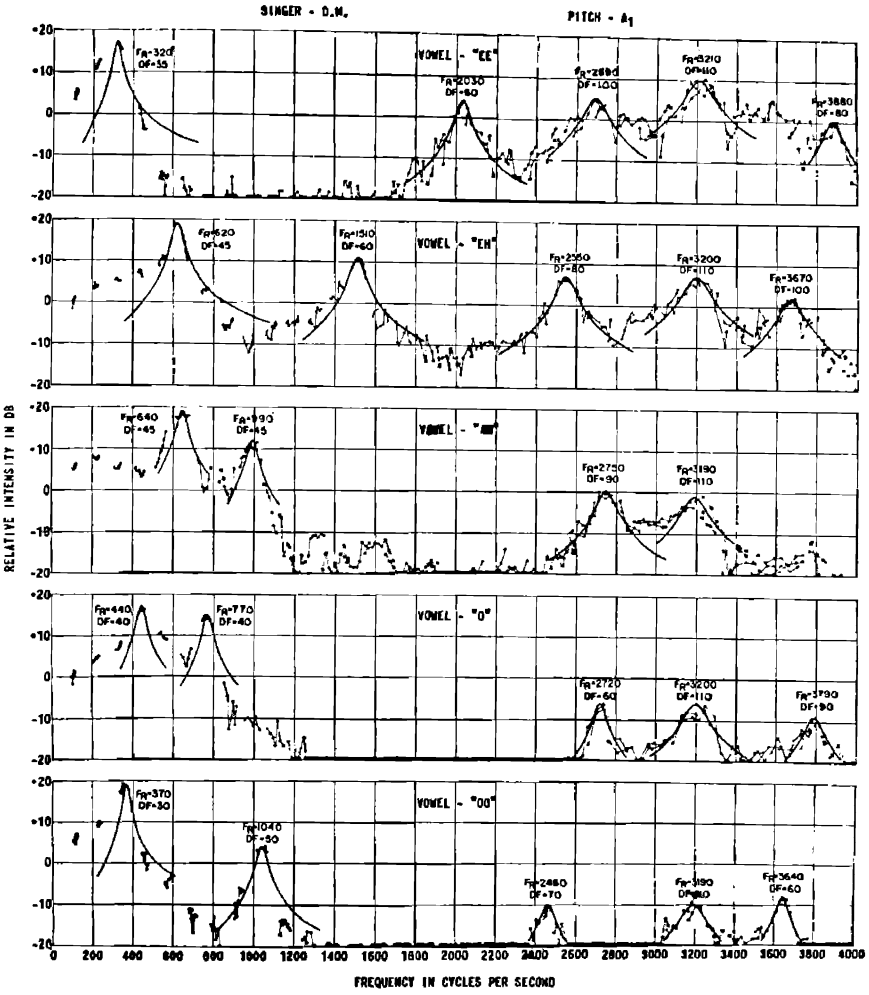


Fig. 3.11 A figure from Lewis [35] showing vowel power spectra drawn with a harmonic analyzer. Formant peaks are hand-drawn over the computed spectra

particular frequencies of the harmonics, their spectral amplitudes followed the same overall outline, which Lewis took to be the resonance spectrum of the vocal tract.

His spectra showed frequency components up to 4 kHz, and four or five formants were clearly visible for each vowel. F_1 and F_2 for [a] were measured to be 675 and 1,150 Hz, respectively. Other vowels' formant values are equally plausible by modern standards. Lewis's landmark study was the first to demonstrate that there are generally five formants for any vowel, and also that the amplitude differences between formants play an important role in the vowel quality. He imagined that each formant was the output of a simple resonator represented in the vocal tract somehow. He also proposed the prescient conjecture that "the *typical* quality of a vowel is determined by the two resonators of lowest frequency, with individual voice differences resulting from the action of the other resonators." Because he studied vowels in a singing voice, Lewis also discovered (without realizing it) what is now often called the singer's formant, which manifested as a 3,200 Hz formant present in most every vowel in the analyses.

Chiba and Kajiyama [10] presented power spectra of the harmonics in a variety of Japanese vowels, though they did not describe their method of computing the spectra. As far as interpreting their results, the nineteenth century debate over the best theory of vowel production was regarded as still unresolved by Chiba and Kajiyama [10]. They were led to follow Scripture in favoring the "inharmonic theory" of Willis and Hermann because they frequently found formant frequencies which did not coincide with any harmonic in their calculated spectra. The nature of this finding and the way it is interpreted reflects the very same misunderstanding of the action of a resonator that had befuddled Scripture before them.

Harvey Fletcher [21, 22] was, to his credit, possibly the first prominent speech scientist to understand that these two competing models of vowel production are not incompatible, and are both correct. It was seemingly Fletcher who first called the Willis theory "inharmonic," since it postulates that the formant frequencies are excited directly by each glottal pulse in turn, and so need have no relation to the series of harmonics of the glottal fundamental frequency. The Helmholtz theory (which had originated with Wheatstone) was then dubbed the "harmonic" theory, since it postulates that the series of harmonics of the glottal fundamental are selectively emphasized, possibly in groups, by the formant frequencies which are in fact resonances. Fletcher made it clear that a formant resonance can act to emphasize harmonics emanating from the voice source no matter whether the formant frequency coincides with any one of them, thus obviating the need for any previously supposed "accommodation" of the vocal tract to the harmonics. This new understanding of the physical basis of vowel acoustics would eventually be called the *source-filter theory*.

Vowel sounds had already been described, in essence, in terms of a source-filter theory by Lewis.

> This theory, which might well be called the cord-tone-resonance theory, states in effect that the vocal cords, during phonation, set up in air immediately adjacent to them a

> complex motion which consists of a fundamental component and a large number of its overtones. This complex motion constitutes the so-called cord-tone. The theory further states that the vocal cavities, on which the cord-tone acts as a force, have the properties of simple resonators and thus serve to modify the spectrum of the energy flowing from the cords. In terms of this theory, a vowel sound, as emitted from the mouth, is due to both selective generation and selective transmission … and it is composed *mainly* of a harmonic series of simple motions, each of which has a determinable magnitude [35].

Lewis sounds like he is summarizing someone else's theory, but he gives no citation; as far as I can tell his discussion is a synthesis of his own devise. The ideas trace back to Wheatstone, but Lewis's summary gives an excellent (albeit early) account of what would come to be accepted as the source-filter theory of speech production, particularly after it was presented to the linguistic phonetics community by Joos [31].

3.2.3 Spectrographic Analysis

Just as speech spectrum analysis was coming to be better developed in both the analog and discrete-time regimes, providing a picture of the distribution of speech energy across the frequency range, the *spectrograph* device was developed as a means of providing "a permanent visual record showing the distribution of energy in both frequency and time" [33]. While the *spectrogram* image output by the spectrograph is now usually described as a logarithmic plot (in dB) of the squared magnitude of the short-time Fourier transform (to be presented in detail in the next chapter), in fact the analog spectrogram bears a closer affinity to other analog signal processors like the Henrici analyzer, and predated the explicit specification of this transform by more than 20 years. The first "sound spectrograph" device was developed at Bell Labs shortly before World War II, was closely held following the outbreak of war, and was finally described in the open literature following the war's end [33]. At that point, the short-time spectra it output were only understood from the perspective of electrical filters, since the original spectrograph was an entirely analog electrical signal processing device.

The spectrogram would ultimately prove to be the most universally valuable representation for the analysis of speech spectra, because speech is such an inherently time-varying signal. It was shown in some of the earliest papers [34] that the time course of vowel formants could be readily imaged by setting the spectrograph filter to a "wide" bandwidth. Kopp and Green called the resulting images of formants "resonance bars," owing to their appearance as thick bands of dark grey on the spectrogram, an image now familiar to thousands of speech scientists around the world (v. Fig. 3.12). Such spectrograms effectively performed a very short-time spectral analysis due to the short impulse response of the filter, and it was possible to observe the impulses generated by the vocal cords as "vertical striations" through the frequency range. It was also possible to perform a "narrowband" analysis using the spectrograph, which yielded a spectrogram with

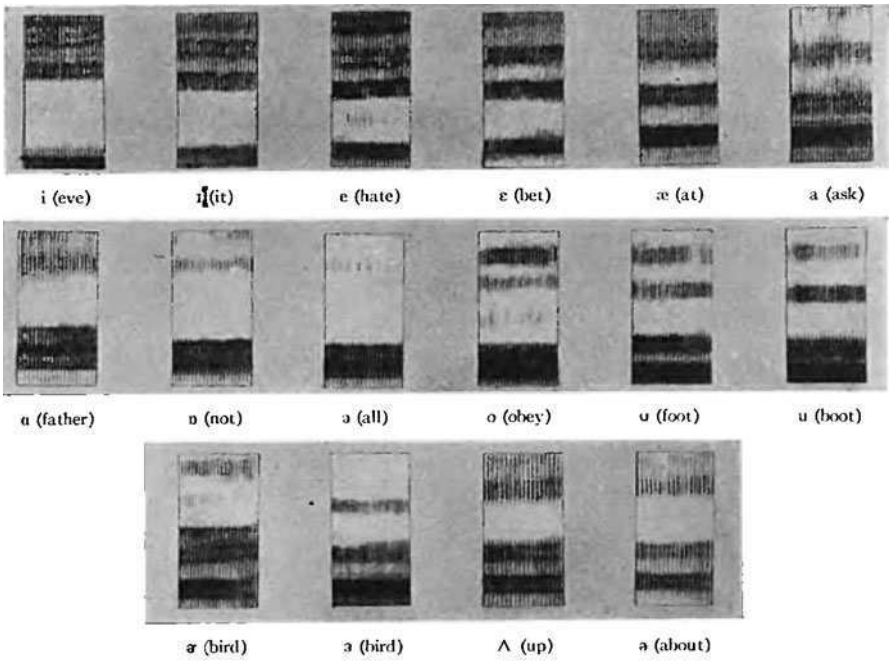


Fig. 3.12 Some of the first speech spectrograms ever published [34]

sufficient frequency resolution to show the fundamental frequency of voicing together with the harmonics; the longer impulse response of the narrowband filter made it impossible to simultaneously resolve the vocal cord impulses, however.

An important publication by Joos [31] helped bring spectrographic analysis of speech to the attention of the phonetics community, as did the book by Potter et al. [46]. In the ensuing years, the spectrographic analysis of speech became a standard approach in phonetics. An archetypical publication by Potter and Steinberg [47] delved into many issues of speech features in spectrograms, and was very instructive as to measurement methods. This paper also demonstrated how to exploit the spectrograph's capability of computing a single power spectrum analysis at a given time point, in conjunction with a spectrogram. Through the 1950s, the spectrograph's analyses became quite frequently used to measure speech features such as vowel formants (e.g. [18, 44]), but there remained debate about the best ways of making the desired measurements.

Near the end of the 1960s, thanks to the re-invention of the fast Fourier transform algorithm (discussed above), digital computers had become powerful enough to be able to compute spectrograms which could simulate the analog output of the spectrograph device. The general algorithm for a digital spectrogram (presented in the next chapter) was widely disseminated by several proponents, including Oppenheim [42]. In recent years, the procedure of making measurements "manually" from a spectrogram has fallen somewhat out of favor, and has been

gradually replaced by parametric spectral analysis, to be discussed in the next section. Current developments in spectrographic, or more generally *time–frequency* analysis, are nevertheless very promising, and it is not unreasonable to suggest that the new methods presented in the upcoming chapters may yet revive the perceived utility of spectrogram reading.

3.3 Parametric Spectral Analysis

By the 1960s, the "source-filter theory" of speech production was coming to the fore, and as a result the vocal tract began to be viewed as a resonant filter which was describable by a few parameters, in particular its resonance frequencies and their bandwidths and relative amplitudes. These in turn are compactly represented using the *transfer function* of the filter, which will be treated more thoroughly in Chap. 7. It is the modeling of the vocal tract as a linear filter which led to a number of related approaches to speech spectrum analysis that have been termed "parametric."

The *analysis-by-synthesis* scheme was outlined as follows by Bell et al. [2]. The speech was first sent through a "filter set," and the outputs were sampled digitally and stored. This yielded a basic spectrum estimate which was like a spectrogram. A "spectrum generator" was then used to synthesize speech according to a linear source-filter model, governed by parameters of a voice source and vocal-tract transfer function (i.e. formants) input by the user. A "comparator" then computed a measure of the difference between the input speech spectra (which are non-parametric but rather imprecise) and those generated by the spectrum generator. When a synthesized spectrum providing minimum error was obtained, the parameters of its transfer function and source characteristics were used to produce the parametric spectral estimate. Since this was a rather involved and cumbersome procedure, it was a priority to develop a means of estimating a vocal tract filter model from some input speech without actually performing analysis by synthesis, and thus a number of approaches to spectral estimation by analysis of the speech "time series" data (i.e. the waveform) were developed during the late 1960s.

In the field of time-series statistics that I have alluded to previously, many decades of the twentieth century and beyond have been spent on the development of models for time series data, and for statistically estimating the parameters of such models. One of the earliest models for certain kinds of time series was christened *linear prediction* by Wiener [58], but this idea appears to have its roots in work by Carl Gauss in 1795 [38]. With the advent of digital speech processing and the analysis-by-synthesis techniques just described, the notion arose that a digital speech waveform can be viewed as time series data, since it is in essence just a sequence of ordinate values at equally spaced time points. The application of linear prediction models to speech "time series" was initiated by Saito and Itakura [49], and also by Atal and Schroeder [1]. The method was developed into a standard approach during the 1970s, particularly through the seminal book by

Markel and Gray [38], who proved many mathematical facts about the various possible means of working with such models, while also teaching the community various algorithms for the method. A different statistical time series model known as the *maximum entropy method* was introduced by Burg [6], but this was soon recognized as yet another algorithm for linear prediction analysis [3]. It is left to the modern era to refine and further improve upon speech spectrum analysis, some of the methods for which are to be described in the upcoming chapters.

References

1. B.S. Atal, M.R. Schroeder, Predictive coding of speech signals. in *Proc. 1967 Conf. Commun. and Process* (1967) pp. 360–361
2. C.G. Bell, H. Fujisaki, J.M. Heinz, K.N. Stevens, A.S. House, Reduction of speech spectra by analysis-by-synthesis techniques. J. Acoust. Soc. Am. **33**(12), 1725–1736 (1961)
3. A. van den Bos, Alternative interpretation of maximum entropy spectral analysis. IEEE Trans. Inform. Theory **17**, 493–494 (1971)
4. H. Bouasse, *Acoustique Générale: Ondes Aériennes* (Delagrave, 1926)
5. L. de Broglie, *Certitudes et Incertitudes de la Science*. (Albin Michel, Paris, 1966)
6. J. P. Burg, A new analysis technique for time series data. in *Modern Spectrum Analysis*, ed. by D.G. Childers (IEEE Press, New York 1978), pp. 42–49. Reprint of a paper presented at the NATO Advanced Study Institute on Signal Processing with Emphasis on Underwater Acoustics, 1968
7. G.A. Carse, J. Urquhart, Harmonic analysis. in *Modern Instruments and Methods of Calculation*, ed. by E.M. Horsburgh (Tomash, Los Angeles, 1914) pp. 220–248 (Reprinted 1982)
8. H.S. Carslaw, *Introduction to the Theory of Fourier's Series and Integrals*, 3rd edn (Dover Publications, New York, 1930)
9. A. Cauchy, Mémoire sur la théorie de la propagation des ondes à la surface d'un fluide pesant. Mémoires des Savans Étranger, pp. 3–313 (1827). Reprinted in Cauchy's Oevres Complètes 1re série tome 1
10. T. Chiba, M. Kajiyama, *The Vowel: Its Nature and Structure* (Phonetic Society of Japan, Tokyo, 1958)
11. J.W. Cooley, J.W. Tukey, An algorithm for the machine calculation of complex Fourier series. Math. Comput. **19**, 297–301 (1965)
12. I.B. Crandall, The sounds of speech. Bell. Sys. Tech. J. **IV**, 586–626 (1925)
13. I.B. Crandall, Dynamical study of the vowel sounds, part II. Bell. Sys. Tech. J. **VI**, 100–116 (1927)
14. I.B. Crandall, C.F. Sacia, A dynamical study of the vowel sounds. Bell. Sys. Tech. J. **III**, 232–237 (1924)
15. G.C. Danielson, C. Lanczos, Some improvements in practical Fourier analysis and their application to X-ray scattering from liquids. J. Franklin Inst. **233**, 365–380, 435–452 (1942)
16. F.C. Donders, Über die Natur der Vokale. Archiv f. d. holländische Beiträge z. Natur- u. Heilkunde **I** (1858)
17. A.J. Ellis, Analysis and synthesis of vowel sounds. In: *On the Sensations of Tone* (*Helmholtz*) [29], pp. 538–543
18. G. Fant, *Acoustic Theory of Speech Production* (The Hague, Mouton, 1960). Reissued 1970
19. P. Flandrin, *Time-Frequency/Time-Scale Analysis*. English edn (Academic Press, San Diego, 1999)
20. H. Fletcher, The nature of speech and its interpretation. J. Franklin Inst. **193**(6), 729–747 (1922)

21. H. Fletcher, *Speech and Hearing* (D. van Nostrand, Princeton, 1929)
22. H. Fletcher, *Speech and Hearing in Communication*. (D. van Nostrand, Princeton, 1953)
23. J.B.J. Fourier, Théorie de la propagation de la chaleur dans les solides (1807). Manuscript first published in [26]
24. C. Godfrey, On the application of Fourier's double integrals to optical problems. Phil. Trans. R. Soc. Lond. Ser. A **195**, 329–362 (1900)
25. M. Gouy, Sur le mouvement lumineux. J. Phys. **5** (1886)
26. I. Grattan-Guinness, *Joseph Fourier 1768–1830* (The MIT Press, Cambridge, 1972)
27. H.H. Hall, Sound analysis. J. Acoust. Soc. Am. **8**, 257–262 (1937)
28. M.T. Heideman, D.H. Johnson, C.S. Burrus, Gauss and the history of the fast Fourier transform. IEEE Acoust. Speech Sig. Proc. Mag. **1**, 14–21 (1984)
29. H. Helmholtz, *On the Sensations of Tone*, 2nd English edn. (Longmans & Co., Oxford, 1885)
30. F. Jenkin, J.A. Ewing, On the harmonic analysis of certain vowel sounds. Trans. R. Soc. Edinb. **28**, 745–777 (1878)
31. M. Joos, *Acoustic Phonetics*. (No. 23 in Language Monographs. Linguistic Society of America, Baltimore, 1948)
32. R. Koenig, Die manometrischen Flammen. Ann. der Phys. Chem. **222**(6), 161–199 (1872)
33. W. Koenig, H.K. Dunn, L.Y. Lacy, The sound spectrograph. J. Acoust. Soc. Am. **18**(1), 19–49 (1946)
34. G.A. Kopp, H.C. Green, Basic phonetic principles of visible speech. J. Acoust. Soc. Am. **18**(1), 74–89 (1946)
35. D. Lewis, Vocal resonance. J. Acoust. Soc. Am. **8**, 91–99 (1936)
36. R.J. Lloyd, Some researches into the nature of vowel-sound. Ph.D. thesis (University of London, 1890)
37. R.J. Lloyd, Speech sounds: their nature and causation. Phonet. Stud. **IV**, 37–67 (1891)
38. J.D. Markel, A.H. Gray Jr, *Linear Prediction of Speech*. (Springer, Berlin, 1976)
39. E. Merritt, On a method of photographing the manometric flame, with applications to the study of the vowel A. Phys. Rev. **I** (1893)
40. D.C. Miller, *The Science of Musical Sounds*, 2nd edn. (MacMillan, New York, 1926)
41. J.G. M'Kendrick, Experimental phonetics. Nature **65**(1678), 182–189 (1901)
42. A. V. Oppenheim, Speech spectrograms using the fast Fourier transform. IEEE Spectrum (8), 57–62 (1970)
43. R.A.S. Paget, The production of artificial vowel sounds. Proc. R. Soc. Lond. A **102**(719), 752–765 (1923)
44. G.E. Peterson, H.L. Barney, Control methods used in a study of the vowels. J. Acoust. Soc. Am. **24**(2), 175–184 (1952)
45. H. Pipping, *Zur Phonetik der finnischen Sprache*. (Helsingfors, 1899)
46. R.K. Potter, G.A. Kopp, H.C. Green, *Visible Speech*. (D. van Nostrand, New York, 1947)
47. R.K. Potter, J.C. Steinberg, Toward the specification of speech. J. Acoust. Soc. Am. **22**(6), 807–820 (1950)
48. C.F. Sacia, Photomechanical wave analyzer applied to inharmonic analysis. J. Opt. Soc. Am. Rev. Sci. Instrum. **9**, 487–494 (1924)
49. S. Saito, F. Itakura, The theoretical consideration of statistically optimum methods for speech spectral density. (Technical Report 3107, Electrical Communication Laboratory, N. T. T., Tokyo, 1966) (in Japanese)
50. Schneebeli: Sur la théorie du timbre et particulièrement des voyelles. Archives des Sciences Physiques et Naturelles de Genève (1879)
51. E.W. Scripture, *The Elements of Experimental Phonetics*. (Charles Scribner's Sons, New York, 1902)
52. J.C. Steinberg, Application of sound measuring instruments to the study of phonetic problems. J Acoustical Soc Am **6**, 16–24 (1934)
53. J.W. Strutt (Baron Rayleigh), *The Theory of Sound*, vol. II, 2nd edn (Macmillan, London, 1896)

54. S.P. Thompson, A new method of approximate harmonic analysis by selected ordinates. Proc. Phys. Soc. Lond. **23**, 334–343 (1911)
55. W. Thomson (Lord Kelvin), Harmonic analyzer. Proc. R. Soc. Lond. **27**, 371–373 (1878)
56. W. Thomson (Lord Kelvin), P.G. Tait, *Treatise on Natural Philosophy*, vol. 1, 2nd edn. (Cambridge University Press, Cambridge, 1912)
57. E. Whittaker, G. Robinson, *The Calculus of Observations*. 4th edn (D. van Nostrand, New York, 1944)
58. N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series, with Engineering Applications*. (Technology Press of M.I.T., Cambridge, 1949)
59. R. Willis, On the vowel sounds, and on reed organ-pipes. Trans. Camb. Phil. Soc **III**, 231–268 (1830)

Chapter 4
The Fourier Power Spectrum and Spectrogram

This chapter covers the traditional speech analysis methods which rely on the discrete Fourier transform and its extension to the ubiquitous time–frequency representation known as the *spectrogram*. The first topic is the power spectrum of a signal window, which is derived from the magnitude of the Fourier transform in the manner explained in Chap. 2. Here, I discuss some of the methods for making power spectra of speech sounds, in an effort to show the best ways of accomplishing the desired imaging. Power spectra may be used to examine the formants of vowels and other resonant sounds, and when treated statistically they may also illuminate aspects of the noise produced during voiceless consonants. A third important application of power spectra is in the analysis and detection of different phonation types such as creaky and breathy voicing. Numerous figures provide examples of power spectra illustrating the points discussed in the text.

The second topic is the spectrogram; since this has not been treated properly here until the present chapter, the discussion is divided into first theoretical and then practical matters. The mathematical definition of the spectrogram is presented with an eye toward both its computational and historical aspects, a simplified algorithm for spectrogram computation is provided, and the ever-present problem of the uncertainty principle is discussed. Turning to more practical concerns of the speech scientist, guidelines are given for setting the various user-defined parameters of the spectrogram in order to obtain the best possible images for showing attributes of speech such as formants, transient consonant events and noise, and finally the fundamental frequency of voicing. The effects of the various parameter settings are illustrated in spectrograms of a variety of synthetic and natural speech sounds.

4.1 The Power Spectrum in Speech Analysis

What speech scientists normally call the *power spectrum* of a digital signal is simply the squared magnitude of the discrete Fourier transform defined in Eq. 2.35, graphed showing a decibel-scaled amplitude as a function of frequency

S. A. Fulop, *Speech Spectrum Analysis*, Signals and Communication Technology,
DOI: 10.1007/978-3-642-17478-0_4,

(v. Figs. 2.7, 2.8). Furthermore, it is important to note that in practice the frequency range of the graph is limited to show only the lower half of the positive frequencies computed by the DFT. As was explained in the previous chapter, neither the negative frequencies nor the upper half-range of the positive frequencies add any new information. This standard form of the graph is displayed automatically by the typical speech software (e.g. Praat) functions for obtaining and showing a power spectrum.

4.1.1 Vowel Spectra

In the previous chapter, the nature of vowels was discussed from an historical perspective on spectrum analysis. We concluded that according to current understanding, a vowel sound can be analyzed in the frequency domain as comprising harmonics of a fundamental (which is the frequency of the phonation) passed through resonant filters called *formants*. On the other hand, vowels and other voiced speech sounds have a second valid model, which is to simply say that the formants are excited by each glottal impulse during phonation. This is what we may call the "harmonic-inharmonic duality" of voiced speech, in analogy to the famous wave-particle duality of light that came to be recognized also during the early 20th century. Just as light can be detected as particles or waves, depending on the type of phenomenon examined, so voiced speech can be measured as having either a harmonic spectrum or an inharmonic spectrum, depending on the type of analysis performed.

A good benchmark from which to test the performance of many spectrum analysis methods is to examine synthesized vowels whose formants and other parameters can be known in advance. An excellent simple vowel synthesizer is included with recent versions of the Praat software [5], and this has been used to create the synthesized vowels examined in this book. The Praat synthesizer operates by invoking the source-filter model; beginning with a fairly simple glottal source vibrating with a user-input fundamental frequency, the algorithm computes the output from a virtual vocal tract having user-input values of four formants. The scheme is a typical implementation of what has come to be known as a "formant synthesizer."

To create a small corpus of synthesized vowel tokens, a number of vowels were synthesized using Praat to have English-like characteristics. These vowels have already been used for a study of formant measurement accuracy [12], but here they are recruited for a variety of tests and illustrations through the remainder of the book. Figure 4.1 shows two kinds of power spectra overlaid for each of three English-style vowels. The vowel [æ] is synthesized with formants $F_1 = 731\,\text{Hz}$ and $F_2 = 1,768\,\text{Hz}$; [i] has $F_1 = 306\,\text{Hz}$ and $F_2 = 2,241\,\text{Hz}$; [ɔ] has $F_1 = 602\,\text{Hz}$ and $F_2 = 884\,\text{Hz}$. Perceptual studies have generally shown that these two formants are by far the most important for determining the perceived vowel quality, so the remaining formants were set to $F_3 = 2,500\,\text{Hz}$ and $F_4 = 3,500\,\text{Hz}$ in all vowel tokens. The standard glottal source spectral amplitude rolloff employed by Praat

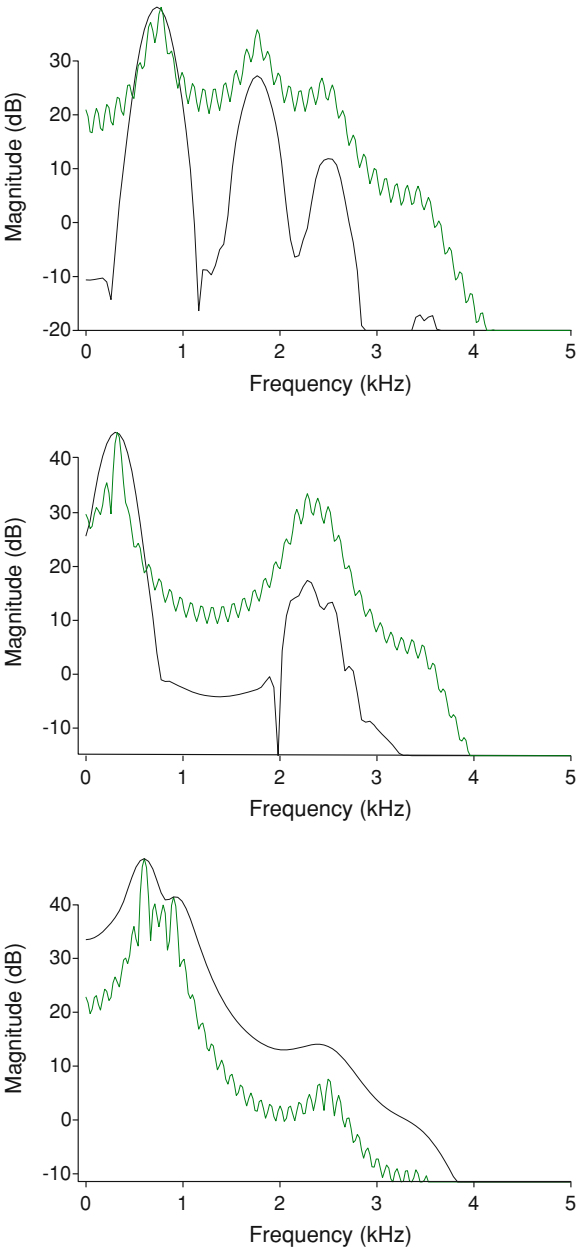


Fig. 4.1 Short (12 ms) and long (40 ms) window power spectra of synthetic vowels [æ] (*top*) [i] (*middle*) [ɔ], using a Gaussian window

makes the fourth formant almost absent from the sounds. All tokens were generated for approximately 300 ms using a glottal source having a frequency beginning at 120 Hz and descending at a rate of 2 octaves per second. This results in a fairly natural baritone falling intonation for English.

The "squiggly" spectra in Fig. 4.1 were computed in Praat with a 40 ms Gaussian window.[1] This window length encompasses several glottal impulses in the synthesized phonation, and as a result the Fourier spectrum shows the fundamental frequency of phonation and its many harmonics as the numerous little peaks in the graph. The harmonics are not well resolved here because the fundamental frequency changed slightly during the course of the analysis window, which normally happens during real vowels. This type of long-window spectrum is for historical reasons called "narrowband," and it is the right sort of frequency analysis to highlight the harmonic nature of vowels. The formants can be seen, just as Helmholtz had stated [18], as regions in the frequency range within which harmonics are emphasized.

The smoother spectra in Fig. 4.1 were computed using a 12 ms Gaussian window, meaning about 6 ms from the signal was effectively used, and so the window encompasses a domain smaller than one complete cycle of the glottal source. This makes it impossible to detect the frequency of phonation, so the single glottal impulse is analyzed as a transient which excites the formants. The window was centered on the portion of the signal waveform lying between glottal impulses, in order to focus on the resonance of the formants and deemphasize the spectrum of the transient impulse which is very broadband. This kind of short window spectrum is for historical reasons called "wideband," and it is the right sort of frequency analysis to highlight the inharmonic nature of vowels.

The three synthesized vowels of Fig. 4.1 were chosen because they represent three commonly encountered extremes of vowel quality. The vowel [æ] has its three major formants well separated in frequency, while [i] has F_2 and F_3 in close proximity, and [ɔ] has F_1 and F_2 in close proximity.

Vowel formants can in principle be measured from either the short or long-window power spectrum. A commonly used procedure at one time [26] was to hand-draw a smooth line around the harmonic peaks in a long-window spectrum, and in this way the actual peaks of the formants can be estimated. It is not wise to select a harmonic and measure its peak as that of a formant, since the two kinds of peaks will not coincide in general (this is a reminder that the Helmholtz "accommodation theory" is not correct). On the other hand, in Fig. 4.1 it is quite plain that the closely spaced F_2 and F_3 of the vowel [i] cannot be discerned at all in the long-window spectrum, while they are able to be measured from the short-window spectrum. This brief example demonstrates that, for synthesized vowels, a short-window power spectrum gives a better look at the formants and enables their estimation from the peaks with relative ease.

Turning our attention to vowels occurring in natural speech, it quickly becomes apparent that each different spectral analysis seems to provide conflicting

[1] It should be noted that a Gaussian window function is strongly tapered, and so the computed spectrum is mostly derived from the center half of the window length. Praat has a number of features intended to compensate for this, to be described in the appendix.

information about the vowel. In particular, the problem of formant estimation cannot so easily be solved by a single power spectrum. Figure 4.2 shows two power spectra from the [i] of *heed*. Both spectra are computed from 12 ms Gaussian windows; the key difference is the position of the window with respect to a phonation cycle. The dark line uses a window centered between glottal impulses, and appears to give a much more usable result, although there nevertheless appears to be a low peak at around 140 Hz that cannot be attributed to a formant. The first peak above this is located at 400 Hz, and a weak peak is visible at 2,170 Hz on the left shoulder of a larger peak located at 2,440 Hz. The speaker is male, so these are at least plausible values of the first three formants for this vowel. The lowest peak is probably a glimpse of the fundamental frequency resulting from the window barely encompassing two glottal impulses.

The lighter line in Fig. 4.2 uses a window centered on a glottal impulse. The formant frequencies are greatly obscured in this spectrum, resulting chiefly from the tremendous amount of broadband energy present in the impulse itself. It takes a millisecond or more for the resonances to appear following the glottal transient excitation, and the spectrum of the excitation itself does not contain well-resolved formants.

Figure 4.3 uses a 41 ms Gaussian window, yielding a look at the harmonics of the voicing. It is plain to see that formants cannot readily be measured or easily separated in this spectrum. The second harmonic is very loud, and so one might guess that the first formant is nearly coincident with it, but its frequency of 287 Hz is not a very plausible formant value. The reinforcement of this harmonic probably results from the lower resonance band known as the *voice bar*, which will be the subject of a later discussion. A comparison of the short and long window spectra reveals some disagreement; there is no evidence in the long frame spectrum of Fig. 4.3 to support the presence of the 400 Hz formant that appears in Fig. 4.2. How could this be? We will see in Chap. 6 that the chief reason for this is the nonstationary nature of the speech signal even at the time scale of a single glottal

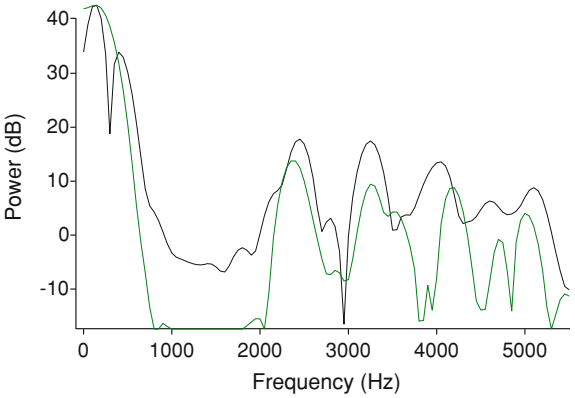


Fig. 4.2 Praat spectral slices with 12 ms Gaussian windows during steady [i] of *heed*. Phonation period is 7 ms. *Dark line* uses a window centered on the time between glottal impulses; *light line* uses a window centered on a glottal impulse

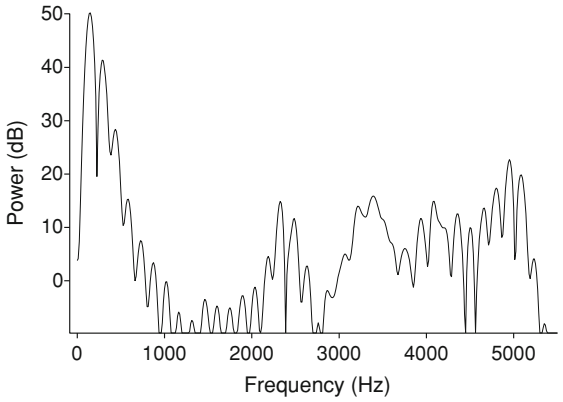


Fig. 4.3 41 ms spectral slice during steady [i] of *heed*. Phonation period is 7 ms

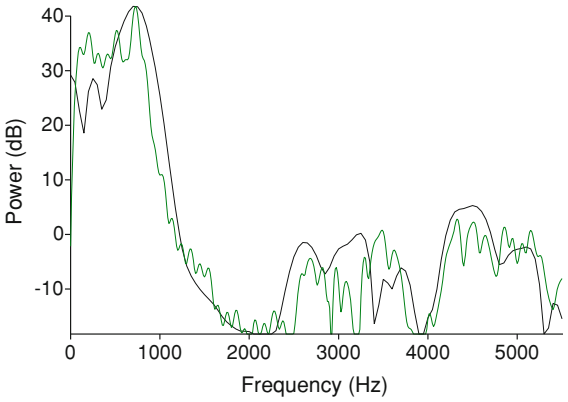


Fig. 4.4 Power spectra from 12 ms (*black*) and 41 ms slices in the English vowel [ɔ] of *hawed*

cycle. The long-window analysis provides a long-term spectrum in which the frequencies resulting from excitation of the formants has been obscured by the glottal impulse spectral information.

Figure 4.4 demonstrates that real speech can present seemingly insurmountable difficulties for measuring formants by simple spectrum analysis. The figure shows two power spectra computed from a 12 ms and a 41 ms Gaussian window during the vowel [ɔ] of *hawed*. With an average male speaker pronouncing this vowel, phonetic theory informs us that the values of F_1 and F_2 are in proximity and must both lie between 500 and 1,000 Hz. Sadly, neither type of power spectrum permits us to locate two formants in this expected range. The short-window spectrum does display an additional peak at 250 Hz, but this cannot be one of the defining vowel formants and is most likely a manifestation of the voice bar. The fundamental frequency of this vowel at the point of analysis is about 104 Hz.

4.1.2 Obstruent Spectra and Averaging Techniques

Obstruent sounds such as fricatives are chiefly characterized by noise, which is by definition a random signal. Normally the long-frame spectrum of such noise will betray the randomness by being completely "filled in" without any fundamental or harmonics apparent, although the noise spectrum of a speech sound is nevertheless shaped by the vocal tract. Example spectra of three English fricatives are shown in Fig. 4.5. That these spectra have different shapes partly accounts for our perception of these as different speech sounds in English, although consonants require being uttered in the context of a syllable in order to be most reliably perceived, and in that context other transitional acoustic features become important as well.

4.1.2.1 Spectral Moments

It has often proven difficult for phoneticians to describe such noisy obstruent spectra in a useful way. This is a result of the randomness; obstruent spectra are not as consistent from utterance to utterance as vowel spectra, and the features that are observable are rather vague, in the vein of noticing that the spectrum of [s] in Fig. 4.5 "has a broad peak around 8 kHz," or that the spectrum of [θ] "is very spread out" (e.g. [16]). These descriptions are not easy to quantify, so phoneticians have sought "metrics" of obstruent noise spectra that are more quantitative. One frequently used set of metrics for speech noise spectra are the *spectral moments*.

Since the power spectrum of a noise is statistically random, it is not unreasonable to treat it literally as a probability density function, which is a mathematical object characterizable by an infinite set of numbers called its moments. The first moment is generally equivalent to the simple mean of the spectrum, thereby providing a basic measure of location for the spectrum along the frequency axis. The second moment is equivalent to the variance (square of the standard deviation), thereby providing a rough measure inverse to how tightly the spectrum is crowded around its mean. Higher moments, if used, should simply be thought of as numbers which further quantify the power spectrum.

I do not wish to discuss the mathematical statistics behind the moments (for this see [33]) since the computation of spectral moments is really a data reduction procedure and not a spectrum analysis method per se; it is sufficient for our purpose here to note that Praat software, among others, has the ability to compute several moments characterizing any power spectrum. The idea of computing statistical moments to quantify noisy speech spectra dates back at least to Forrest et al. [11]. The method was used with some success in later studies of fricatives. Most of the studies which have used spectral moments concluded that the first two are by far the most important for characterizing obstruent power spectra (e.g. [21]).

Fulop et al. [14] used moments to classify the burst spectra of click sounds in Yeyi, and also discussed a number of statistical issues with the interpretation of spectral moments. Figure 4.6 shows example spectra of a dental click burst, which

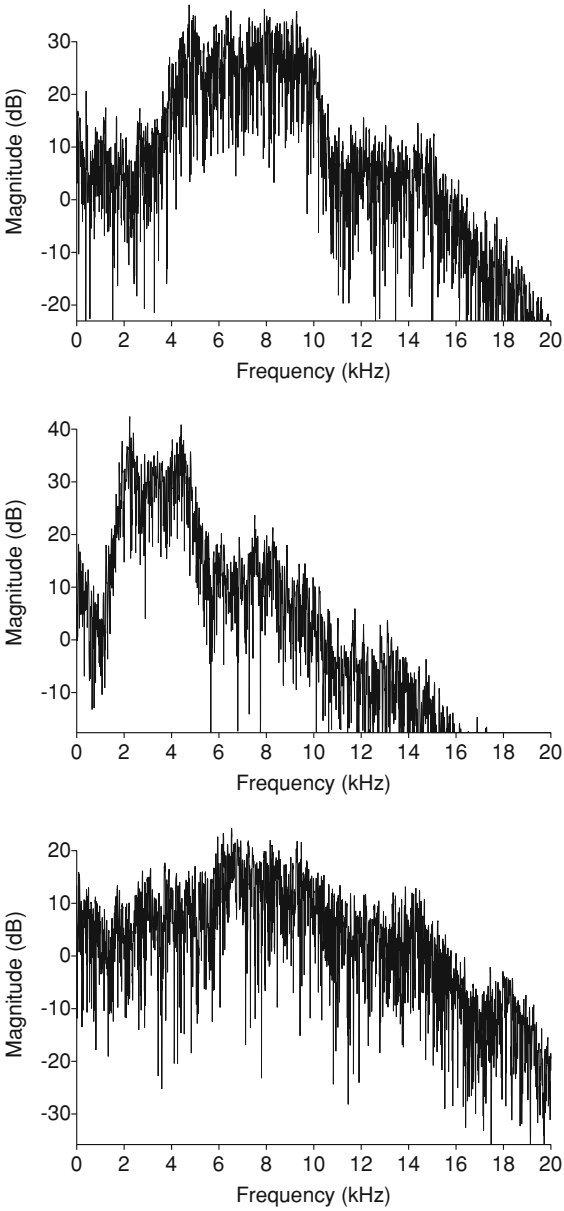


Fig. 4.5 Power spectra of English fricatives, entire length (150–180 ms). [s] (*top*); [ʃ] (*middle*); [θ]

also illustrates the importance of a good window function. The original study [14] used rectangular windows, which was likely not such a good procedure as it may have fatally affected the data to the point of altering the results. An alveolar click burst spectrum is shown in Fig. 4.7, which shows apparent differences from the dental. These differences were quantifiable for good classification using the moments alone, in spite of the rectangular windows.

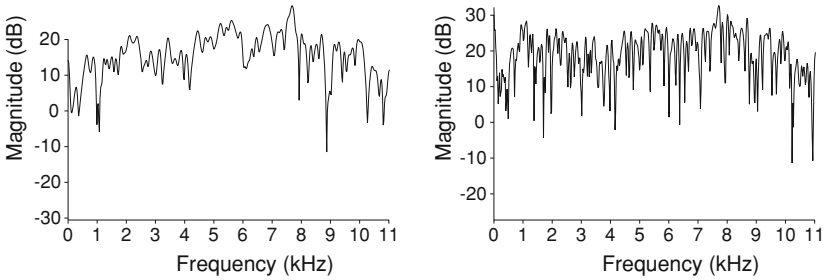


Fig. 4.6 Power spectra computed from a 26 ms dental click burst in Yeyi. *Left panel* uses a Gaussian window (which is the recommended procedure), *right panel* uses a rectangular window

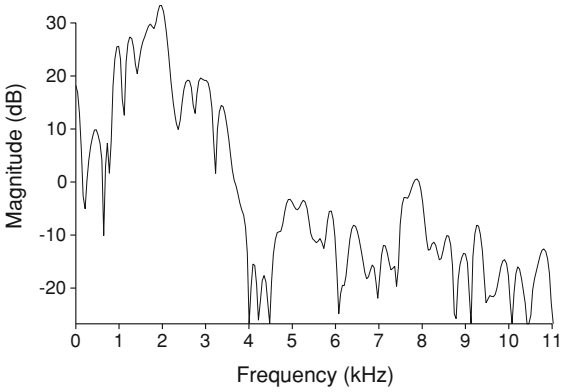


Fig. 4.7 Power spectrum computed from an 18 ms alveolar click burst in Yeyi, using a Gaussian window. Note the difference from the preceding figure showing the dental click

4.1.2.2 Time and Ensemble Average Spectra

Since obstruent noise is inherently random, one way to increase the resolution of pertinent information out of the randomness involves averaging the spectra of a number of windows on the same speech sound. It can be proven that applying this technique (which was first promoted in [35]) to a statistically random process will generally decrease the variance of the resulting average, and provide a more highly resolved picture of the spectrum. To give a conceptual analogy, if you wanted to check that 100 coin tosses yields an expected 50 heads, it is a good idea to perform the 100-toss experiment 10 times and average the results, since this average is much more likely to be very near to 50 than any single run of the experiment.

Two fairly simple ways of implementing this concept have been described for fricatives in the literature [30, 31]. One method is to clip out a number of (possibly overlapping) analysis frames from different time points within a single fricative token, and compute an average of their power spectra. The result is called a *time average* spectrum. An alternative approach requires a number of tokens of the same fricative repeated by the same speaker. We compute a spectrum for one analysis window per token and then compute an average spectrum from these. This is called

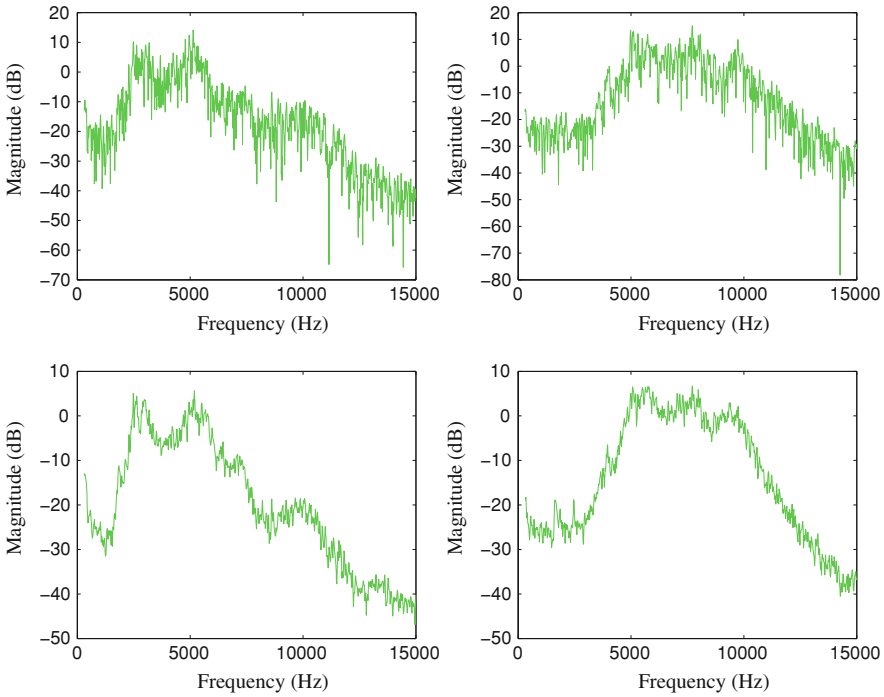


Fig. 4.8 *Top line*: power spectra of 55 ms windows from English fricatives [ʃ, s]; *Bottom line*: ensemble average power spectra of the same two fricatives, each using 10 repetitions from one speaker

an *ensemble average* spectrum, and is illustrated in Fig. 4.8. Average spectra will depend upon the stage at which the averaging operation is performed, and this is not always explicitly stated in the literature. The "raw" spectrum, being the result of a Fourier transform, is a vector of complex numbers, and it would not be meaningful to average these. In the linked Matlab code here, the averaging is performed on the squared magnitude of the Fourier transform, before conversion to a logarithmic (dB) scale. The results would look different if averaging was performed on the magnitude spectrum without first squaring, or after conversion to dB.

4.1.3 Phonation Types

Long frame power spectra showing voicing harmonics have proven to be very useful tools for detecting a key acoustic difference between different phonation types. A number of studies [24, 34] have demonstrated that, all else being equal (meaning the speaker and the other speech sound features), the relative amplitudes of the fundamental harmonic H_1 and the next two harmonics H_2, H_3 in the series usually distinguish two or more different phonation types when they occur in any

given language. Normally, creaky phonation is typified by the lowest amplitude of H_1 relative to H_2 and H_3; modal phonation displays a somewhat higher H_1 amplitude or somewhat lower H_2 or H_3 than in creaky phonation, while breathy phonation is characterized by the highest relative H_1 amplitude overall. These descriptions are illustrated in Fig. 4.9, showing creaky, modal, and breathy voice

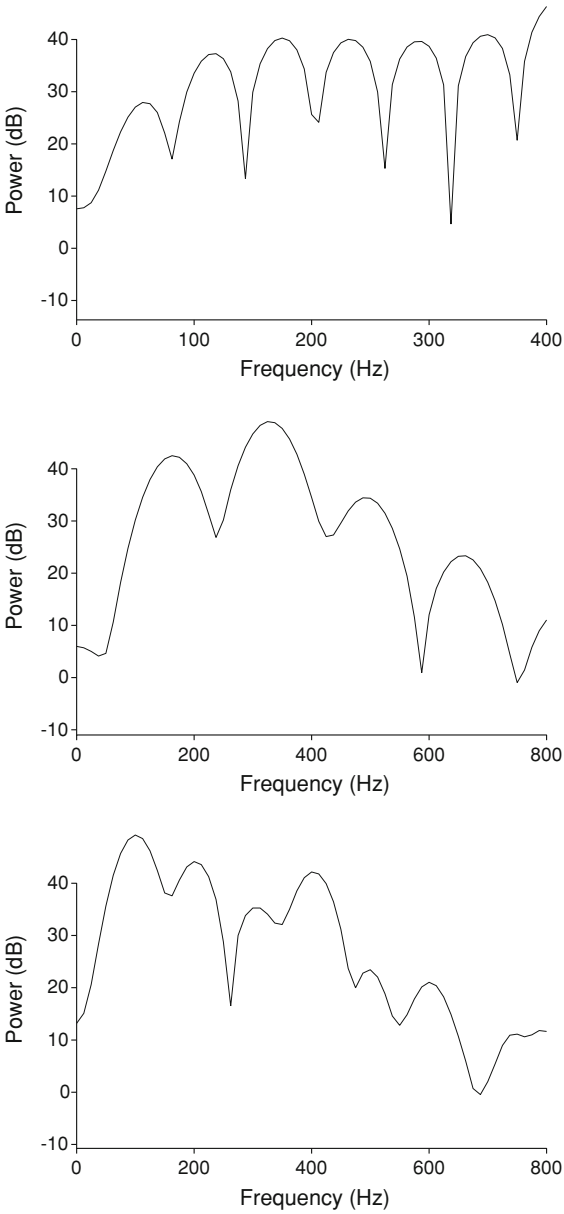


Fig. 4.9 Power spectra computed from 30–40 ms (approx.) Gaussian windows of the vowel [e], pronounced by the author in three different phonation types. *Top to bottom*: creaky, modal, and breathy voice

for the steady vowel [e]. The creaky example is characterized by high amplitudes of several harmonics above the fundamental; the modal example has a relatively high amplitude of H_2 alone, while the breathy example has H_1 as the loudest harmonic.

A recent small study [13] of the Hmong language has demonstrated that breathy and whispery phonation can also be distinguished by the relative harmonic amplitudes in similar fashion to the above. Example power spectra from that study are shown in Fig. 4.10, in which it can be observed that the modal vowel has a smaller relative H_1 amplitude than the breathy vowel, which in turn is less than that of the vowel following a whispery voiced stop release. These differences were found to be so robust in the Hmong words studied that they did not require statistical validation.

Another means of distinguishing phonation types uses a measure of the amplitude of all the harmonics relative to the noise in the speech. There have been a number of methods for computing this *harmonicity* metric (sometimes called the harmonics-to-noise ratio) presented in the literature; an excellent metric of this kind is computable using Praat software. The above mentioned study [13] of Hmong phonation types also demonstrated that the modal, breathy, and whispery phonations of Hmong can be distinguished by means of the harmonicity. It was found that the modal vowels usually had less harmonicity than the breathy vowels (which was surprising), while the whispery phonation had substantially less harmonicity than the modal (which was expected).

4.2 Principles of the Spectrogram

A *spectrogram* is a particular *time–frequency representation*, which is a function of both time and frequency that represents the energy distribution in a signal over the time–frequency plane. It is derived similarly to a power spectrum, from an extension of the Fourier transform to a joint time–frequency setting known as a *short-time* Fourier transform (sometimes also called the *Gabor* transform, which is strictly speaking not quite the same thing).

4.2.1 Definitions of the Spectrogram

The spectrogram will be defined intuitively here, with the details relegated to a math box below.

4.2.1.1 Sequence of Fourier Transforms

Before defining the spectrogram it is necessary to define the short-time Fourier transform (STFT). This can be understood as a time series of Fourier transforms,

where for each time point the Fourier transform is computed for a time-limited analysis window on the signal. The purpose of the window is to localize the Fourier transform's view of the signal to the vicinity of a particular point in time. The STFT $S_w(t, \omega)$ is then parameterized by the window function w (so there is

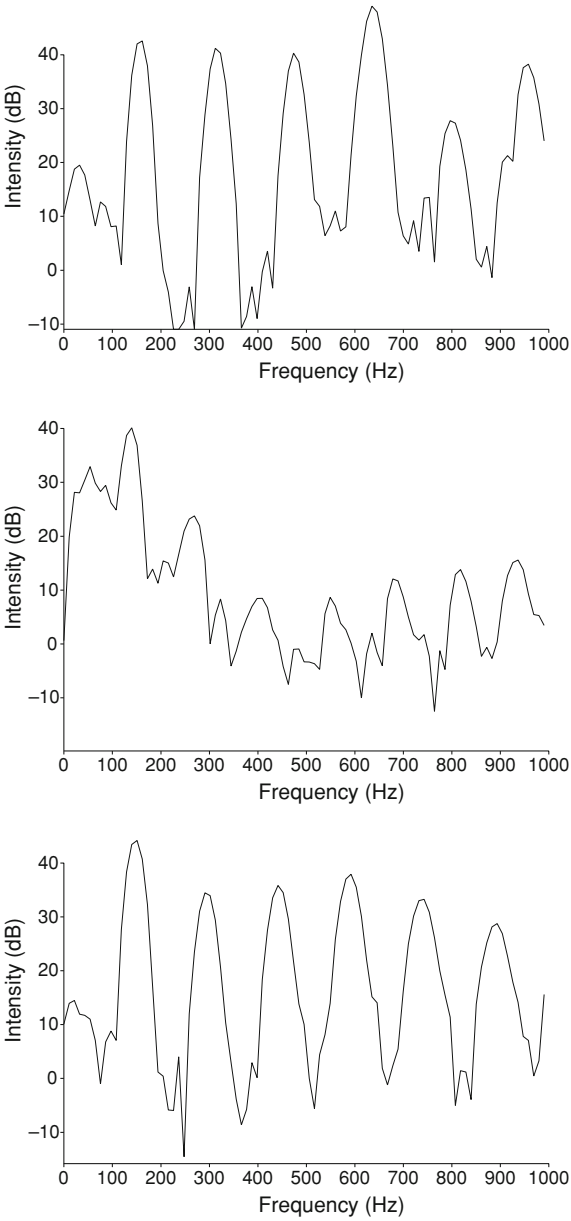


Fig. 4.10 Power spectra from the first 50 ms (approx.) of the vowel following release of stops in three distinct Hmong syllables. *Top*: [dɔ̀] (modal); *Middle*: [dʰɔ̀] (whispery stop release); *Bottom*: [dɔ̤] (breathy)

really an infinite family of such functions), and is a function of both time and frequency.

Since the STFT is derived from the Fourier transform, like the latter it is generally complex-valued and invertible, meaning that the signal itself results from an inverse transform of its STFT. Recall that the (time-independent) energy density spectrum was defined as the squared magnitude of the Fourier transform. The time–frequency spectrogram Spgm_w is then defined as the squared magnitude of the STFT, so that $\mathrm{Spgm}_w(t,\omega) \stackrel{\text{def}}{=} |S(t,\omega)|^2$. The picture it provides is three-dimensional, a sort of "running power spectrum" of successive windowed slices of the signal. It is not difficult to imagine how important the window function is to the particular properties of a spectrogram. To actually compute a short-time Fourier transform, as usual we need to work in a digital setting; the particulars are relegated to the math box, but a simplified algorithm will also be presented below.

In a continuous-time setting, the short-time Fourier transform of a signal $s(\tau)$ is usually defined by the following equation [9]:

$$S_w(t,\omega) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} s(\tau) w(t-\tau) \mathrm{e}^{-\mathrm{i}\omega\tau} \,\mathrm{d}\tau, \tag{4.1}$$

in which the function $w(t)$ is a real-valued window function having finite support (normally this will be one of the window functions considered in Chap. 2). Comparing with Eq. 2.25, it can be seen that at each time point t_0 the function $S_w(t_0,\omega)$ is simply the Fourier transform of a portion of the signal windowed by w around t_0. This definition is equivalent to what much literature calls the *Gabor transform* in the analog regime, but the digital version of the STFT described presently uses a different method than the digital Gabor transform, so it is inaccurate to simply identify the two.

Another form of the STFT which is sometimes used is given by:

$$S'_w(t,\omega) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} s(\tau+t) w(-\tau) \mathrm{e}^{-\mathrm{i}\omega\tau} \,\mathrm{d}\tau \tag{4.2}$$

This alternate form is also called the *moving window transform* [22], and its result is equivalent to that from the previous definition except for a linear frequency term added to the phase [9]. The spectrogram can be equivalently defined as the squared magnitude of the STFT (Eq. 4.1) or of the moving window transform (Eq. 4.2), because the two transforms do not differ in magnitude.

The STFT in either of the above forms is invertible using the same window function in the analog regime; the following inversion formula corresponds to Eq. 4.1:

$$s(\tau) = \frac{1}{2\pi} \int \int_{-\infty}^{\infty} S_w(t, \omega) w(t - \tau) \mathrm{e}^{\mathrm{i}\omega\tau} \, \mathrm{d}t \, \mathrm{d}\omega. \tag{4.3}$$

The inverse STFT can be thought of as providing an "expansion" of the signal using a continuum of "elementary signals" representing the time- and frequency-shifted windows [3]. It is really a signal synthesis formula, while the STFT itself is an analysis formula.

To present the mathematics behind the digital spectrogram I will rely on [28]. The digital STFT method starts from the analysis formula (Eq. 4.1) and samples it in time and frequency. First, let us define the discrete-time STFT at particular time point n_0:

$$S(n_0, \omega) \stackrel{\text{def}}{=} \sum_{m=-\infty}^{\infty} s(m) w(n_0 - m) \mathrm{e}^{-\mathrm{i}\omega m} \tag{4.4}$$

where m takes integer values. Comparing with Eq. 2.29, this is just the discrete-time Fourier transform of the product of the signal with the analysis window. Next, make the time point a variable, thus defining the discrete-time STFT as a function of time point n and frequency ω:

$$S(n, \omega) \stackrel{\text{def}}{=} \sum_{m=-\infty}^{\infty} s(m) w(n - m) \mathrm{e}^{-\mathrm{i}\omega m} \tag{4.5}$$

So much for the discrete-time part; the last step is to move to discrete-frequency as well. Just as the discrete Fourier transform of Eq. 2.35 yields frequency samples of the discrete-time transform, and is thus suitable for digital implementation, there is a similar equation for the discrete (digital) STFT which frequency-samples the discrete-time version above:

$$S(n, \omega_k) \stackrel{\text{def}}{=} \sum_{m=0}^{N-1} s(m) w(n - m) \exp\left[\frac{-2\pi \mathrm{i} k m}{N}\right] \tag{4.6}$$

Computing the result of this equation with a digital signal and specific window function yields a matrix of complex numbers, one for each "cell" in the time–frequency grid. To get a digital spectrogram from this, one should simply compute the magnitude of each number in the STFT matrix and square it.

Equation 4.6 is now the digital version of the analysis formula; one might think that a corresponding inversion formula could likewise be produced by analogy with the continuous regime, but the discretization process has rendered the situation mathematically more complicated, so if we tried to use the sampled inversion formula to invert the sampled analysis formula, it would be improper [1]. Gabor originally approached the problem from the synthesis angle, and his signal expansion yields a discrete version of an inversion

formula rather than an analysis formula. It is possible to derive both a synthesis (now called a *Gabor expansion*) and a corresponding analysis formula (the Gabor transform) in the digital regime, with the snag that each now must utilize different window functions which are mathematically dual in an interesting way that has inspired quite a bit of research on the topic (e.g. [8]).

Comparing the final definition of the digital STFT above with the DFT equation (2.35), it is plain that the digital STFT is simply a sequence of DFTs of signal-window products, in which the window is "sliding" along the signal. The defining equation above assumes that the window is moved by one time sample for each new DFT. It is usual practice to *decimate* the above definition, so that the signal-window products $s(m)w(nL - m)$ are taken only for integer multiples of some integer $L > 1$—i.e. the window is customarily slid along by more than one sample point. The practical STFT parameter L is variously called the "hop size" or "time step" or "frame advance."

It is important to keep the frame advance less than the window length, and preferably less than the *effective* length (about half of the window), or the spectrogram will be decimated too much and the successive windowed frames will not overlap enough to yield a good representation.[2] On the other hand, given the way most digital spectrograms are displayed (to be described below) it is not usually worthwhile to have a very small frame advance, and this does lead to greatly increased computation times. The importance of the proper frame advance is illustrated in Fig. 4.11 showing three spectrograms of the word *dad*. The similarity of the first two spectrograms there is notable given that the second one required far more FFTs to be calculated; both of these are fairly optimal in performance for the parameters employed, but the top example is far more efficient. The lower image is, by contrast, over-decimated and totally undesirable.

4.2.1.2 STFT from a Filter Bank

The earliest spectrograms were computed using analog electronics, without any Fourier transforms being directly performed. The intuition behind this process begins with considering a single fixed frequency within a short-time Fourier transform of a signal. Such a frequency band of an STFT is equal to the signal first modulated by the fixed frequency and passed through a filter whose impulse response is the analysis window. In the digital realm, the STFT has just a certain number of frequency bins, and each frequency row of the STFT matrix can be similarly computed as the signal passed through a bandpass filter whose impulse

[2] From the mathematical theory of time–frequency analysis, a deep result called the Balian–Low theorem [17] establishes that a digital STFT must have overlapping windows in order to completely represent the signal (see also [2] for discussion).

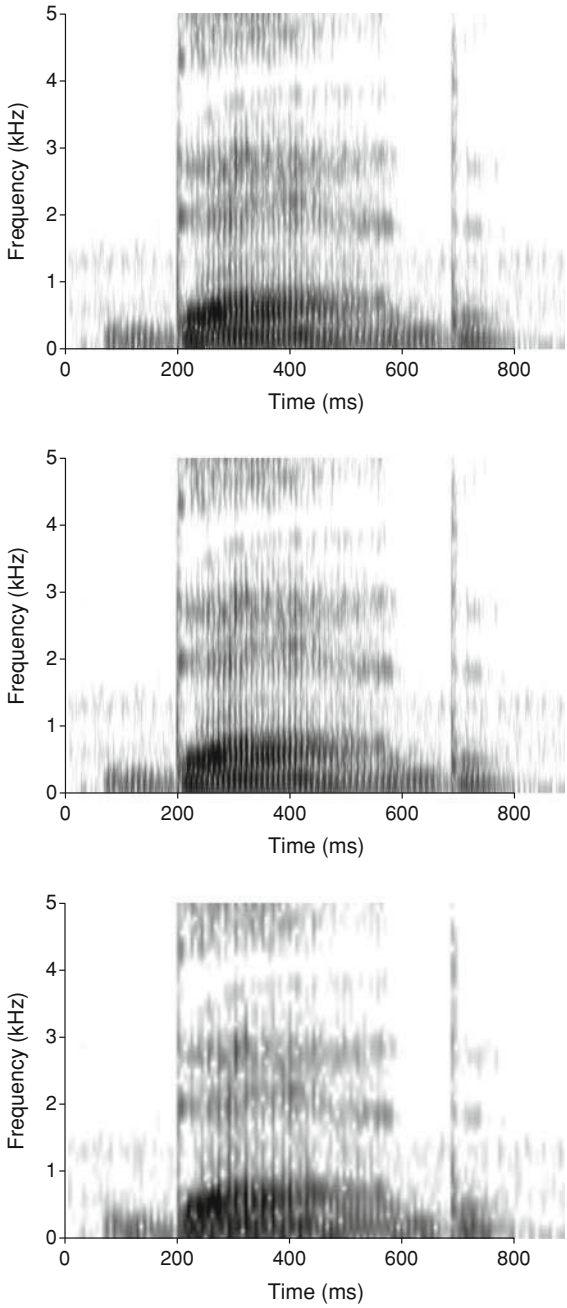


Fig. 4.11 Spectrograms of the English word *dad* [dæd], computed with 12 ms Gaussian window using Praat. *Top panel* uses the Praat standard frame advance of 2 ms; *middle panel* uses a frame advance of 0.1 ms; *lower panel* uses a frame advance of 6 ms

response equals the analysis window. The reader is referred to [28] for details concerning this "filter bank" method of computing the STFT. It will be significant in later chapters that when the STFT is viewed in this fashion, each frequency row in the matrix is itself an analytic signal whose instantaneous frequency can be calculated.

4.2.1.3 Algorithm for the Spectrogram

We can assume that a digital signal is represented as a vector (basically just a list) of ordinate values, one for each time sample. To compute a spectrogram in a high-level environment such as Matlab, the following rough algorithm can be implemented:

1. Divide the signal into "slices" which overlap, and which are each the length of the desired analysis window.
2. Index each slice by the time point at its center. The signal is now represented redundantly as a vector of slices.
3. Multiply each slice by a tapering function (or if you want rectangular windows, skip this step).
4. Compute the FFT of each slice, keeping in mind that the FFT frame length can always be longer than the slice if it is zero-padded properly. Matlab automatically takes care of this.
5. Discard the negative frequency range and the upper half of the positive frequencies. You now have a vector of the discrete Fourier transforms of the signal slices. This is referred to as the STFT matrix; it is a two-dimensional matrix of complex numbers, in which each cell lies at an intersection of time and frequency. The time points are the center points of the signal slices; the frequencies are the values in the FFT. The number of frequencies in the matrix depends on the length of the FFT frame—e.g. a 1,024-sample FFT frame will yield 512 frequency bins.
6. Take the absolute value of each number in the STFT matrix, then square the results of that. You now have the spectrogram matrix.

When applied to sound signals, it is customary to scale the spectrogram magnitude logarithmically, as a decibel scale. When a spectrogram is plotted for a screen view or printable image, it is standard to show the decibel magnitude by linking the values to a colormap. In this way, the color plot can be used to show the magnitudes over the two-dimensional matrix. While full-color spectrograms can be useful with a good choice of colormap, it is a long-standing tradition that spectrograms are shown using a grayscale colormap, thus mimicking the appearance of the old analog spectrograms that were singed electrically onto Teledeltos paper. The highest amplitudes are then plotted in the darkest gray or black, and the gray gets lighter as the amplitudes get smaller. To facilitate readability of a digital spectrogram plot, it is important to use a graphical plotting routine that does "interpolated shading" from cell to cell of the spectrogram matrix. This makes the

digital spectrogram look "analog," which is what humans need to be able to read it. It is possible to plot the digital spectrogram matrix literally as a matrix of blocks of different colors (this is, inexplicably, the way some spectrogram routines work), but the result is almost unreadable.

4.2.2 Development of Spectrogram Theory

I already mentioned the spectrogram's origins, and its ensuing influence on speech science, in the previous chapter. But the success of the spectrogram as an analysis tool in the early going relied solely on a device that produced spectrograms. A complete theoretical understanding of the spectrogram, which would ultimately facilitate the switch to digital computation, took about twenty more years to congeal. The first mathematical attack on this sort of time–frequency analysis was carried out by Gabor [15], in apparent ignorance of the existence of the spectrograph device (although Gabor did mention it in a footnote that was added following acceptance of his paper). Gabor developed a digital form of the inverse STFT formula mentioned above, which is usually called the Gabor signal expansion, but its relationship to the STFT analysis formula was not understood until many others had studied the idea decades later. Gabor did not himself characterize his representation as a short-time power spectrum.

The short-time power spectrum was given its first rigorous mathematical treatment by Fano [7], but this was limited to a particular impractical form of the time window. The first short-time power spectrum allowing an arbitrary (continuous-time) window function was derived by Schroeder and Atal [29], although these authors did not identify their function (which was defined in continuous time essentially as in Eq. 4.1) as an invertible short-time Fourier transform—as Gabor had only developed the synthesis formula in the discrete regime, these authors developed only the analysis formula in the analog regime, so the relationship between the two was still muddled for a while. A further treatment was completed by Montgomery and Reed [27], who generalized Helstrom's [19] work on Gabor's signal expansions and examined both the synthesis and analysis formulae, showing for the first time that an STFT in the analog regime is generally invertible like a plain Fourier transform. By the 1970s, the analysis formula was standardly called the *short-time Fourier transform*, and in its digital "sampled" form this is most frequently said to be equivalent to the Gabor transform, although the two have disparate algorithmic aspects.

4.2.3 Uncertainty Principle

The *uncertainty principle*, often named after the physicist Heisenberg, can be stated in the most general terms as in [10]: *A nonzero function and its Fourier transform cannot both be sharply localized*. Heisenberg introduced this as a

general constraint on observational precision in quantum physics. In that regime, the principle implies, among other things, that the position and momentum in a quantum state cannot both be precisely determined. In the regime of signal analysis, to which Gabor [15] introduced the uncertainty principle, it implies that a signal $s(t)$ and its Fourier spectrum cannot both have a small domain. This is often stated in the literature by saying that a signal cannot be both highly time-limited and highly band-limited in frequency [10]. It is important to emphasize that, owing to the uncertainty principle, "obtaining an intrinsic and infinitely precise description of the 'time–frequency content' of a signal is out of the question" [6].

The above general uncertainty principle is a global constraint on an entire signal; when performing time–frequency analysis, however, we normally are interested in viewing some kind of breakdown of a complicated signal into "events" in time and "components" in frequency—it is local time–frequency content that interests us. Moreover, the usual discussion of uncertainty in signal processing mentions only the global *signal* uncertainty as above; when a signal is analyzed with a spectrogram, additional uncertainty is introduced by the windowing operation to yield what has been termed the *spectrographic* uncertainty [25].

Speaking in general terms, a spectrogram can be thought of as displaying a time–frequency distribution which has a statistical *variance* in both the time and frequency directions—this is why the image is smeared. Operating locally, at each time–frequency cell there will be a *conditional time variance* for that frequency and a *conditional frequency variance* at that time. Each of these will be partly determined from the window function used to compute the spectrogram. The product of these two conditional variances must be locally greater than a certain quantity; this is the local spectrographic uncertainty principle [25], which provides a lower bound on the possible precision of a spectrogram at each time–frequency cell. The spectrographic uncertainty is often described in speech science as the "resolution trade-off" between time and frequency. We will see in the next chapter that it is possible to work with other time–frequency representations that do not have nearly as much uncertainty as the spectrogram.

4.3 Spectrographic Analysis of Speech

"There is no uniqueness of a time–frequency representation: there are many different ways of describing the 'time–frequency content' of a signal" [6]. Owing to the quoted fact, computing a spectrogram of a bit of speech is not something that can be done effectively without taking care to understand and set the values of a number of parameters which have a dramatic effect on the analysis and its appearance in a display. The most important of these are:

- the length of the analysis window;
- the particular window function to be used;

- the number of points used to compute the Fourier transform of each windowed segment;
- the frame advance;
- the dynamic range of the amplitude plot;
- whether to use *pre-emphasis* for the display.

Let us now discuss each of these in turn.

4.3.1 General Guidelines

A spectrogram provides a time–frequency representation of a signal that by nature assumes the signal is "short-time stationary," meaning that the spectrogram is unable to detect any changes in the signal during the span of one analysis window. Accordingly, the most important parameter to be set when computing a speech spectrogram is the window length. A window whose effective length is shorter than one glottal cycle during voiced segments will provide the most generally useful and informative type of speech spectrogram, often called a "wideband" spectrogram in reference to the original filterbank computation method. With such a short analysis window, the spectrographic uncertainty principle causes the frequency resolution and precision to be fairly poor, with the consequence that the individual component frequencies such as formants are "smeared" in the frequency axis. Nevertheless, a wideband spectrogram is essential for observing the formant frequencies which are excited with each glottal impulse, and also for observing the brief events associated with consonants such as stop burst noises.

By contrast, a window whose effective length encompasses more than two glottal cycles will show the fundamental frequency of voicing and the harmonics associated to it, and will no longer show formant frequencies as being directly excited. In this type of "narrowband" spectrogram, formants can be observed only indirectly as groups of harmonics having a higher intensity. The spectrographic uncertainty principle then causes brief events to be less resolved, and smeared in the time axis. The spectrogram with longer window is unable to detect any changes, such as glottal impulses, that occur within the window's span.

From the discussion of window functions for power spectra in the previous chapter, it was concluded that the optimal choices are drawn from either the Gaussian or Kaiser families of functions. It has been noted, however, that window performance for a single spectrum computation does not necessarily translate to the time–frequency regime for spectrograms. One study specifically examining spectrogram performance concluded that a Gaussian window provides optimal time–frequency localization (i.e. minimal smearing) [20]. For reasons of tradition, the Hamming and "Hanning" windows have been perennial favorites in speech analysis, but more recently the Gaussian window has begun to get promoted in our field as well. I would recommend that readers choose either Gaussian or Kaiser windows for creating spectrograms. Figure 4.12 compares spectrograms computed

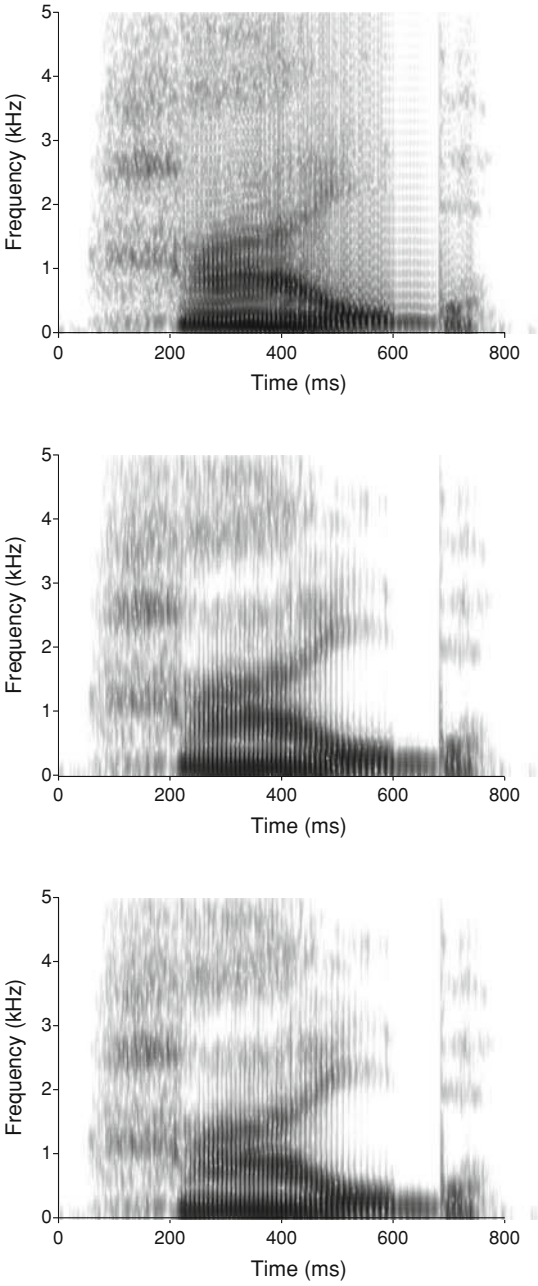


Fig. 4.12 Spectrograms of the English word *hide* [haɪd], computed with 6 ms rectangular window (*top*), Hann window, and Gaussian window (*bottom*) using Praat

with rectangular, Hann, and Gaussian windows. Decreased separation of F_1 from the voice bar is observable in the Hann windowed example versus the Gaussian, as well as slightly broader formants.

The number of points used to compute the Fourier transforms for the spectrogram will normally be at least as long as the analysis window, but beyond this it affects the frequency sampling of the resulting image (see the discussion of zero-padding in the preceding chapter). A 256-point frame, for example, will provide only 128 frequency bins to sample the entire frequency range whether it is plotted or not. Supposing we have a sound whose highest recorded frequency is 22.5 kHz, this is not a very rich frequency sampling. In general, I recommend using at least 1,024 points for the Fourier transforms, and 2,048 points is often useful as well. Unfortunately, many popular programs such as Praat do not allow the user to set the FFT frame size independently of the window length, and will automatically set the frame size to some value like the first power of 2 greater than the window length.

Recall that a spectrogram is plotted by linking the decibel amplitude values over the time–frequency matrix to a colormap, usually a grayscale. We do not normally want to show the lowest amplitudes in the STFT matrix computed from a signal, since these are frequently just unwanted noise. The spectrographic parameter commonly called the *dynamic range* has to be set by the user; this is the value that sets how large an amplitude value needs to be to "make the cut" and get plotted in the spectrogram. It is standard for the dynamic range to be a decibel value that refers to the difference between the loudest and quietest amplitudes to be plotted. A good value for speech is usually around 50 dB; this means that time–frequency cells with amplitude more than 50 dB down from the loudest cells will simply be "clipped" out of the plot. A value of 30 dB will usually show too little of the time–frequency content of the signal, losing a lot of valuable information, while a value of 70 dB or greater would only be suitable for high-quality sound recordings made with a soundbooth and an expensive microphone, since otherwise a lot of noise will be shown as light gray in the plot.

In the old days of analog spectrograms, the best dynamic range that could be obtained using the Teledeltos paper was around 35 dB. This presented a problem for speech applications, since the average person's glottal source sound decreases in amplitude as the frequencies get higher, at a rate of −6 dB/octave. As a result, everyone's F_1 is generally their loudest formant, with F_2 somewhat quieter, and so on up the frequency scale. A spectrogram of a vowel that is plotted using 35 dB of dynamic range will only show F_3 very faintly, and higher formants may be missing from the plot. A solution to this that was employed in the early spectrograph devices was called *pre-emphasis*, and it involves artificially changing the spectral amplitude slope by some amount to counteract the natural roll-off of the voice source. A common standard pre-emphasis was 6 dB/octave, the precise opposite of the average person's natural roll-off. The intended result was to even out the amplitudes of the formants across the frequency range. Pre-emphasis is generally effective for this purpose, but I find that in the modern world of spectrograms with much better dynamic range, it can also be detrimental to an accurate picture of the time–frequency content of speech. There is no longer quite as much reason to even out the amplitudes of speech formants artificially; a spectrogram with at least 50 dB dynamic range is usually good enough to show all the formants reasonably

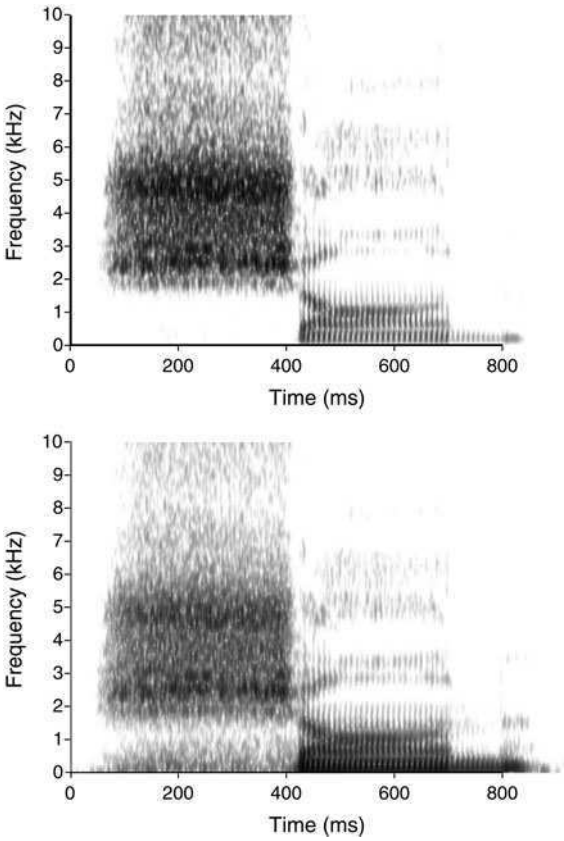


Fig. 4.13 Spectrograms of the author's name *Sean* [ʃɑn], computed with 12 ms Gaussian window using Praat. *Upper panel* uses 40 dB of dynamic range with 6 dB/octave pre-emphasis, while the *lower panel* uses 55 dB of dynamic range with no pre-emphasis

well. The down side of pre-emphasis is chiefly that it increases the amplitude of high-frequency consonant noise far too much, preventing a realistic picture of the speech time–frequency content. For the most realistic spectrograms, I recommend against pre-emphasis, but anyone following this advice should turn up the dynamic range slightly so that high formants are not too quiet. Figure 4.13 shows two spectrograms of the same utterance, with and without pre-emphasis.

4.3.2 Short Window (Wideband) Analysis

Wideband (short window) spectrograms have often been described in the literature as displaying poor frequency resolution and precision, which is true. When applied in speech science, however, it has often been written that such spectrograms are unable to resolve the harmonics of the glottal source, and hence the formants look like fat dark bars. That is to say, formants are supposedly so fat because there are a number of harmonics resonating within each one, and these are all mushed

together in the spectrogram [23]. This way of understanding a wideband spectrogram of voiced speech is not correct.

A better way of understanding wideband versus narrowband speech spectrograms is to think in terms of the "harmonic–inharmonic duality" of speech that was discussed earlier. A wideband spectrogram is the right sort of time–frequency analysis for showing the inharmonic aspect of voiced speech. Each formant is shown as it is excited by the glottal impulses, albeit as a fat smeared bar due to the spectrographic uncertainty principle. An important consequence of this view is that an often-stated "problem" stemming from higher fundamental voicing frequency is not necessarily a problem at all:

> The higher the fundamental frequency, the fewer [are] the harmonics which define the formant and the greater [is] the probability that the most prominent harmonic will be distant from the center frequency of the formant. [26]

The above quoted statement of the problem surely does apply if one elects to examine formants using a narrowband analysis showing the harmonics, but it need not apply to a wideband analysis because the latter does not literally portray a smeared version of the former. From the inharmonic model of voiced speech, we can recognize that each formant is excited by the glottal impulses nonetheless, whether the fundamental frequency is low or high. It is worth commenting that the supposed "problem" resurrects the forgotten debate over the accommodation theory that was chronicled in Chap. 3.

4.3.2.1 Vowels

Wideband spectrograms of vowels are chiefly useful for showing the formants. Such images were also widely used at one time to measure the formant frequencies, but this is no mean feat because of the smearing. As a result, "manual" measurement of formants from spectrograms has fallen somewhat into disuse in recent years, at least for research studies where a large number of formants have to be measured.[3] For the best possible localization of formants in a spectrogram, a window should be used whose effective length is only about 1 or 2 ms shorter than one glottal cycle. Shorter windows yield fatter-looking formants, while longer windows will begin to show harmonics.

Figure 4.14 shows wideband spectrograms of two synthesized vowels, computed using 14 ms Gaussian windows (the effective length is about 7 ms). It is essentially impossible to distinguish F_2 from F_3 in [i], and F_1 is difficult to discern from F_2 in [ɔ]. The low F_1 of [i] is also very poorly localized. Since formants are excited with each glottal closure, one might think to try zooming in the analysis to show only a few glottal cycles. An attempt at this is shown for [ɔ] in the figure, and

[3] One anonymous grant reviewer once wrote to me that nobody uses spectrograms anymore, except for showing speech sounds to phonetics students.

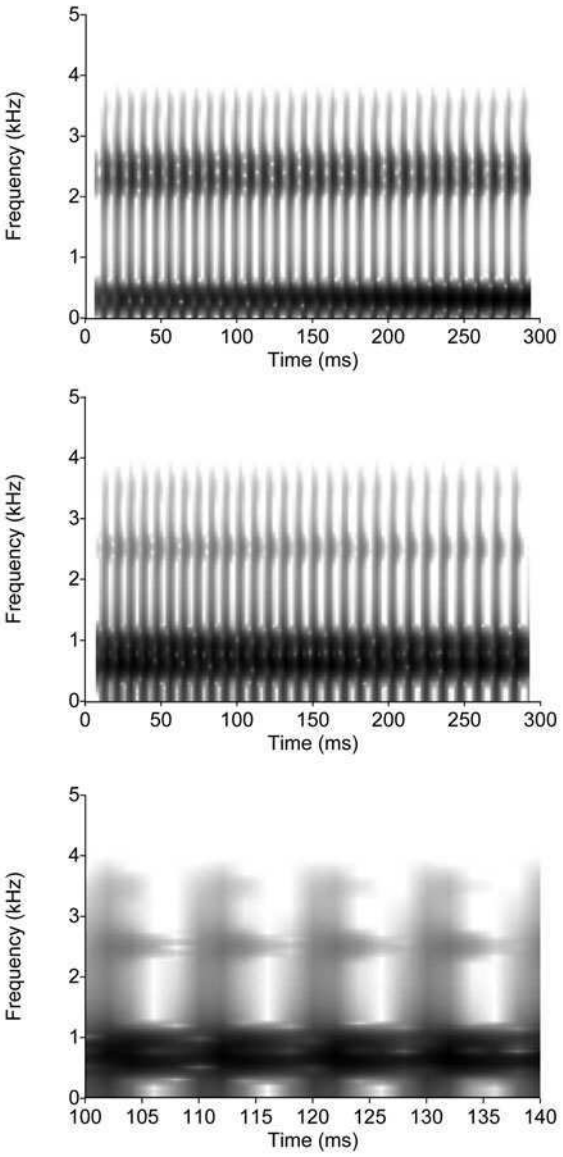


Fig. 4.14 Spectrograms of synthetic vowels [i] (*top*), [ɔ] (*middle*), [ɔ] 40 ms segment, computed with 14 ms Gaussian window using Praat

it is somewhat surprising how much it does help in discerning the two lowest formants, in spite of the poor frequency localization.

We have just seen that wideband spectrograms, while they do show formants, are not much good for measuring them precisely even in clean synthesized vowels which exactly obey the source-filter theory of speech production. Knowing this, one can only dread the prospect of measuring formants from spectrograms of real

speech, and indeed the values are even less clear for in vivo vowels. The reasons for this are chiefly these two: first, real vowels generally have a voice bar, which is a low resonance around 200–250 Hz (higher for smaller vocal tracts) that is difficult to resolve from F_1 in high vowels[4]; second, real phonation and speech does not precisely obey the source-filter theory because the air flow makes aeroacoustic processes relevant, and this in turn introduces complicating features into the spectrum.

Figure 4.15 shows spectrograms of English words containing the vowels [æ, i, ɔ]. The formants of [æ] are maximally separated from each other and from the voice bar, so in this case the challenge is minimal. The real trouble is evident in the other two spectrograms. F_1 of [i] is difficult or impossible to separate from the voice bar, while F_1 and F_2 of [ɔ] cannot be discerned from each other.

4.3.2.2 Obstruents

Stop consonants have a number of attributes that are visible and measurable with the aid of a wideband spectrogram. These include the burst event, prevoicing during the closure, aspiration following release, breathy voicing following release, as well as vowel formant transitions going in to the closure and emerging from the release. To best observe these attributes, the wideband spectrogram should be set to have similar parameters to those most useful for vowels. For a given speaker, the window duration should be set to slightly less than one glottal cycle. Pre-emphasis is likely to be a very bad option for any obstruent spectrogram, since it will artifically increase the amplitude of all the high-frequency noise, making the stop burst and aspiration appear louder than it actually is. Measuring the formant transitions going in to and coming out of a stop is at least as problematic as measuring vowel formants in the first place, but a spectrogram can at least give some impression of the transitional values, and these are generally indicative of the place of articulation. A spectrogram is not the best thing for analyzing stop burst spectra; for this, a carefully windowed power spectrum (shown in the previous section) is probably the best tool.

Figure 4.16 shows spectrograms of English words pronounced [k^hɑt] and [gɑt], in which many of the attributes mentioned above can be observed. The initial [k^h] appears to begin with two closely spaced burst events, not an uncommon occurrence for velar stops. The bursts are followed by a period of broadband aspiration noise, during which it is possible to detect formants in the noise. The initial [g] begins with a voiced closure, which is visible as a voice bar. The vowel begins immediately after the stop burst, and so in this case the formant transitions out of the stop are easier to measure. This example shows a common characteristic of

[4] The controversial notion that the voice bar is a separate resonance will be discussed in Chap. 6.

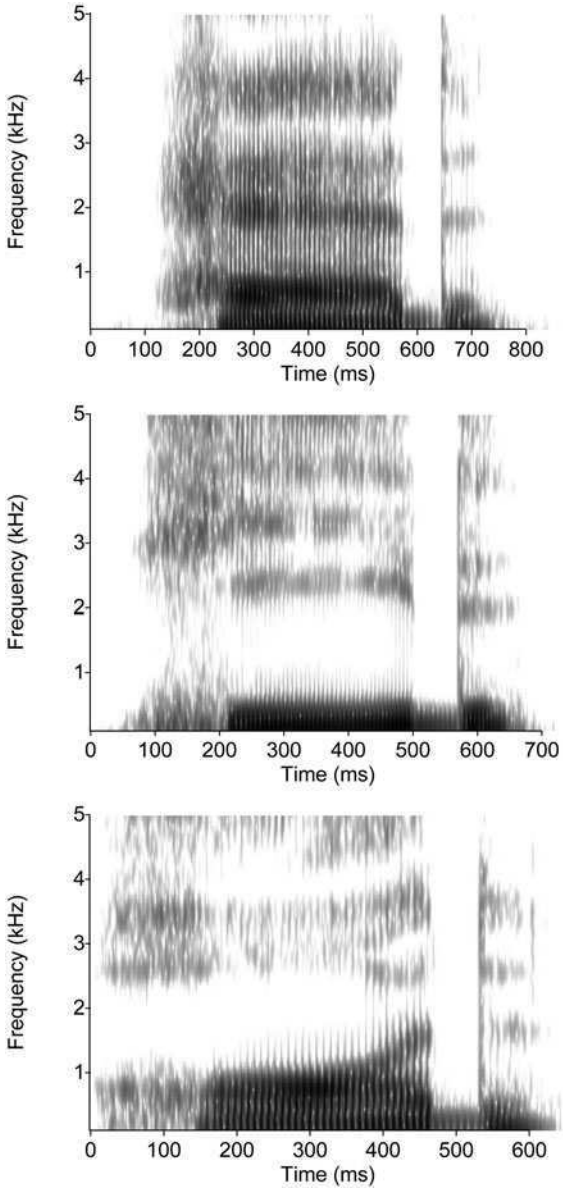


Fig. 4.15 Spectrograms of English words *had* [hæd], *heed* [hid], *hawed* [hɔəd] spoken by the author in an American English dialect. Computed with 12 ms Gaussian window using Praat

velar stops, with F_2 and F_3 appearing to emerge from origin points very close to each other. Formant transitions can also be observed heading into the final [t] of both these words.

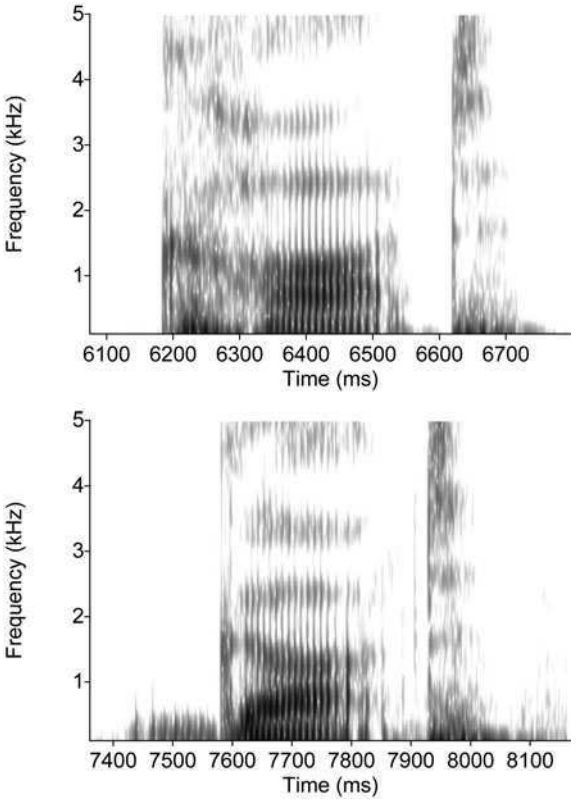


Fig. 4.16 Spectrograms of English words *caught* [k^h ɑt], *got* [gɑt] spoken by the author in an American English dialect. Computed with 12 ms Gaussian window using Praat

While fricative noise spectra are best quantified using the power spectrum techniques discussed previously, a wideband spectrogram can provide a general look at a fricative that allows one to check whether it is voiced, whether the noise spectrum changes appreciably through the duration (this will not be detected with a single power spectrum), and also permits examination of the formant transitions in surrounding vowels, which can help confirm a place of articulation. When examining fricatives, it is useful to use a frequency range with good extension in to the high values. The examples in Fig. 4.17 show that English [f] has noise extending over 14 kHz, and [s] extends above 11 kHz. It is also easy to observe the large differences between [f] and [v], resulting from the voicing diminishing the airflow power that is available to generate a noise with labiodental origin.

4.3.2.3 Sonorants

Approximants and semivowels are chiefly characterized by formants, in similar fashion to vowels, although they may be very brief sounds. Accordingly, they can be observed using the same spectrographic parameters normally employed for

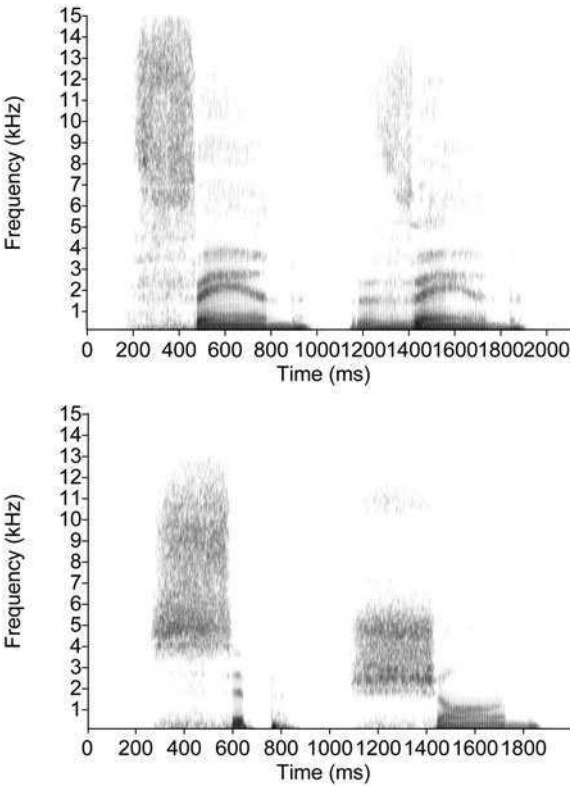


Fig. 4.17 Spectrograms of English word pairs *fan* [fæn], *van* [væn] (*upper*) and *sip* [sɪp], *shone* [ʃan], spoken by the author in an American English dialect. Computed with 12 ms Gaussian window using Praat

vowels. Figure 4.18 shows spectrograms illustrating word-initial [ɹ] and [l] in English. This variety of [l] is characterized by a very large frequency spread between its lower and upper formants. The English [ɹ] sound is quite uncommon in the world's languages; it is characterized by a very low F_3 (below 2,000 Hz), which can be seen rising rapidly out into the vowel in this example.

The nasals have formants which can appear quite different from those of oral sounds. They are generated by the coupled oral and nasal cavities, and are generally more damped as a result, increasing their bandwidth; this can make nasal formants appear even more smeared in frequency than oral formants. The coupled cavities also cause the spectrum to contain at least one *zero*, or frequency at which energy is absent. It is impossible to directly observe a zero in a spectrogram, although it may be evident in a single power spectrum of a nasal. Figure 4.19 shows spectrograms illustrating word-initial [m] and [n] in English. There it can be observed that the different nasals do have slightly different upper formants (which are invariably very faint); nasals in general always have a strong resonance around 250 Hz, however [32], which is often called the "nasal murmur." The different nasals are distinguished spectrally by their distinctive upper formants, but also by formant transitions in nearby vowels in similar fashion to the oral stops.

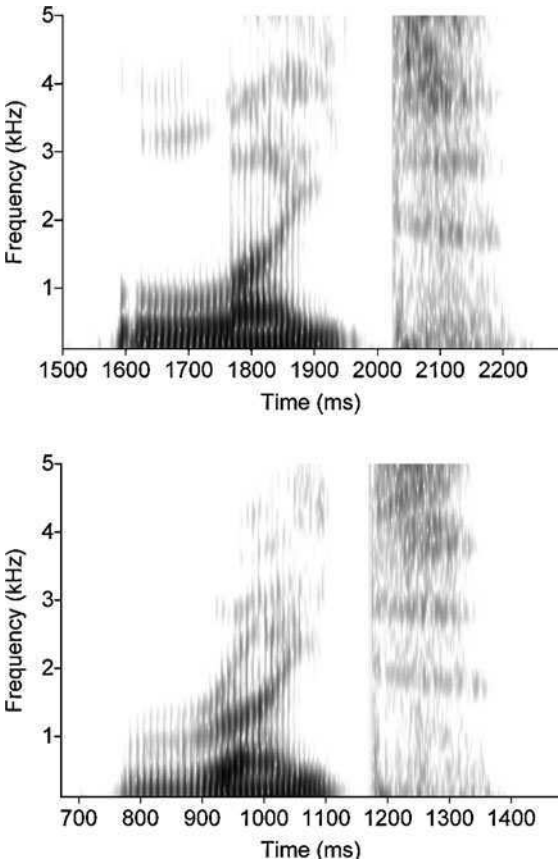


Fig. 4.18 Spectrograms of English words *light* [laɪt], *right* [ɹaɪt] spoken by the author in an American English dialect. Computed with 12 ms Gaussian window using Praat

4.3.3 Long Window (Narrowband) Analysis

So far, only wideband speech spectrograms have been closely examined; these are generally the most useful for observing the acoustic correlates of phonetic articulations during speech. The term "narrowband" is traditionally given to spectrograms computed using an analysis window that is at least two glottal periods in duration. This increased time of assumed signal stationarity now "observes" the glottal periodicity in each window, and then the Fourier transform of each window contains the fundamental voicing frequency and its co-occurring harmonics. The spectrographic uncertainty induces the "time–frequency resolution tradeoff," whereby the longer analysis window prevents good time localization of any brief events or transitions. Figure 4.20 compares a wideband (12 ms Gaussian window) with a narrowband (120 ms Gaussian) spectrogram of the English word *syllable*. The wideband image as usual shows the glottal impulses as vertical lines, the time precision is good, and the frequency precision is not good. In contrast, the narrowband image cannot resolve the glottal impulses, but the fundamental and its

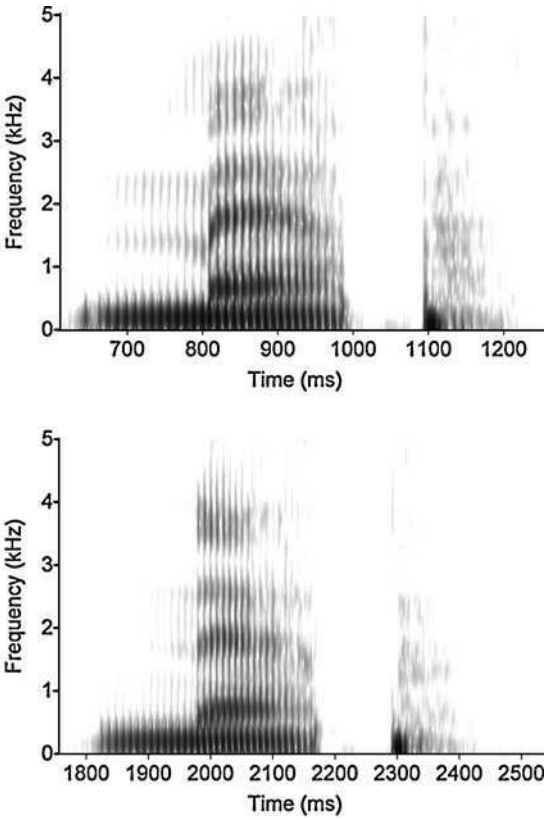


Fig. 4.19 Spectrograms of English words *map* [mæp], *nap* [næp] spoken by the author in an American English dialect. Computed with 12 ms Gaussian window using Praat

harmonics are shown. Here the time precision is poor, however, and formants are shown only indirectly as groups of louder harmonics.

It is evident from any example such as Fig. 4.20 that a narrowband spectrogram is not preferable to a wideband one for measuring formants. However, since the fundamental frequency and harmonics are shown relatively well, one can observe in this example that the pitch of the voice goes up and down over the course of the word. The narrowband spectrogram is, in fact, a reasonably good way to track and measure the pitch of the voice (usually equated with the fundamental frequency, but I will discuss this more at a later point), which is useful for studies of stress or intonation. A standard type of tool known as a pitch-tracking algorithm is most commonly employed for this purpose, but such algorithms are partly probabilistic, have to be heuristically guided, and have many parameters that must be tinkered with to achieve even modest performance. Figure 4.21 shows just the low-frequency range of a sentence portion in which the speaker says *a stressed syllable is usually*, viewed as a narrowband spectrogram with the pitch track provided from Praat's algorithm overlaid as a white line. While the fundamental frequency is

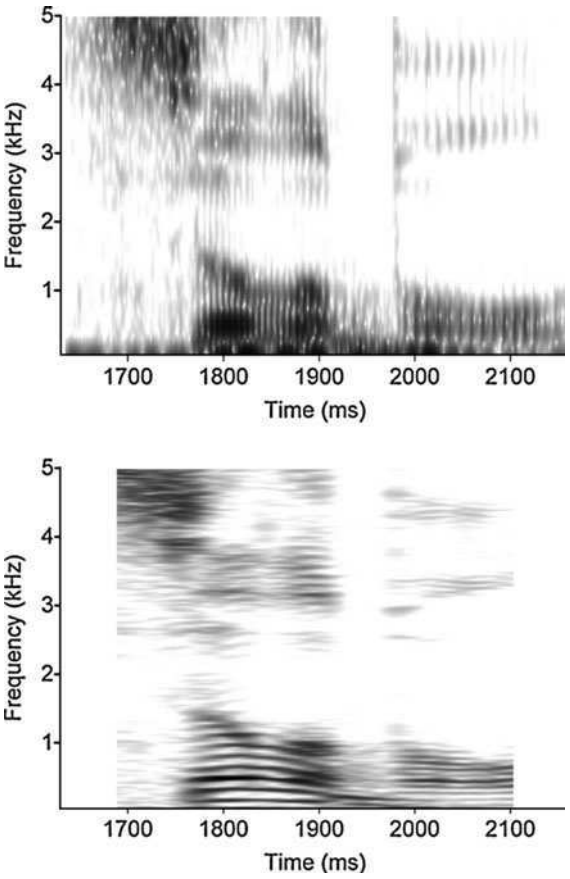


Fig. 4.20 Spectrograms of the English word *syllable* spoken by the author in a Canadian English sentence context. *Top* with 12 ms Gaussian window; *Bottom* with 120 ms Gaussian window

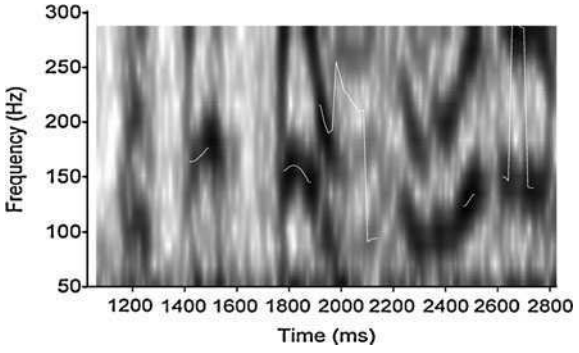


Fig. 4.21 Low-frequency range of narrowband spectrogram (120 ms Gaussian windows), showing the annotated phrase uttered in a sentence context. Praat pitch track is overlaid in white

observable in the spectrogram during voiced segments, it is not very well localized, which makes careful manual measurement of the center of the gray line a required procedure for measuring the voice pitch from this image. The pitch tracking

algorithm is semi-automated and the data can be stored as soon as it is calculated; nevertheless, anyone would agree that the results in this case (an unremarkable example) are far from excellent. In fact, the computed pitch track (which I tried to make as good as possible by tinkering with the settings) is only occasionally in agreement with the fundamental in the spectrogram; indeed, it is only occasionally able to be computed at all in this example. The spectrogram must be viewed as the gold standard here because it is just a time–frequency analysis and not a heuristic method, so this shows that a narrowband spectrogram is a more accurate and reliable pitch-tracking method than a state of the art algorithm.

4.4 Appendix: Praat and Matlab Techniques

4.4.1 Praat Functions

To create vowel sounds using a formant synthesizer, select the following from the `Object` view:

`New` $-\,-\,>$ `Sound` $-\,-\,>$ `Create sound from Vowel Editor`

The vowel editor allows one to set four formants manually; the first two can also be set by positioning the cursor in a vowel space and listening for the result. The user also sets the duration and intonation of the synthesized sound.

There are a few ways to get a power spectrum in Praat. The simplest is from the `Edit sound` view, where under the `Spectrum` menu there is a function to `View spectral slice` from a selected area. The window function will be whatever the spectrogram viewing windows are set to, under the spectrogram settings described below.

Often it will be useful to obtain a spectrum object rather than just a view. First, the target sound segment may need to be extracted from a longer sound, and for this the function `Extract selected sound (windowed)` should be used from the `File` menu in the sound editing view. Upon selecting this function, a dialog box appears allowing the window function to be set; one of the Gaussian or Kaiser options is always recommended.

There are two options for a spectrum object to be computed from a sound, using the `Spectrum` menu in the `Object` view. If a raw complex Fourier spectrum is desired or measurements are to be taken, then `To spectrum` should be selected. The resulting spectrum object can be exported to use the raw values, it can be viewed and measured using an edit view (in which it is shown as a power spectrum on a dB scale), or it can be drawn as a power spectrum in the `Picture` area. If only a power spectrum picture will be needed, a useful selection is `To ltas`, which stands for *long-term average spectrum*. As far as I can tell, the resulting power spectrum is the same as the one computed from a `Spectrum` object, but the `ltas` object can only be drawn in the picture area and cannot be measured using an edit view.

Spectral moments can be computed from a `Spectrum` object under the `Query` menu. Any of the central moments (which are normally used) can be precisely obtained from the selection `Get central moment`; alternatively, various statistical quantities related to the moments can be obtained individually, such as the mean (center of gravity, equal to the first "central" moment),[5] skewness (a sort of normalized 3rd moment), and kurtosis (a sort of normalized 4th moment).

The harmonicity of a target sound can be obtained directly from the sound object under the `Periodicity` menu, by selecting any of the harmonicity functions which compute a harmonicity object. The harmonicity object is a vector of dB values computed at a sequence of time points from the beginning to the end of the target sound, which may be queried at particular points, drawn as a graph, or exported as text. The values measure the harmonics-to-noise ratio using Boersma's published algorithm [4].

A spectrogram is optionally displayed in Praat's edit view of a sound, under the waveform. The appearance of this spectrogram is set by the spectrogram settings under the `Spectrum` menu of the edit view. The regular settings allow the user to set the frequency range shown, the analysis window length, and the dynamic range of the view. The window function is set to a Gaussian by default for this view; users should be aware that the Gaussian window length is automatically doubled for the computation, so that the length setting corresponds more closely to the effective length for the analysis. If 10 ms windows are really wanted, then 5 ms should be used for the setting. The advanced settings allow the user to make further adjustments including the maximum number of time points and frequency points for which the spectrogram is computed, but these settings should not normally require adjustment (v. the Praat manual for more information). The window tapering function can be changed here if desired, and the pre-emphasis function can be set to a specific amount of roll-off. Setting this to zero turns off pre-emphasis.

It is also possible, under the menu obtained by pressing `Spectrum` from the object list, to compute a Spectrogram object from any sound object. The spectrogram object can then be optimally displayed in the picture section. When the object is first created, the user sets the window length and tapering function, as well as the highest frequency computed in the spectrogram. The default time and frequency steps are generally adequate here. Once the spectrogram object exists, then one selects `Draw --> Paint` to display an image in the Praat picture section. It is at this point that a number of important spectrogram display options are set, including the time and frequency range to be displayed, and the dynamic range and pre-emphasis settings. All of the spectrograms presented in this chapter were created using this procedure.

[5] Although commonly phrased this way, strictly speaking there is no such thing as a first central moment; the quantity commonly so called is technically known as the *first moment about zero*, with the *n*th central moment being generally defined in terms of the $n - 1$th moment [33].

4.4.2 Matlab Code

I have provided several m-files (a term for Matlab programs) for computing and displaying a traditional spectrum and spectrogram. The m-files discussed here also depend upon the presence of other ancillary m-files which are provided as part of the code package. A user should first learn to use the Matlab built-in function `wavread` to read sound files in the `.wav` format into a Matlab vector variable.

$$[\texttt{signal}, \texttt{Fs}, \texttt{bits}] = \texttt{wavread}(\text{`}\texttt{yourfile.wav}\text{'});$$

The above Matlab command will read the contents of the named file into the vector `signal`, and will also extract the sampling rate as `Fs` and the quantization bit depth as `bits`. The bit depth is usually irrelevant for our purposes, but the sampling rate is usually of critical importance.

The file `powerspec.m` provides a basic Fourier power spectrum from an FFT, and is invoked using

$$[\texttt{PS}, \texttt{f}] = \texttt{powerspec}(\texttt{signal}, \texttt{Fs}, \texttt{low}, \texttt{high});$$

where the spectrum vector can be stored as `PS` and the frequency axis vector as `f`. Using these variables is optional, so the functions here can always be called using only the command to the right of the equals sign if an image is all that is desired. In the above command, `signal` should be replaced with the name of the target signal vector stored in the workspace, `Fs` should be the correct sampling rate, and `low, high` should be the low and high limits of the frequency range to be shown.

The file `powerspec_ensemble.m` provides an average power spectrum from a target matrix of signals, one in each column (so the signals must all be the same length). It is invoked using

$$[\texttt{PS}, \texttt{f}] = \texttt{powerspec\_ensemble}(\texttt{signals}, \texttt{Fs}, \texttt{low}, \texttt{high});$$

The file `specgram_2010.m` provides a conventional spectrogram of the sort provided by Praat, and is invoked using

$$[\texttt{STFTpos}, \texttt{f}, \texttt{tforspgm}] = \texttt{specgram\_2010}(\texttt{signal}, \texttt{Fs}, \texttt{window}, \texttt{overlap}, \texttt{fftn}, \texttt{low}, \texttt{high}, \texttt{clip});$$

in which the analysis window length (in samples) is set in the 3rd argument, the number of samples by which successive windows overlap is set in the 4th argument, the FFT frame size is set in the 5th argument, and the dynamic range is set as a negative dB value in the 8th argument. For example assuming a 44.1 kHz sampling rate, to compute a spectrogram with 1,024 frequency bins using 10 ms windows overlapped by 7 ms, use 441 for `window`, 309 for `overlap`, and 2,048 for `fftn`. The window function is set to Kaiser by default, but this can be changed by altering one line of the code. The optional storage variables can keep the displayed portion of the complex STFT, the frequency axis vector, and the time axis vector.

References

1. L. Auslander, C. Buffalano, R. Orr, R. Tolimieri, A comparison of the Gabor and short-time Fourier transforms for signal detection and feature extraction in noisy environments, in *Proceedings of the SPIE Advanced Signal Processing*: *Algorithms*, *Architectures* and *Implementations*, vol. 1348, pp. 230–247 (1990)
2. L. Auslander, I.C. Gertner, R. Tolimieri, The discrete Zak transform application to time–frequency analysis and synthesis of nonstationary signals. IEEE Trans. Signal Process. **39**(4), 825–835 (1991)
3. M.J. Bastiaans, A sampling theorem for the complex spectrogram, and Gabor's expansion of a signal in Gaussian elementary signals. Opt. Eng. **20**(4), 594–598 (1981)
4. P. Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, in *Proceedings of the Institute of Phonetic Sciences*, vol. 17, pp. 97–110, University of Amsterdam (1993)
5. P. Boersma, D. Weenink, Praat: doing phonetics by computer. Computer software (2009)
6. R. Carmona, W.L. Hwang, B. Torrésani, *Practical Time–Frequency Analysis*: *Gabor and Wavelet Transforms*, *with an Implementation in S* (Academic Press, San Diego, 1998)
7. R.M. Fano, Short-time autocorrelation functions and power spectra. J. Acoust. Soc. Am. **22**(5), 546–550 (1950)
8. H.G. Feichtinger, T. Strohmer (eds.), *Gabor Analysis and Algorithms* (Birkhäuser, Boston, 1998)
9. K.R. Fitz, S.A. Fulop, A unified theory of time–frequency reassignment. Preprint posted on arXiv.org (2005)
10. G.B. Folland, A. Sitaram, The uncertainty principle: a mathematical survey. J. Fourier Anal. Appl. **3**(3), 207–238 (1997)
11. K. Forrest, G. Weismer, P. Milenkovic, R.N. Dougall, Statistical analysis of word-initial voiceless obstruents: preliminary data. J. Acoust. Soc. Am. **84**(1), 115–123 (1988)
12. S.A. Fulop, Accuracy of formant measurement for synthesized vowels using the reassigned spectrogram and comparison with linear prediction. J. Acoust. Soc. Am. **127**(4), 2114–2117 (2010)
13. S.A. Fulop, C. Golston, Breathy and whispery voicing in White Hmong, in *Proceedings of Meetings on Acoustics*, vol. 4. Acoustical Society of America (2008)
14. S.A. Fulop, P. Ladefoged, F. Liu, R. Vossen, Yeyi clicks: acoustic description and analysis. Phonetica **60**(4), 231–260 (2003)
15. D. Gabor, Theory of communication. J. IEE Part III **93**(26), 429–457 (1946)
16. M. Gordon, P. Barthmaier, K. Sands, A cross-linguistic acoustic study of voiceless fricatives. J. Int. Phonetic Assoc. **32**(2), 141–174 (2002)
17. K. Gröchenig, *Foundations of Time–Frequency Analysis* (Birkhäuser, Boston, 2001)
18. H. Helmholtz, *On the Sensations of Tone*, 2nd English edn. (Longmans & Co., London, 1885)
19. C.W. Helstrom, An expansion of a signal in Gaussian elementary signals. IEEE Trans. Inf. Theory **IT-12**, 81–82 (1966)
20. A.J.E.M. Janssen, Optimality property of the Gaussian window spectrogram. IEEE Trans. Signal Process. **39**(1), 202–204 (1991)
21. A. Jongman, R. Wayland, S. Wong, Acoustic characteristics of English fricatives. J. Acoust. Soc. Am. **108**(3), 1252–1263 (2000)
22. K. Kodera, R. Gendrin, C. de Villedary, Analysis of time-varying signals with small *BT* values. IEEE Trans. Acoust. Speech Signal Process. **ASSP-26**(1), 64–76 (1978)
23. P. Ladefoged, *A Course in Phonetics*, 5th edn. (Thomson, Boston, 2006)
24. P. Ladefoged, I. Maddieson, M. Jackson, Investigating phonation types in different languages, in *Vocal Physiology*: *Voice Production*, *Mechanisms*, *and Functions* ed. by O. Fujimura (Raven Press, New York, 1988)
25. P.J. Loughlin, L. Cohen, The uncertainty principle: global, local, or both? IEEE Trans. Signal Process. **52**(5), 1218–1227 (2004)

26. R.B. Monsen, A.M. Engebretson, The accuracy of formant frequency measurements: a comparison of spectrographic analysis and linear prediction. J. Speech Hearing Res. **26**(3), 89–97 (1983)
27. L.K. Montgomery, I.S. Reed, A generalization of the Gabor–Helstrom transform. IEEE Trans. Inf. Theory **IT-13**, 344–345 (1967)
28. S.H. Nawab, T.F. Quatieri, Short-time Fourier transform, in *Advanced Topics in Signal Processing*, Chap. 6, ed. by J.S. Lim, A.V. Oppenheim (Prentice-Hall, Upper Saddle River, 1988)
29. M.R. Schroeder, B.S. Atal, Generalized short-time power spectra and autocorrelation functions. J. Acoust. Soc. Am. **34**(11), 1679–1683 (1962)
30. C.H. Shadle, Phonetics, acoustic, in *Encyclopedia of Language and Linguistics*, vol. 9, 2nd edn., pp. 442–460, ed. by K. Brown (Elsevier, Amsterdam, 2006)
31. C.H. Shadle, C.U. Dobelke, C. Scully, Spectral analysis of fricatives in vowel context. J. Phys. IV **2**(Colloque C1), 295–298 (1992)
32. K.N. Stevens, *Acoustic Phonetics* (The MIT Press, Cambridge, 1998)
33. A. Stuart, J.K. Ord, *Distribution Theory*, *Kendall's Advanced Theory of Statistics*, vol. 1 (Edward Arnold, London, 1994)
34. R. Wayland, A. Jongman, Acoustic correlates of breathy and clear vowels: the case of Khmer. J. Phonetics **31**, 181–201 (2003)
35. P.D. Welch, The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans. Audio Electroacoust. **15**(2), 70–73 (1967)

Chapter 5
Alternative Time–Frequency Representations

The spectrogram is a well-studied time–frequency representation, but there are numerous others. There has been a rich literature on this subject, and many different time–frequency representations have been devised, studied, and applied to various signal analysis problems (e.g. [1]). Unfortunately, the subject has never to my knowledge been made accessible to speech scientists, with the result that we have rarely availed ourselves of any such representations other than the spectrogram. This chapter is an attempt to rectify this situation somewhat, although the presentation takes on a more advanced mathematical character at certain points.

Any time–frequency representation can be viewed pragmatically in two ways. Firstly, it can be thought of as an attempt to show how the energy in a signal is distributed in the time–frequency plane; certainly, a spectrogram can be viewed in this way. This view gave rise to the term *distribution* being used in place of *representation* much of the time. Quite often, however, a subject signal is in some physical sense known to consist of a number of distinct "components," each having its own frequency that may be changing through time. Thus, just as a Fourier spectrum is often viewed as a "decomposition" of a stationary signal showing its frequency components, a time–frequency representation of a non-stationary signal may often be viewed as a decomposition showing how the various component signals change in frequency with the passage of time. In this context, the distinct elements of a multicomponent signal are called *line components*. What is really sought in many cases is then the instantaneous frequencies of the line components in a signal, although this is not a notion that is rigorously definable [13].

This chapter introduces the general theory of quadratic, or bilinear, time–frequency representations in the most gentle way I could come up with. In order to do this, it is necessary to discuss the Wigner–Ville distribution, which plays a central part in defining the class of quadratic distributions. My ultimate goal, however, is to promote a particular quadratic time–frequency representation called the Zhao–Atlas–Marks distribution, as a possible alternative to the spectrogram for speech analysis. The idea here is to describe the representation, show a few

S. A. Fulop, *Speech Spectrum Analysis*, Signals and Communication Technology,
DOI: 10.1007/978-3-642-17478-0_5, © Springer-Verlag Berlin Heidelberg 2011

examples comparing it to spectrograms, and enable the reader to be able to go ahead and try the ZAM distribution in various situations where a spectrogram might be applied. It is certainly not necessary to acquire a deep understanding of the theoretical points, any more than most speech scientists have ever understood the spectrograms which they have employed nevertheless.

5.1 Wigner–Ville Distribution

The Wigner–Ville distribution (WVD) is of tremendous theoretical importance in time–frequency analysis [6]. It has many interesting and contradictory properties which make it at once appealing for applied signal analysis, but also virtually useless for this purpose. Understanding the WVD is important as an entrée into the realm of the "quadratic" time–frequency distributions, all of which are definable and understandable in terms of the WVD. It is useful to note at first that, while the STFT is a linear transform and is not a member of the quadratic class, the spectrogram derived from it is a quadratic distribution which can be defined using the WVD instead of the STFT. Considering the good points, the WVD of a linear frequency modulated sinusoid (i.e. a linear FM chirp) *exactly* shows the instantaneous frequency of the chirp, which is enticing for those of us interested in determining the instantaneous frequencies of signal components. Secondly, the theoretical WVD does not depend on a choice of window; it is in this sense the "canonical" time–frequency analysis of a signal, depending only upon the signal itself.

On the minus side, unlike a spectrogram the WVD's real values are quite often negative, which spoils the physical interpretability, and generally the images of a WVD are displayed to show only the positive part. Secondly, while the WVD can precisely follow the instantaneous frequency of certain simple test signals, for more complicated modulations or multicomponent signals, the bilinearity of the transform introduces interferences and "cross-terms" in the time–frequency distribution which make it virtually unreadable as an analysis of speech. Nevertheless, we will observe later in the chapter that a close relative of the WVD is in many respects an improvement upon the spectrogram for speech spectrum analysis.

5.1.1 Definition and Theory

The Wigner–Ville distribution of a signal $z(t)$ is defined using the following integral in continuous time [5]:

$$W_z(t,\omega) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} z\left(t+\frac{\tau}{2}\right) z^*\left(t-\frac{\tau}{2}\right) \mathrm{e}^{-\mathrm{i}\omega\tau}\,\mathrm{d}\tau. \tag{5.1}$$

It was first derived by the physicist Eugene Wigner for use in quantum mechanics, where he attempted to represent a probability distribution over paired physical properties which are not simultaneously measurable [21]. The distribution was thereafter derived in a signal analysis setting by Ville [20], who provided no evidence that he was aware of Wigner's quasi-probability distribution of a similar nature. In the context of signals, time and frequency play the roles of the physical properties which are not simultaneously measurable. This distribution, like most other relatives including the Zhao–Atlas–Marks representation, works best for signal analysis when $z(t)$ is the analytic signal associated to a real signal of interest.

The core of the expression for the WVD clearly resembles the definition of autocorrelation, except that there are two time variables, τ for the lag and t for the signal time. This function:

$$K_z(t,\tau) = z\left(t+\frac{\tau}{2}\right)z^*\left(t-\frac{\tau}{2}\right) \tag{5.2}$$

is also known as the *instantaneous autocorrelation*, and we can now see that the WVD is its Fourier transform from *lag* to frequency. Since I already mentioned how a simple power spectrum is the Fourier transform of a non-normalized autocorrelation (i.e., an autocovariance), it is at least plausible that the WVD will serve as a kind of "instantaneous" spectrum, being as it is the Fourier transform of the instantaneous autocorrelation.

5.1.2 Discrete Implementation

While the STFT was in some sense originally conceived as a discrete transform by Gabor, a fair amount of ingenuity was involved in the initial formulation of a discrete Wigner–Ville distribution by Claasen and Mecklenbräuker [10]. The discrete version of the WVD can be obtained by sampling the continuous distribution, but some snags have to be dealt with.

The method for N samples of a digital signal $z(n)$ with sampling rate f_s involves sampling the instantaneous autocorrelation function at the core of the transform, using integer index m to represent the lag samples and integer index k to represent the discrete frequency bins as in a DFT. The correctness of the following expression is proven in [8]:

$$W_z\left(\frac{n}{f_s},\frac{kf_s}{2N}\right) = 2\sum_{|m|<N/2} z\left(\frac{n+m}{f_s}\right)z^*\left(\frac{n-m}{f_s}\right)\exp\left[\frac{\mathrm{i}2\pi km}{N}\right] \tag{5.3}$$

Changing variables to get rid of the sampling rate yields the following expression for the *discrete Wigner–Ville distribution* of $z(n)$:

$$W_z(n,k) \stackrel{\text{def}}{=} 2 \sum_{|m|<N/2} z(n+m)z^*(n-m) \exp\left[\frac{\mathrm{i}2\pi km}{N}\right] \tag{5.4}$$

By comparing with Eq. 2.35 for the DFT, it can be seen that the above expression involves a DFT of the sampled instantaneous autocorrelation (assuming the necessary periodicity of the DFT), and so the above definition can be slightly recast in a format friendly to computation [8]:

$$W_z(n,k) = 2 \underset{m \to k}{\mathrm{DFT}}\{z(n+m)z^*(n-m)\}. \tag{5.5}$$

An improvement upon the above specification of a discrete WVD has recently been published [19], but I hope I am correct in my judgement that the practical differences for our purposes will be slight.

In order to implement the above equation a number of finer algorithmic points need to be dealt with, which unfortunately seem never to have been discussed at length in the signal processing literature. To expound on this I am relying on a few tips scattered around [2, 9], together with my own analysis of some public domain Matlab code released by Auger and colleagues [4]. Given a digital signal $s(n)$, the following steps can be taken to compute its discrete WVD. The reader is referred to the linked Matlab code for exact details.

1. Compute the associated analytic signal $z(n)$;
2. Initialize a vector `t` of numbers from 1 to the length of the signal (in samples), stepping uniformly by some number `step`.
3. The length `N` of the lag window must be set; it is normally input by the user.
4. Initialize the discrete Wigner–Ville matrix with `N` frequency bins and `length(t)` time columns.
5. For each time column `ti` positioned at a sample point in time vector `t`, set up the lag window vector `tau` to be centered around 0 with length `N`, or some smaller length as needed to prevent the lag window from running beyond the ends of the signal.
6. Optionally, initialize a tapered window function (e.g. Kaiser, Gaussian) the same length as the lag window.
7. Using the current lag window, compute the current time column of the instantaneous autocorrelation matrix as $\mathtt{z(ti+tau)z^*(ti-tau)}$, optionally multiplying pointwise by the tapered window function; using a tapered window will produce a discrete *pseudo-Wigner–Ville* distribution that is smoothed in frequency.
8. When the autocorrelation matrix has been constructed column-by-column in this fashion, the discrete WVD is computed as the real part of the FFT of the matrix.

The above outlined procedure contains a number of parameters not present in the "pure" Wigner–Ville distribution. First of all, we set a time step of a certain number of samples; this amounts to a time frame of samples over which the signal is assumed to be unchanging. In practice, this "time window" decimates the

transform to reduce computational requirements, and should be "smaller than the local time of stationarity" [9]; to approximate the pure WVD, the value of `step` should be set to 1 (i.e. not decimated). Secondly, the procedure sets up a "lag window" that is possibly shorter than the entire signal; again this is to facilitate computation, and to approximate the pure WVD the value `N` should be set to the entire signal. Strictly, the WVD with a lag window is called the *pseudo-Wigner–Ville* distribution [17]; as with the spectrogram, a rectangular window is not the best choice, and it is conventional to multiply by a tapered window function $w(\tau)$ in the lag domain. A window in the lag domain has the effect of smoothing in the frequency dimension, and is often referred to as a frequency-smoothing window.

The action of a lag window as a frequency smoothing window derives from the following identity for a signal $z(t)$ [12]:

$$\int_{-\infty}^{\infty} h(\tau) z\left(t+\frac{\tau}{2}\right) z^*\left(t-\frac{\tau}{2}\right) \mathrm{e}^{-\mathrm{i}2\pi\nu\tau}\,\mathrm{d}\tau = \int_{-\infty}^{\infty} H(\nu-f) W_z(t,f)\,\mathrm{d}f, \quad (5.6)$$

in which $h(\tau)$ is a lag window, $H(\nu-f)$ is its Fourier transform involving a second frequency variable ν known as the *Doppler* (it is formally dual to the lag as frequency is dual to the time), and $W_z(t,f)$ is the Wigner–Ville distribution of the signal.

5.1.3 Features of the Wigner–Ville Distribution

The Wigner–Ville distribution is of interest here chiefly to allow the introduction of other quadratic time–frequency distributions, but let me illustrate some of its features in contrast with the spectrogram. The spectrogram is fundamentally a short-time analysis, meaning the choice of time window is its most important parameter. The WVD, by contrast, does not necessarily involve a window as such, except where needed to make computation tractable. Fundamentally, in its unvarnished form the WVD provides a joint time–frequency representation of an entire signal at once, by making use of the instantaneous autocorrelation function. These factors are illustrated in Fig. 5.1, in which spectrograms are compared with a WVD of a simple double sine wave. The upper panel, with rectangular windows, shows the extreme importance of a tapering function for a spectrogram; there is a great deal of spectral leakage everywhere. The middle panel, with 100 ms Gaussian windows, shows how even a long window like this does not produce a spectrogram with fabulous frequency localization; the sine waves are rendered as fat bands. The Wigner–Ville distribution in the bottom panel, by contrast, has remarkable frequency localization on the sine waves, but it is contaminated by the presence of a "false component" at a frequency midway between the true components. Such artifacts are called *cross-terms*, and their presence in every WVD is what makes the representation all but

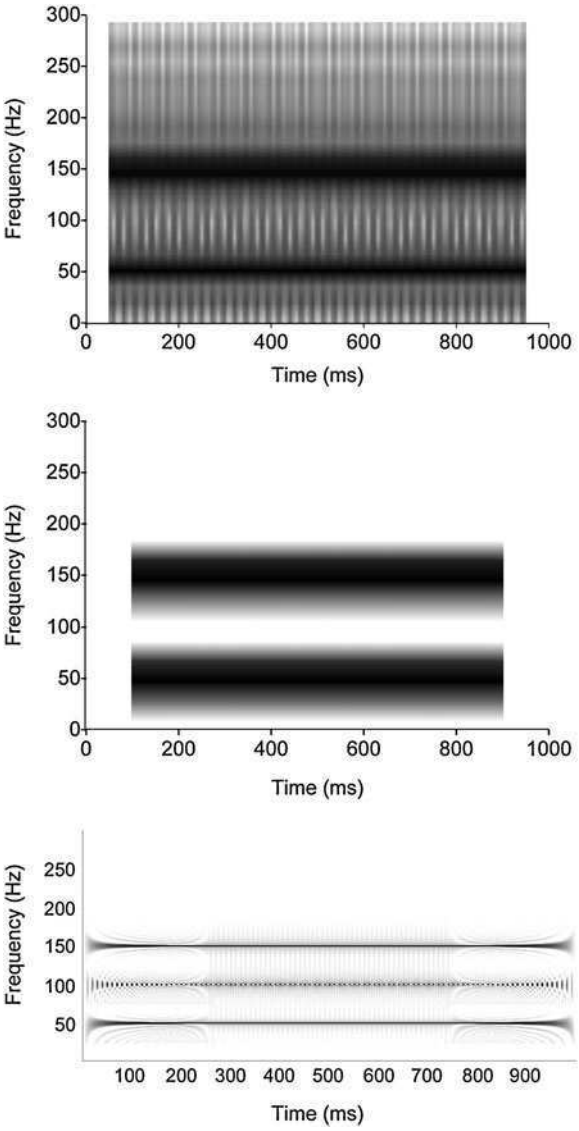


Fig. 5.1 The *top two panels* show spectrograms of a signal consisting of two sine waves (50 and 150 Hz); both are computed using 100 ms analysis windows, only the *middle panel* uses a Gaussian taper function. The *lower panel* shows a Wigner–Ville distribution of the same signal, computed with 0.512 s rectangular lag windows, and 1.56 ms time steps. This somewhat approximates the "pure" WVD

useless for speech analysis. Technically, a spectrogram still has cross-terms, but they only arise between components or events whose STFTs overlap [15]—i.e., when the components (or events) are too close in frequency (or time) to be fully resolved because of the spectrographic uncertainty principle.

Figure 5.2 shows four Wigner–Ville distributions of a sine wave with an up and down frequency modulation in the middle. The top left is a "pure" WVD (approximated as closely as possible by the discrete algorithm here). This representation is highly localized on the signal itself, but is once again contaminated by interference.

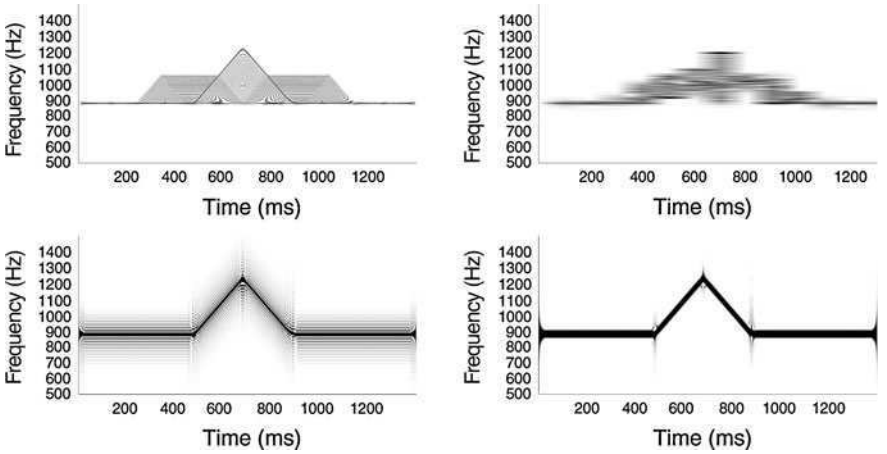


Fig. 5.2 Four Wigner–Ville distributions showing a sine wave with up and down linear frequency modulation. *Top row*: 1 ms time step, full signal lag window (approximates pure WVD); 100 ms time step, 410 ms rectangular lag window. *Bottom row*: 1 ms time step, 5 ms rectangular lag window; 1 ms time step, 5 ms Kaiser lag window

These are not cross-terms since there is only one component here, but are instead known as *inner interferences* [12] resulting from a modulating component interfering with itself; either way, they are a most undesirable feature of the WVD.

The top right panel uses 100 ms time steps, which as in a spectrogram has the effect of assuming the signal is unchanging over 100 ms increments. You can see that this assumption is totally unwarranted for the signal at hand, so that the linear frequency modulation is no longer properly located. The bottom left goes back to the fine 1 ms time steps, only this time a 5 ms rectangular window in the lag domain is employed, thus making a pseudo-Wigner–Ville distribution. This has a smoothing effect in the frequency domain, which eliminates a large amount of the interference. The bottom right shows a pseudo-Wigner–Ville distribution which is equivalent to the left panel, but with a Kaiser window function applied to the 5 ms lag frames. It compares very favorably to the two spectrograms of the same signal shown in Fig. 5.3. You can see that the pseudo-WVD seems to be quite a useful representation because of the tapering function. In general, however, this simple approach to smoothing the distribution is not very effective at eliminating cross-terms, so we next move to the more successful and more complicated approach involving generalized quadratic distributions.

5.2 Zhao–Atlas–Marks Distribution

We have so far seen that the Wigner–Ville distribution has excellent theoretical time–frequency resolution, but even in "pseudo-" form with a tapered lag window, it displays too much interference and cross terms to be a useful analytical tool.

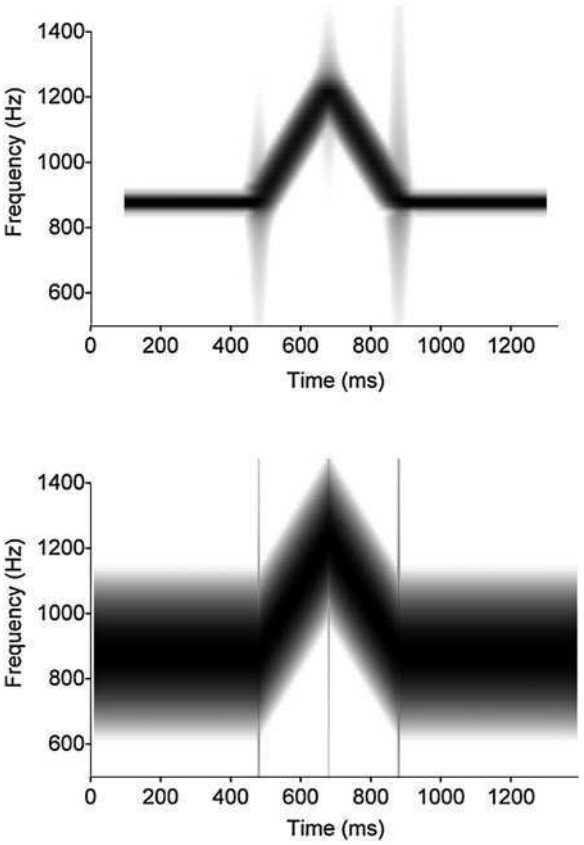


Fig. 5.3 Two spectrograms of the same signal from the previous figure. *Top*: 200 ms Gaussian window; *bottom*: 20 ms Gaussian window

The spectrogram displays very little interference and cross terms, but its time–frequency resolution is not so good. The Zhao–Atlas–Marks distribution family [25] (also known as the "cone-shaped kernel" distributions) is one example of a good compromise; its resolution is "quite superior to the spectrogram and the interference is quite insignificant" [3]. The ZAM has properties that are well-suited to examining speech signals [3], and which can improve upon spectrograms in many cases. "Heuristically, we can say that the time–frequency smoothing by a spectrogram is based on only one 'degree of freedom,' as it employs a unique short-time window" [12]. With the ZAM, however, it is possible to take both the time and frequency dimensions into consideration, and proceed to develop smoothing that has two degrees of freedom.

5.2.1 Quadratic Distributions

I showed in the previous section that the Wigner–Ville distribution derives directly from the instantaneous autocorrelation function by a Fourier transform in the lag

variable. It was also shown that the pseudo-Wigner–Ville distribution in turn derives from the pure WVD through multiplying by a lag window $w(\tau)$ inside the Fourier transform, and that this can smooth the frequency dimension of the WVD. It was first shown by Cohen [11] in a quantum mechanics context that a more general approach to smoothing the WVD is both possible and desirable; instead of just multiplying by a lag window, one can perform a *time convolution*[1] of the instantaneous autocorrelation with a two-dimensional function $G(t,\tau)$ called the *time-lag kernel*, and this has the effect of smoothing both the time and frequency dimensions of the final distribution. The following expression defines the general class of quadratic time–frequency distributions which can be computed in this fashion:

$$\rho_z(t,f) = \underset{\tau\to f}{\mathscr{F}}\left\{G(t,\tau)\underset{t}{*}\left[z\left(t+\frac{\tau}{2}\right)z^*\left(t-\frac{\tau}{2}\right)\right]\right\} \tag{5.7}$$

where the symbol $\mathscr{F}$ is the Fourier transform operation and $\underset{t}{*}$ indicates a time convolution. A great deal of research in time–frequency analysis has been devoted to the development of useful kernels within this so-called "Cohen class." It is also important to remember that the spectrogram itself can be defined and computed in a similar fashion to the WVD, thus showing that the spectrogram is a special member of Cohen's class—in particular it is a type of quadratic distribution which is always positive and has maximally suppressed interference and cross-terms.

5.2.2 Discrete Implementation

Converting the expression (5.7) for a general quadratic time–frequency distribution into discrete time–frequency requires a number of steps similar to the derivation of the discrete WVD; the reader is referred to [6] for an account of the steps which lead from the above equation to the discrete time–frequency version:

$$\rho_z(n,k) = 2\underset{m\to k}{\mathrm{DFT}}\{G(n,m)\underset{n}{*}[z(n+m)z^*(n-m)]\} \tag{5.8}$$

where n is the sampled time, m is the sampled lag, and k is the sampled frequency. The above definition is intended to be friendly to computation; a rough algorithm for computing any particular quadratic time–frequency distribution from a digital signal $s(n)$ is equal to the earlier algorithm for the discrete WVD, with the addition of one step following (or during) computation of the instantaneous autocorrelation. That step is to perform a discrete convolution in n (time) with the discrete form of the smoothing kernel $G(n,m)$. In some cases, an approximation to the discrete convolution may be the best option, owing to difficulties with the discrete-time version of the smoothing function. Detailed considerations on this point are presented in the math box below.

[1] The process of convolution will be defined below in the discrete-time context.

When the smoothing kernel for the Zhao–Atlas–Marks distribution family is converted to its discrete counterpart, it takes the following form parameterized by a [8]:

$$G(n,m) = \left[w(m)\,\mathrm{rect}\left(\frac{an}{4m}\right)\right] * *\mathrm{sinc}(n)\mathrm{sinc}(m), \tag{5.9}$$

in which $w(m)$ is a typical kind of tapering function (Kaiser, Gaussian), $\mathrm{rect}(x)$ is the standard rectangular step function [23], $\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$ is the standard (normalized) cardinal sine function [24], and where the double convolution is required in continuous-time *before* sampling the lag vector. This is computationally impossible, and can only be approximated using digital oversampling. A less computationally intensive way is to devise a different discrete time-lag kernel which approximates the effects of the convolutions in the above expression. Two possible approximate ZAM kernels are [7]:

$$G(n,m) \overset{\text{def}}{=} \begin{cases} w(m) & \text{if } |an| \le |2m| \\ 0 & \text{otherwise;} \end{cases} \tag{5.10}$$

$$G(n,m) \overset{\text{def}}{=} \frac{w(m)}{2}\left[1 + \tanh(|4m| - |2an|)\right] \tag{5.11}$$

The algorithm due to Auger et al. [4] employed in the linked Matlab code computes the ZAM distribution using discrete convolution in time only (see below for a definition) of the instantaneous autocorrelation with the first approximate kernel above. This method also adds a separate time window using a convolution operation, so the implementation is really a time-smoothed ZAM.

The discrete convolution of two complex-valued functions f, g on the integers (e.g. sampling indices) is given by [22]:

$$(f * g)(n) \overset{\text{def}}{=} \sum_{m=-\infty}^{\infty} f(m)g(n-m) = \sum_{m=-\infty}^{\infty} f(n-m)g(m) \tag{5.12}$$

The algorithm employed here for the ZAM distribution employs both a time window and a lag window, in contrast to the original ZAM distribution which has a lag window only. The two windows contribute to smoothing the interferences in the time and frequency dimensions respectively; the user can potentially change the tapering function that is applied in each window, a flexibility that generalizes the original Zhao–Atlas–Marks distribution into what we might call the *smoothed* ZAM. In addition to changing the smoothing window, a user also has independent control over the time frame used to smooth in time versus frequency; this is in contrast to the spectrogram where one always suffers from the trade-off between time and frequency smoothing/resolution. Since there is only one window for us to

set when computing a spectrogram, it affects both time and frequency localization. When the spectrogram time window is long, one sees good frequency localization but poor time localization; when the window is short, the opposite is true. With the ZAM, the time window does not affect the apparent localization of frequencies, and the lag window does not affect the time localization.

Figure 5.4 illustrates four ZAM images of an artificial signal composed of two sine waves, one of which includes a linear frequency modulation up and down. The different window values illustrate the effects of varying degrees of smoothing/resolution in each dimension. A shorter time window yields better time resolution in the ZAM, without any trade-off in the frequency dimension. A longer lag window yields better frequency resolution, with no loss of time resolution. It should be noted that, as with the WVD images presented above, these ZAM images discard negative amplitudes which may be present in the raw transform.

The top left panel of Fig. 5.4 shows the ZAM of the signal with a very short time window and a moderate lag window. Frequency localization is acceptable, but the interference between components is too spread out. This type of interference is in a sense "physically real," since it results from the phenomenon of *beating* between components. The top right panel of Fig. 5.4 shows the ZAM with longer 22 ms time window and shorter lag window; indeed it is easy to see that the resolution is poorer in both dimensions. The bottom left panel uses the 22 ms time window with a very long 205 ms lag window, which greatly improves the frequency localization. These three images were all computed using Kaiser window functions. It is important to note that the ZAM distribution can only be computed with a time window that is not longer than the lag window. The bottom right image

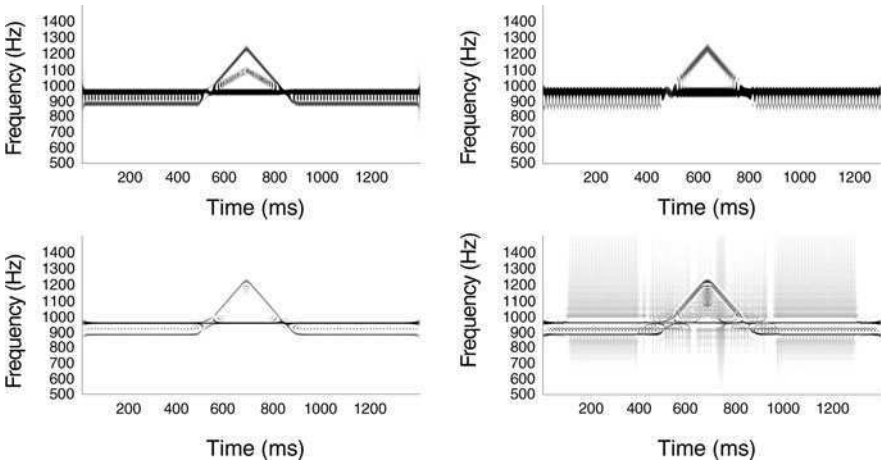


Fig. 5.4 ZAM distributions of a signal comprised of a sine wave and a second sinusoid with linear frequency modulations. *Top row*: 3 ms time, 51.3 ms lag windows; equal 22 ms time and lag windows. *Bottom row*: 22 ms time, 204.7 ms lag windows; the same but with rectangular windows (no taper)

is computed with the same parameters as the bottom left, only using rectangular windows instead of Kaiser; the need for the tapered windows is apparent from this.

In the realm of quadratic time–frequency representations, the trade-offs among time–frequency resolution, cross-term attenuation, and positivity presents something of a paradox. It is not possible for an alternative distribution to significantly improve upon the resolution of a spectrogram without introducing more cross-terms and/or more negative values. The ZAM distribution pushes a significant amount of the signal energy into the negative amplitudes, where it will not be plotted because it has no meaningful physical interpretation. The overall effect of this can rob a ZAM image of some of the signal energy that we would expect to see; in speech signals, the effect can lead to ZAM displays which show components having improperly small bandwidths [14], or improperly small amplitudes. From a certain perspective, at least the first of these problems can be an advantage, since it means that speech components can appear to be quite concentrated around their instantaneous frequencies, rendering these easier to measure. Nonetheless, the various tradeoffs negate the utility of the ZAM for measuring the amplitudes of the components in a multicomponent signal such as speech, as could be discerned from the detailed study by Hlawatsch et al. [14].

5.2.3 Speech Analysis with ZAM

For our first speech example, let us take a look at the ZAM distributions in Fig. 5.5, showing the same English utterances depicted as spectrograms in Fig. 4.15 in the previous chapter. I don't think it is outrageous to judge that much of the relevant spectral information is more clearly presented in the ZAM distributions here, although this distribution does tend to eliminate some of the signal information that we might wish to see (probably the missing energy has been pushed over to false components with negative amplitudes which are not displayed). The ZAM images appear to favor the frequency components of the signal rather more than the impulsive events or the noise, for one thing, and this may or may not be desirable in every application. When it comes to depicting the formant frequencies, the ZAM images do seem to come out ahead of the spectrograms. The formants are narrower in the ZAM images, and thus more easily separated from each other and from the voice bar—in spite of the ZAM images having been computed with shorter smoothing windows, whose effect generally would be to decrease the frequency localization. Note in particular that F_1 in *heed* is well-separated just above the much louder voice bar, and also that F_1 and F_2 in *hawed* can be discerned below and above 500 Hz respectively. The improved formant localization is due to the ZAM distribution being governed by its own variation of the uncertainty principle, which is more favorable to time–frequency localization than the spectrographic uncertainty. There is no way to "beat" the uncertainty principle entirely, but it is evidently possible to improve upon the spectrogram in this respect.

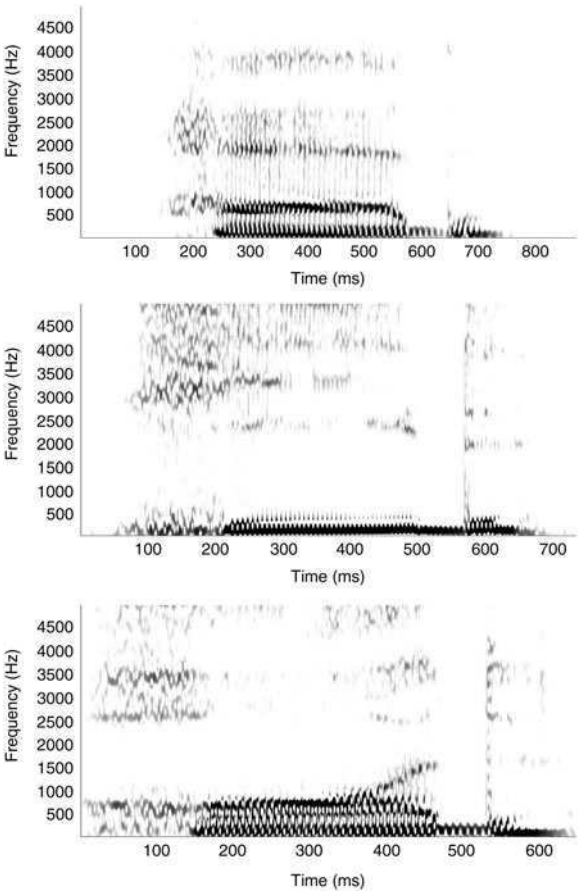


Fig. 5.5 Zhao–Atlas–Marks distributions of English words *had, heed, hawed* (*top to bottom*). Computed with 8 ms time and lag window dimensions, using a Kaiser tapering function

Figure 5.6 illustrates the flexibility of the smoothed ZAM for speech, stemming from the two window parameters. The top panel shows an image computed with 8 ms time and lag windows, producing something reminiscent of a wideband spectrogram but with superior frequency localization. The middle panel uses 40 ms for both windows, which yields something similar to a narrowband spectrogram. The lower panel mixes the two conditions; the 40 ms lag window is long enough to resolve the harmonics of the voice, while the 8 ms time window is short enough that we can observe the vocal cord impulses represented as specks between harmonics, and also show the release burst of the final [d] with excellent time localization. This kind of mixed "wide and narrowband" time frequency representation cannot be contemplated, much less approximated, by means of a spectrogram.

Turning to our little corpus of synthesized vowels, Fig. 5.7 revisits two of the problematic vowels that were shown in spectrograms in Fig. 4.14. In the ZAM images, there is significantly improved formant localization. Moreover, in these cases it is of considerable utility to zoom in on a small number of glottal

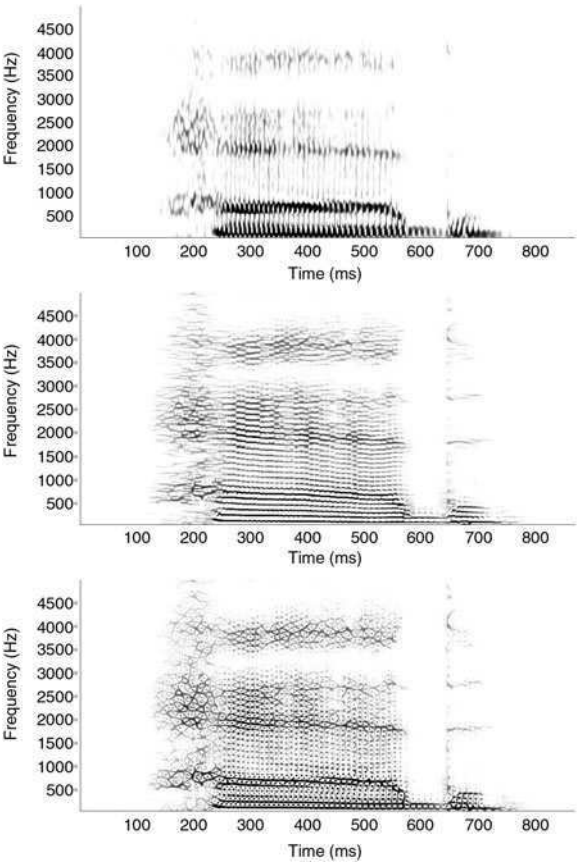


Fig. 5.6 ZAM distributions of English word *had*. *Top image* computed using 8 ms time and lag smoothing windows; *middle image* computed with 40 ms time and lag windows; *bottom image* computed with 8 ms time window and 40 ms lag window

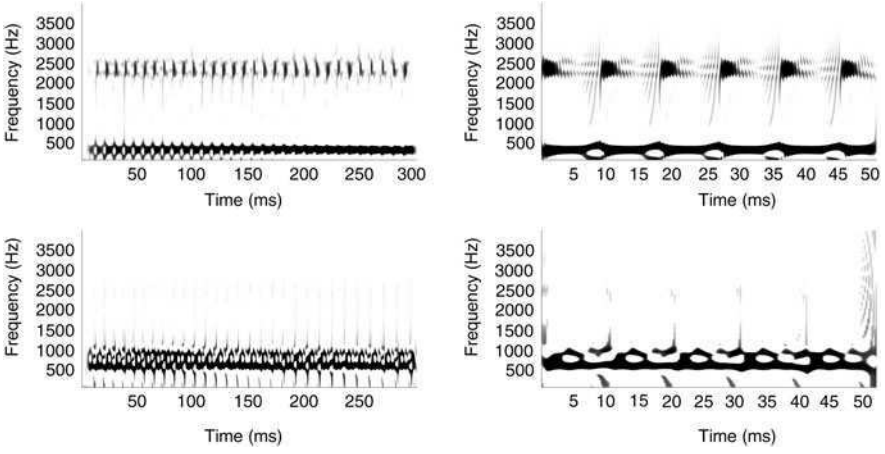


Fig. 5.7 ZAM distributions of synthesized vowels, all computed with 7 ms Kaiser windows in time and lag. *Top row*: [i] full length and close-up of a few glottal pulses. *Bottom row*: [ɔ] full length and close-up of glottal pulses

pulsations. F_2 and F_3 can be observed separately following the initial impulse excitations during the vowel [i], while F_1 and F_2 of [ɔ] can be quite readily separated and measured with reasonable precision. It is also worth keeping in mind that these synthesized vowels do not, by design, include a voice bar, which is so often observed in natural speech (v. the discussion in the next chapter).

One notable feature is the apparent "splitting" of F_1 that is evident in the magnified analysis of [i]. This feature has been noted in ZAM distributions of speech signals [16], where it was explained that the apparent splitting of a component can be indicative of a sudden phase shift (also explained in greater detail in [12]), and the authors speculated that the feature could indicate a phase jump that is theoretically predicted to occur at the instant of glottal closure. This glottal phase jump is expected because the fundamental frequency of phonation is in general not commensurate with the formants, and so each new glottal impulse effectively "cuts" the formant off abruptly when it begins anew. The Wigner–Ville and ZAM distributions are both much more sensitive to phase shifts in a component than the spectrogram.

In the spectrograms of these vowels in Fig. 4.14, it was observed that the formants are too poorly localized to be measured reliably. F_2 and F_3 were smeared together in the vowel [i], F_1 and F_2 were smeared together in the vowel [ɔ], and zooming in to show a few glottal pulsations showed extremely fat formants that would be very difficult to accurately measure. The difference in time–frequency

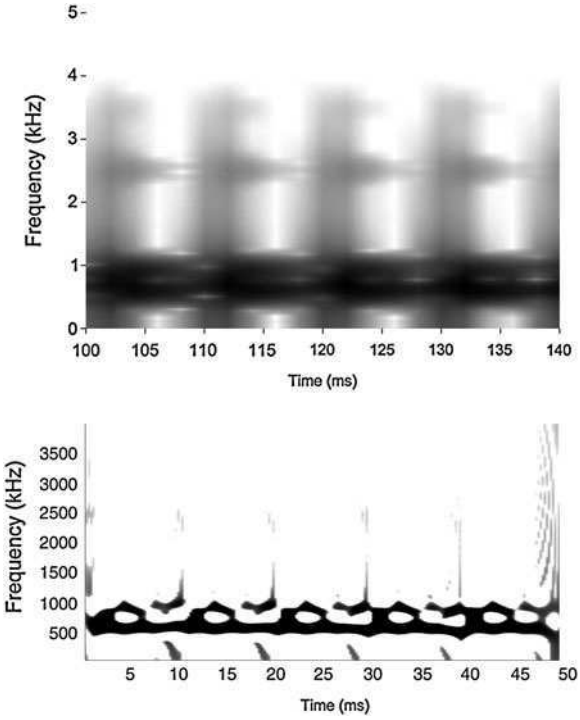


Fig. 5.8 Comparing spectrogram (*top*) with ZAM distribution of a few glottal pulses from synthesized [ɔ]

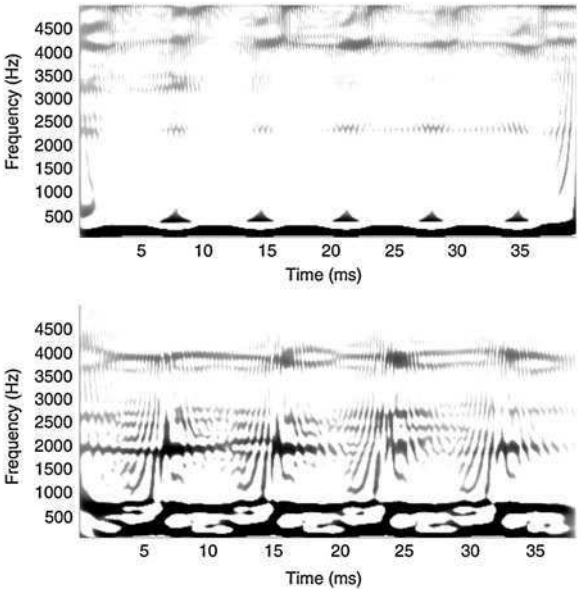


Fig. 5.9 ZAM distributions showing natural English vowels [i] (*top*) and [æ], computing using 9.3 ms Kaiser windows in the time and lag domains

localization between spectrogram and ZAM is starkly illustrated in Fig. 5.8, although the ZAM image has the drawback that faint signal components are attenuated too much, so that they can hardly be found. The ZAM in the figure was computed with 100 dB (i.e. complete) dynamic range in an effort to bring out the faint F_3 of this vowel.

The superior resolution and precision of the ZAM distribution makes it more amenable to zooming in on small portions of a speech signal. Figure 5.9 shows two ZAM close-ups of the author's phonation during the vowels of English *heed* and *had*. While these "magnified" images of phonation show formants more clearly than the spectrograms of Fig. 4.15, it is still not obvious that one could make a good measurement of all of the relevant formants. In particular, F_1 of [i] is still indiscernible from a voice bar. With these kinds of close-up images, however, it is possible to view the fine structure of phonation in much greater detail than has previously been typical. Such images lead to new questions concerning the very nature of the time–frequency structure of phonation, questions which I will attempt to address by applying yet a new approach to time–frequency analysis in the next chapter.

5.3 Appendix: Matlab Routines

When using the routines provided for the WVD and ZAM distributions, it is extremely important to first convert the subject signal to its associated analytic signal, as this will prevent a large amount of the aliasing that is inherent from the discretization of the schemes. The mfile `newhilbert.m` has been written to

incorporate some very recent developments to refine the algorithm for computing the discrete analytic signal by means of the Hilbert transform [18]. It is invoked simply by:

$$\texttt{output} = \texttt{newhilbert}\,(\texttt{xr}, \texttt{n});$$

where `output` names the new analytic signal, `xr` is the input real signal (optionally a matrix of signal columns can be used to transform multiple signals at once) and optional parameter `n` is the frame length for the transform, which is set equal to the signal length if the parameter is missing.

The basic Wigner–Ville distribution is produced using mfile `WignerVille.m` according to the following template:

$$[\texttt{wvd}, \texttt{f}, \texttt{t}] = \texttt{WignerVille}(\texttt{signal}, \texttt{Fs}, \texttt{step}, \texttt{fftn}, \texttt{low}, \texttt{high}, \texttt{clip})$$

where the `step` parameter sets the hop size from one analysis point to the next (if it is greater than one, the resulting distribution is decimated), and `fftn` sets the WVD overall length (and also the frequency sampling), which is the maximum lag window length (the lag window is altered throughout the distribution, so that it stays within the confines of the signal). The hop size effectively introduces a kind of time window that does no smoothing; a larger step assumes longer-term stationarity of the signal. All such parameters in the code which set a particular length within the signal are currently in units of samples.

A form of pseudo-Wigner–Ville distribution is carried out by `WinWignerVille.m`, using a Kaiser tapering function applied to the lag window. As usual in all these mfiles, the functions can be called with or without the value-passing variables and the equals sign. For the Wigner–Ville and Zhao–Atlas–Marks functions, the core algorithm is taken from the public-domain Matlab code released by Auger and colleagues [4].

The smoothed Zhao–Atlas–Marks distribution can be created using `ZAM.m` for a full-color plot, or `ZAMgray.m` for a spectrogram-like grayscale. All color time–frequency routines here use my customized colormap `myjet.m` to plot amplitudes. This color scheme uses a "rainbow" progression from dark red (loudest) through red, orange, and yellow, to green (quietest). The two ZAM functions are called according to the following:

$$[\texttt{zam}, \texttt{f}, \texttt{t}] = \texttt{ZAM}(\texttt{signal}, \texttt{Fs}, \texttt{window}, \texttt{step}, \texttt{fwin}, \texttt{fftn}, \texttt{low}, \texttt{high}, \texttt{clip})$$

in which there are numerous parameters to be provided. This modified form of ZAM involves fully separated smoothing in time and frequency; the lengths of these smoothing windows (tapered by a Kaiser function in the code) are set by `window` and `fwin`, respectively. Note that this algorithm requires the windows to have an odd number of samples. The value of `step` determines the hop size between points of analysis, and `fftn` is the Fourier transform frame size which determines the frequency sampling. Should the intrepid or knowledgeable reader choose to examine my code, it is useful to remember that these functions which compute a bilinear time–frequency representation invariably perform a Fourier

transform of an instantaneous autocorrelation matrix, and all the action is in the creation of the latter matrix by discrete convolution.

One may choose to try measuring frequencies of components, such as formants, using a ZAM image. To facilitate this, an additional function `ZAMm.m` has been included, which adds a small figure in the lower left of the screen that displays the precise frequency under the mouse pointer in the main ZAM figure. This frequency is still reported after changing the zoom view of the main image.

References

1. M. Akay (ed.), *Time Frequency and Wavelets in Biomedical Signal Processing* (IEEE Press, Piscataway, 1998)
2. M.G. Amin, Time and lag window selection in Wigner–Ville distribution, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (IEEE, New York, 1987), pp. 1529–1532
3. L.E. Atlas, P.J. Loughlin, J.W. Pitton, Signal analysis with cone kernel time–frequency distributions and their application to speech, in *Time–Frequency Signal Analysis*: *Methods and Applications*, Chap.16, ed. by B. Boashash (Halsted Press, New York, 1992)
4. F. Auger, P. Flandrin, P. Gonçalvès, O. Lemoine, Time–Frequency Toolbox Reference Guide (1996)
5. B. Boashash, Heuristic formulation of time–frequency distributions, in *Time Frequency Signal Analysis and Processing*, Chap. 2, ed. by B. Boashash (Elsevier, Amsterdam, 2003)
6. B. Boashash, Theory of quadratic TFDs, in *Time Frequency Signal Analysis and Processing*, Chap. 3, ed. by B. Boashash (Elsevier, Amsterdam, 2003)
7. B. Boashash, G.R. Putland, Computation of discrete quadratic TFDs, in *Time Frequency Signal Analysis and Processing*, Chap. 6.5, ed. by B. Boashash (Elsevier, Amsterdam, 2003)
8. B. Boashash, G.R. Putland, Discrete time–frequency distributions, in *Time Frequency Signal Analysis and Processing*, Chap. 6.1, ed. by B. Boashash (Elsevier, Amsterdam, 2003)
9. B. Boashash, L. White, J. Imberger, Wigner–Ville analysis of non-stationary random signals, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (IEEE, New York, 1986), pp. 2323–2326
10. T.A.C.M. Claasen, W.F.G. Mecklenbräuker, The Wigner distribution—a tool for time–frequency signal analysis, part II: discrete-time signals. Philips J. Res. **35**(4/5), 276–300 (1980)
11. L. Cohen, Generalized phase-space distribution functions. J. Math. Phys. **7**(5), 781–786 (1966)
12. P. Flandrin, *Time–Frequency/Time-Scale Analysis*, English edn. (Academic Press, San Diego, 1999)
13. K. Gröchenig, *Foundations of Time–Frequency Analysis* (Birkhäuser, Boston, 2001)
14. F. Hlawatsch, T.G. Manickam, R.L. Urbanke, W. Jones, Smoothed pseudo-Wigner distribution, Choi–Williams distribution, and cone-kernel representation: ambiguity-domain analysis and experimental comparison. Signal Process. **43**, 149–168 (1995)
15. S. Kadambe, G.F. Boudreaux-Bartels, A comparison of the existence of "cross-terms" in the Wigner distribution and the squared magnitude of the wavelet transform and the short time Fourier transform. IEEE Trans. Signal Process. **40**(10), 2498–2517 (1992)
16. P.J. Loughlin, J.W. Pitton, L.E. Atlas, Bilinear time–frequency representations: new insights and properties. IEEE Trans. Signal Process. **41**(2), 750–767 (1993)
17. W. Martin, P. Flandrin, Wigner–Ville spectral analysis of nonstationary processes. IEEE Trans. Acoust. Speech Signal Process. **33**(6), 1461–1470 (1985)

18. J.M. O'Toole, M. Mesbah, B. Boashash, A new discrete analytic signal for reducing aliasing in the discrete Wigner–Ville distribution. IEEE Trans. Signal Process. **56**(11), 5427–5434 (2008)
19. J.M. O'Toole, M. Mesbah, B. Boashash, Improved discrete definition of quadratic time–frequency distributions. IEEE Trans. Signal Process. **58**(2), 906–911 (2010)
20. J. Ville, Théorie et applications de la notion de signal analytique. Cables et Transmission **2A**(1), 61–77 (1948)
21. E.P. Wigner, On the quantum correction for thermodynamic equilibrium. Phys. Rev. **40**(5), 749–759 (1932)
22. Wikipedia (2009), Convolution. http://www.wikipedia.org
23. Wikipedia (2009), Rectangular function. http://www.wikipedia.org
24. Wikipedia (2009), Sinc function. http://www.wikipedia.org
25. Y. Zhao, L.E. Atlas, R.J. Marks II, The use of cone-shaped kernels for generalized time–frequency representations of nonstationary signals. IEEE Trans. Acoust. Speech Signal Process. **38**(7), 1084–1091 (1990)

Chapter 6
The Reassigned Spectrogram

This chapter introduces a relatively new modified form of the spectrogram which has variously been described by the term *reassigned*, or by the phrase *time-corrected instantaneous frequency*. While the latter is more descriptive of the scheme, the former is shorter and seems to have gained supremacy in the (still rather sparse) literature on the subject. The reassignment process yields a modification of a spectrogram which effectively "sharpens" it, concentrating the smeared out spectrographic points around tighter lines in both the frequency and time dimensions. This approach relies on the understanding of the spectrogram as showing instantaneous frequencies of line components, and sharpens that view of things while doing away with the idea of showing how energy is distributed in the time–frequency plane.

In the first section, I describe how reassignment of the spectrogram works from a theoretical and historical perspective. Next, an algorithm for reassignment is presented in some detail. Following this, I describe a further modification of a reassigned spectrogram using a process I call *pruning*; this allows one to selectively eliminate "spurious" points from the image which are not likely to be associated to either a line component or an impulse in the signal. Such extra points arise from randomness that afflicts the reassignment scheme at low amplitudes, and also from the interference terms that are a natural element of the spectrogram prior to reassignment.

Having described all of the processing methods, I turn next to the applications. It will first be observed that reassigned spectrograms can provide unprecedented imaging of the fine time–frequency structure of the phonation process at high magnification. Indeed, so many new aspects of phonation are revealed with the technique that our current understanding of the phonation process itself is inadequate to explain them all. After discussing phonation as such, I turn to the problem of formant measurement, which is now found to be both easier because of the clarity of the image, and harder because the detailed view of phonation spotlights the shortcomings of simple speech production models. The chapter goes on to show reassigned spectrograms applied to consonants, and finally to pitch tracking;

S. A. Fulop, *Speech Spectrum Analysis*, Signals and Communication Technology,
DOI: 10.1007/978-3-642-17478-0_6,

enormous success can be realized with this last technique. This chapter is partly a revision of a previously published paper [5]. It has been rewritten to suit the current purpose, with updated discussion and all new figures.

6.1 Reassignment: History and Definitions

The historical development of the spectrogram and its mathematical representation as the short-time Fourier transform have been treated in Chap. 4. It was just following the complete derivation of the short-time Fourier transform in the 1960s [20, 27], when a paper was published by Rihaczek [26] that set the stage for the development of the reassignment method. Rihaczek, in fact, was far from thrilled by the STFT (or the spectrogram, presumably); he dismissed it as a relative curiosity because of its smearing, which prevents it from capturing the time–frequency energy distribution with precision. Rihaczek tried to do better than the STFT, and derived a complex energy density of the signal over the time–frequency plane. Integrating over each time–frequency cell of this density yields an energy distribution which is distinct from a spectrogram; it is today understood as another of the quadratic distributions considered in the preceding chapter. The important thing about it is the time–frequency points where Rihaczek's distribution gets most of its energy from, and how this pertains to the short-time Fourier transform.

A digital STFT provides, in effect, a "stack" of complex analytic signals, one for each frequency bin in the matrix. This perspective on the STFT was first highlighted by Kodera et al. in 1976 [17]; these authors recognized that most of the energy in Rihaczek's distribution (in its discrete form) is concentrated around the instantaneous frequencies of these "signals" provided at each frequency bin. So while the STFT frequency bins quantize the frequency range at regular intervals, the actual instantaneous frequency of a signal component within a particular bin can be computed more accurately using Rihaczek's equations—assuming there is only one significant component in a bin. Kodera et al. showed that the instantaneous frequency of a particular row in the STFT matrix can be computed from the time derivative of the complex argument (also called the complex phase) of the STFT, as is implied from Rihaczek's paper. I believe this marks the first point in history when anyone found much use for the complex phase in the STFT—prior to this, the STFT was seen purely as a means of defining the spectrogram, which is computed using the magnitude of the STFT and discarding the phase. It turns out that there is a great deal of important information hidden in the STFT complex phase, and this fact is what makes reassignment of the spectrogram possible.

The instantaneous frequencies as a function of time, which correspond to the stack of frequency bins in the STFT matrix, are altogether called the *channelized instantaneous frequency* (CIF) of the signal [22], and this is defined as the time derivative of the STFT phase. If there is just one signal

component dominant in the neighborhood of a frequency bin, then the CIF spectrum will show the instantaneous frequency of that component with arbitrary precision (i.e. not quantized by the discrete time–frequency grid). The CIF thus provides us with a holy grail of time–frequency analysis, viz. the time-varying instantaneous frequencies of the line components in a multicomponent signal.

An analogous (indeed, a mathematically dual) relationship holds for the quantized time axis of the digital STFT. First, one must recognize that the dual value to the instantaneous frequency in the time domain is the *group delay*, which can be thought of as the transit time or time delay introduced by the transmission of a signal through whatever system we are studying, such as the vocal tract. In order to get the output times exactly right, they need to be corrected by using the group delay, which can be thought of profitably as a "local time correction" for each time point. One can treat each time index in an STFT matrix (which corresponds to a "column vector" of values across frequency bins in the matrix) as a signal in the frequency domain, whose group delay for each frequency can also be computed using Rihaczek's equations, thus yielding a new vector of corrected event times for each matrix cell. The entire matrix of these time corrections was termed the *local group delay* (LGD) by Nelson [22], and is defined as the frequency derivative of the STFT phase.

For this chapter, the short-time Fourier transform is defined in continuous time in the following way for any window function $h(t)$:

$$\mathrm{STFT}_h(\omega, T) = \int_{-\infty}^{\infty} f(t+T)h(-t)e^{-i\omega t}\,dt. \tag{6.1}$$

The signal time variable is now represented with capital T. The channelized instantaneous frequency and local group delay are defined with the following equations in continuous time:

$$\mathrm{CIF}(\omega, T) = \frac{\partial}{\partial T}\arg(\mathrm{STFT}_h(\omega, T)) \tag{6.2}$$

$$\mathrm{LGD}(\omega, T) = -\frac{\partial}{\partial \omega}\arg(\mathrm{STFT}_h(\omega, T)) \tag{6.3}$$

These definitions of the CIF and LGD were first put into practice digitally by Kodera et al. [16, 17]. Since both definitions are just derivatives of the complex phase of a function that we already would have in digital form (viz. the STFT matrix), it is sufficient to use the finite differences of the phases in the matrix to get digital versions of the CIF and LGD. For example, the CIF can be defined as follows using a difference expression in place of the derivative; this involves the values of the complex phase at

times offset by ε, which should be something small such as two sample points.

$$\mathrm{CIF}(\omega, T) \stackrel{\text{def}}{=} \frac{1}{\varepsilon}\left[\phi\left(T + \frac{\varepsilon}{2}, \omega\right) - \phi\left(T - \frac{\varepsilon}{2}, \omega\right)\right], \tag{6.4}$$

where the symbol ϕ stands for the argument (phase) of the STFT.
A possible problem with the simple finite difference procedure just shown is that the complex phase is a quantity that "rotates" modulo 2π. Being as it is an angular measurement it is inherently confined to a circle; the quantity may take any value from 0 to 2π but then there is in general a discontinuity point, so that all integer multiples of $2\pi = 0$ as the circle repeats. When the phase is allowed to have arbitrary values, it is said to be in its *unwrapped* form without explicitly being confined to an interval $[0, 2\pi)$; this is the best form to compute finite differences from because there is not an explicit discontinuity to contend with, but normally when the complex STFT is computed, the argument is *wrapped*, i.e. confined to an interval of length 2π.
Another way to compute the finite difference approximation to the STFT phase derivatives was devised by Nelson [21, 22], who made use of the following "self-cross-spectral" surfaces defined using the STFT matrix:

$$C(\omega, T, \varepsilon) \stackrel{\text{def}}{=} \mathrm{STFT}\left(\omega, T + \frac{\varepsilon}{2}\right)\mathrm{STFT}^*\left(\omega, T - \frac{\varepsilon}{2}\right), \tag{6.5}$$

$$L(\omega, T, \varepsilon) \stackrel{\text{def}}{=} \mathrm{STFT}\left(\omega + \frac{\varepsilon}{2}, T\right)\mathrm{STFT}^*\left(\omega - \frac{\varepsilon}{2}, T\right). \tag{6.6}$$

Nelson noted that the surface C encodes the channelized instantaneous frequency in its complex argument, while L encodes the local group delay in its argument. It can be proven [8] that the complex phase of these discrete functions provides precisely what would be obtained from the finite difference procedure described above. The advantage of Nelson's innovation is that there is no longer any concern about the wrapped phase causing problems, because the finite phase difference is indirectly calculated.

It was Kodera, de Villedary and Gendrin [17] who first realized that the channelized instantaneous frequency and local group delay can form the basis of a remapping procedure, whereby the time–frequency matrix values of a digital spectrogram can be repositioned to new locations given by the corresponding instantaneous frequencies (found in the CIF matrix) and the time index points corrected by the LGD matrix values. The reassignment of the spectrogram matrix is then defined as a mapping of each magnitude value from its original frequency–time location at (ω, T) to a new location in the frequency–time plane defined by $[\mathrm{CIF}(\omega, T), T + \mathrm{LGD}(\omega, T)]$.

As a first illustration of the results of the above considerations, Fig. 6.1 shows a spectrogram of a signal composed of two sine waves, together with its reassigned

version. The reassigned spectrogram was actually computed using Nelson's method (detailed below).

6.2 Reassigning the Spectrogram

6.2.1 Nelson's Algorithm

The equations behind two approaches to the reassignment method are given in the box above, as developed by Kodera et al. and by Nelson. Other methods for computing reassigned spectrograms have also been developed; the three primary methods are described and compared by Fulop and Fitz [8]. For the purposes of this book, I have settled on the method of Nelson as the simplest both to compute and describe. Let me now lay out an algorithm for Nelson's reassignment

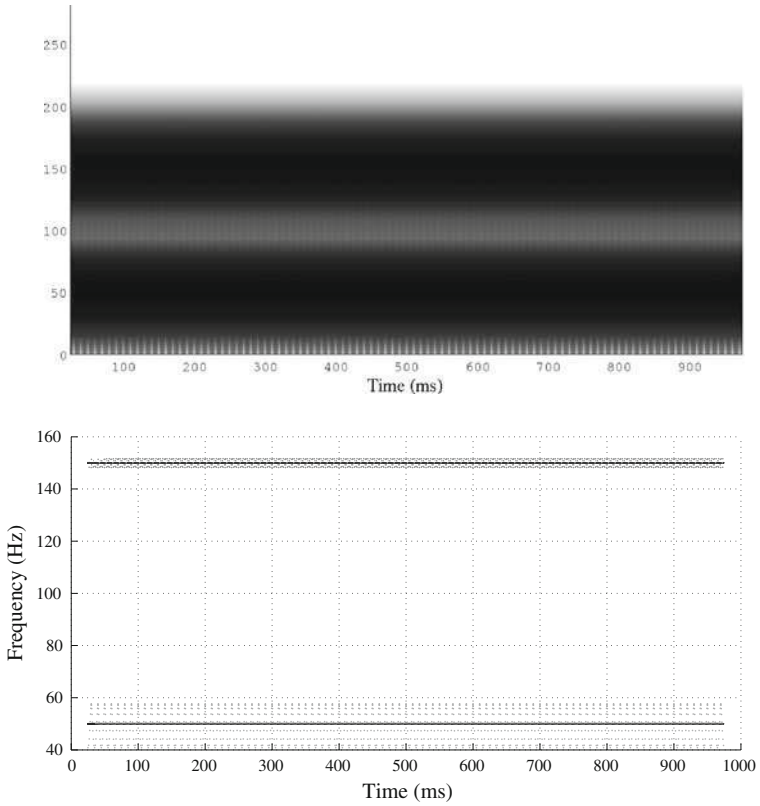


Fig. 6.1 The figure contrasts the conventional spectrogram (*upper panel*) with the reassigned spectrogram of a signal composed from a 50 and 150 Hz sine wave for 1 s, sampled at 32 kHz. Some interference artifacts can be noted. Both images computed using 1,600 point (50 ms) Kaiser windows and 20 point frame advance

technique in considerable detail. The input to the procedure is a signal sampled at `Fs` Hz, together with a number of user-defined parameters as needed below.

1. First, one builds two matrices `S` and `Sdel` of tapered signal windows of length `win_size` (user-supplied) time samples. `Sdel` has windows that are delayed by one sample with respect to `S`. A Kaiser window function should be used (see below). The signal windows must overlap by a user-input amount, which should in practice be around 95% of the window for the best display.
2. One next computes three short-time Fourier transform matrices; each column is an `fftn`-length DFT of a signal window, computed with a fast Fourier transform function here called `fft`. The length `fftn` is user-input; if it is longer than `win_size` (often an excellent idea), it is then zero-padded beyond that.

   `STFTdel = fft(Sdel)`
   `STFT = fft(S)`
   `STFTfreqdel` is just `STFT` "frequency delayed," which is to say rotated by one frequency bin—shift the rows of `STFT` up by one step and move the former last row to the new first row.

3. Now it is time to compute Nelson's cross-spectral matrix:

   $\mathtt{C} = \mathtt{STFT} \times \mathtt{STFTdel}^*$, where the notation $A \times B$ denotes a pointwise product among the elements of the matrices, and the asterisk denotes the complex conjugate of the matrix, pointwise. The resulting C is now a matrix of complex numbers; each row's phase angles encode the CIF values in the "channel" or bin indexed by that row.

4. So now one can compute the channelized instantaneous frequency:

$$\mathtt{CIF} = \frac{\mathtt{Fs}}{2\pi} \times \arg(\mathtt{C}). \tag{6.7}$$

5. For the local group delay, one first computes Nelson's other cross-spectral matrix: $\mathtt{L} = \mathtt{STFT} \times \mathtt{STFTfreqdel}^*$. The result is also a matrix of complex numbers; each column's phase angles encode the LGD values over all frequencies at the time index of the column.
6. Now one can compute the local group delay:

$$\mathtt{LGD} = \frac{-\mathtt{fftn}}{2\pi\mathtt{Fs}} \times \arg(\mathtt{L}). \tag{6.8}$$

We now have what is needed to reassign the spectrogram. To perform the image plot as shown in Fig. 6.1, compute the log of the square of the magnitude of each value in the original matrix `STFT`, thereby obtaining the ordinary spectrogram matrix. Then each spectrogram amplitude value is repositioned on the time axis at its new corrected time by adding to its signal time the coindexed value in the `LGD` matrix, plus an additional time offset of $\frac{\mathtt{win\_size}}{2\mathtt{Fs}}$. The offset is required because the LGD computation corrects the time relative to the leading edge of the analysis

window, but it is conventional to reference the signal time to the center of the window. For the frequency repositioning of each point in the spectrogram, one uses the coindexed instantaneous frequency in the `CIF` matrix. Note that the reassigned spectrogram plotted in this fashion can have multiple points of different amplitudes plotted at the same location on the time–frequency axes; it isn't any longer a "matrix" that is being plotted. A useful Matlab plotting routine that handles this type of 3-dimensional data with arbitrary positions is the 3D scatterplot, which is used for the reassigned spectrograms in this book.

The earlier discussions of windowing functions prove to be very handy at this juncture, because perhaps surprisingly, the selection of an optimal window function is even more critical for the reassigned spectrogram than it is for the conventional one. This is because the regular spectrogram is quite smeared due to its inherent uncertainty, with the result that only large amounts of spectral leakage will be noticeable (such as if rectangular windows are used). The reassignment procedure, on the other hand, effectively moves all the spectrogram points surrounding a component's instantaneous frequency to that instantaneous frequency. If there are any spurious peaks nearby resulting from spectral leakage when

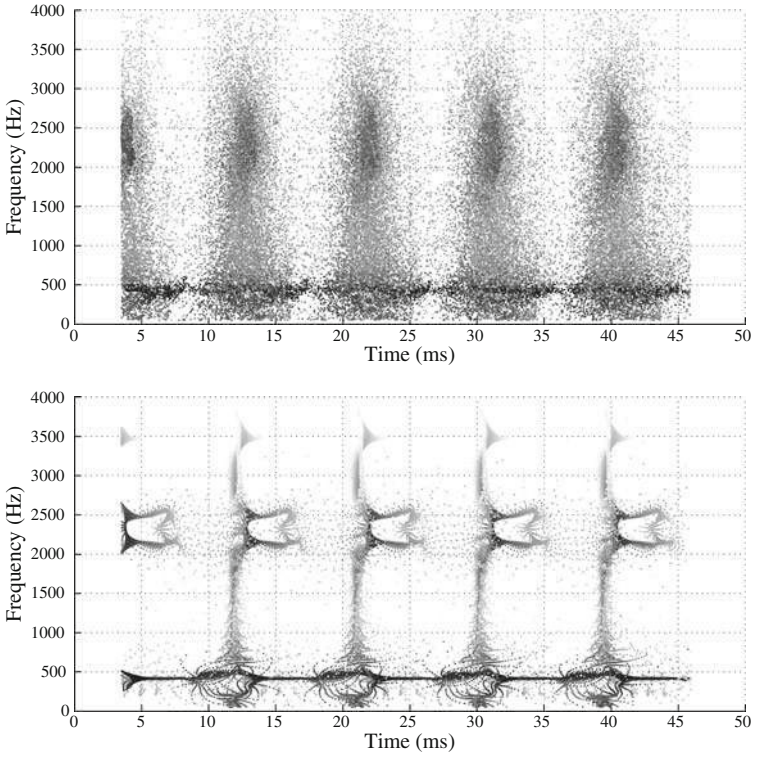


Fig. 6.2 Reassigned spectrograms of glottal pulses from the synthesized [e] vowel. Computed with 7 ms windows, rectangular in the *upper panel* and Hann function in the *lower*

suboptimal tapering is used, it will reassign points to those peak instantaneous frequencies as well. Moreover, leakage around a spectral peak can also act to move the location of the peak slightly. Figures 6.2 and 6.3 show several reassigned spectrograms of our synthesized vowel [e], and it is quite evident that the Kaiser

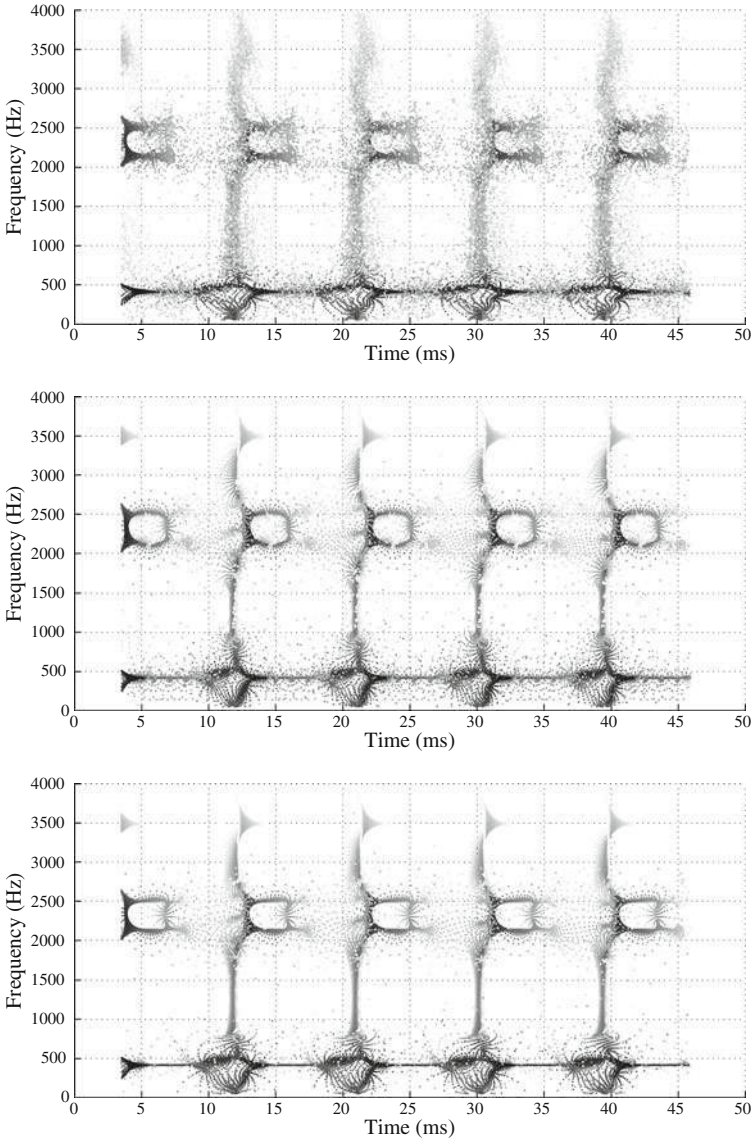


Fig. 6.3 Reassigned spectrograms of glottal pulses from [e] as in the previous figure. *Top panel*: 7 ms Gaussian windows with $\alpha = 2.5$; *middle panel*: Gaussian with $\alpha = 3.5$; *bottom panel*: Kaiser windows with $\beta = 3$

window is the best performing (as was originally concluded by Harris [14]). All further reassigned spectrograms shown in this book will be computed using Kaiser windows with $\beta = 3$.

Since the reassignment method will be unfamiliar to most speech researchers, it is important to be assured that the resulting spectrograms are indeed completely accurate. The reassigned spectrogram is in effect a "sharpening" of the conventional spectrogram which eliminates the time–frequency smearing that is induced by the spectrographic uncertainty principle. Any point in the original spectrogram having a significant amplitude will be reassigned closer to some time–frequency "ridge" that it was associated to. The reassigned spectrogram is thus completely faithful to the original and does not introduce anything spurious, as was clearly demonstrated by Gardner and Magnasco [12]. A drawback of this faithfulness, however, is that the reassignment also serves to sharpen spectrographic points and ridges that arose from interference terms in the first place, and these merely clutter the resulting image because they do not represent anything of interest in the signal. The next section shows how this problem can be greatly ameliorated.

6.2.2 *Reassigned Power Spectrum*

As a brief addendum, I will mention that it is possible to perform reassignment in the frequency domain by computing the channelized instantaneous frequency vector corresponding to a single-frame FFT of a signal, and in such a way obtain a single reassigned power spectrum showing instantaneous frequencies of signal components, instead of Fourier frequencies. Such representations can potentially replace the traditional Fourier power spectra of Chap. 4 with more precise renderings. Figure 6.4 shows CIF power spectra of both a short (one glottal cycle) window and a long window from an English vowel [ɔ]. The short window analysis

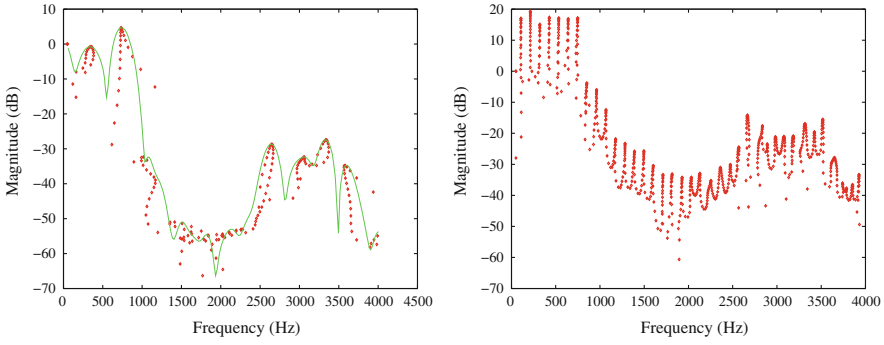


Fig. 6.4 CIF spectra (scatterplots) of natural English vowel [ɔ], using a 10 ms analysis window (*left*, with Fourier spectrum overlaid as a smooth line) and a 40 ms window

shows the formant peak locations with greater precision than the overlaid Fourier spectrum, and the long-window analysis shows the harmonic spectrum of the phonation with great precision. It is useful to keep in mind, however, that the time-varying nature of most speech attributes such as formants militates against relying upon any single-frame frequency analysis.

6.3 Pruning the Reassigned Spectrogram

The reassigned spectrograms shown in the Figs. 6.1–6.3 have been computed using the algorithm given above, without any further processing applied. A major problem just mentioned above that arises with "pure" reassigned spectrograms is that some of the plotted points are not affiliated to anything of interest in the signal, and neither were they so affiliated in the original spectrogram. The post-processing method I call *pruning* was first outlined by Nelson [22, 23], and involves the computation of the second-order mixed partial derivative of the STFT complex phase (the reassignment method itself involves computing just the first derivatives of this phase). The purpose of pruning is to selectively eliminate points from the plot of a reassigned spectrogram which are not associated to signal elements of particular interest. There are in general two different kinds of signal elements which can be retained and thus effectively emphasized by means of this pruning, namely quasisinusoidal line components possibly having some frequency or amplitude modulation (often called *AM/FM components*), and impulse-like events. One can choose to keep points affiliated to either or both of these kinds of elements using the procedure now to be outlined.

6.3.1 General Definitions

The second-order mixed derivative of the STFT phase is equivalent to either the frequency derivative of the CIF, or the time derivative of the LGD, since it is a fundamental theorem of calculus that the mixed partial derivative can be taken in either order. Nelson [22, 23] argued that the nearly stationary AM/FM components of a signal $x(T)$ should have a second-order mixed phase derivative near zero. By plotting just those points in a reassigned spectrogram meeting this condition on the phase derivative to within a threshold, a spectrogram showing chiefly the line components can be drawn. For further explication of this fact, the reader is referred to other papers [9]. The numerical threshold can be empirically determined, and will in practice depend on the degree of deviation from a pure sinusoid that is tolerable in the application at hand. This means that greater tolerance in this threshold will be required where line components having high AM/FM rates are expected—for speech signals an absolute value of the derivative on the order of 0.2 is often a reasonable threshold, but a smaller value can often be useful as well.

On the other hand, a numerical derivative threshold value which is several orders of magnitude smaller can be used to eliminate nearly every point that does not represent a pure sinusoid with no frequency modulation, as is illustrated in Fig. 6.5.

Nelson [23] further asserted that the impulses in a signal $x(T)$ should have a mixed phase derivative close to 1. By plotting just those points meeting this condition to within a threshold, a spectrogram showing chiefly the impulsive events in a signal can alternatively be drawn. For display purposes it is appropriate to be quite tolerant in this threshold, depending on what sort of signal content we desire to regard as "impulsive." A derivative value between 0.75 and 1.25 usually yields good results for speech signals, without straying too far from identifiably impulse-like events. Plotting all points meeting the disjunction of the above conditions results in a spectrogram showing quasisinusoidal components and impulses together, to the exclusion of most everything else.

Nelson's conditions on the second-order mixed partial STFT phase derivative take the following general forms in continuous time. The first expression holds of line components, while the second holds of impulses.

$$\frac{\partial^2}{\partial\omega\partial T}\arg(\mathrm{STFT}_h(\omega,T)) = \frac{\partial}{\partial\omega}\mathrm{CIF}_x(\omega,T) \approx 0. \tag{6.9}$$

$$\frac{\partial^2}{\partial T\partial\omega}\arg(\mathrm{STFT}_h(\omega,T)) = \frac{\partial}{\partial T}\mathrm{LGD}_x(\omega,T) \approx 1. \tag{6.10}$$

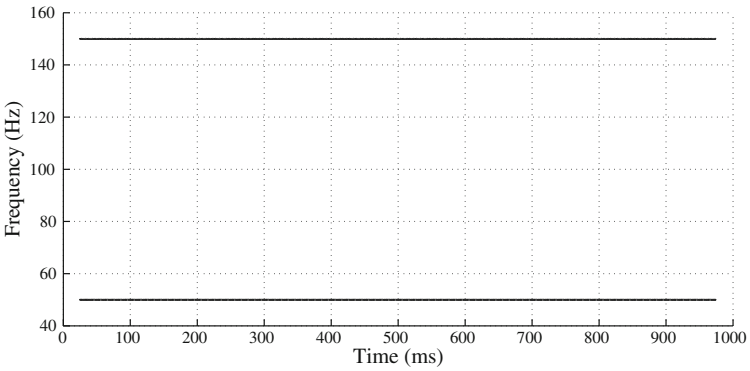


Fig. 6.5 Another view of the double sine wave shown in Fig. 6.1, this time showing only those points whose absolute value second-order mixed partial derivative of the STFT phase is less than 10^{-4}

6.3.2 Cross-Spectral Method

The mathematical theory behind cross-spectral expressions for all higher-order partial derivatives of the STFT phase is completely presented in prior literature [23], and we have relied on this in developing the particular algorithm for the second-order mixed partial derivative that is presented below. The steps in the computational method will be based on Nelson's algorithm for the reassigned spectrogram described above. Readers are invited to refer to that algorithm to complement that presented below. It is important to note that, just as with the cross-spectral method for computing the first-order STFT phase derivatives (and thereby the reassigned spectrogram) discussed above, the method presented here will compute an *approximation* of the second-order mixed partial STFT phase derivative. This approximation is generally so close that it might not matter for practical purposes. An alternative method which is more computationally complex has also been published [9], which computes an exact value of the mixed partial derivative.

1. Referring to the algorithm for the Nelson method that was outlined above, compute the three STFT matrices:

   `STFTdel = fft(Sdel)`
   `STFT = fft(S)`
   `STFTfreqdel` is `STFT` rotated by one frequency bin—this can be accomplished by shifting the rows in `STFT` up by one step and moving the former last row to the new first row.

2. Additionally compute one more STFT matrix:

   `STFTfrtimedel` is `STFTdel` similarly rotated by one frequency bin.

3. Next compute a cross-spectral surface by applying Nelson's theory:

$$\texttt{MixCIF} = \texttt{STFT} \times \texttt{STFTdel}^{*} \times (\texttt{STFTfreqdel} \times \texttt{STFTfrtimedel}^{*})^{*} \tag{6.11}$$

4. Now the partial frequency derivative of the channelized instantaneous frequency can be computed:

$$\texttt{CIFderiv} = \frac{\texttt{fftn} \cdot \texttt{Fs}}{2\pi \cdot \texttt{win\_size}} \times \arg(\texttt{MixCIF}) \times \arg(\texttt{MixCIF}) \tag{6.12}$$

   where the $\arg(\cdot)$ function is valued in the range $(0, 2\pi)$, and `Fs` is the sampling rate (in Hz) of the signal.

The final quantity computed by the above algorithm is equivalent to the partial time derivative of the local group delay, and either of these represents the (unique) second-order mixed partial derivative of the STFT phase. It is then simple enough, depending upon the plotting routine, to plot only those points in a reassigned spectrogram whose coindexed values in the `CIFderiv` matrix are within whatever threshold of 0 (for highlighting components), or 1 (for highlighting impulses).

6.3.3 Justifying the Interpretation of the Phase Derivative

The following brief discussion is taken from [9]. In regions where the CIF is not changing with frequency, the spectrum is dominated by a single component that is highly concentrated in frequency (i.e. a sinusoid). In these regions, all nearby spectral data is mapped to the frequency of the dominant sinusoid, so that the variation (partial derivative) with respect to frequency is near zero. Similarly, in regions in which all spectral data is mapped to the time of a dominant component that is highly concentrated in time (i.e. an impulse), the variation (partial derivative) of the reassigned time with respect to time is near zero. Since the reassigned time is computed by adding the LGD to the nominal time, t,

$$0 \simeq \frac{\partial}{\partial t}[t + \mathrm{LGD}(t, \omega)] = 1 + \frac{\partial}{\partial t}\mathrm{LGD}(t, \omega) \tag{6.13}$$

so

$$1 \simeq -\frac{\partial}{\partial t}\mathrm{LGD}(t, \omega) \tag{6.14}$$

That is, as the nominal time, t, *increases*, the time correction (LGD) for data in the neighborhood of a dominant impulse *decreases* proportionally.

6.3.4 Separation of Formants from Glottal Impulses

The features of a voiced speech sound which are of primary interest in a spectrogram are the line components and the glottal impulses. It is therefore quite useful to apply the pruning procedure in order to highlight these signal elements while excluding other spectrographic points that are likely insignificant or which result from interferences. In a wideband (short window) spectrogram, the significant line components will normally be formants and other resonances. By pruning a reassigned spectrogram of a portion of voiced speech, it is possible to retain only points affiliated to resonances, or only points affiliated to impulses. It is also possible to retain both kinds of points, which results in a spectrogram that shows all the significant signal elements while eliminating unimportant clutter. Figures 6.6 and 6.7 show several examples of pruned reassigned spectograms of our synthesized vowels [æ] and [ɔ]. Pruning parameters for spectrograms are reported as the threshold value for each kind of signal element that is retained. That is, a spectrogram which highlights components with a "threshold of 0.1" retains only points whose mixed STFT phase derivative lies between -0.1 and 0.1. Similarly, a spectrogram which

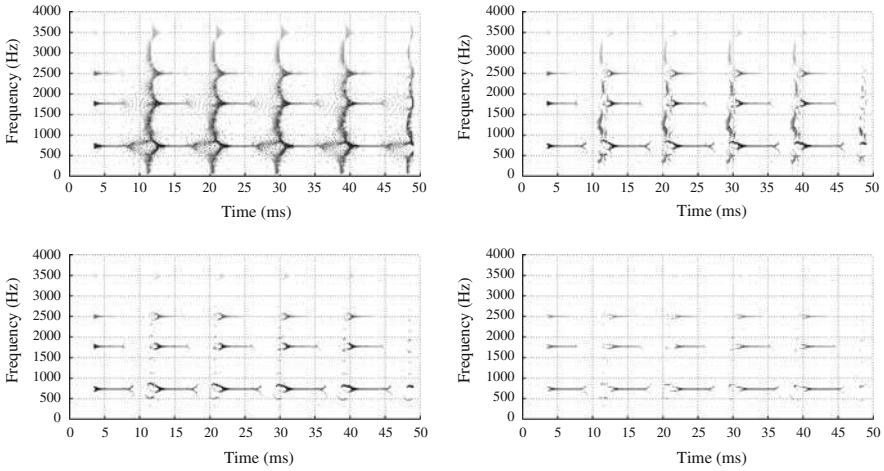


Fig. 6.6 Four reassigned spectrograms showing several glottal cycles of synthesized [æ], all computed with 7 ms Kaiser windows. *Top line* shows an unprocessed reassigned spectrogram, and one pruned using thresholds of 0.1 for components and 0.25 for impulses. *Bottom line* shows two spectrograms pruned to show components only; the *left image* uses a component threshold of 0.1 (so separating out the components from the right hand plot on the upper line); the *right image* uses a tighter threshold of 0.01

highlights impulses with a "threshold of 0.2" retains only points whose mixed STFT phase derivative lies between 0.8 and 1.2.

The improvement in image quality resulting from the pruning procedure is evident in the real speech examples shown in Figs. 6.8 and 6.9. Even a tight partial derivative threshold of 0.01 can be quite useful for highlighting formants, as shown in the lower panel of Fig. 6.9.

6.4 Analyzing Phonation

6.4.1 Beyond Source-Filter Theory

The source-filter theory of speech production models phonation as a purely acoustic process, the excitation of a linear filter by an impulse train. The vocal cord source is modeled as a pure volume velocity source that provides an acoustic excitation when the vocal cords come together and produce an abrupt change in the pressure and volume velocity. Key to the source-filter approximation is that possible acoustic effects from the flowing air itself, known as *aeroacoustic* effects, are neglected. Such a pure acoustic source is known as a *monopole* source in current aeroacoustic theory [19]; this is what Lord Rayleigh termed a "simple

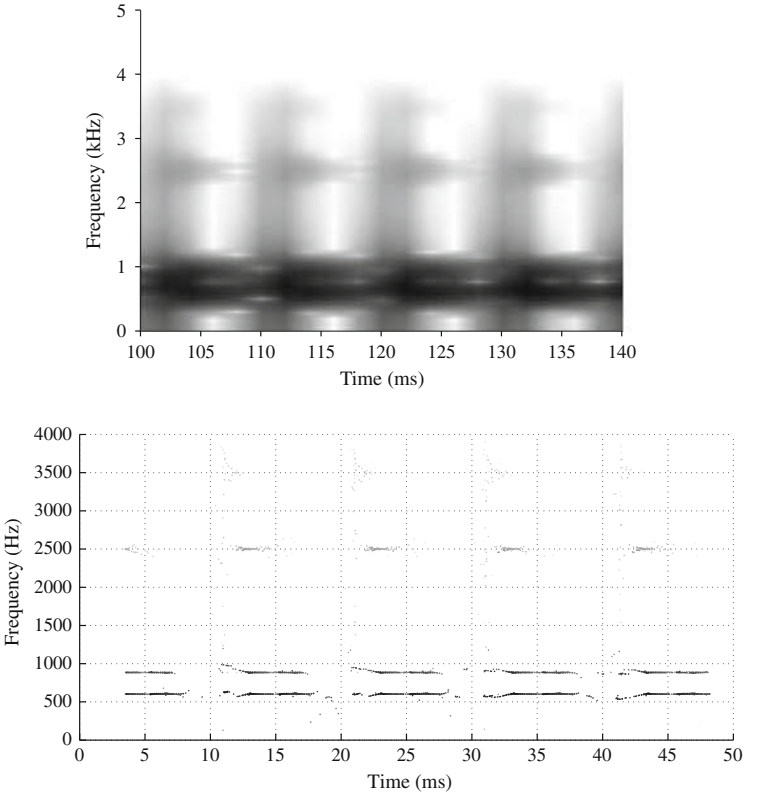


Fig. 6.7 The figure compares a previously shown spectrogram of several glottal cycles from the synthesized [ɔ] with a reassigned spectrogram pruned to show components only using a tight threshold of 0.01

source" in his seminal discussion [30]. The filter resonances of the mouth are then predicted to "ring" after excitation by each pulse, yielding the formants of a speech sound.

In reality, the natural phonation process does not conform to these conditions, and can only be modeled this way to a first approximation [4]. The degree to which ordinary modal phonation deviates from the ideal monopole source has been the subject of some recent findings and debate, but it is by now clear that sources of sound in phonation cannot be completely characterized by volume velocity at the glottis [19]. McGowan's theoretical efforts predicted the presence of an aeroacoustic *dipole* source due to an oscillating rotational flow in the glottis, along with the monopole volume velocity source, and the dipole source was identified as the main source of random noise in breathy voicing. McGowan did not suggest that the dipole source would be dominant in modal phonation, but predicted that it

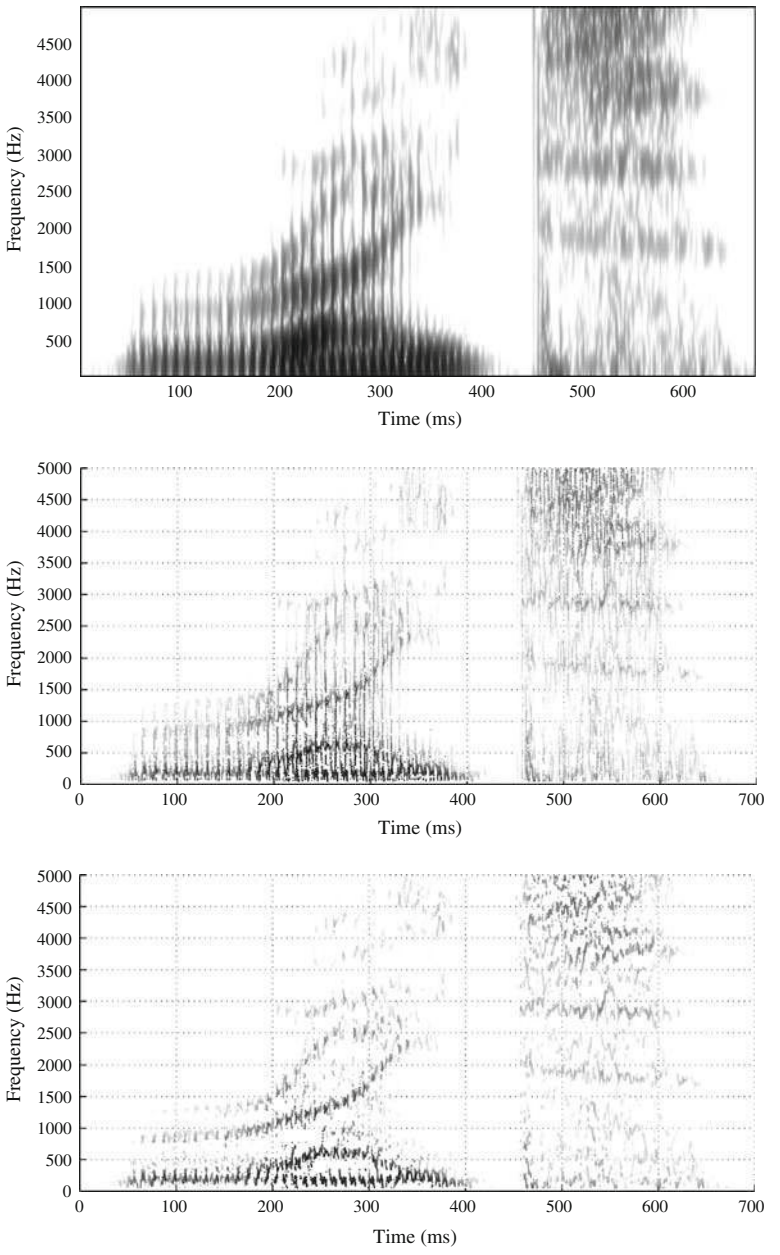


Fig. 6.8 Conventional and reassigned spectrograms of the English word *right*. The *middle reassigned image* is pruned to show components and impulses, while the *lower image* is pruned to components only

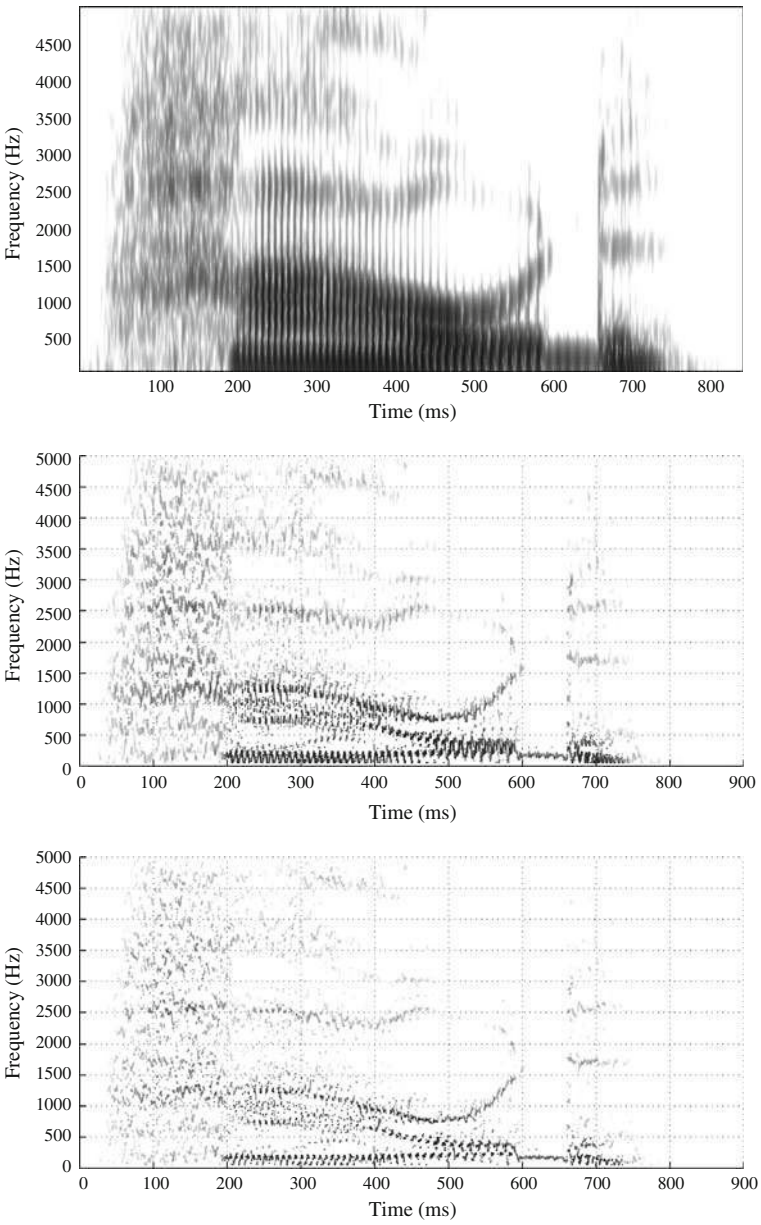


Fig. 6.9 Conventional and reassigned spectrograms of the English word *how'd*. The *middle reassigned image* is pruned to show components only with a loose threshold for the partial derivative of <0.1, while the *lower image* is pruned to show components meeting a tight threshold of <0.01

would become more important at higher frequencies and with greater airflow through the glottis.

Stevens [29] was careful to note that the acoustic source at the glottis can be approximated by a monopole only on the condition that "the mode of vocal fold vibration is such that the glottis is closed or remains sufficiently narrow over a cycle of vibration." The validity of this approximation for modal phonation was tested by Zhao et al. [31], and it was not found to accord with the detailed picture of phonation. These authors performed a rigorous computational aeroacoustic study of the acoustic effects of airflow through oscillating (virtual) vocal cords, finding as expected that the glottal source was indeed an aeroacoustic monopole *when there was zero airflow*. However, when a more realistic mean flow was provided through the glottis model, it was found that glottal motion in the flow induced unsteady vortex shedding, and the unsteady force thereby exerted on the glottis produced a dipole source which dominated all other sources by orders of magnitude. The monopole source was not only not a valid approximation at higher airflows, it was actually negligible.

In spite of this, Zhao et al. [31] did find that it is still possible to approximate the glottal dipole source with a pure volume velocity source in order to model the gross features of phonation. However, the so-called "fine structure" of the glottal wave—the features other than the gross periodicity—were not successfully modeled by invoking this commonplace approximation. In particular, the additional sound sources downstream of the glottis are missing. These include the unsteady forces on the downstream side of the vocal folds (a dipole source) and the unsteady flow downstream of the glottis (a quadrupole source) [31]. Quatieri [25] has pointed to evidence of such "secondary sources" in real speech signals, which can appear in a glottal waveform as a renewed excitation approximately 2/3 of the way through the glottal cycle. This feature is consistent with the notion that it is a chiefly aeroacoustic source due to vortex impaction downstream from the glottis [25]. Moreover, the secondary sources, being (in Quatieri's terms) "nonacoustic," have different travel times in general from the glottal acoustic source and excite different regions of the vocal tract. The sum of this finding is that different sources can excite different formants, an explanation put forth by Shadle et al. [28] to explain inconsistencies in the glottal waveform estimated using inverse filtering based on linear source-filter theory.

Apart from the chiefly aeroacoustic nature of the glottal source, it has also been recognized that there is significant source–tract interaction over the course of a glottal cycle, particularly near the first formant frequency where the input impedance of the vocal tract is nearest to the resistance of the glottal source [4]. Flanagan pointed to experimental observations of pitch-synchronous variations in the tuning and damping of F_1, which he attributed to significant source–tract interaction. Quatieri [25] explained that the first formant frequency is expected to rise at the onset of the glottal open phase, and fall near the termination of this phase, owing to the changing glottal impedance. Source-filter theory, in contrast, generally assumes that the glottal impedance can be approximated by a fixed value equal to that of a closed glottis.

The sum of all this is that only vowels whose production closely conforms to the source-filter model can be expected to display well-defined vertical impulses and well-defined horizontal formants in a reassigned spectrogram. Our synthesized vowels have been created strictly according to the source-filter model, and so can serve as a kind of benchmark in this respect. Vowels of real speech can be expected to yield reassigned spectrograms which deviate somewhat from the appearance of those provided for the synthesized vowels. The degree of deviation in the image from that of a synthesized vowel can be regarded in a loose sense as a qualitative metric of the degree of deviation of the real speech production from the source-filter theory. It will next be observed that the degree of deviation from the source-filter model depends chiefly upon the phonation type employed during speech production, and also upon various idiosyncratic characteristics of the phonation process in a particular speaker.

6.4.2 Observations on Phonation Types

A steady vowel [e] was produced by the author in four phonation types: creaky, stiff (vocal cords pressed together firmly), modal, and breathy. To discuss the fine details of the different phonation types, we refer to the reassigned spectrograms of Fig. 6.10. The figure shows the four segments of [e] in each of the four phonation types, post-processed by the pruning technique to eliminate all points that are not affiliated to either a line component or an impulsive event. The different character of the phonation in each panel is easily noted, particularly in the breathy voiced example, which appears to have very indistinct glottal impulses and very unsteady formants.

Creaky phonation is of great interest as a baseline case because it has extremely small airflow volume, and the closed phase takes up a maximally large proportion of the phonatory cycle. These properties render the creaky phonation process as purely acoustic as possible in human speech—i.e. aeroacoustic effects are expected to be negligible. As would be expected then, the creaky [e] spectrogram of Fig. 6.10 displays glottal impulses which are very straight and sharply defined, and formants which are cleanly excited and which do not change in frequency—except for some aberrations which occur during the very brief open phase immediately preceding the impulse (which arises from the closing event).

During the open phase, F_1 is observed to suddenly change, splitting into a discernible component whose frequency increases and a second, more robust component whose frequency decreases. The increase of formant frequencies during the glottal open phase, as discussed above, has been widely predicted [4], but the appearance of a lower frequency component during this phase (labeled “voice bar” in the figure) is unexpected. Nevertheless, this low-frequency component will appear as a band traditionally called a voice bar on a spectrogram showing a much longer segment of a vowel. The nature and origin of the voice bar remains mysterious at this point in time, since it has never been addressed clearly in the

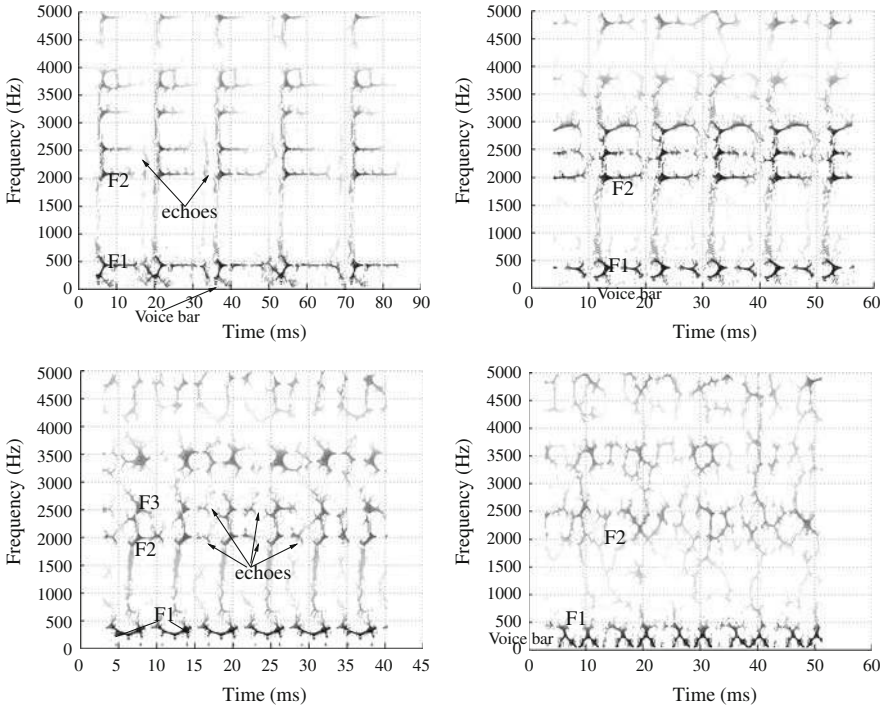


Fig. 6.10 Vowel [e] spoken with four phonation types, imaged using reassigned spectrograms which highlight both the line components and the impulsive events. *Top row* shows creaky and stiff voice (computed using a 7.8 ms window), *bottom row* shows modal and breathy voice (computed using a 5.9 ms window)

literature. It was identified by Fant [3] (p. 109) as an "extra voice bar formant," but this remark was not explained, and its physiological origin is not obvious.

Some suggestive findings in this connection have been published by Castelli et al. [1], which were further validated by Dang and Honda [2]. These groups of authors both found that a low resonance at around 250 Hz can appear in a vowel spectrum as a result of coupling with the nasal sinus, even through the closed velum in a nonnasalized vowel. This phenomenon is referred to as "transvelar coupling," and it might be a plausible explanation of the voice bar's origin. As discussed by Dang and Honda, "the 250 Hz peak is clearly seen for open vowels since their first vowel formant is higher" [2]. The effect of the voice bar on the first formant of high vowels can be to obscure its true value; these effects will be further explored below.

The voice bar is commonly observable in the speech of men as well as women (in whom it has a higher frequency), but it is not normally used in synthesized vowels, where its absence probably relates to the stereotypical "reedy" sound of a computer voice. Nevertheless, in Chap. 5, it was noted that a high-resolution analysis of synthesized phonation using the ZAM distribution can often display an

apparent splitting of F_1 in the vicinity of the glottal closure, and this feature of the analysis can appear remarkably similar to a voice bar resonance. Indeed, it is often difficult or impossible to ascertain from a single time–frequency representation (e.g. a reassigned spectrogram) of a vowel whether a voice bar is present, or whether a low F_1 is splitting due to a phase shift at the moment of glottal closure. In cases where this is not clear, corroborating evidence of a voice bar's presence or absence can be provided from a linear prediction analysis of the subject vowel, as will be elaborated in Chap. 7.

In spite of the very long closed phase and very brief open phase manifested by the creaky [e] segment, it is nevertheless easy to see evidence of the secondary excitations at higher formants during the open phase (labeled "echoes" in the panel) that were discussed by Quatieri [25], which he attributed to aeroacoustic sources. At any rate, these echoes are not predicted by the source-filter model.

Under more natural airflow conditions, the phonation process is expected to become more significantly aeroacoustic (following the computational simulations of [31]), meaning that the higher airflow cannot be neglected and may have clearly observable effects on the excitation of resonances. Several effects can be observed in the stiff-voiced and modal-voiced [e] panels of Fig. 6.10, and these include random disturbances and back-and-forth undulation of the impulses, particularly in the upper frequency range. The increasing randomness of the impulse events due to the increasing importance of the aeroacoustic dipole source as the frequency increases was a prediction of McGowan [19].

The stiff-voiced [e] panel of the figure displays a longer open phase, during which the low-frequency voice bar formant is excited for a longer time period than in the creaky voiced case. It appears that this voice bar is "out of phase" with the first formant; the former dominates the open phase, while the latter dominates the closed phase. In the modal-voiced panel, the spectral randomness induced by the strong airflow is now beginning to dominate the image, although the principal formants are still relatively easy to track. The voice bar is now no longer apparent in this example, which has a much higher F_0, but the effect of a voice bar (i.e. a resonance sounding lower than F_1) can be seen to arise from the rising and falling of F_1 during each glottal cycle. Meanwhile, the formants above F_3 in this example are greatly disturbed by random effects, and the F_2–F_3 region shows clearly observable secondary excitations approximately 2/3 of the way through each cycle (as was discussed in [25]). These echo impulses are also observable as a second, roughly vertical, event in between the primary impulses near the top of the frequency range.

The panel of Fig. 6.10 showing a breathy-voiced [e] segment displays a great deal of random fluctuations in the formants, which are now impossible to track precisely. The voice bar itself now seems more perturbed, and dominates the low range for an interval of each cycle about equal to that in which F_1 is dominant. The glottal "impulses" are now completely perturbed by random fluctuation, as well. It is also fairly easy to see the "classic" acoustic correlates of the different phonation types emergent in the panels of Fig. 6.10, namely relative spectral tilt. A number of studies have shown both theoretically (e.g. [29]) and through direct

measurement (e.g. [18]) that more pressed or creaky phonation yields less relative amplitude of the lowest spectral components (lessening the spectral tilt), while more slack or breathy phonation yields greater relative amplitude of the lowest components (increasing the spectral tilt). Stevens' calculations of spectra from various glottal airflow waveforms correlate the observed higher amplitude in the upper components in creaky and stiff voicing with the decreased duration of the airflow pulse. The use of spectral tilt in some form as a metric of phonation type was also discussed in Chap. 4.

Further examples of these detailed features of phonation are shown in Fig. 6.11, which displays a few glottal cycles drawn from naturally produced vowels [ɛ] in *head* and [æ] in *had*. Once again one can observe the secondary excitations of higher formants in the middle of each cycle, as well as the trading between F_1 and voice bar as the glottis shifts from closed to open phase. This figure also

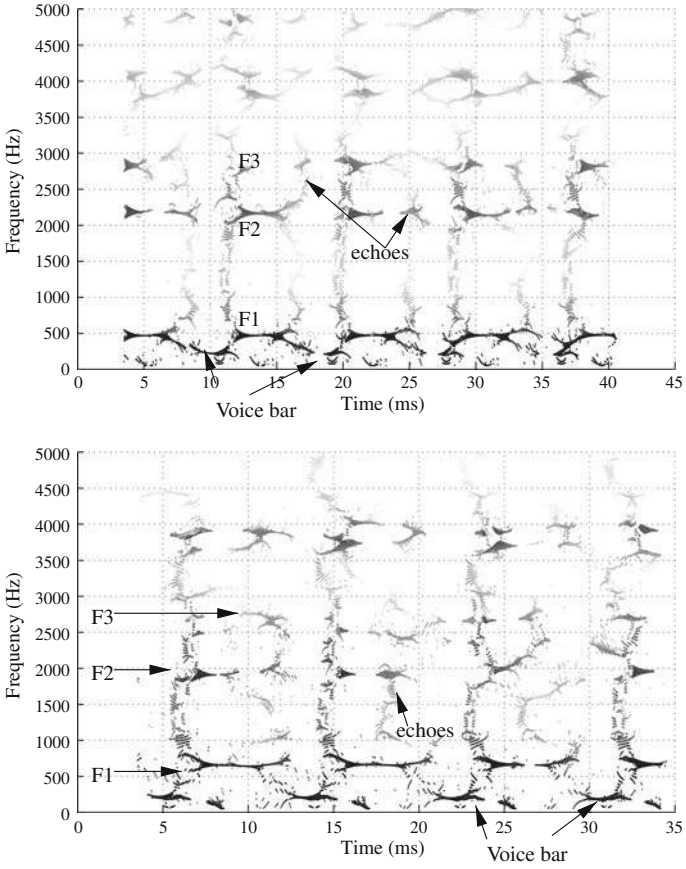


Fig. 6.11 The vowels [ɛ] (*upper*) and [æ] are shown, with the partial derivative thresholding set to highlight both components and impulses. Annotations in the images indicate the formants, voice bar, and the appearance of secondary excitations (echoes)

demonstrates that the voice bar is extremely low (below 250 Hz) even for a low vowel such as [æ] with a higher F_1. This demonstrates that whatever its source, the voice bar does not appear to bear any relationship to the value of F_1.

6.4.3 Phonation as a Biometric

It has been noted (e.g. [24]) that glottal vibration videos show large variations in the movement patterns of the vocal folds from one person to another. These idiosyncratic features of phonation have different effects on the aeroacoustics of the glottal output, which results in each person's phonation presenting a notably distinct reassigned spectrogram in a close-up view. These images, however, are largely consistent for one person's voice in any given phonation type and vowel, leading to the prospect that they are individuating and can be matched within the same speaker [7].

Research into the use of reassigned spectrograms of phonation as biometrics is currently just beginning, so only a few remarks will be made here as this is by no means a proven technique. At the very least, however, reassigned spectrograms of the phonation of different speakers are obviously very different, as shown in Fig. 6.12. The differences are manifold, and include the shape of the undulating vertical impulse representing the glottal closure pressure pulse, the excitation of different formants at different times relative to this impulse, the presence or absence of a voice bar and its phase relationship to F_1, patterns of formant movement and "splitting" within each glottal cycle, and the location of echo impulses between the primary glottal pressure pulses.

Figure 6.13 shows the potential for such spectrograms to be "matched" for a particular speaker's vowels. The two male speakers shown in the figure are saying the same word two times each, but the similarity between the speakers is evidently much less than the self-similarity of each speaker's repetitions. This degree of similarity in the fine structure of a given speaker's phonation has been found (anecdotally) to remain stable across time spans of up to two years between repetitions [7].

6.5 Dynamics of Formants

One study [6] has recently established that the known formants of synthesized vowels can be measured manually from reassigned spectrograms with excellent accuracy. However, the study failed to notice or compensate for the expected systematic differences between known production formant values for the vowels, and the values of formant spectral peaks which were directly compared to them (this issue will be discussed in detail in Chap. 7). In spite of this flawed procedure, formant measurements performed on a set of synthesized vowels (the same tokens

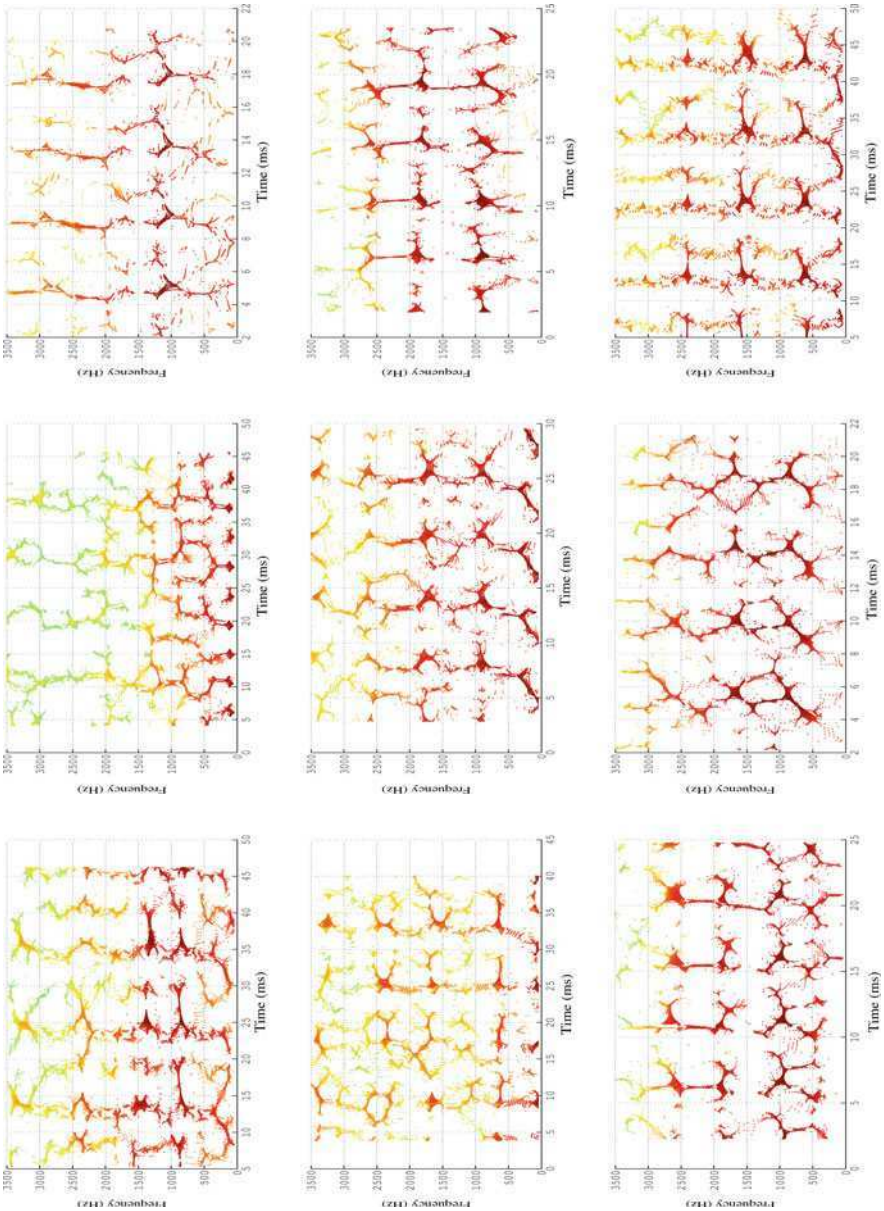


Fig. 6.12 Phonation spectrograms (vowel [æ]) for nine speakers

used in this book) were usually within 10 Hz of the production value, and the greatest difference noted was 1.6%.

The question which presents itself now concerns the actual method of measuring a formant "manually" from a reassigned spectrogram. Figure 6.14 shows

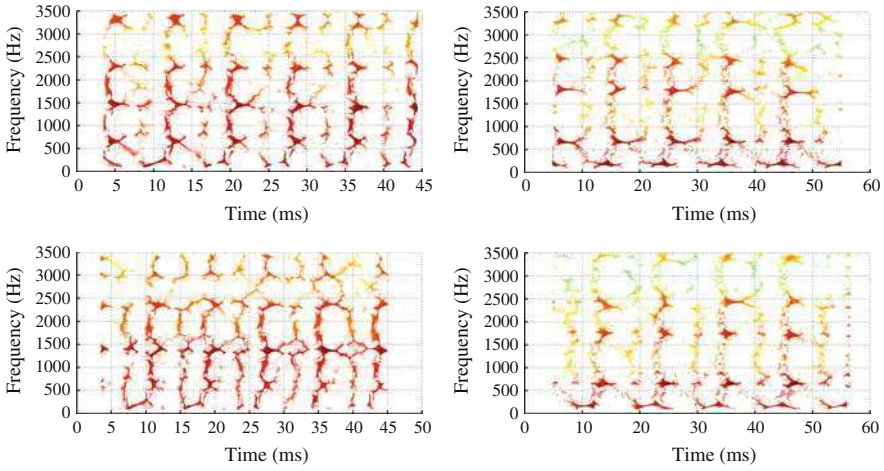


Fig. 6.13 Matching phonation spectrograms of vowel [æ], one male speaker each column

that for a synthesized vowel, such measurement is very easy. I used a version of the reassigned spectrogram routine (included with the linked Matlab code) which is augmented with a tool that automatically reports the precise frequency at the mouse pointer on the spectrogram image. I made a reassigned spectrogram of a few glottal cycles in the vowel, using a very large window overlap (corresponding to a frame advance of four samples) and a reasonably large FFT frame (the length of the FFT performed was 2,048 in this and most other examples). These parameter settings allow the image to retain good resolution when zoomed in (v. lower panel of the figure), although they do not improve the fundamental accuracy of the reassigned spectrogram. By zooming in on each formant and positioning the mouse pointer carefully on the obvious line component representing the formant after the initial excitation has tapered, a measurement accurate to within 1 Hz is easily obtained.

When a real vowel is examined, the simplicity of the synthesized phonation and formants shown above is replaced by a much more complicated picture, v. Fig. 6.15. The reassigned spectrograms in the figure were made in much the same fashion as for the synthesized vowels, although for each set of glottal cycles examined, it is important to customize the length of the analysis window so that it is between three quarters of a cycle and one cycle. A longer window will always provide the best possible resolution of close formants; it is possible to use a window slightly longer than one cycle to maximize resolution, because the tapering function will shorten the effective window length.

Figure 6.15 shows three particular features which are typical of real phonation that are not evident in synthesized vowels, all of which militate against accurate formant measurement. Firstly, there is the voice bar, rumbling along below F_1 and often quite a bit louder as well. It is often difficult to clearly separate these two low resonances; one aspect that helps here is their phase difference, since the voice bar

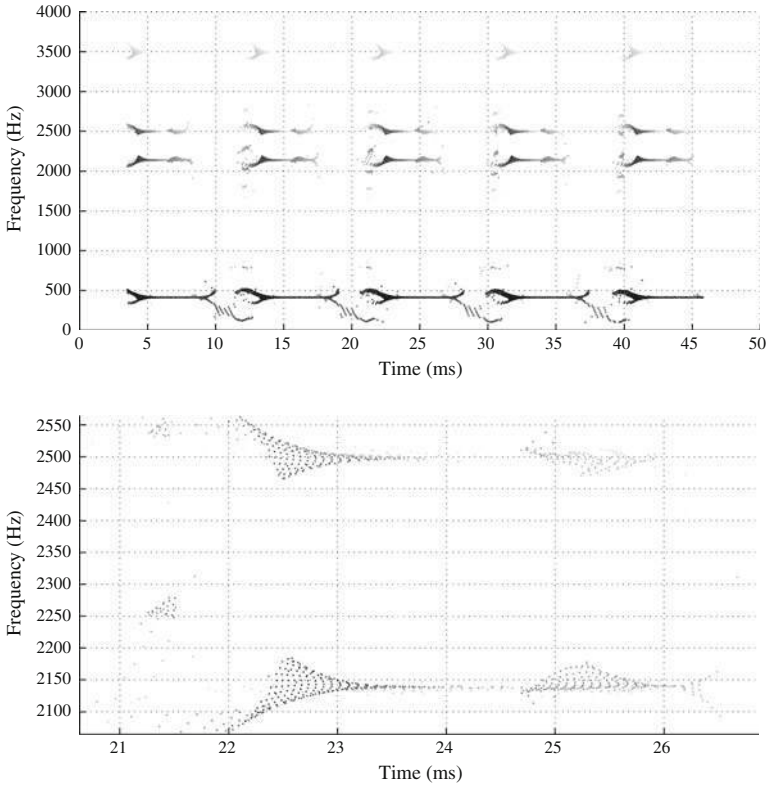


Fig. 6.14 Illustration of how to measure formants in a synthesized vowel [e]. *Top panel* shows a reassigned spectrogram of several glottal cycles, computed using 7 ms windows and pruned to show components with a threshold of 0.1. *Bottom panel* simply zooms in the upper figure to magnify F_2 and F_3 in a single cycle. Each formant would be measured by positioning the mouse pointer along the apparent line component after the initial formant excitation has tapered

seems more prominent during the open phase (although it may also be observed during the closed phase), while F_1 is clearly excited by the pressure pulse that initiates the cycle's closed phase. Secondly, each formant may be observed to change in frequency during the course of one glottal cycle. This leads to the obvious question whether it even makes sense to report a single value for such a dynamic component. Thirdly, different formants may apparently be excited at different times during the cycle (as predicted, [28]), making it impossible to apply a simple standard such as reporting the value at a certain time relative to the closing pulse.

The only thing clear from all this is that the clearer view of phonation and formants provided by the reassigned spectrogram is opening a "can of worms." Although these images show the fine structure of phonation more clearly than ever, they raise hitherto unasked questions about just how a given formant ought to be measured. It seems to me that no one is currently in a position to offer definitive

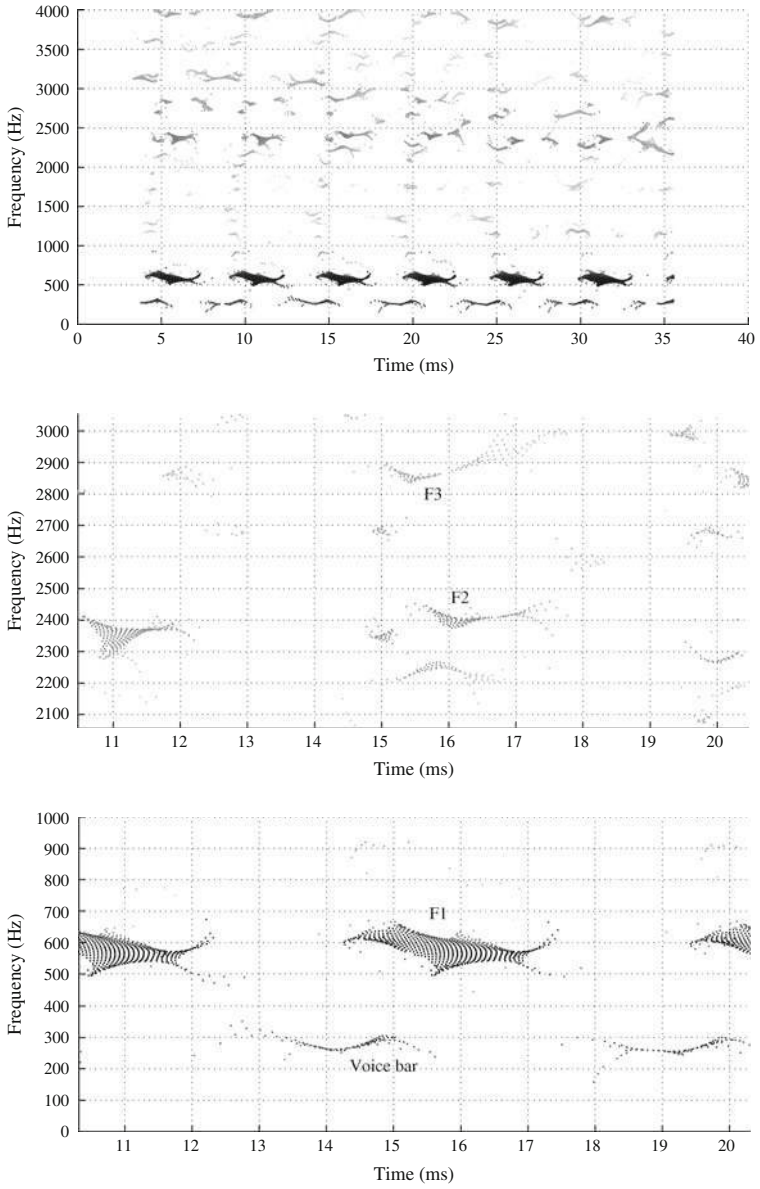


Fig. 6.15 Three reassigned spectrograms of the first part of natural English [eɪ]. The *middle panel* zooms in for measuring the F_2–F_3 region; *lower panel* shows the F_1 region including the voice bar. Computed using 9.4 ms windows, and a pruning threshold to highlight components only

Table 6.1 Formant values for natural English vowels spoken by the author, manually measured from reassigned spectrograms

Vowel	Voice bar	F_1	F_2	F_3
[i]	195	380	2,363	3,220
[ɪ]	227	479	2,290	2,771
[eɪ]	197	412	2,527	2,988
[ɛ]	221	471	2,164	2,782
[æ]	205	667	1,917	2,516
[ɔ]	175	505	712	2,655
[oʊ]	184	364	705	2,428
[ʊ]	291	350	879	2,387
[u]	130	350	715	3,034

Table 6.2 Formant values for some Finnish long vowels, female speaker, manually measured from reassigned spectrograms

Vowel	Voice bar	F_1	F_2	F_3
[aː]	280	725	1,544	2,683
[æː]	247	771	1,808	2,704
[eː]	271	621	2,403	2,860
[iː]	271	473	2,452	2,952
[yː]	326	449	1,753	2,489
[oː]	262	621	1,403	2,703

answers to these questions; no doubt, considerable further study and experience with reassigned spectrograms of speech will be essential to progress in this area. In spite of the present state of ignorance, I have gone ahead with reporting a number of formant measurements obtained using the above techniques. Some formant values obtained for my own English vowels are provided in Table 6.1, and only one value is reported for each, following tradition (perhaps unwisely). Generally, I report the value of a formant where it is most clearly separated from other proximal resonances. At other times the same formant may change in frequency, and this dynamic aspect is not reported in the tables. Most formants have a "fatter" segment in the spectrograms when they are strongly excited, which tapers to a point in one direction or the other; in these cases, the frequency of the tapered point is reported, since this strategy proves to be the most accurate method for synthesized vowels.

Table 6.2 reports some formant values for Finnish long vowels, measured in the same way as above with a female subject. Figure 6.16 shows some exemplary reassigned spectrograms from this subject's vowel [eː]. One interesting finding that emerges from this data is the female's voice bar being consistently higher than that of the male English speaker.

The most important general finding about formants that has emerged from working with reassigned spectrograms is that they are more dynamic and mysterious than is often recognized. Given this, could such mysterious properties of formants be used as the primary phonetic features behind a linguistic sound contrast? Preliminary investigation of the rare and peculiar vowel feature known as

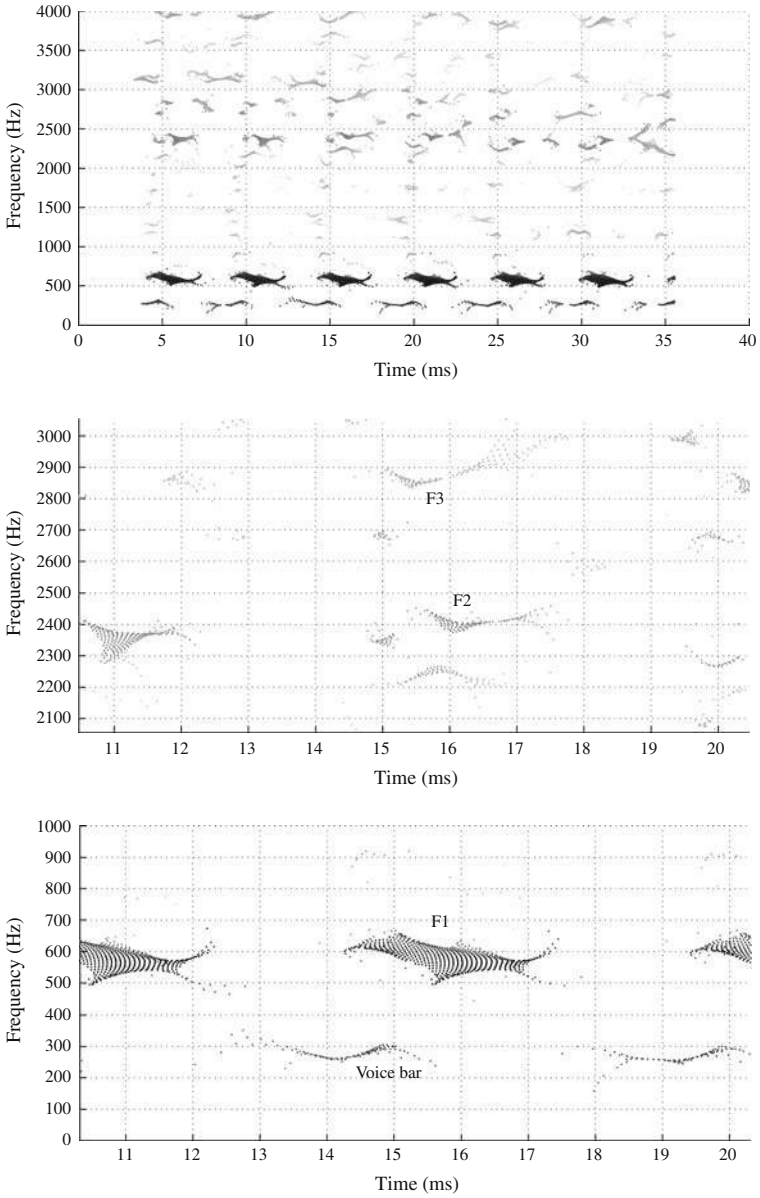


Fig. 6.16 Three reassigned spectrograms of Finnish vowel [e:]. The *middle panel* zooms in for measuring the F_2–F_3 region, *lower panel* shows the F_1 region including the voice bar. Computed using 6.3 ms windows, and a pruning threshold to highlight components only

Advanced Tongue Root has provided some data suggesting that, indeed, some vowel contrasts could be mediated by unexpected aspects of the dynamic formants.

Some languages, almost all within Africa, possess a system of vowels which includes several pairs of similar vowels that are distinguished by positioning of the tongue *root* (i.e. the base of the tongue muscle primarily forming the anterior pharynx wall). Generally in these cases, one vowel of a pair is produced with a more advanced (forward) tongue root posture (and thus an expanded pharynx cavity), while the other is produced with a more retracted tongue root (and thus a more constricted pharynx). These articulatory facts have been verified occasionally using imaging technology. Acoustic studies (e.g. [10, 13]) have uniformly demonstrated that most tongue root vowel pairs are discriminable chiefly by the value of F_1; in particular, the more advanced vowel usually has a lower F_1. This acoustic manifestation of the contrast was, however, not found in the case of the Degema low vowels [a̟, a̠] (the first of these diacritics under the vowels shows advanced tongue root, the second retracted) [10]. In fact, the cited study of Degema tongue root contrasts could find no consistent acoustic difference between these two allegedly distinct vowels, and one of the study's conclusions was that the contrast between the low vowels could currently be neutralized.

Revisiting some of the recordings of Degema low vowels which were examined in the earlier study, I noticed some indications that the "alleged" contrast between the two low vowels could perhaps be mediated by more complicated differences in the resonance dynamics. A representative example showing the lower formants of the two low vowels is shown in Fig. 6.17. It should be noted that the dynamics of F_1, together with a mysterious additional resonance between F_1 and F_2 (possibly nasal in origin), are completely different between the two vowels. Measurements from these data show that for the advanced vowel, the apparent F_1 begins at 983 Hz and descends to 769 Hz, and reappears during the open phase at 677 Hz. Showing a different pattern, the retracted vowel initially excites resonances at 932 Hz and 541 Hz, which later "merge" to an apparent single resonance at 782 Hz. This example is provided only as an indication of the complexity of facts about vowels which can be revealed using reassigned spectrograms. There is not yet a complete analysis of the situation in all our Degema recordings.

6.6 Nasals and Nasalization

Nasalization has long posed a problem when measuring formants, owing to a number of factors which include the introduction of nasal zeros which further damp (and thus smear) vowel formants, as well as nasal formants which are often proximal to the oral formants. The reassigned spectrogram is relatively immune to such effects, since it shows only the instantaneous frequencies of components and does not show their overall energy distribution (i.e., bandwidth). It must be recognized, however, that a reassigned spectrogram can only show components, so it cannot show the location of zeros. Of course, the conventional spectrogram can only show a zero implicitly, by the absence of energy, and this can be an ambiguous feature.

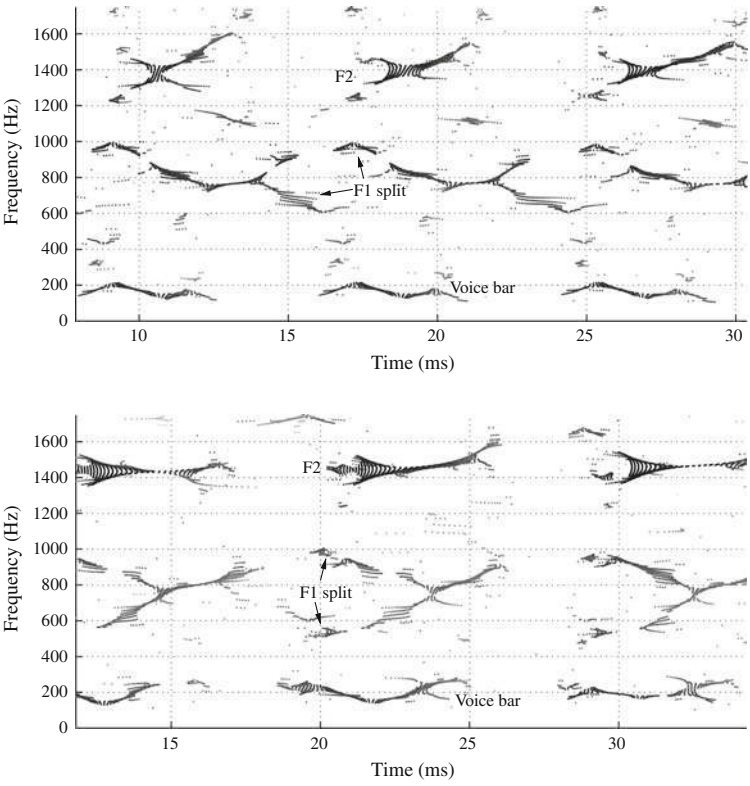


Fig. 6.17 Reassigned spectrograms of two Degema vowels, zoomed in to the F_1–F_2 region. *Upper panel* shows advanced tongue root [a̟]; lower panel shows retracted tongue root [a̠]. Computed using 8 ms windows, and a pruning threshold to highlight components only

Figure 6.18 shows a reassigned spectrogram highlighting just the line components of an oral vowel [æ] found in *had* spliced beside a nasalized [æ̃] found in *clan*. These are expected to have very similar oral articulation. The formants in the nasalized vowel are not any more difficult to observe in the reassigned spectrogram, and the imaging of numerous other distracting resonances helps to provide a clear picture of what happens as a result of the nasal coupling. Theoretical considerations of the two coupled resonance cavities in a nasalized low front vowel [29] predict that the original oral formants will shift. In particular, F_1 is calculated to shift downward considerably, while F_2 is calculated to shift downward slightly, and the formant values observed and labeled in the figure accord reasonably well with this. It can be further noted that both F_1 and F_2 appear to be "split" when first excited by each pulse in the nasalized vowel, and these split components then later converge during the glottal cycle. The upper part of the split F_1 is labeled "nasal formant" because such a resonance just above F_1 is predicted to occur from Stevens' calculations, induced by the coupling of the nasal and oral cavities. It may be postulated that the split F_2 also results from a proximal nasal formant.

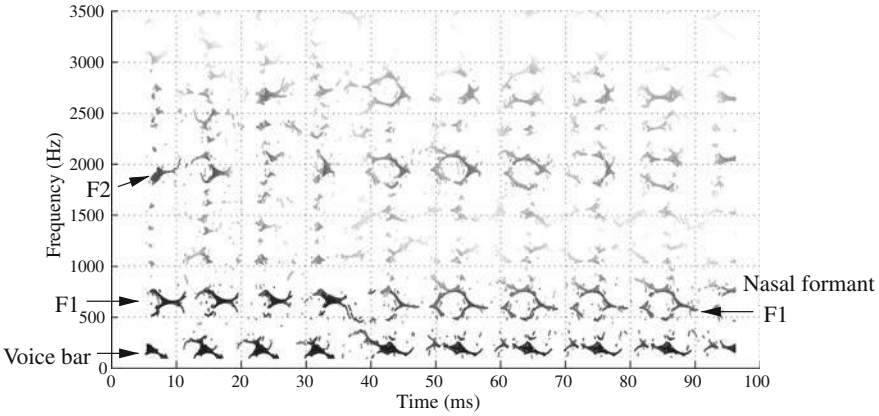


Fig. 6.18 Reassigned spectrogram imaging glottal pulsations in a spliced signal of the vowel [æ] for the first few pulses and nasalized [æ̃] for the second part. Computed using 10.5 ms analysis windows, and pruned to show line components only

Figure 6.19 shows reassigned spectrograms of the vowel-nasal transition period in the English words *clam* and *clan*. The vowel formants are somewhat perturbed during the transition, but the formants of the nasal consonant are reasonably well highlighted in the reassigned spectrograms. The clear differences between the labial and alveolar nasal formants are easier to pinpoint than in a conventional spectrogram.

6.7 Stops and Fricatives

Obstruents (stops and fricatives) usually have a significant noise event associated with them, and this is where a weakness of the reassigned spectrogram is revealed. Because the representation is designed to decompose every signal into components, it will still do so during random noise. Since noise does not contain any actual components, the representation of noise in this setting looks like a random walk involving numerous ever-changing "components," all of which are physically unreal. Gardner and Magnasco [12] explained that the apparent components in the noise are actually lines which separate spectrographic zeros. This takes a little getting used to, and cannot be regarded as an advantage of the reassigned spectrogram, but it is worth living with given the manifold advantages. In the end, however, a conventional spectrogram usually provides an equally useful image of a fricative noise.

In spite of the failure of the reassigned spectrogram to provide a better picture of noise, it can be very useful for probing various aspects of obstruents aside from the noise itself. Figure 6.20 shows the formant transitions coming out of an English [b], for instance. Sounds with multiple or "sloppy" release events can be

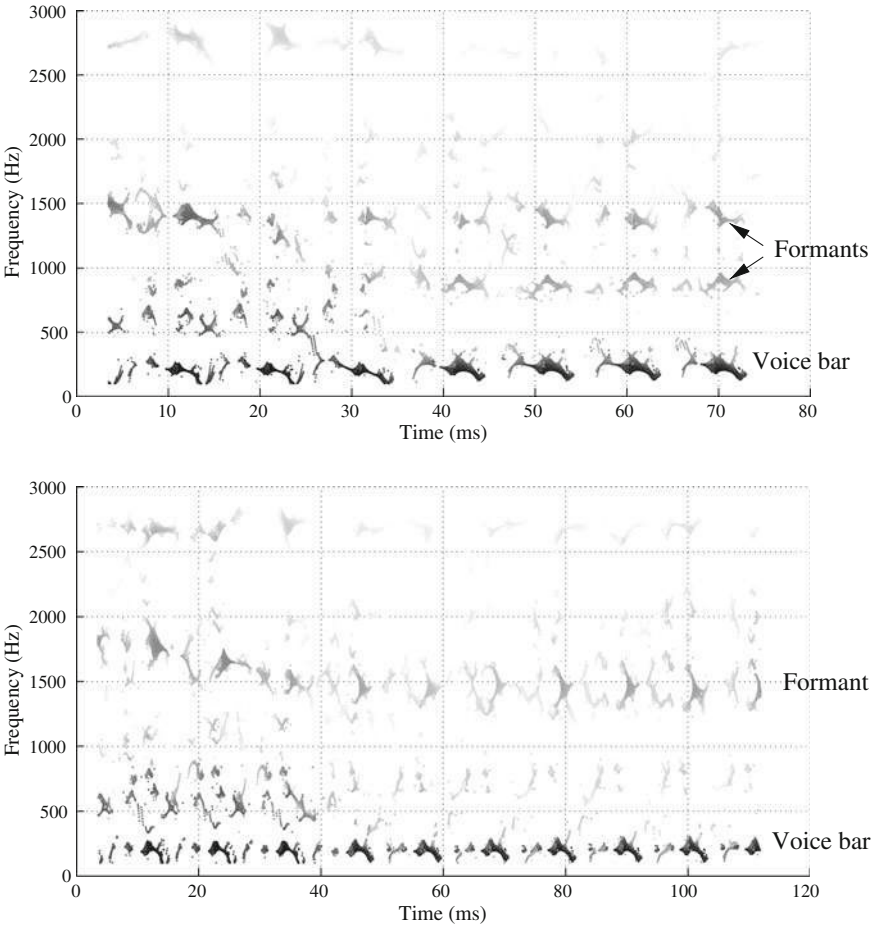


Fig. 6.19 Glottal pulsations in the portion of [æm] (*top*) and [æn] at the closing of the vowel and the beginning of the nasal. Computed using 7 ms analysis windows, pruned to show components with threshold of 0.15

imaged with remarkable time precision using the reassigned spectrogram, owing to the time reassignment using local group delay. Figure 6.21 shows a dental ejective click (which is probably uvular) in the Yeyi language.[1] Click sounds characteristically involve two stop releases; the first *anterior ingressive* release uses the front of the tongue in a sucking action from a dental point of articulation in this instance, while the subsequent ejective release here involves forcing air out with the rising larynx with the back of the tongue releasing from the uvula. Both of

[1] The click sound shown is taken from recordings that were described by Fulop, Ladefoged, Liu and Vossen [11], although the original judgement that this click was not uvular has been revised.

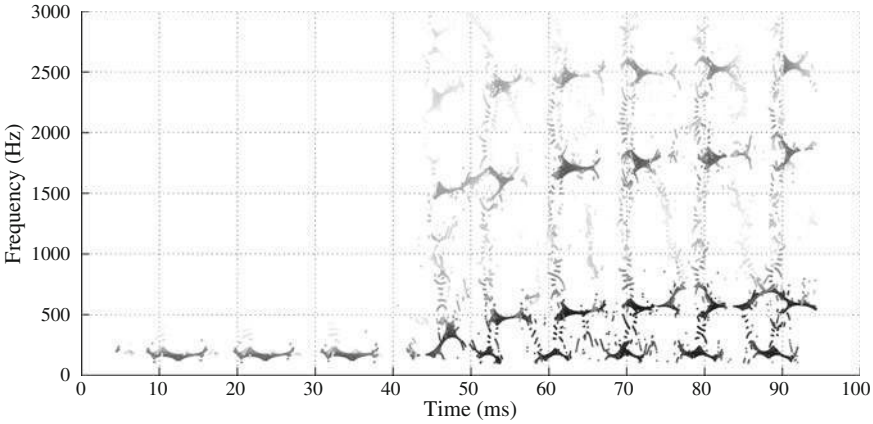


Fig. 6.20 Reassigned spectrogram showing the first portion of *bad*. Pruning thresholds were set to 0.25 for both components and impulses; 8.9 ms windows were used

these are quite sloppy, and it is easy to observe as closely spaced impulses the multiple release events in both the anterior and the dorsal release. A comparable level of detail is not seen in the conventional spectrogram.

6.8 Pitch Tracking

All of the spectrograms presented so far for speech analysis have been of the "wideband" variety using short windows, since the focus has been on the imaging of resonances and impulses. Moving to a long analysis window allows the display of a narrowband spectrogram, which can be useful for pitch tracking. Indeed, the narrowband reassigned spectrogram with pruning applied to show only the line components (which are the harmonics in a long-window analysis) appears to be a much more informative and robust method for pitch tracking than dedicated pitch tracking algorithms. In fact, the reassigned spectrogram has recently been shown to significantly outperform standard methods on the task of fundamental frequency tracking [15], mainly because the usual techniques (namely autocorrelation or the cepstrum) rely on what the cited authors call the "local stationarity assumption," and this is violated by any frequency-modulated signal.

Figure 6.22 compares a conventional pitch track with a reassigned spectrogram, applied to a portion of a naturally spoken sentence. The pitch track provided by Praat software is completely inaccurate and worthless in this example, although efforts were made to tweak the parameters of the tracking algorithm.[2] The

[2] The unusually poor performance of the Praat pitch tracker here may be due to an instability of the baseline voltage which is a common shortcoming of cheap microphones. Such "wavering" of the voltage spoils the values used by Praat's autocorrelation-based algorithm.

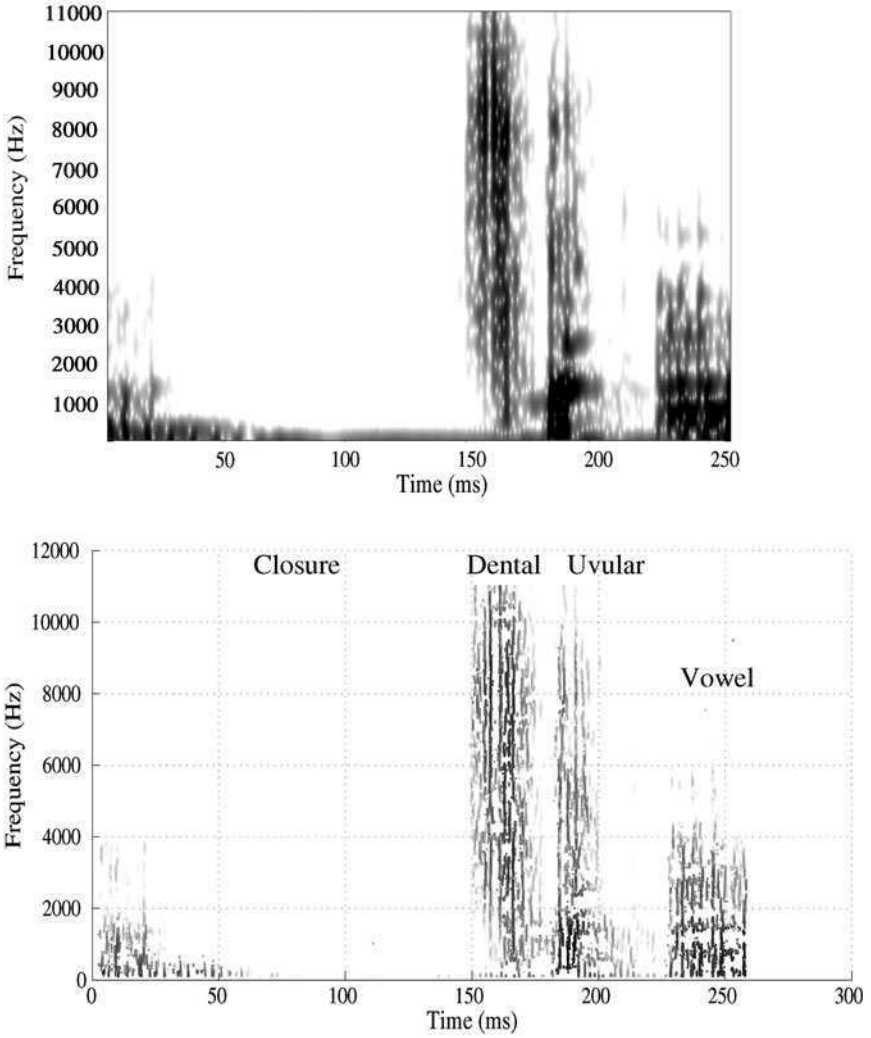


Fig. 6.21 Conventional and reassigned spectrograms, the latter highlighting just impulses using a pruning threshold of 0.25, showing the dental ejective uvular click in the Yeyi words [ku q|'akasa] (to drizzle). Computed using 5 ms analysis windows

reassigned spectrogram, on the other hand, shows the fundamental frequency quite well throughout every voiced sound, except at the point of lowest excursion during the final [l]-sound of *syllable*. The dropout of F_0 here may be due to the limits of the low-fidelity microphone having been reached; higher harmonics are also shown in the image, and they can easily be used to measure the pitch. If the first harmonic above the fundamental is tracked, one need only divide the measured values by 2.

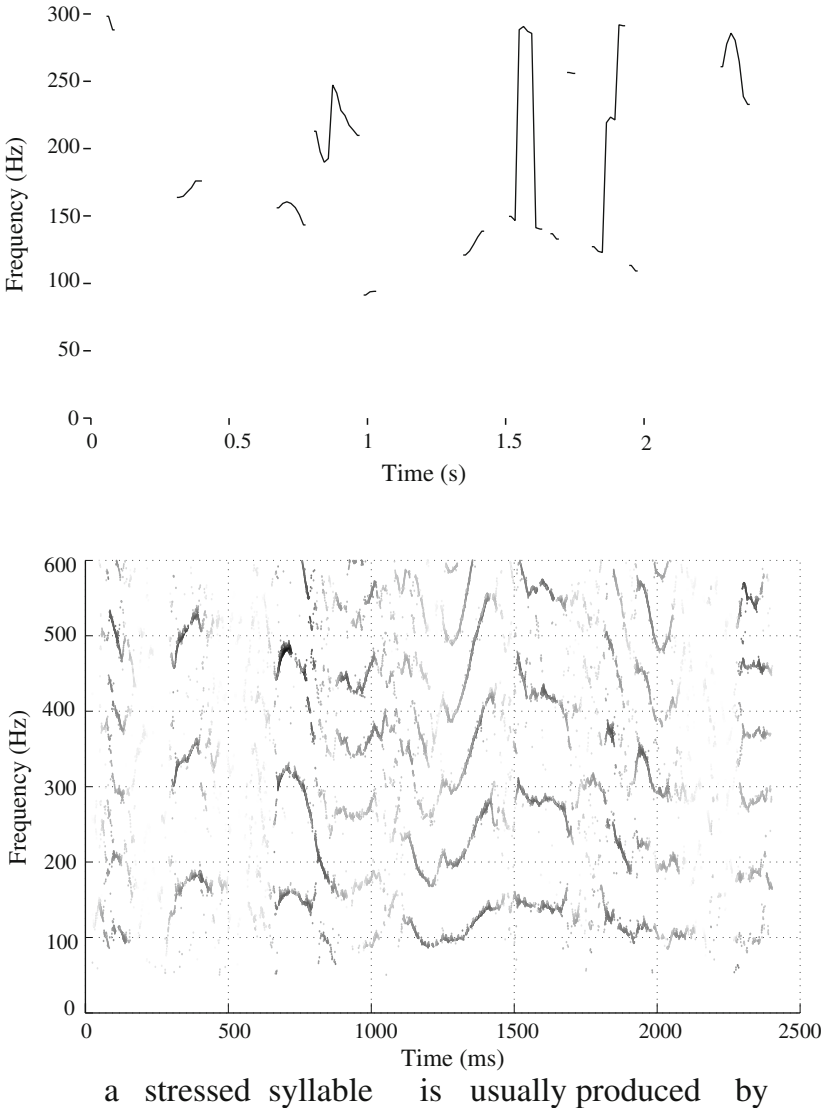


Fig. 6.22 Standard pitch tracking algorithm (Praat software, upper panel) performance contrasted with reassigned narrowband spectrogram computed with 40 ms windows and pruned to highlight harmonics. The subject signal is excised from a naturally produced English sentence, aligned with the analyses above

6.9 Appendix: Matlab Techniques

There are six mfiles included for computing and displaying reassigned spectrograms. Three of these are for displaying complete images, and the remaining three are for displaying pruned images using the partial derivative threshold technique.

All six are invoked using nearly the same template including a mandatory eight arguments and an optional five (or six) output variables. The file `Nelsonspec.m` is the most basic of the group; it computes and displays a grayscale reassigned spectrogram according to the template:

$$[\texttt{STFTpos}, \texttt{CIFpos}, \texttt{LGDpos}, \texttt{f}, \texttt{t}] = \texttt{Nelsonspec}(\texttt{signal}, \texttt{Fs}, \texttt{window}, \texttt{overlap}, \texttt{fftn}, \texttt{low}, \texttt{high}, \texttt{clip})$$

where the inputs are the same as for the conventional spectrogram routine; `window` and `overlap` set respectively the length of the STFT analysis window (in samples) and the number of samples by which successive windows overlap—so this is the complement of the hop size, which is `window–overlap`. The main difference in the display (apart from the reassignment of the values) is that here I use a 3D scatterplot instead of an interpolated shaded plot; this makes it important to use Nelsonspec and relatives with a very large overlap, usually at least 90% of the window, while this much overlap is just wasted computation in a conventional spectrogram.

The output variables (if used) will send back respectively the displayed portion of the complex STFT, the coindexed channelized instantaneous frequency values (Hz), the coindexed local group delay values (seconds), and the frequency (Hz) and time (seconds) axis vectors. Other files called with the same syntax and options are as follows: `Nelsonspecjet.m` is the same as above, except the image is displayed using my custom colormap. `Nelsonspecm.m` is also the same, but adds a small window in the lower left of the screen where the frequency under the mouse pointer is reported. This feature is intended to facilitate manual measurement, and still functions if the zoom is changed. All three of these functions include the option, after displaying the reassigned spectrogram, of also displaying a conventional spectrogram in a separate figure. The tapering window is set to Kaiser by default in the code.

The three related functions which display a pruned reassigned spectrogram are called using the following template:

$$[\texttt{STFTpos}, \texttt{CIFpos}, \texttt{LGDpos}, \texttt{CIFderiv}, \texttt{f}, \texttt{t}] = \texttt{Nelsonspec\_both}(\texttt{signal}, \texttt{Fs}, \texttt{window}, \texttt{overlap}, \texttt{fftn}, \texttt{low}, \texttt{high}, \texttt{clip})$$

This is identical to the above functions except for the additional output variable `CIFderiv`, which (if used) will return the matrix of mixed partial derivative values corresponding to each reassigned data point in the spectrogram. There are two similar functions:

`Nelsonspecjet_both.m` displays the pruned image using my custom colormap.
`Nelsonspecjet_bothm.m` includes the frequency measurement tool in the color version. All three of these functions have the degree of pruning set in the code; the relevant line sets the variable named `plot_these`, and includes numerical thresholds for the mixed partial derivative of the STFT phase in the following expression:

$$(\texttt{abs(CIFderiv)} < \texttt{x} \mid \texttt{abs(CIFderiv} - 1) < \texttt{y})$$

In the actual code, the placeholders `x` and `y` are filled by numbers between 0 and 1 which set the thresholds for line components and impulses, respectively. The default settings are $x = 0.1$ and $y = 0.25$, which I have found works well for highlighting components and impulses in speech while removing most everything else. If the user prefers to see components only, simply set $y = 0$ in the code directly.

As an added bonus, two mfiles are included which plot a reassigned power spectrum, which is a channelized instantaneous frequency spectrum computed from a single FFT of a signal using no time reassignment. `Nelsonpower.m` is invoked with the following command template:

$$[\texttt{PS, CIF, f}] = \texttt{Nelsonpower(signal, Fs, low, high)}$$

The optional output variables provide the spectral amplitudes, the vector of channelized instantaneous frequencies, and the frequency axis vector respectively. `Nelsonpowerm.m` is essentially the same function, with the addition of a tool which reports the frequency and intensity values at the mouse pointer.

References

1. E. Castelli, P. Perrier, P. Badin, Acoustic considerations upon the low nasal formant based on nasopharyngeal tract transfer function measurements, in *Eurospeech* '89, pp. 2412–2415 (ISCA, Kolkata, 1989)
2. J. Dang, K. Honda, An improved vocal tract model of vowel production implementing piriform resonance and transvelar nasal coupling, in: *ICSLP*-1996, pp. 965–968 (ISCA, Kolkata, 1996)
3. G. Fant, *Acoustic Theory of Speech Production* (Mouton, The Hague, 1960). Reissued 1970
4. J.L. Flanagan, Speech Analysis, Synthesis and Perception, 2nd edn. (Springer, Berlin , 1972)
5. S.A. Fulop, Phonetic applications of the time-corrected instantaneous frequency spectrogram. Phonetica **64**, 237–62 (2007)
6. S.A. Fulop, Accuracy of formant measurement for synthesized vowels using the reassigned spectrogram and comparison with linear prediction. J. Acoust. Soc. Am. **127**(4), 2114–7 (2010)
7. S.A. Fulop, S.F. Disner, Advanced time–frequency displays applied to forensic speaker identification, in *Proceedings of Meetings on Acoustics*, vol.6. Acoustical Society of America (2009)
8. S.A. Fulop, K. Fitz, Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. J. Acoust. Soc. Am. **119**(1), 360–71 (2006)
9. S.A. Fulop, K. Fitz, Separation of components from impulses in reassigned spectrograms. J. Acoust. Soc. Am. **121**(3), 1510–8 (2007)
10. S.A. Fulop, E. Kari, P. Ladefoged, An acoustic study of the tongue root contrast in Degema vowels. Phonetica **55**(1–2), 80–98 (1998)
11. S.A. Fulop, P. Ladefoged, F. Liu, R. Vossen, Yeyi clicks: acoustic description and analysis. Phonetica. **60**(4), 231–60 (2003)
12. T.J. Gardner, M.O. Magnasco, Sparse time–frequency representations. Proc. Nat. Acad. Sci. **103**(16), 6094–9 (2006)

13. S.G. Guion, M.W. Post, D.L. Payne, Phonetic correlates of tongue root vowel contrasts in Maa. J. Phonetics **32**, 517–42 (2004)
14. F.J. Harris, On the use of windows for harmonic analysis with the discrete Fourier transform. Proc. IEEE **66**(1), 51–83 (1978)
15. M. Ito, M. Yano, Sinusoidal modeling for nonstationary voiced speech based on a local vector transform. J. Acoust. Soc. Am. **121**(3), 1717–27 (2007)
16. K. Kodera, R. Gendrin, C. de Villedary, Analysis of time-varying signals with small *BT* values. IEEE Trans. Acoust. Speech Signal Process. **ASSP-26**(1), 64–76 (1978)
17. K. Kodera, C. de Villedary, R. Gendrin, A new method for the numerical analysis of non-stationary signals. Phys. Earth Planet. Inter. **12**, 142–50 (1976)
18. P. Ladefoged, I. Maddieson, M. Jackson, Investigating phonation types in different languages. In: O. Fujimura (eds) Vocal Physiology: Voice Production, Mechanisms, and Functions, (Raven Press, New York, 1988)
19. R.S. McGowan, An aeroacoustic approach to phonation. J. Acoust. Soc. Am. **83**(2), 696–704 (1988)
20. L.K. Montgomery, I.S. Reed, A generalization of the Gabor–Helstrom transform. IEEE Trans. Inf. Theory **IT-13**, 344–345 (1967)
21. D.J. Nelson, Special purpose correlation functions for improved signal detection and parameter estimation, in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, pp. 73–76 (1993)
22. D.J. Nelson, Cross-spectral methods for processing speech. J. Acoust. Soc. Am. **110**(5), 2575–2592 (2001)
23. D.J. Nelson, Instantaneous higher order phase derivatives. Digital Signal Process. **12**, 416–428 (2002)
24. M.D. Plumpe, T.F. Quatieri, D.A. Reynolds, Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Trans. Speech Audio Process. **7**(5), 569–86 (1999)
25. T.F. Quatieri, Discrete-Time Speech Signal Processing. (Prentice Hall, Upper Saddle River, 2002)
26. A.W. Rihaczek, Signal energy distribution in time and frequency. IEEE Trans. Inf. Theory **IT-14**(3), 369–74 (1968)
27. M.R. Schroeder, B.S. Atal, Generalized short-time power spectra and autocorrelation functions. J. Acoust. Soc. Am. **34**(11), 1679–1683 (1962)
28. C.H. Shadle, A. Barney, P.O.A.L. Davies, Fluid flow in a dynamical mechanical model of the vocal folds and tract II: Implications for speech production studies. J. Acoust. Soc. Am. **105**(1), 456–66 (1999)
29. K.N. Stevens, *Acoustic Phonetics* (The MIT Press, Cambridge, 1998)
30. J.W. Strutt (Baron Rayleigh), *The Theory of Sound*, vol. II, 2nd edn. (1896)
31. W. Zhao, C. Zhang, S.H. Frankel, L. Mongeau, Computational aeroacoustics of phonation, part I: computational methods and sound generation mechanisms. J. Acoust. Soc. Am. **112**(5), 2134–46 (2002)

Chapter 7
Linear Prediction and ARMA Spectrum Estimation

Up to this point, I have presented speech analysis methods which obtain spectral or time–frequency information from the signal data directly by means of some kind of transformation. It has been shown how different processing schemes can extract such information from the signal in different ways, but none has relied on any special assumptions about a speech signal beyond the most general and widely-held sort. In this chapter, I introduce an entirely different approach to what is often called "spectral estimation," in which the signal is explicitly assumed to conform to the outlines of a model. The parameters of the assumed model are then estimated from the signal data, and the values of the parameters are used as a kind of proxy estimate of corresponding signal properties. This general type of spectral analysis is often called *parametric*, in opposition to the *nonparametric* methods to which I have limited the presentation thus far.

Confronted with such a general description of the parametric approach, a reader might well ask "why would I want to do that?" This question really does need to be taken seriously, it will be seen. Parametric spectrum estimation came on the scene in the 1960s, whereupon it was frequently argued as being capable of providing superior spectral estimates over Fourier analysis for many different kinds of signals (e.g. [8, 31]). In reality, parametric spectra are only superior to Fourier spectra when the signal conforms quite well to the assumptions made by the underlying model [20], to say nothing of the more recent nonparametric methods like reassignment. This caveat notwithstanding, the parametric method known here as linear prediction analysis (and by many other names in the wider literature) has found wide acceptance in the speech analysis and processing community. This is partly because speech signals, we will observe, often do conform reasonably well to the assumptions of the underlying model. Another reason, however, has more to do with expedience than prudence. Parametric methods, having relatively few parameters to estimate, lend themselves more readily to automation. After all, why peer at a computer screen all day making tedious measurements when a computer can do it for you?

S. A. Fulop, *Speech Spectrum Analysis*, Signals and Communication Technology,
DOI: 10.1007/978-3-642-17478-0_7, © Springer-Verlag Berlin Heidelberg 2011

Like most of the present book, this chapter has both an expository and a methodological goal. Carrying out the first goal involves explaining the fundamentals of linear time-invariant systems (i.e. filters), and how these are commonly represented on the complex number plane with the aid of the z-transform (v. Sect. 7.3). I also explain the relationship between filters and linear predictive (i.e. autoregressive) processes. The connection to speech modeling involves the source-filter theory, wherein the speech signal is assumed to emanate from a source (the vocal cords) and is subsequently filtered through some resonances. With the vocal tract being modeled as a linear filter, it becomes possible to model a speech signal as a linear predictive process (v. Sect. 7.2). The computable parameters of such a process are a set of coefficients of a polynomial, which are commonly known as either the linear prediction coefficients or the filter coefficients. After computing these parameters, it is then possible to either compute a power spectrum of the resulting speech model, or to compute each resonance frequency (and bandwidth) of the model in turn. In this way, the speech model resonances can be used as "proxy estimates" of the resonances in the vocal tract (v. Sect. 7.3).

Given a desire to estimate speech resonances using linear prediction analysis, it is important for the practitioner to make informed decisions concerning a variety of methodological variables during the procedure. Accordingly, I discuss the pros and cons of a number of possible procedures which strongly affect the results of linear prediction analysis (v. Sect. 7.4). Examples of linear prediction results for a variety of procedures are provided for synthesized vowels as well as real speech, so the reader can judge for him or herself how the accuracy of the modeling can be optimized through judicious choices of methodology (v. Sect. 7.5). Finally, it is shown how the methods of linear prediction can be extended to the realm of more complicated autoregressive moving average modeling of speech (v. Sect. 7.6). As usual, an appendix is provided in which the use of both Praat functions and the supplied Matlab code are detailed.

7.1 Preliminaries

In order to make the subject of linear prediction understandable, it is necessary to first go through some preliminaries that were not presented in Chap. 2. The whole subject is quite mathematical, but I have tried to minimize the number of equations or relegate them to the math boxes. I first go through the theory of linear filters, and show how a digital filter can be specified with the aid of the z-transform, which is a relative of the Fourier transform. The essential equivalence between a filter and an *autoregressive process* (the basis of linear prediction) is also shown. I next introduce the source-filter theory of speech in this connection, showing how speech can be modeled as an autoregressive process. For this section I have relied on some standard sources such as [14, 24, 29, 30] for the mathematics and facts.

7.1.1 Linear Filters and the z-Transform

A discrete-time system (filter) is *linear* just when it satisfies the superposition principle, meaning that if we input a weighted sum $a_1x_1(n) + a_2x_2(n)$ of two or more digital signals, the output can be decomposed into a weighted sum of the system responses to the component signals. Another important property in our current setting is *time-invariance*. A discrete-time system is time-invariant (also called shift-invariant) just in the case that if we delay our input signal by k points in time to provide $x(n-k)$, the resulting output is simply delayed by the same k time points, but is otherwise equal to the output function $y(n)$ that results from an undelayed input $x(n)$.

It can be shown how any discrete-time linear and time-invariant (LTI) system with input function x and output function y obeys a difference equation of the following form:

$$y(n) = -\sum_{k=1}^{N} a_k y(n-k) + \sum_{\ell=0}^{M} b_\ell x(n-\ell) \tag{7.1}$$

in which the sets $\{a_k\}$ and $\{b_\ell\}$ of coefficients contain only constants which are independent of the input and output. An LTI system is completely characterized by its *impulse response* $h(n)$, which is its output when provided with a unit sample sequence. Specifically, the output of a system given arbitrary input function x is computed by the *discrete convolution* of x with h:

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k) \tag{7.2}$$

The form of a typical impulse response function is not very illuminating about the system possessing it. Furthermore, computing convolutions directly is inconvenient. A more convenient setting for the analysis of discrete LTI systems and their input–output relations results from application of the *z-transform*. The z-transform of a discrete-time signal $x(n)$ is defined:

$$X(z) \stackrel{\text{def}}{=} \sum_{n=-\infty}^{\infty} x(n)z^{-n} \tag{7.3}$$

where z is a complex variable. Note first of all that this is a power series expansion quite similar in form to a discrete Fourier transform, only there is no exponential function involved. Here a signal is effectively represented in the complex number plane.

The z-transform is actually a well-studied type of complex power series called a *Laurent series expansion* [1]. It was introduced in discrete-time signal processing (and given its name) by Ragazzini and Zadeh [32]. To complete the representation of a signal using a z-transform, one must keep account of the *region of convergence* of the transform, i.e. the part of the complex plane within which the power series does not go infinite. If the unit

circle $|z| = 1$ is within the region of convergence, then we can evaluate the z-transform on the circle, as

$$X(z)|_{z=e^{i\omega}} = \sum_{n=-\infty}^{\infty} x(n)e^{-i\omega n}. \tag{7.4}$$

The astute reader will recognize this as being the discrete Fourier transform of $x(n)$, so this shows the relation between the two transforms. The DFT of a signal is precisely its z-transform evaluated on the unit circle.

The z-transform has many useful properties which make computation in the z domain easier to manage; perhaps the most important of these is the convolution property, which states that the z-transform of a convolution of two functions is just the product of their two z-transforms. Thanks to this fact, the input–output relation of Eq. 7.2 describing an LTI system has the following form in the z domain:

$$Y(z) = H(z)X(z) \tag{7.5}$$

where the functions Y, H, X are the respective z-transforms of the output, impulse response, and input functions. The function $H(z)$ is commonly called the *system function* characterizing the LTI system at hand.

By computing the z-transforms of both sides of the difference equation (7.1) describing an LTI system, the system function can be shown to take the following form:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{\ell=0}^{M} b_\ell z^{-\ell}}{1 - \sum_{k=1}^{N} a_k z^{-k}} \tag{7.6}$$

which is known as a rational system function. The points in the complex plane at which such a function goes to zero are precisely those points at which the numerator is zero. The zeros, or roots, of the numerator polynomial[1] are therefore called the zeros of the system. On the other hand, the zeros of the denominator expression are points in the z-plane where the system function goes infinite; these points are known as the *poles* of the system, following the terminology of complex analysis. Such systems can exist in some important particular forms; the most important for our purposes currently is the case in which the numerator is a trivial constant, so that $H(z)$ is completely characterized by the N poles whose values are determined by the parameters $\{a_k\}$. This sort of case is called an *all-pole* system or filter.

This way of viewing an LTI system is mathematically equivalent to viewing the output function as a random process. In particular, if we go back to Eq. 7.1, observe that in the all-pole case just described we can deal only with the first

[1] The expressions in the numerator and denominator are not strictly polynomials, but are actually a more general sort of creature called a *Laurent polynomial*, in which the indeterminate z is allowed to take negative powers.

summation, since the second reduces to a constant. Then we can write the equation in a slightly different form [29]:

$$y(n) - a_1 y(n-1) - \cdots - a_N y(n-N) = \varepsilon_n \tag{7.7}$$

where now we have introduced a representative value of a random process $\{\varepsilon_n\}$ to serve as the input function. This equation is nothing but a multiple linear regression equation from the theory of statistics, in which ε_n plays the part of an error term. It is thus shown how the output of an all-pole system can be regarded as a linear regression on its own past values, which in the terminology of time series statistics is called an *autoregressive* (*AR*) *process*.

Another view of the same fact derives from noticing that in Eq. 7.7, the error term ε_n is just the difference between the signal's value $y(n)$ at time n and the weighted sum of its N past values. Thus, the signal can be approximated, or "predicted," to within a certain error by a linear combination of its past values. This explains the commonplace term *linear prediction* being invoked when this kind of signal model is employed.

7.1.2 Source-Filter Model of Speech

The source-filter model of speech production decomposes the speech process into a number of component systems through which the input signal from the vocal cords must pass before being emitted from the body. In the z domain, the linear "source-filter" speech production model can be written very succinctly indeed [24]:

$$S(z) = E(z)G(z)V(z)L(z) \tag{7.8}$$

in which the vocal tract output system function $S(z)$ is given (thanks to the work of Fant [13]) as a product of four component systems. $E(z)$ here stands for the input system, a sequence of simple impulses from the vocal cords separated by the fundamental period of voicing. Since the vocal cords in fact emit filtered impulses, the system function $G(z)$ represents a "glottal shaping model" which is typically implemented as a simple low-pass filter. The main vocal tract system $V(z)$ is a fairly complicated resonating filter, which models the formant frequencies. Finally, $L(z)$ represents the effects of the lip opening on the ultimately radiated sound. All of these filters have been theoretically argued [14] to be well represented by all-pole system functions.

It has been shown in a number of sources (e.g. [14]) how the above model can be approximately implemented using a single resonant filter with an all-pole system function, in place of the three all-pole system functions G, V, L. This source-filter model of speech is standardly written in the form of the first equation below, in which the denominator is a Laurent polynomial whose roots are the poles of the vocal tract system including the glottal shaping and the lip radiation:

$$S(z) = \frac{E(z)}{1 - \sum_{k=1}^{N} a_k z^{-k}} \tag{7.9}$$

$$\hat{S}(z) = \frac{\text{Gain}}{1 - \sum_{k=1}^{N} a_k z^{-k}} \tag{7.10}$$

The second of the two equations is an approximation of the first that relies on the explicit separation of the source function $E(z)$ from the filter [22]. To write the second equation, the source is replaced by a value called the *gain factor*, so then the estimated vocal tract output $\hat{S}(z)$ only approximates $S(z)$ by taking the glottal source out of the picture.

Here is the situation we have arrived at. The Eq. 7.7 for a linear predictive (autoregressive) process was derived from Eq. 7.1 describing the input–output relation of a linear time-invariant filter. The coefficients $\{a_k\}$ are generally called the *prediction* coefficients, and their number N is called the prediction order. Then, Eq. 7.10 was derived from an approximate model of the speech production process. Notice that it has the form of a simple input constant multiplied by a rational system function as in (7.6), but with only the coefficients $\{a_k\}$ in the denominator. These coefficients may be called the *filter coefficients*, and their number N is called the filter order. Now recall that Eq. 7.6 was also derived from Eq. 7.1 by moving to the z domain. The two sets of coefficients $\{a_k\}$ are really the same set. So this demonstrates that an approximate speech production model represents the vocal tract as an LTI filter, whose z-transform is of the all-pole variety, and in which the coefficients of the denominator polynomial are in fact a set of linear prediction coefficients modeling the output of the filter—i.e., the speech. The gain factor is indeed just a form of the prediction error ε_n.

This confluence of the two mathematical streams—filters and linear prediction of a process—is certainly interesting. Since when do filters have anything to do with prediction of a signal from its own past? Indeed, cautionary tales lurk here. It must be remembered that the process of linear prediction is really just a type of parametric signal modeling, also called autoregressive modeling, and filters need not be used to develop this concept. A wide range of signals can be profitably modeled using linear prediction, including random noise and musical sounds, and there is no filter represented directly in the autoregressive process definition of Eq. 7.7. The prediction coefficients can only be sensibly interpreted as poles in the transfer function of a filter under the happy circumstance that some sort of filtering of a source is actually taking place. In the case of speech, this is a reasonable view of the process, but only when the number of coefficients equals the number of poles in the physical vocal tract filter. In the case of, say, an electric guitar, it is not likely to be a profitable perspective in any event. Yet a linear prediction model of an electric guitar note can be quite a good representation, depending on the parameters and what is desired from them, as will be shown in the next section. The moral here is that, just because we can represent a signal using a linear prediction model which is *mathematically* equivalent to an all-pole filtering

model, this alone cannot guarantee that actual filters in the real world are busy filtering the signal. In simple terms, mathematical equivalence among models is not the same as actual equivalence among the things being modeled.

7.2 Speech Spectra from Linear Prediction

In speech science, linear prediction is commonly referred to as "LPC analysis," where the abbreviation is for "linear predictive coding." This terminology is somewhat inaccurate, because although linear prediction is commonly used for speech coding and processing, phoneticians and speech scientists who use linear prediction are not normally doing any speech coding. Let us instead use the phrase "linear prediction analysis," or the shorter "LP analysis," to refer to the use of linear prediction methods in a typical phonetic application. The ultimate goal of linear prediction analysis is normally to obtain an approximation of the speech spectrum with the glottal source removed, including estimates of key values such as the frequencies of resonances. An expression (7.10) for the approximate transfer function $\hat{S}(z)$ of the signal was obtained above in the z domain. From this one can derive the approximate Fourier transform of the signal, which yields an approximation, or "smoothing," of the spectrum.

Given the facts presented in the math box in the preceding section, the approximate Fourier transform representation $\hat{S}(\omega)$ of the signal can be derived from $\hat{S}(z)$ by setting $z = e^{i\omega T_s}$, with T_s equal to the sampling period of the digital signal (i.e. the reciprocal of the sampling frequency). From the discussion of Chap. 2, we may recall that the power spectrum we are after is the square of the magnitude of $\hat{S}(\omega)$, which may be written in terms of the discrete Fourier transform in the following way [22]:

$$|\hat{S}(\omega)|^2 = \frac{\text{Gain}^2}{\left|1 - \sum_{k=1}^{p} a_k e^{-ik\omega T_s}\right|^2} \tag{7.11}$$

From the foregoing considerations it may be seen that in a practical algorithm for linear predictive speech analysis there will be three general steps. The first thing is to compute the set of p predictor (filter) coefficients $\{a_1, \ldots, a_p\}$, for a user-selected number p. The second step is to compute the gain factor; this together with the coefficients is sufficient for many purposes. If desired, the estimated power spectrum can then be computed from these parameters using Eq. 7.11.

7.2.1 Computing the LP Coefficients

The computation of the predictor coefficients is the most technical part, and for our purposes here it would be too long-winded to discuss the specific methods for

computing these in any detail. Indeed, entire books [24] have been devoted to this topic alone, so I am not ashamed to simply refer the interested reader to the rich array of literature. In modern practice there are a number of different techniques available which will compute the p predictor coefficients from a window of the signal. Praat software and Matlab (with Signal Processing toolbox installed) both provide three choices for the algorithm which are usually termed the "autocorrelation," "covariance," and "Burg" methods. These are now classic methods which have been discussed in many papers and books; the autocorrelation and covariance algorithms are treated in [23, 24], while Burg's "maximum entropy method" dating from [9] is given a fully fledged algorithm in [2]. As the Matlab signal processing toolbox costs extra, I also included a Matlab procedure from a freely available toolbox that uses an alternative method for computing the linear prediction coefficients from the statistical *cumulants* of the signal window [16]. All of these methods work well when a tapered speech signal window encompasses a few glottal cycles; anywhere from 25 to 40 ms is a commonly chosen window range. This general approach is usually called a *pitch-asynchronous* analysis, because no special care is taken to align the analysis window with the glottal cycles.

The essential equality of the autocorrelation and covariance methods over sufficiently long windows was shown by Markel and Gray [24], while the identity between the Burg and autocorrelation methods was shown by van den Bos [5]. Thus in practice, any of the three methods applied to a 40 ms tapered speech window will all yield identical results. The cumulant-based method is the newest, and can actually depart significantly from the others and may improve upon them under certain conditions. How it works depends on a parameter to the procedure called the "cumulant order," which sets how the statistical cumulant series is calculated and used to estimate the linear prediction model. If the cumulant order is 2, then only second-order cumulants are used. These are simply autocorrelations, which renders the technique identical to the autocorrelation algorithm in this case. The cumulant order may also be set to 3 or 4, however, and using higher-order cumulants in the calculation will practically always yield a different model of the speech.

The covariance method can also be applied to good effect when a window shorter than one glottal cycle is selected [24]. This is known as a *closed phase* analysis, and it is trickier to do properly because the window on the signal must be untapered (rectangular) and has to start precisely at the glottal impulse generated by the closing of the vocal cords. This technique theoretically provides the most accurate formant estimates for speech analysis [24], but in practice it has some pitfalls. All the above methods will be compared anecdotally with some examples in a subsequent section.

7.2.2 Computing the Gain and Power Spectrum

The simplest way to compute the gain factor is discussed by Markel and Gray [24], who present it in terms of the signal autocorrelation function r as defined in Eq. 2.10:

$$\text{Gain} = \sqrt{\sum_{i=0}^{p} a_i r(i)} \tag{7.12}$$

in which the "zeroth" prediction coefficient $a_0 = 1$. The linear prediction coefficients by themselves are sufficient to estimate the shape of the signal spectrum. The purpose of the gain factor is to bring the absolute power of the estimated spectrum into close alignment with the raw power spectrum that results from a Fourier transform. In fact, the resulting estimated spectrum will always have a slightly higher average value than the power spectrum from an FFT [24]. The accompanying Matlab code provides precise details of the gain computation by the above means.

It is fairly simple to compute the power spectrum determined by the linear prediction coefficients and the gain factor. As given in Eq. 7.11, Makhoul [23] states that one may proceed by "dividing Gain^2 by the magnitude squared of the FFT of the sequence: $1, a_1, a_2, \ldots, a_p$. Arbitrary frequency resolution can be obtained by simply appending an appropriate number of zeros to this sequence before taking the FFT." This last remark is very important, since for a typical speech prediction order of $p = 10$, the FFT of the coefficient sequence alone would be an 11-point FFT, which would then provide just 5 frequency bins to cover the relevant frequency range. The computation of the linear prediction power spectrum thus provides one of the most striking demonstrations of the utility of zero-padding, belying Hamming's qualms [18]—one can use an 11-point coefficient sequence followed by 2,037 zeros to compute a perfectly nice 2,048-point FFT for the spectrum, which will then be quantized into 1,024 frequency bins.

7.3 Interpretation as Filter Spectrum

Linear prediction spectrum analysis came on the scene in the 1960s, but not every developer was interested in speech spectra. Burg, for instance, came up with his algorithm to examine other kinds of natural signals [9]. As was mentioned above, a linear prediction model can be used to generate a smoothed power spectrum of a wide range of signals—the smaller the number of coefficients, the smoother the spectrum. The limit as the number of coefficients increases is just the precise Fourier transform spectrum of the signal window. Figure 7.1 illustrates the smoothing of the power spectrum of a 70 ms Kaiser-windowed snippet of an electric guitar note. The 14-coefficient model shows a broad outline of the spectral shape, while the 60-coefficient model shows every harmonic peak while still smoothing away the aperiodic components found between peaks. There is no profitable way to view these smoothed spectra as literally showing resonance peaks of a filter that has been applied.

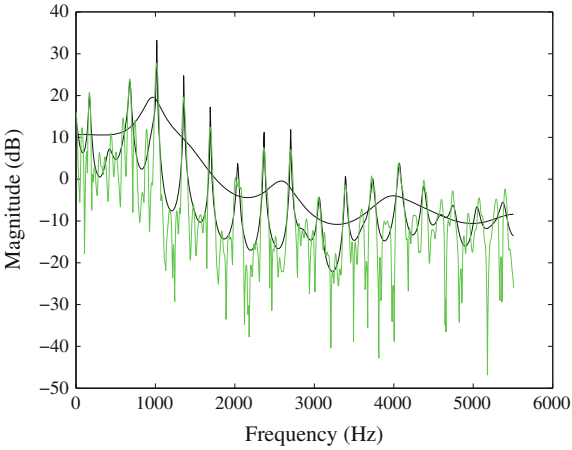


Fig. 7.1 Power spectrum of electric guitar note (*green/gray* line) smoothed using a 14-coefficient LP analysis (*smooth black line*) and a 60-coefficient analysis (*peaked black line*)

7.3.1 Poles and Resonances

Notwithstanding the above, it has been shown that the linear prediction model can be represented as a source-filter model in Eq. 7.10. If the model is applied to speech, then the all-pole filter can faithfully model the speech "filter" (i.e. the vocal tract) separately from the glottal source. We should thus be able to get a good look at the formants if we choose our parameters well. The ultimate question of the "reality" of the resonances in a linear prediction filter model for a signal is analogous to the question of the reality of components revealed in the Fourier spectrum. In each case, the mathematical soundness of the representation is beyond doubt, but the resulting representation may not directly show things that are "real" in a physical sense.

This is really the main problem with the application of linear prediction analysis to speech, because speech is a special sort of signal that is actually produced using a filter that has discoverable physical properties. We therefore wish to select a number of coefficients (poles) for our model that will allow the results to have a physical interpretation in terms of actual vocal tract resonances. Turning to the z-transform view in which the $\{a_k\}$ act as coefficients in the polynomial determining the filter, it can be shown how each root of this polynomial has a particular role to play in the smooth LP power spectrum. From this one can deduce approximately the number of coefficients that will be "just right" for a model, as this number must equal the number of roots.

The theory of filters is well-understood in terms of system functions expressed in the z domain, so I will just provide a simple explanation and refer readers interested in details of this theory to a standard signal processing textbook [30]. An all-pole filter (Eq. 7.10) of the kind emulating a vocal tract must be some sort of linear resonating filter, having a number of resonances which are emphasized in its frequency response. The roots of the denominator polynomial (i.e. poles of the filter) could in principle be either real or complex numbers, but it turns out that for a well-behaved resonating filter any complex roots must come in conjugate pairs,

while real roots may be singletons. It is a further fact of filter theory that a pair of complex conjugate roots determines a resonance in the filter's frequency response spectrum (in fact the root in the upper half of the z-plane determines the positive frequency, while its conjugate root simply reflects this information as a negative frequency). The proximity of a root to the unit circle determines the sharpness (i.e. bandwidth) of a corresponding resonance peak. A pair of roots which are not particularly near the unit circle will determine a very broad resonance that is not recognizable in the spectrum as a peak, because it may simply "shape" the sides of other sharper peaks. Similarly, a real root (one is present whenever an odd number of poles is used) may change the shape of the spectrum but cannot add a peak of its own.

The precise relationship between a complex pole and a resonance was first presented in a linear prediction context by Atal and Hanauer [3]. Christensen et al. [10] gave similar formulae which express the resonance frequency in terms of the argument (angle) of the pole and the bandwidth in terms of the magnitude. So for a complex pole $z = x + iy = re^{i\theta}$ one has the following expressions for the frequency F and bandwidth B of the resonance:

$$F = \frac{\theta}{2\pi T_s} = \frac{\text{atan2}(\text{y}, \text{x})}{2\pi T_s} \tag{7.13}$$

$$B = \frac{\ln(r)}{\pi T_s} = \frac{\ln(x^2 + y^2)}{2\pi T_s}, \tag{7.14}$$

where T_s is the sampling period and atan2 is the four-quadrant arctangent function [36]. Observe how the frequency of the resonance is determined completely from the argument (angle) of the pole, while the bandwidth is determined completely by the magnitude of the pole.

Figure 7.2 illustrates the relationship between the complex roots (poles) of the filter transfer function and the spectrum, both derived from linear prediction analysis with 10 poles of the synthesized vowel [æ]. The peaks of the spectrum were found using a typical sort of peak-picking algorithm, while the poles of the transfer function were determined by solving the roots of the Laurent polynomial in the denominator (Eq. 7.10) using a Matlab root solving routine. The two sets of values are presented in Table 7.1 for two different methods of computing the filter coefficients, alongside the

Table 7.1 Linear prediction 10-pole spectrum peaks (Hz) for vowel [æ] compared to known values and solved roots, cf. Fig. 7.2

Known	731	1,768	2,500	3,500	
Peaks (Burg)	732	1,787	2,471	3,291	
Peaks (rcest)	723	1,777	2,481	3,369	
Roots (Burg)	728	1,791	2,480	3,344	1,012
Roots (rcest)	723	1,777	2,503	3,452	1,305

The Burg LP algorithm results are compared with those from the cumulant-based estimates labeled "rcest" applied to the same 50 ms Kaiser window of the vowel

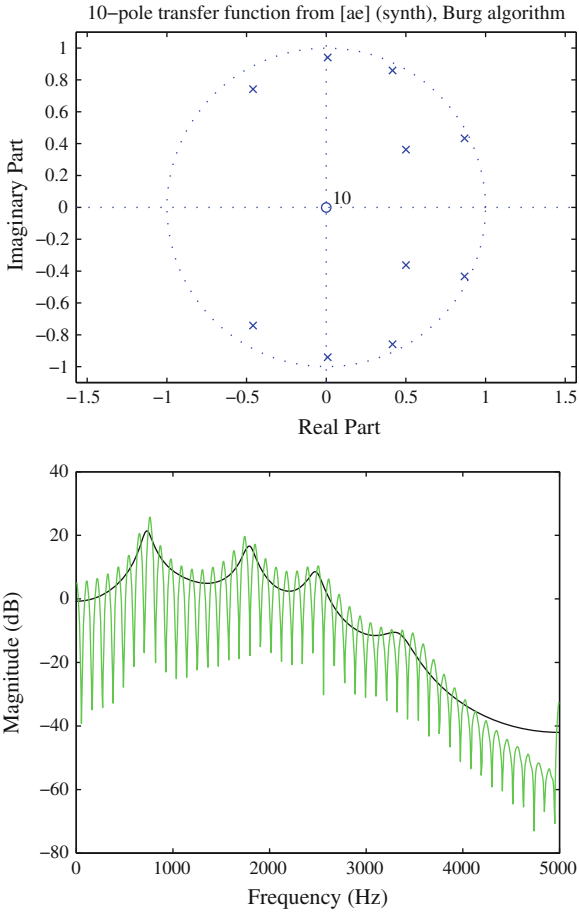


Fig. 7.2 *Upper panel* shows the ten poles of an LP model of synthesized [æ]; *lower panel* shows the corresponding power spectrum (*smooth line*) together with the Fourier spectrum of the analysis window

resonance values which were used to synthesize the vowel. In this context where one has actual resonances, it is common to think of LP analysis as a tool for estimating, or "measuring," the resonance frequencies and bandwidths.[2]

Peak-picking of the spectrum in Fig. 7.2 yields reasonably accurate measurements of the lower two formants, but the accuracy decreases for the higher formants which are lower in amplitude (and also less important for vowel characterization). Meanwhile, the analysis using ten poles presents the problem that there are five resonance frequencies which correspond to the pole pairs. Three of these are at least as accurate as those provided by peak-picking, while the extra resonance from each LP method represents a "shaping" resonance rather than a

[2] The view of LP analysis as a measurement tool is fraught with difficulty; should a scientist accept a measurement from a tool that would provide "wrong" measurements if it were not specially set up using partial foreknowledge of the desired outcome?

spectral peak; this fact can be recognized from its much higher computed bandwidth (not shown here) than those of the peak resonances. The frequencies of the spectral peaks are generally different from the values found by solving the roots; the reason for these discrepancies will now be discussed.

7.3.2 Picking Peaks Versus Solving Roots

Given any linear prediction analysis of speech, we have seen two ways of extracting the resulting estimates of the resonance frequencies: direct solution of the polynomial roots (a method which does not even require a power spectrum to have been calculated), and picking peak values from the corresponding power spectrum. McCandless [25] discussed pros and cons of each method, and settled on peak-picking for reasons of computational tractability and reliability. Polynomial root-solving is indeed an intensive numerical procedure, but its intractability is less of an issue with today's faster computers. In truth, however, there is a subtlety which distinguishes the two methods that is often overlooked, as it was by McCandless (op. cit.) and also by Christensen et al. [10] in their comparison of LP formant extraction methods.

It is important to recognize that the resonances that are computed directly from the LP filter polynomial are estimated values of assumed physical resonances in the filter—one might say they are the "production resonances." The filter spectrum, meanwhile, can be understood as a superposition of the spectra of the various resonances. This superposition introduces the prospect for the individual resonance peaks to influence each other's location, with the result that the locations of peaks in the combined frequency response may be slightly different from the production resonances. Nevertheless, it must be admitted that humans can only detect resonances by hearing them, which in some fashion involves hearing the combined frequency response spectrum. So, the peaks in the filter spectrum are essentially like "auditory resonances," whose values may differ somewhat from the production resonances.[3] It is therefore effectively impossible for us to hear, or otherwise detect acoustically, the "true" formants that are involved in speech.

To evaluate the accuracy of the formants estimated with linear prediction, I often rely on synthesized vowels whose formants are known. In doing so, the above discussion must be kept in mind; the known "true" formants of the synthesized vowels are in fact the production formants, which may differ somewhat (just how much depends on the spectral shape) from the formant peaks in the resulting spectrum. The formants estimated from a speech spectrum using any means including reassigned spectrograms etc. cannot be expected to be exactly equal to the true production formants in general. Table 7.2 provides another set of values comparing the estimated formants from root-solving and peak-picking the

[3] It was Paul Boersma who pointed out this important fact to me.

Table 7.2 Linear prediction (Burg) 10-pole spectrum peaks (Hz) for synthesized vowel [i] compared to known values and solved roots

Known	306	2,241	2,500	3,500	
Peaks	317	2,217	2,485	3,291	
Roots	318	2,207	2,499	3,343	995

same linear prediction analysis of a synthesized vowel. The "extra" pole at 995 Hz has a large bandwidth, which could be used to identify it as a shaping resonance rather than as a peak.

It seems that there are quite a number of factors that need to be weighed in order to decide upon a favorite method for determining the filter resonances from an LP analysis. It is also true that our response to these factors could well be different today from what it would have been in the 1970s, which was the last time the matter was seriously examined in the literature. So let us lay down our hand and look at the cards, as it were.

The method of peak-picking the LP spectrum has a couple of strong points: (1) the peaks thus found will realistically correspond to spectral peaks detectable by the auditory process; (2) it is computationally tractable (indeed very fast) to pick peaks automatically from a spectrum. This method is, however, seriously hampered by a severe weakness, which is that picking peaks may miss actual peaks because of apparent merging in the spectrum, even when they have in fact been separated by the LP analysis into two distinct pole pairs. In practice, two distinct resonances have to be farther apart than the minimum resolvability distance in order to be split into two obvious peaks in the power spectrum, and this seriously hinders the resolution of the peak-picking procedure.

The method of root-solving the transfer function denominator has the advantage that, with a good root-solving algorithm, pole pairs (and thus, resonances) that are quite proximal can nevertheless be resolved. The potential formant resolving power is indeed superior to that of even the reassigned spectrogram for higher-pitched voices. The computational intractability of the scheme should not deter us in the present and future eras, as it deterred our predecessors in the 1970s. Nevertheless, this technique has two negative features, only one of which can be mitigated through careful methodology.

The first problem is that a good LP model of speech has to provide at least one or two more pole pairs than the number of expected resonances (formants), because of the need to include the glottal shaping and lip radiation filters in the model. This means that a good model should have shaping resonances in the spectrum as well as resonances which produce peaks. But, the root-solving technique will in general find all of the poles, so some pole pairs will indicate shaping resonances (as has already been seen in examples above).

Fortunately, the bandwidth of a resonance can be computed from its poles using Eq. 7.14, and this value can be used as a threshold to decide a shaping resonance from a peak. Experience has shown that the great majority of computed resonances with bandwidth greater than about 300–400 Hz are not peaks, but are shaping resonances that should not be reported as formants or anything else. The precise

value of the "resonance peak threshold" can be empirically adjusted in each circumstance, but the general application of a bandwidth condition is a simple way to report only spectral peaks. It is a further matter to decide which peaks are formants as against other kinds of resonances; there is no generally applicable way of doing this other than guesswork.

The second problem with using root-solving to report spectral peaks is that, as discussed above, the peaks thus found are the "production resonances" rather than the auditory ones actually present in the power spectrum. There is no way to mitigate this, but one accuracy study [34] found that the size of the systematic difference between peaks derived from poles and the corresponding peaks in the LP spectrum rarely exceeded the formant perception difference limens found by Kewley-Port and Watson [21]. In my judgement, the superior ability to find and resolve spectral peaks using the poles of the LP filter model gives the root-solving methodology a definite advantage overall in comparison with spectral peak-picking, even in view of the former method's being limited to locating production resonances. This is naturally true for accuracy studies on synthesized vowels, in which the known formants are actually the production resonances.

7.4 Practical Spectrum Analysis and Formant Extraction

A point in favor of using linear prediction analysis for locating formants is that the values can be reported by automatic procedures which examine the LP transfer function (in the case of root-solving) or spectrum (in the case of peak-picking). Having decided to employ linear prediction analysis, one is confronted with a number of options governing the overall methodology. Some of these are more critical than others, and some have been dealt with above. The goal of this section is to indicate the general accuracy that can be achieved with LP analysis, while offering suggestions as to the best choices for the various parameters.

7.4.1 Linear Prediction Accuracy Studies

The most complete study of formant measurement accuracy for synthesized vowels seems to be that of Monsen and Engebretson [26]. Their study pitted spectrographic manual measurement against a linear predictive analysis procedure of a sort typically used in speech science at that time. The spectrographic measurement procedure was rather flexible but did not use wideband analysis directly; three experts were provided with a wideband spectrogram "for orientation," paired with a single narrowband power spectrum for each vowel token. The readers could only use the narrowband power spectrum to measure the formant values, which they would have to do by manually smoothing the spectrum to locate the peaks by eye.

The linear predictive analysis was performed manually as well. The procedure was inadequately described, but seems to have involved deliberately overfitting the power spectrum using a linear prediction filter having 22 poles. The values of the first three formants would have to have been extracted from the expected 10 or 11 computed resonances by manual inspection. No mention was made of the specific means for deciding on the formant values from the overfit analysis, although it was mentioned that smoother linear prediction spectra having a smaller number of poles yielded decreased measurement accuracy.

Monsen and Engebretson found that the linear prediction-estimated F_1, F_2 and F_3 of their synthesized vowels were only accurate to within ± 60 Hz, and that manual spectrographic measurement performed comparably on the lower two formants, and less well on F_3. They further suggested that this margin of error is contained within the difference limens for formant perception, but this is apparently not correct. Kewley-Port and Watson's [21] later comprehensive study of formant perception in synthesized vowels reports difference limens of 14 Hz for formants below 800 Hz and 1.5% for higher formants.

The inaccuracies of linear prediction have occasionally been more systematically studied in the literature. Vallabha and Tuller [34] distinguished several sources of systematic errors in linear prediction-estimated formants using a typical pitch-asynchronous methodology, which include the quantization of the frequency range by the harmonics of phonation, as well as the "error"[4] introduced by approximating the speech spectrum peaks by using the resonance values computed directly from the roots of the predictor polynomial. They found the order of magnitude of all of these errors to be approximately the same as the difference limens of perception found by Kewley-Port and Watson (op. cit.). We should keep in mind, however, that these kinds of "systematic" errors do not tell the whole story of the actual measurement errors, which may not in general be the result of anything systematic. My own small study of nine synthesized vowel tokens [15] did not rigorously quantify measurement accuracy, but found that pitch-asynchronous LP roots yielded F_1 and F_2 values which were usually too high by a considerable amount, up to 17%.

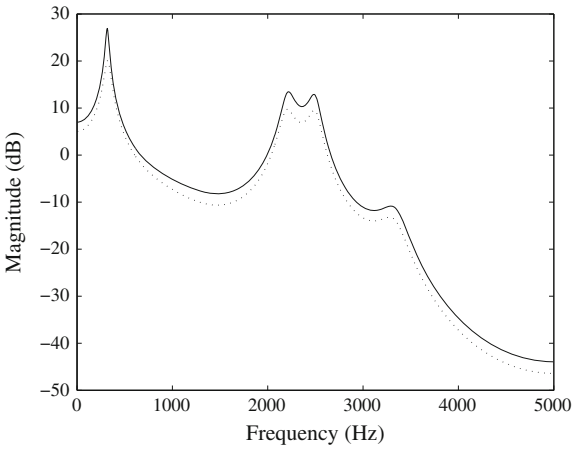
7.4.2 Analysis Windows

The parameters that are the simplest to assign are the length of the analysis windows which are used to develop the LP model, and the tapering function that is applied to these windows prior to the modeling. For a typical pitch-asynchronous analysis, window lengths ranging from 15 to 50 ms have been profitably used. The length must at least be longer than one glottal period to get a reasonable spectrum

[4] The discussion in the previous section clarifies how this systematic difference is not really an error.

estimate, but be short enough to ensure that the speech formants do not vary "all that much" within one window. The LP model will in any event compute only one set of coefficients for one window, and this means that if the formants change from the beginning to the end of the window, the estimated formants will end up being some sort of compromise. This is then a strong deterrent against choosing a much longer window when speech is examined, but as usual there is a countering argument that the formant estimates are generally more accurate when longer windows are used. The analyses of the synthesized vowels that are shown are usually computed from a longer window of around 40–50 ms, but there is no danger in this since these vowels have been constructed so that their formants do not change.

As for the choice of window function, the discussion of windows in Chap. 2 suggests the same advice for linear prediction: try a Gaussian or Kaiser window for best results. The analyses presented here from the Matlab code use a Kaiser window and Praat software is currently limited to a Gaussian, but in the literature, tradition has more often led to the use of the suboptimal Hamming or Hann windows. In truth, the length of the window has a much greater effect on the results of linear prediction than the choice of tapering function. Figure 7.3 compares 10-pole LP spectra computed from the synthesized vowel [i]; one spectrum uses a 50 ms Kaiser window, while another uses a 16 ms Hamming window (adopted for a large study of American English vowels [19]). Observe that the formant values estimated with the shorter Hamming window are slightly less accurate, an effect largely due to the decreased window length. The discussion of windows here



[i] Known (Hz)	306 2241 2500 3500
Peaks (50 ms)	317 2217 2485 3291
Peaks (16 ms)	322 2207 2485 3276

Fig. 7.3 Ten-pole LP spectra of synthesized [i], computed using 50 ms Kaiser window (*solid line*) and 16 ms Hamming window (*dotted*)

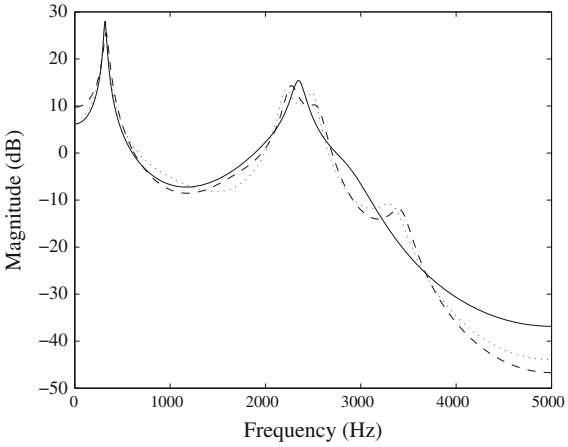
pertains only to a pitch-asynchronous approach; the special methods required for closed phase LP analysis will be discussed separately below.

7.4.3 Filter Order and Pre-Emphasis

The number of poles that are expected to provide an adequate smoothed spectrum showing the "actual" vocal tract filter depends first of all on the total frequency range that is available, which is in turn determined directly by the sampling rate. Since the vocal tract filter is expected from theoretical considerations (e.g. [13]) to have a clear resonant structure only in the range below 5.5 kHz or so, it is common to use a digital speech signal that is sampled at around 10 or 11 kHz. For signals that have been recorded at a higher rate, this is commonly achieved by a *resampling* procedure which can be performed by available software such as Praat. Resampling (also called *downsampling*) is then a frequently necessary first step when performing linear predictive analysis.

After resampling, the number of expected resonances in the spectrum must be theoretically considered. For our synthesized vowels here it is precisely 4, but real speech often provides 5 or even 6 formants together with other possible resonances (e.g. the voice bar, subglottal resonances, nasal resonances). If just 4 formants are expected, this dictates that 8 LP coefficients (poles) must be in the model. Moreover, the original argument discussed earlier which determined that the speech filter could be approximated by a linear all-pole filter included some additional transfer functions besides the vocal tract formants. There were also simpler filters representing the glottal system and the lip radiation system. It is generally argued (e.g. [24]) that something like two additional poles should be added to the linear prediction model to account for these systems, although the effects of these additional elements are hoped to be more like a broad shaping point in the spectrum rather than a sharp peak, since neither the glottal model nor the lip radiation model is supposed to introduce additional peaks. If these further considerations are on the right track, we should get a better smooth spectrum of our synthetic vowel from a linear prediction model using 10 or 11 coefficients than is obtained using 8 coefficients, and this is indeed the case. Figure 7.4 shows three LP spectra for the vowel [i] together with a table of resonance values computed from the poles of the transfer function. The 8-pole analysis fails to resolve the closely spaced F_2 and F_3, and so even the root-solving routine only finds two clear peaks. The other two resonances have bandwidths greater than 500 Hz, making them extremely poor candidates for actual peak values. The 10-pole spectrum yields the most accurate estimates of F_1 and F_3, yet the 11-pole spectrum yields the most accurate estimates of F_2 and F_4.

When LP analysis is applied in speech science, a working principle is often applied to the effect that if one LP spectrum fails to resolve presumed closely-spaced formants, simply redo the procedure adding more poles until the formants are resolved (e.g. Hillenbrand et al. [19]). This generally works up to a point, but



Spectral resonances (Hz, from poles) for different filter orders:

Known	306	2241	2500	3500
8 pole	319	2348	–	–
10 pole	318	2207	2499	3343
11 pole	330	2265	2555	3425

Fig. 7.4 LP (Burg) spectra for synthesized vowel [i] using 8 (*solid line*), 10 (*dotted*) and 11 poles (*dashed*)

care must be taken with the number of poles in a deliberately "overfit" filter model. What is the reason a linear prediction model with more poles can get worse rather than better? In accord with the earlier discussion of modeling, a linear prediction spectrum can be expected to estimate the vocal tract filter well only under the condition that it is an optimal model of the physical facts. If we add more poles willy–nilly, the resulting model eventually loses all correspondence with reality, and the spectrum will revert to a simple smoothing of the Fourier power spectrum. For a reasonable number of poles, such a smoothing still corresponds to a resonance spectrum fairly well, but the limit of this process for a large number of poles just *is* the Fourier power spectrum of the window. In a pitch-asynchronous analysis, where the window is normally quite long, the Fourier spectrum shows all of the harmonics.

A signal which is to be modeled using LP analysis may optionally be subjected to pre-emphasis (recall from Chap. 4) prior to analysis. From the beginning there has been a theoretical argument in favor of this procedure [24], though if pre-emphasis is performed the resulting spectrum must be de-emphasized at the conclusion of the procedure to avoid skewing it. This procedure has not been provided with any Matlab code here, but it is possible to do it in Praat. Figure 7.5 shows the 10-pole LP spectra for synthesized vowels [ae, i] which have proven to be quite accurate, together with 10-pole spectra computed using a pre-emphasis algorithm (followed by de-emphasis). The decreased accuracy of the peak locations is easy to observe in the pre-emphasized spectra. In my experience I have not gained any advantage from pre-emphasis as part of LP analysis, theoretical arguments in its favor notwithstanding.

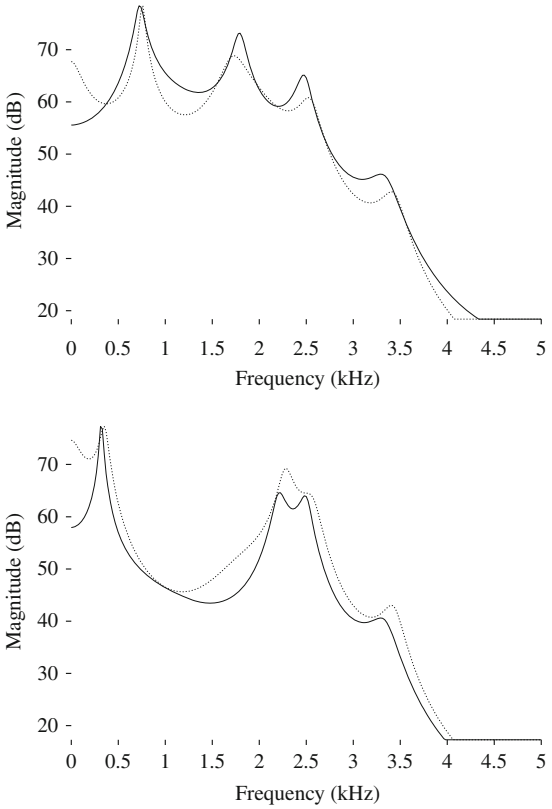


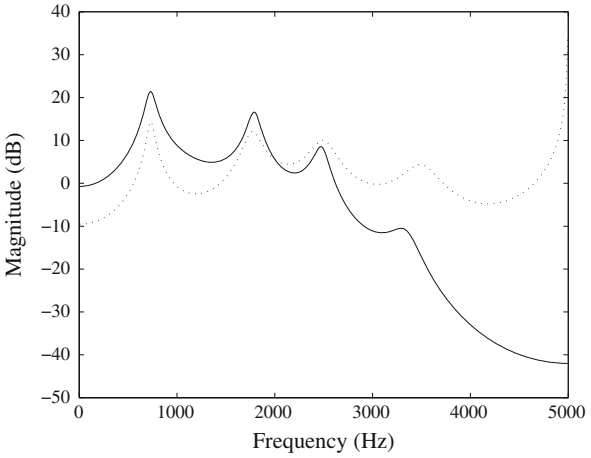
Fig. 7.5 10-pole LP spectra for synthesized vowels [æ] (*upper panel*) and [i]. *Dotted line* spectra use a standard 6 dB/octave pre-emphasis procedure, to no good end

7.4.4 Pitch-Asynchronous Versus Closed Phase

The closed phase procedure for LP speech analysis was argued early on to be potentially superior [24]. To carry it out, one must use an untapered signal window (rectangular window) which begins at the moment of glottal closure, and which extends through to nearly the end of one glottal cycle. For mathematical reasons which will not be explained here, only the covariance algorithm for computing the LP coefficients can be used with such a short analysis window. The number of poles in the model should be the same as would be used for a pitch-asynchronous analysis, so I still use 10 poles for the synthesized vowels. Figure 7.6 realizes the potential of this method, providing the most accurate estimates yet shown of the formants in a synthesized [æ] vowel. The roots of the resulting filter polynomial are exactly equal to the formants which were used in creating the vowel.[5]

Figure 7.7 provides another demonstration of the "perfection" of this method, at least for synthesized vowels. When applied to a vowel having closely spaced

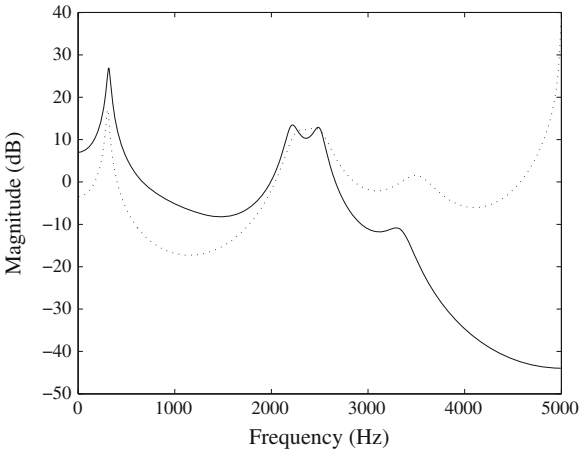
[5] A spurious peak in the spectrum has been introduced at 5 kHz, which is perhaps a result of resampling to 10 kHz.



Comparing peaks computed from poles of the LP filter by the closed phase method with the known production formants:

Known	731 1768 2500 3500
Closed poles	731 1768 2500 3500 5000

Fig. 7.6 Synthesized [æ] spectra comparing the closed phase covariance method (*dotted*) with the pitch-asynchronous Burg method, both using ten poles



Comparing spectral peaks from poles with the known formants:

Known	306 2241 2500 3500
Closed poles	306 2242 2501 3508 5000

Fig. 7.7 Synthesized [i] spectra comparing the closed phase covariance method (*dotted*) with the pitch-asynchronous Burg method, both using ten poles

formants such as synthesized [i], however, these can appear merged in the computed spectrum even when the calculated roots are correct, making it impossible to use a peak-picking routine to uncover them. Clearly the closed phase method is superior for synthesized vowels, although it must be carried out with care. Another advantage is that it is possible to increase the number of poles in the LP model without worrying about reverting to the harmonic spectrum—there is no harmonic spectrum from a single glottal cycle.

7.5 Application to Real Speech

The analysis of real speech introduces numerous challenges for an investigator seeking to "measure" formants using LP procedures. Many do not involve difficulties with linear prediction itself, but derive instead from the numerous ways in which real speech distinguishes itself from the output of a formant synthesizer. Briefly:

- While vowels are synthesized with Praat (or similar systems) in perfect adherence to the source-filter model of speech, real speech departs significantly from this model due to its aeroacoustic aspects and changing impedance at the glottis.
- The only resonances in a synthesized vowel are the formants; in real speech one may also find a voice bar, a nasal resonance or two, and two or three subglottal resonances.
- The values of all resonances in real speech generally change within the course of a single glottal cycle.
- Any amount of breathiness in the voice is the result of pure aeroacoustic noise generation from the glottal outflow, which is expected to hinder the linear prediction model fit to the vowel.
- In the case of real speech, there is no way of knowing what the production formants are in fact, and hence we are prevented from so easily evaluating the accuracy of the formants found by LP analysis.

Before performing LP analysis on a real speech signal, it should be resampled to 10 or 11 kHz; this is the simplest way to ensure that all resonance peaks found in the LP model could possibly correspond to actual physical resonances of the vocal system, since no such resonances are generally expected above at most 5.5 kHz. Most examples here are sampled at 10 kHz, so 5 kHz becomes the cutoff frequency of the power spectrum; 11 kHz is a better sampling rate for female voices, since the formants are expected to be somewhat higher. Most voices exhibit five formants in the range up to 5 kHz; there is also likely to be a voice bar, and there is the possibility of additional resonances which are not formants of the oral tract. This dictates that an LP model should have at least 14 poles, including an extra pair for the glottal and lip radiation systems. This was indeed the number of poles used by Hillenbrand et al. [19] in their study of American English vowels, although they would sometimes increase the number of poles in search of better

resolution. Pre-emphasis was demonstrated in the preceding section, but was not shown to be of much use.

It was already decided above that peaks in the LP model spectrum should be determined by solving the roots of the polynomial and applying a bandwidth threshold to weed out the resonance frequencies which are not peaks. This will provide superior resolution over computing the LP power spectrum and picking its peaks directly. Beyond this, we also need to decide whether to use a pitch-asynchronous analysis (this seems to be the most common method in the literature for LP analysis of real speech) or a closed-phase analysis. The results presented above for synthesized vowels suggest that a closed-phase analysis is the most promising.

While this is not the place to present an exhaustive study, it is not possible to prescribe a specific linear prediction methodology without at least checking a few possibilities for relative accuracy. Table 7.3 presents values measured for the voice bar and first three formants in a natural English vowel [æ] spoken by the author, and resampled to 10 kHz. Recall that the LP peak-picked values are directly comparable to those measured from a reassigned spectrogram, while the LP root-derived values are expected to be somewhat different from the spectral peaks. In spite of this, it appears that the overall best LP method in this case was the 14-pole closed phase root computation.

Since we have settled on root-solving over LP spectrum peak-picking to find resonance values for a number of reasons I will stop including picked peak values in the further comparisons below. A considerable number of the author's own English vowels are used to compare some different LP parameter settings which use root-solving in Table 7.4. Estimates labeled "Burg" use a pitch-asynchronous scheme with a Kaiser window of approximately 50 ms, while estimates labeled "closed" use the covariance algorithm with a short rectangular analysis window containing one glottal cycle. Only root-derived resonance values whose computed bandwidths are less than 300 Hz are reported in the table; some resonances whose presence is expected are not discovered by some analyses, or do not have a bandwidth meeting the condition.

Let me discuss some of the discrepancies and difficulties which are apparent in Table 7.4. A number of vowels (notably [i, ɪ, e, u]) have F_1 measured considerably higher using the reassigned spectrogram. This could be due to two different

Table 7.3 Comparing resonance values for English vowel [æ] found from several LP methods; the number of poles in each LP analysis is given

	Voice bar	F_1	F_2	F_3 (Hz)
Reassigned	205	667	1,917	2,516
12 Burg peaks	132	693	1,880	2,456
12 Burg roots	186	707	1,879	2,479
12 closed peaks	98	649	1,919	2,568
14 Burg peaks	171	689	1,934	2,559
14 Burg roots	191	694	1,940	2,600
14 closed roots	126	661	1,915	2,561

Values measured from reassigned spectrogram provide a benchmark

Table 7.4 Resonance values for nine natural English vowel tokens derived from pitch-asynchronous and closed phase LP analysis are compared with values measured from reassigned spectrograms

	Voice bar	F_1	F_2	F_3 (Hz)
[i] Reassigned	195	380	2363	3220
14 Burg	158	354	2359	3390
14 closed	155	305	2352	3286
[ɪ] Reassigned	227	479	2290	2771
14 Burg	267	–	2294	2752
16 Burg	250	345	2310	2767
14 closed	153	295	2306	2793
[e] Reassigned	197	412	2527	2988
14 Burg	245	303	2528	2864
16 Burg	231	320	2520	2845
14 closed	186	313	2518	2909
[ɛ] Reassigned	221	471	2164	2782
14 Burg	223	477	2164	2832
14 closed	175	476	2162	2820
[æ] Reassigned	205	667	1917	2516
14 Burg	191	694	1940	2600
14 closed	126	661	1915	2561
[ɔ] Reassigned	175	505	712	2655
14 Burg	228	533	737	2706
14 closed	66	487	729	2685
[o] Reassigned	184	364	705	2428
14 Burg	–	363	728	2428
14 closed	–	378	725	2434
[ʊ] Reassigned	291	350	879	2387
14 Burg	–	332	870	2437
14 closed	–	349	869	2405
[u] Reassigned	130	350	715	3034
14 Burg	204	–	704	3001
18 Burg[a]	156	318	715	3008
18 closed[a]	–	–	710	3016

The number of poles in each LP analysis is given

[a] There is a spurious formant in each of these analyses having a bandwidth that meets the condition

factors. Firstly, due to the dynamics of real speech formants within a single glottal cycle, it is difficult to decide what value to report from the reassigned spectrogram. My standard procedure here is to report the value of F_1 that typically stabilizes 1 or 2 ms after the closure impulse. A second factor, however, may be that LP analysis tends to report a value of F_1 that is too low (as was noted by Di Benedetto [11]). While the specific reasons for this tendency are not certain, it may be due to proximity of F_1 with the voice bar, which would explain why the problem is chiefly observed in high vowels (having lower F_1 values).

Another thing to note from Table 7.4 is the wide range of formant values discovered using different LP methods for the same vowel—some of the F_1 values

vary by 10% or more. Let me finally point out that in a few vowels, LP analyses of reasonable filter order (i.e. 14) were unable to detect or separate expected formants (e.g. the case of [u]). In some cases, when the number of poles is increased in order to resolve resonances, extra resonances can be introduced which may or may not reflect anything physically real (such as a subglottal resonance). These problems should inspire considerable skepticism concerning the degree of accuracy that can reasonably be ascribed to formant values measured using linear prediction techniques.

To test performance on a female voice, six Finnish long vowels from a female native speaker have their formant estimates presented in Table 7.5. These tokens were resampled to 11.025 kHz for the LP analyses, since the resulting higher frequency range is more appropriate for a female subject. Compared to the formant and voice bar values measured from reassigned spectrograms, the LP estimates betray a number of problems. For one thing, the pitch-asynchronous and closed phase estimates do not agree with each other very well at all. It appears that in the case of this speaker, the closed phase LP methodology has not been so successful. One possible reason for this is that the female phonation typically has a "softer" glottal closing gesture than that produced by males. Beyond the fact that the closed phase values are not very reliable, numerous formants and also the voice bar were not discovered at all in several cases of LP analysis. When F_1 was estimated, its value is lower than that measured from reassigned spectrograms.

Table 7.5 Resonance values for six Finnish long vowel tokens from a female native speaker

	Voice bar	F_1	F_2	F_3 (Hz)
[a:] Reassigned	280	725	1,544	2,683
14 Burg	–	715	1,576	2,710
14 closed	–	616	1,528	2,752
[æ:] Reassigned	247	771	1,808	2,704
14 Burg	–	756	1,816	2,753
14 closed	–	736	1,809	2,770
[e:] Reassigned	271	621	2,403	2,860
14 Burg	367	589	2,308	2,881
14 closed	119	562	2,403	2,757
[i:] Reassigned	271	473	2,452	2,952
14 Burg	232	522	2,650	–
14 closed	235	–	–	2,856
[y:] Reassigned	326	449	1,753	2,489
16 Burg	–	370	1,756	2,472
18 Burg	333	398	1,776	2,469
14 closed	321	–	1,808	2,538
[o:] Reassigned	262	621	1,403	2,703
14 Burg	291	553	–	2,667
16 Burg	–	566	1,433	2,638
14 closed	276	501	1,277	2,722

LP formant estimates (by root-solving) from pitch-asynchronous and closed phase analyses are compared to values measured using reassigned spectrograms

7.6 Autoregressive Moving Average Modeling

Looking back at Eq. 7.6 defining a rational LTI system in general, we should now recall that the linear prediction (autoregressive) model assumes that the system has no zeros, so that the numerator polynomial in the equation is replaced by a constant known as the gain factor (Eq. 7.10). Such an all-pole model is perfect for modeling a linear system involving a resonant filter with no zeros. Now, while the speech production system can often be profitably modeled as such a system, circumstances arise in which the speech system is theoretically thought to involve some zeros in the system function. In such a circumstance, no all-pole model is going to be terribly good [17], so when linear prediction analysis does not work very well for speech spectrum estimation, one possible cause is the presence of zeros in the system. In this section I will briefly show how to estimate speech spectra by invoking a *pole-zero* model, in which both the numerator and denominator polynomials in Eq. 7.6 are nontrivial.

7.6.1 A Little ARMA Theory

Looking back at Eq. 7.7 defining an autoregressive process, output series $y(n)$ is there represented as a linear combination of its own past values together with the current value of a random process. In the math box below it is shown how an output series $y(n)$ can alternatively be represented as a linear combination of current and past values of a supposed random process, rather than of its own past. Such a process is known as a *moving average* (*MA*) *process*. In the language of LTI systems, this is equivalent to a system with a rational function (Eq. 7.6) in which only the numerator is nontrivial; the system has only zeros, instead of only poles. Combining the two types of process into a single representation yields a "mixed" autoregressive moving average (ARMA) process, which was originally called an autoregressive process with moving average errors. This corresponds to an LTI system with a rational system function having poles and zeros—a pole-zero model.

The following is Priestley's [29] equation for a moving average process:

$$y(n) = \varepsilon_n + b_1\varepsilon_{n-1} + \cdots + b_M\varepsilon_{n-M} \tag{7.15}$$

in which the set $\{b_1, \ldots, b_M\}$ are the MA coefficients. This can be combined with Eq. 7.7 to yield:

$$y(n) = \varepsilon_n - a_1 y(n-1) - \cdots - a_N y(n-N) + b_1\varepsilon_{n-1} + \cdots + b_M\varepsilon_{n-M} \tag{7.16}$$

as an equation for an ARMA process, where I have changed the (arbitrary) signs on the coefficients $\{a_k\}$ to match the treatment in [17]. The LTI system

function for such a process has been given earlier, in Eq. 7.6; it involves both poles and zeros. Just as it was shown earlier how the AR (linear prediction) coefficients $\{a_k\}$ are also the filter coefficients in the denominator of the system function, it can also be shown that the MA coefficients $\{b_\ell\}$ are precisely the filter coefficients in the numerator of the system function.

7.6.2 ARMA Computation

While the computation of the predictor (i.e. pole series) coefficients in the service of LP analysis is basically a linear algebra problem, the computation of both the pole and zero series coefficients together is a nonlinear problem, and is therefore considerably more problematic [33]. As with the LP method, I do not wish to go into the details of how the coefficients can be computed. Suffice it to say that a number of methods for computing the ARMA model coefficients have been proposed over the years, and unfortunately there appears to be very poor agreement among the results of the methods. This is quite in contrast to the situation for linear prediction, where most of the disparate computational algorithms amount to the same solution in practice. Since we cannot here embark on an effort to validate methods of estimating ARMA model parameters, I have provided two methods which, while they give different results, both appear to be sound, are backed up by refereed publications, and have the advantage of Matlab code already freely available.

One ARMA computation method included here is an extension of the cumulant-based LP computation method due to Giannakis and Mendel [16] that was mentioned in an earlier section. The ARMA coefficients are calculated by first obtaining an autoregressive solution for the denominator coefficients $\{a_k\}$ (Eq. 7.6) using the earlier described technique, and then using this solution to write a so-called "residual time series" that is then solved as a moving average problem to obtain the numerator coefficients $\{b_\ell\}$. The other included method called ARMASA is documented by Broersen [7] (whose implementation is used), but it is originally due to Durbin [12]. It is similar to the first method in that it first solves an autoregressive sequence and then uses the result to solve a moving-average problem, but the two techniques clearly differ in their details.

Whichever ARMA algorithm is selected, once both coefficient sequences have been obtained, it is then possible to proceed similarly to the case of LP analysis, where only the sequence $\{a_k\}$ was used. In particular, the poles of the system function in the z-domain are obtained as the roots of the (Laurent) polynomial in the denominator just as before, while the zeros of the system function are obtained analogously from the roots of the numerator polynomial. The frequency and bandwidth of each resonance determined by the poles can still be computed using Eqs. 7.13 and 7.14; Eq. 7.13 can be also be applied to compute the frequency of each spectral zero determined by the system zeros. The power spectrum itself can be computed using an extension of Eq. 7.11 (see box).

The theoretical equation for the power spectrum of an ARMA process can be found in a number of sources [6, 29], but it is more difficult (and more useful) to find an equation in digital form suitable for implementation. The power spectrum $\left|\hat{S}(\omega)\right|^2$ of an ARMA model can be computed from the two coefficient sequences using discrete Fourier transforms in the following way [28]:

$$\left|\hat{S}(\omega)\right|^2 = \frac{\sigma^2 \left|1 - \sum_{\ell=1}^{M} b_\ell e^{-i\ell\omega T_s}\right|^2}{\Delta f \left|1 - \sum_{k=1}^{N} a_k e^{-ik\omega T_s}\right|^2}, \tag{7.17}$$

in which T_s is the sampling period of the signal, Δf is the frequency bin width, and σ^2 is the variance of the ARMA process (analogous to the squared LP gain factor). I had trouble finding a good method for computing this variance. Broersen's ARMASA Matlab routines include a method which he shows how to use, but when tested it appeared to give incorrect results. The notion that the squared LP gain (which does not take into account the MA part of the ARMA model) might serve as a suitable estimate of the ARMA variance is implied in [35], so this is how it is estimated in the plots shown.

7.6.3 Applications to Speech

It has been established a number of times (e.g. [17]) that using an LP model to find the spectral peaks of a process containing zeros can lead to serious errors in the location of the peaks. It is also well-known that speech sounds can contain spectral zeros, particularly when nasalization of any kind is involved. It therefore behooves the would-be speech analyst hoping to rely on a parametric spectral estimate to at least try an ARMA model in many circumstances. The main reason is not to find the locations of spectral zeros, since this is frequently not valuable information about a speech spectrum, but rather to simply get a better estimate of the peak locations. One study, for example, demonstrated the superiority of ARMA modeling applied to synthesized nasals involving 8 poles and 5 zeros [27].

First of all, let us see how ARMA modeling compares with linear prediction. I have computed the ARMA parameters using both supplied methods for two of the Finnish vowels analyzed above, which were not well-represented by their LP models. Table 7.6 compares the resonance values found earlier for these vowels using LP analysis, with the values found using the ARMA "residual time series" (rts) method and the Durbin method (supplied with the software name "ARMASA" by Broersen). Both methods are applied using 14 poles and 2 zeros, under the assumption that there are few, if any, zeros expected in an oral vowel, but also to test the idea that modeling a couple of zeros might be more accurate than no zeros. Power spectra from each of the ARMA models for [o:] are plotted in Fig. 7.8, where it is plain to see how different the models are (and that they are plotted with a somewhat incorrect power gain,

Table 7.6 Formants of two Finnish vowels compared from reassigned spectrograms, Burg LP analysis, and two ARMA estimation methods (cf. Fig. 7.8)

	Voice bar	F_1	F_2	F_3 (Hz)
[a:] Reassigned	280	725	1,544	2,683
14 Burg	–	715	1,576	2,710
(14, 2) ARMArts	–	744	1,544	2,677
(14, 2) ARMASA	–	683	1,539	2,702
[o:] Reassigned	262	621	1,403	2,703
14 Burg	291	553	–	2,667
(14, 2) ARMArts	243	535	738	–
(14, 2)ARMASA	300	556	–	2,695

Table 7.7 Resonance values for English nasals as measured using a reassigned spectrogram, Burg LP analysis, and cumulant-based ARMA modeling

[m] Reassigned	172	371	893	1,146	1,430	1,752	2,020	2,695	3,393	3,900
22 Burg	169	257	–	1011	1,398	1,756	2,138	2,689	3,404	4,123
(22, 5) ARMArts	149	338	–	–	1,350	–	2,043	2,688	3,426	4,141
(22, 5) ARMASA	134	248	990	–	1,406	–	2,048	2,671	3,428	3,978
[n] Reassigned	222	339	1,590–1,482	2,680	3,944					
14 Burg	195	–	1491	2,681	3,919					
(14, 2) ARMArts	173	365	1491	–	3,857					

Table 7.8 Degema advanced versus retracted tongue root low vowels following [m], as measured using reassigned spectrograms, LP analysis, and (24, 2) ARMA models

[m a̟] Reassigned	186	983–769	1429	2017	2441
24 Burg	185	821	1408	1954	2615
(24, 2) ARMArts	158	822	1415	1958	2607
(24, 2) ARMASA	172	864	1434	1945	2607
[m a̠] Reassigned	210	538	933	1445	2334
24 Burg	203	615	999	1422	2335
(24, 2) ARMArts	163	576	973	1421	2337
(24, 2) ARMASA	198	592	989	1418	2337

which does not affect the pole locations). It turns out that neither modeling method is obviously a great improvement over LP modeling for these vowels.

Given the ARMA modeling success with synthesized nasals cited above, we should examine some nasals. As an example, the nasals [m, n] recorded in the English words *clam, clan* have been analyzed to find out the resonance spectra. It is notable that there are many more resonances for the bilabial nasal, so the number of poles (and zeros) in the models have been adjusted to meet expectations. The LP method and the ARMA models are about equally mediocre as far as the accuracy of the located peaks, and all of these parametrics have missed some resonance peaks (Table 7.7). The ARMA model for [m] was set to have 22 poles and 5 zeros while the model for [n] was set to have 14 poles and 2 zeros, but the number of zeros is really just a guess based upon the number of resonances, and the location of the poles is not improved from the pure LP analysis.

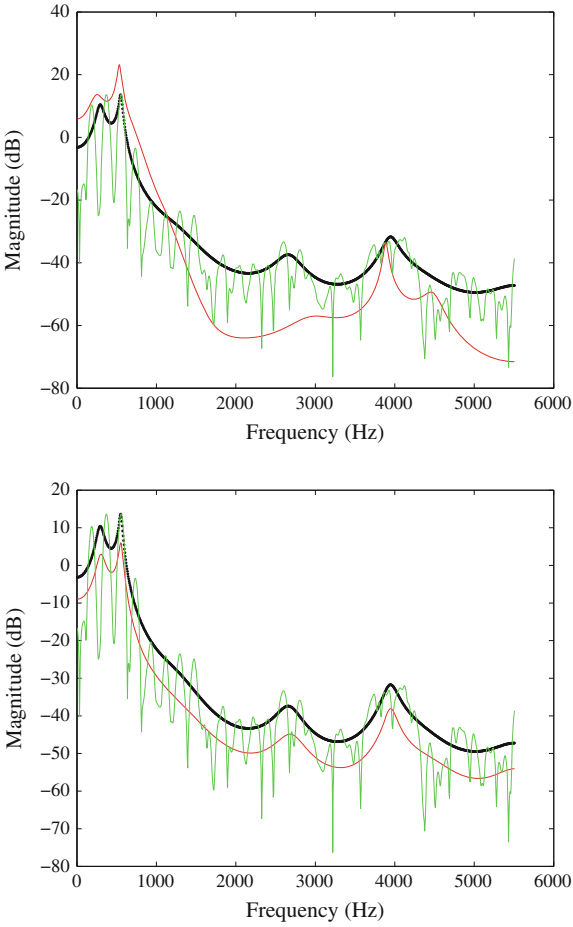


Fig. 7.8 Spectra of Finnish [o:] comparing ARMA estimation method of [16] (*upper panel*) to the method of [12] (*lower panel*). Both images show the Burg LP model spectrum as a *heavy dotted line*, together with the Fourier power spectrum in *green*/*gray* (showing harmonics)

For another example, resonances were measured in the purportedly distinct low vowels of Degema (discussed in the previous chapter) in syllables initiated with a nasal. This environment provides the possibility of slight nasalization, and there is also reason to suspect a more convoluted resonance cavity during vowels produced using tongue root manipulation. Thus, there is the prospect that an ARMA model could be better than LP for such vowels. 24-pole models were tried, in view of the large number of observed resonances below 5 kHz; there were 5 resonances easily measured below 2,500 Hz, and the data here focusses on these only (Table 7.8). In this case, some moderate success with the ARMA models is observed. Particularly in the case of the retracted tongue root [a̘], the ARMA models' added zeros appear to have improved the accuracy of some peak locations. It should also be kept in mind that, because these procedures inherently use one set of peaks to represent a highly dynamic process as resonance properties of a single span of time on the order of 45 ms, there is not any guarantee that even a "perfect" model would

return the values reported from the reassigned spectrogram, since the latter analysis uses much shorter analysis windows.

I hope that readers feel ready to try some ARMA modeling on sounds that have proven problematic for LP analysis. One obvious disadvantage, however, is the need to guess at the number of zeros for the model, in addition to the poles. Mitigating this difficulty is the relatively low sensitivity of the pole locations to the zero locations in these solution methods, as can be verified by increasing the number of zeros in the modeling of the Degema vowels (Table 7.8).

7.7 Appendix: Praat and Matlab Techniques

7.7.1 Praat Functions

When a sound object in the list is selected, you will see a button for "Formants & LPC" which accesses all LP analysis features of Praat. Depressing this button brings a pop-up menu with the LP options. Praat can compute a raw LPC object, using any one of four algorithms which are all essentially equivalent for pitch-asynchronous analysis. Praat is only useful for pitch-asynchronous LP analysis, since it forces the use of a Gaussian taper on the analysis window. Also on the menu is a "To Formant" function, which further automates some of the steps in LP formant estimation.

It may be necessary to downsample the sound object before using an LPC function. As mentioned above, for a good LP model of sonorant speech sounds, the sampling rate should be set to 10 kHz for males and around 11 kHz for female speakers. Resampling is available in Praat when a sound object is selected, as one of the functions accessible by the "Convert" button. Upon selecting one of the functions "To LPC," a dialog opens which allows you to set the prediction order (number of poles), the analysis window length (automatically doubled with a Gaussian taper applied), the time step between successive LP coefficient computations, and the frequency at which pre-emphasis begins. Setting the latter to some frequency higher than the Nyquist frequency will turn off pre-emphasis (e.g. 6,000 Hz for a sound sampled at 10 kHz).

The result of the above LPC function will be an LPC object in the list, which represents filter coefficients as a function of time, stored in successive frames with a constant sampling period [4]. It is important to remember the actual window length (double what was entered above), the time step, and pre-emphasis frequency in order to be able to work with an LPC object effectively. The LPC object can be acted upon in a number of ways, not all of which I can address here. The actual coefficients can be accessed for each analysis window using the "Inspect" button below the object list. The function "To Spectrum" will compute a spectrum object from the LPC object; it is important to enter the de-emphasis frequency equal to whatever pre-emphasis frequency was entered before. The time point of

the analysis window within the LPC object that is desired for the spectral slice must also be entered. Leaving the value set to zero will choose the first analysis frame. Once a spectrum object has been created, it can be viewed using "Edit" or drawn as a graph in the picture area, just as with a Fourier power spectrum object.

Starting from an LPC object, there is also the function "To Formant," which is useful for getting the frequency and bandwidth values of the various resonances determined by the LP coefficients. The function yields a formant object added to the list, which provides frequency and bandwidth information (by root solving) about discovered formants for each analysis frame in the corresponding LPC object. The formant values can be read for each successive analysis frame by first computing the Formant object and then using the "Inspect" button. The formant function tries to heuristically eliminate certain resonances which are not peaks; if every resonance and its bandwidth is desired (which will include any shaping resonances that are not peaks), use "To Formant (keep all)" instead.

Starting from a sound object which has not been downsampled, it is also possible to bypass some of the details of LP analysis by selecting "To Formant" from the "Formants & LPC" functions. This function brings up a dialog requiring you to enter the number of formants sought or expected; the number entered will simply be doubled to determine the number of poles for the analysis. From the discussion in this chapter, it can be gleaned that this is not generally the optimal number of poles, since something like two additional poles should be included in order to model the lip radiation and glottal production filters. Therefore, you might try increasing the number of formants you enter by one, to get a better LP model. The dialog also asks for the maximum frequency of the formants, which should be 5,500 Hz for a female and 5,000 Hz for a male. The function will then automatically resample the sound before computing the LP coefficients and formant properties therefrom. Once again, the analysis window length (automatically doubled by Praat) and the desired starting frequency for any pre-emphasis should also be entered in this dialog. The result of this function will be a formant object, as described above.

7.7.2 Matlab Functions

Readers who wish to use the supplied mfiles for this chapter will need to have installed the Signal Processing Toolbox for Matlab, which costs extra. Some of the routines (in particular, all of the ARMA routines) also rely on two free toolboxes, Higher Order Spectral Analysis and ARMASA, both available through the Matlab Central file exchange service. Rather than using the functions in these toolboxes directly, I have written "wrapper" functions which in turn call functions from the toolboxes, and in this way the output and plots precisely suit the purposes to which they have been put here.

The three functions to be described are intended for LP or ARMA modeling of a "single slice" of a speech signal. The signal provided to the functions for pitch-

asynchronous analysis should be around 20–60 ms in length, not windowed. Praat is useful for cutting out the desired slice of a speech sound. For closed-phase analysis, the provided signal should instead comprise a single glottal cycle from beginning to end. Praat can also be used to downsample the signal in preparation for the parametric procedures provided here. For modeling of the vocal tract resonances in sonorant speech sounds, male speakers should be resampled at 10 kHz, and females at around 11 kHz.

The main function for computing a linear prediction model from a signal and plotting its power spectrum is `lpcspectrum.m`, which is called using the following template:

$$[\texttt{peaks}, \texttt{amps}] = \texttt{lpcspectrum}(\texttt{signal}, \texttt{order}, \texttt{Fs}, \texttt{taper}, \texttt{alg}, \texttt{linestyle})$$

The argument `signal` is a vectorized signal that need not have had a tapered window applied, `order` is the number of LP coefficients in the model, and `Fs` is the signal sampling rate. The argument `taper` should be entered as `'yes'` or `'no'` (including the single quotes, per Matlab syntax). The argument `alg` should be entered as one of `'autocor'`, `'cov'`, `'burg'`, or `'rcest'`. These values will compute the LP model using, respectively, the autocorrelation method, the covariance method, the Burg method, and the third-order cumulant-based method of [16]. For computing an LP model using the closed phase of a glottal cycle, set `alg` to covariance with no taper applied. All pitch-asynchronous models should have tapering applied, unless the provided signal was already tapered by other software. The argument `linestyle` is useful for overlaying plots with different lines; it should be one of the Matlab LineStyle property specifiers, such as `'-'` for a plain line or `':'` for a dotted line.

When lpcspectrum is called, a set of LP coefficients is computed using the specified algorithm, the resulting model power spectrum is plotted, and a number of peak frequencies are automatically picked using a routine from the HOSA toolbox and reported in the Matlab command window. The desired number of peaks to be sought is set in the code, as the value of `npeaks`, which equals 6 by default. The discovered peaks and their power amplitudes can also be returned using the so named output variables in the template, which are optional. At the conclusion of the routine, the user is asked whether to overlay the Fourier power spectrum, where the answer 'h' will not overlay but will "hold" the plot so that future runs through the routine will overlay the next LP spectrum. This is useful for comparing spectra of LP models with differing numbers of poles, for instance.

The function `lpcroots.m` is used for computing an LP model and getting the resonances by solving the polynomial roots, instead of peak-picking a spectrum. It is invoked using the following template:

$$[\texttt{output}, \texttt{amps}] = \texttt{lpcroots}(\texttt{signal}, \texttt{order}, \texttt{Fs}, \texttt{taper}, \texttt{alg})$$

where the arguments have the same meaning as with the lpcspectrum routine above. Running the function automatically prints all solved roots (positive and negative) as frequency values together with bandwidths of the resonances. The

negative-frequency roots simply repeat the information contained in the positive roots. If desired, a two-column matrix of resonance frequencies and bandwidths can be output using the variable of that name. The variable named `amps` will return a vector of the power amplitudes of the resonances; as usual the routine can be called without using output variables.

The main function for ARMA modeling provided is `armaspectrum.m`, and it combines the features of the above LP routines. The function is invoked using the following template:

$$[\texttt{formants}, \texttt{amps}, \texttt{zeros}] = \texttt{armaspectrum}(\texttt{signal}, \texttt{arorder}, \texttt{maorder}, \texttt{Fs}, \texttt{alg})$$

in which the arguments have the same meaning as in the LP routines above, except that now two orders (for the autoregressive part and the moving average part of the model) have to be specified, giving the number of poles and zeros respectively. The argument `alg` should be either `'rts'` (for the HOSA toolbox algorithm) or `'ARMASA'`. Running the function produces a plot of the model power spectrum, although the gain has been estimated using the LP gain factor which is known to be incorrect. The frequencies of all poles (with their bandwidths) and zeros found by root-solving are automatically printed to the command window. The output variable `formants` returns a two-column matrix of resonance frequencies and bandwidths, `amps` returns a vector of the resonance amplitudes, and `zeros` returns a vector of the frequencies of spectral zeros.

References

1. R.L. Allen, D.W. Mills, *Signal Analysis: Time, Frequency, Scale, and Structure* (Wiley, New York, 2004)
2. N. Andersen, On the calculation of filter coefficients for maximum entropy spectral analysis. Geophysics. **39**, 69–72 (1974)
3. B.S. Atal, S.L. Hanauer, Speech analysis and synthesis by linear prediction of the speech wave. J. Acoust. Soc. Am. **50**(2 part 2), 637–655 (1971)
4. P. Boersma, D. Weenink, Praat: Doing phonetics by computer. Comp. Softw. (2009)
5. A. van den Bos, Alternative interpretation of maximum entropy spectral analysis. IEEE Trans. Inform. Theory. **17**, 493–494 (1971)
6. G.E.P. Box, G.M. Jenkins, *Time Series Analysis: Forecasting and Control*, rev. edn. (Holden-Day, San Francisco, 1976)
7. P.M.T. Broersen, *Automatic Autocorrelation and Spectral Analysis* (Springer, Berlin, 2006)
8. H.P. Bucker, Comparison of FFT and Prony algorithms for bearing estimation of narrow-band signals in a realistic ocean environment. J. Acoust. Soc. Am. **61**(3), 756–762 (1977)
9. J.P. Burg, A new analysis technique for time series data. in *Modern Spectrum Analysis*, ed. by D.G. Childers (IEEE Press, New York, 1978), pp. 42–49. Reprint of a paper presented at the NATO Advanced Study Institute on Signal Processing with Emphasis on Underwater Acoustics (1968)
10. R.L. Christensen, W.J. Strong, E.P. Palmer, A comparison of three methods of extracting resonance information from predictor-coefficient coded speech. IEEE Trans. Acoust. Speech Sig. Proc. **24**(1), 8–14 (1976)

11. M.G. Di Benedetto, Vowel representation: some observations on temporal and spectral properties of the first formant frequency. J. Acoust. Soc. Am. **86**(1), 55–66 (1989)
12. J. Durbin, The fitting of time-series models. Rev. de l'Institut International de Stat. **28**(3), 233–244 (1960)
13. G. Fant, *Acoustic Theory of Speech Production* (Mouton, The Hague, 1960) Reissued 1970
14. J.L. Flanagan, *Speech Analysis, Synthesis and Perception*, 2nd edn. (Springer, Berlin, 1972)
15. S.A. Fulop, Accuracy of formant measurement for synthesized vowels using the reassigned spectrogram and comparison with linear prediction. J. Acoust. Soc. Am. **127**(4), 2114–2117 (2010)
16. G.B. Giannakis, J.M. Mendel, Identification of nonminimum phase systems using higher order statistics. IEEE Trans. Acoust. Speech Sig. Proc. **37**(3), 360–377 (1989)
17. P.R. Gutowski, E.A. Robinson, S. Treitel, Spectral estimation: fact or fiction. IEEE Trans. Geosci. Elect. **16**(2), 80–84 (1978)
18. R.W. Hamming, *Digital Filters*. 3rd edn. (Prentice-Hall, Englewood Cliffs, 1989)
19. J. Hillenbrand, L.A. Getty, M.J. Clark, K. Wheeler, Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. **97**(5), 3099–3111 (1995)
20. S.M. Kay, S.L. Marple Jr., Spectrum analysis—a modern perspective. Proc. IEEE. **69**(11), 1380–1419 (1981)
21. D. Kewley-Port, C.S. Watson, Formant-frequency discrimination for isolated English vowels. J. Acoust. Soc. Am. **95**(1), 485–496 (1994)
22. J. Makhoul, Spectral analysis of speech by linear prediction. IEEE Trans. Audio Electroacoust. **21**(3), 140–148 (1973)
23. J. Makhoul, Spectral linear prediction: properties and applications. IEEE Trans. Acoust. Speech Sig. Proc. **23**(3), 283–296 (1975)
24. J.D. Markel, A.H. Gray Jr., *Linear Prediction of Speech* (Springer, Berlin, 1976)
25. S.S. McCandless, An algorithm for automatic formant extraction using linear prediction spectra. IEEE Trans. Acoust. Speech Sig. Proc. **22**(2), 135–141 (1974)
26. R.B. Monsen, A.M. Engebretson, The accuracy of formant frequency measurements: a comparison of spectrographic analysis and linear prediction. J. Speech Hearing Res. **26**(3), 89–97 (1983)
27. H. Morikawa, H. Fujisaki, System identification of the speech production process based on a state-space representation. IEEE Trans. Acoust. Speech Sig. Proc. **32**(2), 252–262 (1984)
28. National Instruments: LabVIEW 8.6 Advanced Signal Processing Toolkit Help (2008). Available online
29. M.B. Priestley, *Spectral Analysis and Time Series*, vol. 1. (Academic Press, London, 1981)
30. J.G. Proakis, D.G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, 2nd edn. 2nd edn. (Macmillan, New York, 1992)
31. H.R. Radoski, E.J. Zawalick, P.F. Fougere, The superiority of maximum entropy power spectrum techniques applied to geomagnetic micropulsations. Phys. Earth Planet. Interiors. **12**, 298–216 (1976)
32. J.R. Ragazzini, L.A. Zadeh, Analysis of sampled-data systems. Trans. Am. Inst. Elec. Eng. **71**, 225–234 (1952)
33. K. Steiglitz, On the simultaneous estimation of poles and zeros in speech analysis. IEEE Trans. Acoust. Speech Sig. Proc. **25**(3), 229–234 (1977)
34. G.K. Vallabha, B. Tuller, Systematic errors in the formant analysis of steady-state vowels. Speech Commun. **38**, 141–160 (2002)
35. W.W.S. Wei, *Time Series Analysis: Univariate and Multivariate Methods*. (Addison-Wesley, Redwood City, 1990)
36. Wikipedia: atan2. http://www.wikipedia.org (2010)

Index

A
Accommodation theory, 56, 57
Aeroacoustics, 140
Affricate, 11
Analysis-by-synthesis, 65
Approximant, 11
ARMA model *see* autoregressive moving average process
Aspiration, 8, 95
Autocorrelation, 15, 174
  instantaneous, 109, 110, 115
Autocovariance, 16
Autoregressive moving average process, 192, 193
  ARMASA method, 193
  coefficients, 193
  poles of, 193
  residual time series method, 193
  spectrum, 193
  zeros of, 193
Autoregressive process, 171, 192

B
Bernoulli, Daniel, 42
Bouasse, Henri, 45
de Broglie, Louis

C
Carslaw, H. S., 51
Cauchy, Augustin, 43
Chiba, Tsutomu, 62
Click sound, 159, 162
Complex conjugate, 18, 132
Cone-shaped kernel distribution *see* Zhao-Atlas-Marks distribution
Consonant, 8
  voicing, 10
Convolution, 115, 169, 170
Crandall, I. B, 57
Cross-spectral surface, 130

D
dB (decibel unit), 19
Degema, 156, 157, 196
Dipole source, 141
Dirichlet, Johann, 43
Doppler frequency, 111
Downsampling *see* resampling
Dynamic range, 91, 122

E
Edison, Thomas, 46
Euler's relation, 14
Euler, Leonhard, 41
Ewing, J. A, 46, 51, 53

F
Filter, 34, 63
  band-pass, 35
  glottal shaping, 171
  linear time-invariant, 169
  lip radiation, 171
  low-pass, 35
  resonance bandwidth, 180
  resonant, 35, 65, 69
Filter bank, 84
Finnish, 154, 191, 194
Flap *see* tap
Fletcher, Harvey, 62

F (*cont.*)
Formant, 7, 52, 52, 53, 54, 70, 72, 89, 118, 121, 136, 139, 141, 144, 149, 171, 176
  measurement of, 72, 73, 74, 93, 122, 149, 178, 188
  nasal, 157, 158, 188, 194
Fourier transform, 22, 43, 109, 115
  discrete, 25, 26, 27, 32, 57, 110, 170, 194
    frame, 25, 26, 27, 32, 33, 34, 151
  discrete short-time
  discrete-time, 24
  fast, 26, 51, 132
  inverse, 23
  inverse short-time
  short-time, 80, 87, 111, 128
    complex phase, 128
    phase derivatives, 130, 136, 137, 138
    phase unwrapping
Fourier's series, 18, 42, 43, 44, 45, 49
  discrete-time, 21, 50, 51, 55, 59
Fourier, J.-B. J, 42
Frame advance, 84, 151
Frequency
  angular, 14
  component, 107, 118, 128, 135, 136, 138, 139, 145, 151, 156, 158, 160
  fundamental, 6, 12, 13, 18, 20, 28, 53, 56, 62, 64, 70, 72, 73, 74, 78, 89, 93, 99, 121, 160
  harmonic, 18, 20, 27, 50, 51, 52, 53, 56, 57, 61, 64, 70, 72, 73, 78, 89, 92, 93, 99, 119, 160, 185
    physical reality of, 43
    relative amplitudes, 78
  instantaneous, 38, 108, 118, 128, 133
    channelized, 128, 132, 135, 138
  Nyquist, 15, 197
Frequency bin, 26, 86, 91, 109, 110, 128, 194
Frequency response, 34
Fricative, 10, 75, 75, 77, 97

G
Gabor transform, 80, 82, 84, 87
Gabor, Denis, 87
Gauss, Carl, 42, 65
Glide *see* semivowel
Glottal impulse *see* phonation, impulse
Glottis, 5, 141, 144, 148
Godfrey, Charles, 43
Gouy, M, 43
Group delay, 129

H
Hall, Harry, 50
Harmonic *see* frequency, harmonic
Harmonic analyzer, 50, 57
Harmonicity, 80
Harmonics-to-noise ratio *see* harmonicity
Helmholtz, Hermann, 52
Hermann, Ludimar, 52, 54, 62
Hilbert transform, 36, 37
Hop size *see* frame advance
Hz (Hertz unit), 13

I
Impulse response, 84, 169
Intonation, 12

J
Jenkin, Fleeming, 46, 51, 53

K
Kajiyama, Masato, 62
Kelvin, Lord *see* Thomson, W. (Lord Kelvin)
Koenig, Rudolph, 47

L
Lag time, 16
Laurent polynomial, 170, 171
Laurent series, 169
Lewis, Don, 60, 62
Linear prediction, 65, 171, 173
  accuracy of, 179
  autocorrelation method, 174
  Burg method, 174
  closed phase analysis, 174
  coefficients, 172, 173, 173, 176, 184
  covariance method, 174
  cumulant-based, 174
  gain factor, 172, 173, 174, 194
  pitch-asynchronous, 174, 182, 186
  poles *see* system function, poles of polynomial roots, 179
  spectrum, 173, 175
    peaks of, 177
    shaping resonance, 178, 180
Lloyd, R. J, 52, 55
Local group delay, 129, 132

M
Manner of articulation, 11
Manometric flame, 47, 53
Maximum entropy method *see* linear prediction, Burg method
Merritt, Ernest, 53
Miller, Dayton, 44, 48, 57
Monopole source, 140
Moving average process, 192

N
Nasal, 11, 98, 157, 194
Nasal murmur, 98

O
Obstruent, 75, 158
Oscillograph, 48

P
Parseval's relation, 19, 23
Period, 13, 73, 74
Periodogram, 23
Phase, 14, 82, 121
Phonation, 6, 62, 70, 72–73, 95, 121, 140
  impulse, 139, 140, 145, 147, 171, 174
Phonation type, 6, 78, 145
  breathy, 6, 79, 80, 95, 147
  creaky, 6, 79, 145
  modal, 6, 79, 80, 146
  slack, 6
  stiff, 6, 146
  whispery, 6, 80
Phonautograph, 46
Phone *see* segment
Phonetics, 5
Phonodeik, 48
Phonograph, 46
Pitch, 12, 52, 100, 160
  measurement of, 101, 161
  tracking algorithm, 100, 160
Pre-emphasis, 91, 185
Prosody, 12

R
Rayleigh, Lord *see* Strutt, J. W. (Lord Rayleigh)
Reassignment method, 128
  pruning, 136
Register *see* phonation type
Resampling, 184

S
Sacia, C. F, 57
Sampling, 12, 15, 50, 91
  quantizing, 15
  rate, 15, 104
Scott, Leon, 46
Scripture, Edward, 54, 55
Segment, 8
Semivowel, 9
Signal, 12
  analog *see* signal, continuous-time
  analytic, 35, 86, 109, 110, 128
  aperiodic, 13, 44
  continuous-time, 12
  deterministic, 12
  digital, 12, 69, 86, 169
  discrete-time, 14
  periodic, 13, 43, 49, 51
  random, 12, 75
Simple harmonic motion, 14, 50
Source-filter theory, 7, 62, 65, 70, 94, 140, 171, 188
Spectrogram, 63, 80, 108, 111, 114–116
  narrowband, 63, 99, 160
  reassigned, 130, 131, 188
    biometric, 149
  wideband, 63, 92
Spectrograph (device), 63
Spectrum
  energy density, 23
  ensemble average, 80
  moments of, 75
  parametric estimation, 167
  power, 19, 20, 27, 28, 31, 33–35, 51, 57, 59, 64, 69, 75, 77–78, 87
    reassigned, 135
  time average, 77
  zero of, 98, 156
Steinberg, John C, 59
Stop, 10, 80, 89, 95, 158
Strutt, J. W. (Lord Rayleigh), 52, 56
Subglottal resonance, 188
Syllable, 7
  coda, 8
  nucleus, 8
  onset, 8
System function, 170
  all-pole, 170–172, 176, 192
  poles of, 170–171, 176

S (*cont.*)
rational, 170, 192
zeros of, 170

T
Tap, 10
Thomson, W. (Lord Kelvin), 50
Time series, 16, 68
Time-frequency representation, 80, 107, 128
cross-terms in, 111, 113, 115, 118
impulse in, 137, 139
interference in, 112, 135
quadratic, 115
smoothing kernel, 115
smoothing window, 111, 115, 116
Tone, 12
Transvelar coupling, 146

U
Uncertainty principle, 87
spectrographic, 88, 111, 118, 135

V
Vocal cords *see* vocal folds
Vocal folds, 5, 52, 62, 63, 144, 171
Vocoid, 9
Voice bar, 55, 95, 118, 121, 122, 145, 147, 151, 188
Voicing *see* phonation
Vowel, 8, 52
advanced tongue root, 155, 156, 195
front-back, 10
height, 9
nasalized, 157, 158
rhotic, 10, 57
synthetic, 70, 119

W
Waveform, 46, 48, 51, 53, 55, 57, 59
Whisper, 6
Wigner-Ville distribution, 108, 112
discrete, 109
pseudo, 110, 113
Willis, Robert, 52
Window, 25, 29, 33, 34, 81, 86
length of, 25, 33, 89, 93, 151, 182
Window function, 28, 110, 116, 133, 182
Gaussian, 29, 32, 89
Hamming, 30, 32, 89
Hann, 30, 32, 90
Kaiser, 31, 32, 89, 133
rectangular, 186

Y
Yeyi, 159, 161

Z
z-transform, 169
Zero-padding, 25, 32, 86, 132, 175
Zhao-Atlas-Marks distribution, 113
discrete, 115