

Analysis, Synthesis, and Perception of Musical Sounds *The Sound of Music* contains a detailed treatment of basic methods for analysis and synthesis of musical sounds, including the phase vocoder method, the McAuley-Obregon frequency-tracking method, the constant-Q transform, and methods for pitch tracking with several examples shown. Various aspects of musical sound spectra such as spectral envelopes, spectral centroid, spectral flux, and spectral irregularity are defined and discussed. One chapter is devoted to the control and synthesis of spectral envelopes. Two advanced methods of analysis/synthesis are given: "Notes Plus Transients Plus Notes" and "Spectromorphological Resynthesis"; are covered. Methods for feature matching are given. The last two chapters discuss the perception of musical sounds based on discrimination and multidimensional scaling listener models.

"In this book, Dr. Beauchamp has assembled lucid accounts of the most important digital techniques applied to the contemporary analysis and synthesis of musical sound. The special value of studying these techniques is that methods of analysis and synthesis largely determine our ways of thinking about sound—especially the perception of it."

—William M. Hartmann, Michigan State University

"Through the years, James Beauchamp has made many excellent contributions to the written literature dealing with electronic music. For anyone who is interested in achieving a sophisticated understanding of the techniques of computer music, this book will be essential reading."

—Jed Chabada, Electronic Music Foundation

Editor

Analysis, Synthesis, and Perception of Musical Sounds *The Sound of Music*

Analysis, Synthesis, and Perception of Musical Sounds

The Sound of Music

James W. Beauchamp

Editor

Pearson



springer.com

 Springer



Modern Acoustics and Signal Processing

Analysis, Synthesis, and Perception of Musical Sounds

Modern Acoustics and Signal Processing

Editors-in-Chief

ROBERT T. BEYER

Department of Physics, Brown University, Providence, Rhode Island

WILLIAM HARTMANN

Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan

Editorial Board

YOICHI ANDO, Graduate School of Science and Technology, Kobe University, Kobe, Japan

ARTHUR B. BAGGEROER, Department of Ocean Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts

NEVILLE H. FLETCHER, Research School of Physical Science and Engineering, Australian National University, Canberra, Australia

CHRISTOPHER R. FULLER, Department of Mechanical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia

WILLIAM M. HARTMANN, Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan

JOANNE L. MILLER, Department of Psychology, Northeastern University, Boston, Massachusetts

JULIA DOSWELL ROYSTER, Environmental Noise Consultants, Raleigh, North Carolina

LARRY ROYSTER, Department of Mechanical and Aerospace Engineering, North Carolina State University, Raleigh, North Carolina

MANFRED R. SCHRÖDER, Göttingen, Germany

ALEXANDRA I. TOLSTOY, ATolstoy Sciences, Annandale, Virginia

WILLIAM A. VON WINKLE, New London, Connecticut

Books In The Series

Producing Speech: Contemporary Issues for Katherine Safford Harris, edited by Fredericka Bell-Berti and Lawrence J. Raphael

Signals, Sound, and Sensation, by William M. Hartmann

Computational Ocean Acoustics, by Finn B. Jensen, William A. Kuperman, Michael B. Porter, and Henrik Schmidt

Pattern Recognition and Prediction with Applications to Signal Characterization, by David H. Kil and Frances B. Shin

Oceanography and Acoustics: Prediction and Propagation Models, edited by Alan R. Robinson and Ding Lee

Handbook of Condenser Microphones, edited by George S.K. Wong and Tony F.W. Embleton

(continued after index)

Analysis, Synthesis, and Perception of Musical Sounds

The Sound of Music

James W. Beauchamp

Editor

University of Illinois at Urbana, USA



Springer

James W. Beauchamp
Professor Emeritus
School of Music
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801
USA
jwbeauch@uiuc.edu

Cover illustration: Analysis and resynthesis of a piano tone.

Library of Congress Control Number: 2006920599

ISBN-10: 0-387-32496-8 e-ISBN-10: 0-387-32576-X
ISBN-13: 978-0387-32496-8 e-ISBN-13: 978-0387-32576-7

Printed on acid-free paper.

© 2007 Springer Science+Business Media, LLC

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

9 8 7 6 5 4 3 2 1

springer.com

To Karen Fuchs-Beauchamp and Nathan Charles Beauchamp

Preface

The title of this book, *Analysis, Synthesis, and Perception of Musical Sounds*, has been the subject of many conference sessions (for example, at the 127th Meeting of the Acoustical Society of America at Cambridge, Massachusetts in May, 1994, which originally inspired this book) and journal papers, but there has been little to date which combines these subjects into a single volume. Traditionally, dating back to Helmholtz (1877), the subject of analysis of musical sounds consisted solely of harmonic analysis of sustained-tone instruments. However, many other applications have been developed during the last several decades, and the topics of analysis, synthesis, and perception (AS&P) are very representative of these applications.

It almost goes without saying that the principal tool that has facilitated AS&P is the digital computer, and all of the projects described in this book have used this indispensable tool. Another common thread is that all of these projects have used a form of time-varying spectral analysis [usually implemented using a form of the short-time Fourier transform (STFT)], which models signals as sums of sine waves (sinusoids).

Indisputably, the first time-varying spectral analysis and synthesis of musical sounds by a digital computer was accomplished in Melville Clark Jr.'s lab at MIT (Luce, 1963, 1975; Luce and Clark, 1967; Strong and Clark, 1967a, 1967b). Projects by Beauchamp and Fornango (1966), Freedman (1967, 1968), and Beauchamp (1969, 1974, 1975) at the University of Illinois at Urbana-Champaign, Risset and Mathews (1969) at Bell Telephone Laboratories, and Keeler (1972) at the University of Waterloo soon followed. Some of these projects were described in the book *Music by Computers* (von Forester and Beauchamp, eds., 1969). Strong and Clark's project (1967a, 1967b) was the first to incorporate listening tests in publications on musical sound synthesis derived from spectral analysis. Luce, Strong, and Clark were also first to emphasize the importance of musical instrument *spectral envelopes*, which are smoothed versions of sound spectra. Later, John Grey, James A. Moorer, and John Gordon at Stanford University completed a much more extensive series of perceptual studies based on spectral analysis/synthesis in the mid-1970s (Grey, 1975, 1977; Grey and Moorer, 1977; Grey and Gordon, 1978), including the use of the multidimensional scaling (MDS) method to determine a

“space” of musical timbres. These were preceded by similar timbre space studies by Wedin and Goude (1972), Wessel (1973), and Miller and Carterette (1975), which also used the MDS method but only employed original acoustic sounds or artificial sounds not obtained by analysis/synthesis.

The *phase vocoder*, a method of time-varying analysis/synthesis similar to that used by the early music researchers, was first employed for speech applications by Flanagan and Golden (1966) and Portnoff (1976) and later extended for music by Moorer (1978) and Dolson (1986). Again for speech, McAulay and Quatieri (1986) introduced the spectral frequency tracking (SFT) method, and a similar method (called PARSHL) was developed for music applications by Smith and Serra (1987). This method (now called SMS) was extended by Serra and Smith (1990) with the additional feature of extracting a time-varying noise residual from the sound signal. Separate control of the noise residual offered advantages such as reduction of artifacts when time-scaling is employed. A freely downloadable source-code package (called SNDAN) which combines a tunable phase vocoder and the SFT method was described by Beauchamp (1993). Since then, many new music analysis/synthesis methods have been developed. A comparison of current methods was given in Wright et al. (2001).

Other aspects of the history of analysis/synthesis are discussed in the chapter by Levine and Smith (Chapter 4).

This book consists of eight chapters. In the first chapter James Beauchamp discusses basic methods of time-varying spectral analysis and synthesis and gives examples of the analysis of various musical instruments. The two analysis/synthesis methods presented are the Harmonic Filter Bank (HFB, aka phase vocoder) and the Spectral Frequency-Tracking (SFT) methods. The HFB method, where the frequencies of analysis can be aligned with frequencies of a harmonic sound, works best for sounds that are quasiperiodic, i.e., they have nearly constant pitch (i.e., fundamental frequency). The SFT method works best for sounds with variable pitch. Both methods can be used for sounds with inharmonic partials, although the HFB has the advantage of avoiding problems of excessive amplitude thresholding and partial frequency mistracking. This chapter also defines several “higher-level” measures of spectra, which may be useful for classifying instruments. These are the *spectral centroid* (associated with “perceptual brightness”), *spectral irregularity*, *inharmonicity*, *decay rate*, *spectrotemporal incoherence*, and *inverse spectral density*, and examples for different instruments are given. Beauchamp concludes by showing how the SFT method can be used to track the fundamental frequency as well as to separate the harmonics of a signal with substantial time-varying pitch.

While the traditional Fourier transform yields frequencies that are uniformly spaced, it is possible to define a variation on this transform, called the constant-Q transform, which yields an analysis at logarithmically spaced frequencies. In Chapter 2, Judith Brown looks at methods of analysis using this transform. She then shows how fundamental-frequency (pitch) tracking can be based on pattern matching of the constant-Q transform output, giving examples of violin performance analysis. Next, a high-resolution pitch analyzer is described, which is based on the phase changes of spectral components, to improve the precision of pitch tracking. This pitch analyzer was applied to the problem of resolving the frequency

ratios of musical instrument partials in order to determine the degree to which they were, or were not, harmonic. Finally, a listening experiment was conducted to determine the perceived pitch center of viola vibrato tones, and results for relatively experienced and inexperienced listeners are compared. This also yielded an estimate of the pitch JND for these listeners.

In Chapter 3, Lippold Haken, Kelly Fitz, and Paul Christensen describe a novel analysis/synthesis method and how it can be used as a synthesis engine for a “fingerboard” musical instrument. The method is an extension of the SFT method described in Chapter 1. The two extensions are *noise enhancement* and *spectral reassignment*. Rather than separate additive noise into a residual as has been done by Serra and Smith (1990), noise is treated in terms of separable “noise-factor” signals that are modulated onto individual partials during synthesis. Thus, each partial is represented by three parameters: amplitude, frequency, and noise factor. With spectral reassignment, the time and frequency for each time frame and partial within the frame are reestimated by utilizing centroids of the windowed time function and its Fourier transform. The overall method results in improved analysis/synthesis of complex sounds having sharp transients and inharmonic partials. The result is parameter streams that can be easily manipulated in time and frequency. The method has been used as the synthesis engine of a new “fingerboard” musical instrument, called the *Continuum*, which, in addition to pitch and loudness control, affords timbral control by morphing between two target instrument sounds appropriate for each pitch.

Another method of processing complex, even polyphonic, sounds with increased perceptual accuracy is described by Scott Levine and Julius Smith in Chapter 4. Their method builds on the sinusoids-plus-noise model developed by Serra and Smith (1990). The new method divides the signal into three parts: time-varying sinusoids, time-varying noise, and transients. The signal is first segmented into attack-transient and nontransient time regions. The transient segments are coded using a variation on an MPEG audio transient coder. Nontransient time regions are analyzed as “multiresolution sinusoids” and noise. “Multiresolution” means that frequencies below 5000 Hz are analyzed as time-varying sinusoids for the frequency ranges 0–1250 Hz, 1250–2500 Hz, and 2500–5000 Hz with different time resolutions of 46 ms, 23 ms, and 11.5 ms, respectively. Overlap regions between transient and sinusoids are phase-matched to avoid discontinuities. Noise is modeled in terms of Bark bands, which are critical bands varying in bandwidth across the spectrum (Zwicker, 1961). Below 5000 Hz noise is based on the residual between the signal and the sum of analyzed sinusoids. Above 5000 Hz noise is based on the entire signal. Time variation of the noise is given in terms of a piecewise linear curve for the amplitude of each Bark-band noise. The method allows time expansion and other modifications (such as frequency tuning) without loss of fidelity, including the preservation of sharp attack transients.

In Chapter 5, Xavier Rodet and Diemo Schwarz describe various methods for representing signals in terms of time-varying spectral envelopes. A tacit assumption is that the spectral envelope provides appropriate spectral variation as the fundamental frequency (pitch) varies. It is also useful for morphing between different vocal or instrumental spectra. The chapter outlines the importance of the

source/filter model, especially for speech signals, and the importance of *formants*, which are pronounced maxima within spectra or filter response functions at particular frequencies, usually higher than the fundamental. Source spectra generally have no formants, but they can vary with time and with intensity; in the latter case, usually the tilt (i.e., average slope) of the spectrum varies with intensity. Three important properties of a spectral envelope are given: (1) It should envelope the spectral maxima; (2) it should be smooth; and (3) it should adapt to fast variation. Later, properties of exactness and robustness are added. Then, various spectral-envelope estimation methods are given, including methods that are derived by *autoregression* (AR) [also called *linear predictive coding* (LPC)], *cepstrum*, *discrete cepstrum*, and several enhancements of the discrete cepstrum method. The spectral envelope of the residual signal is treated as a special case, because this is assumed to be nonsinusoidal. Other topics covered are concerned with synthesis: filter coefficients, geometric representations, formants, spectral-envelope manipulation, morphing, sine-wave additive synthesis, and inverse-FFT synthesis.

In Chapter 6 Andrew Horner discusses methods of data reduction for multiple wavetable and frequency-modulation (FM) resynthesis based on matching the time-varying spectral analysis of harmonic (or approximately harmonic) fixed-pitch musical instrument tones. A relative-amplitude spectral error formula is defined, and the use of a genetic algorithm combined with the well-known least-squares method to compute a set of near-optimum spectra and associated amplitude-vs-time envelopes for resynthesis is described. Several different methods of resynthesis are examined: wavetable indexing, wavetable interpolation, group additive, formant FM, double FM, and nested FM. Results are shown for trumpet, tenor voice, and Chinese pipa tone matches using each of the methods. Wavetable indexing and wavetable interpolation are found to give the best matches. However, wavetable indexing is found to require the least memory, while wavetable interpolation is found to be the most computationally efficient of the two methods.

John Hajda reviews recent research on the salience of various timbre-related parameters in Chapter 7. Two basic methods for studying timbre are *classification* and *relational measures*. Some spectrotemporal parameters that may impact timbre are time-envelope (attack, steady-state, decay), spectral centroid, spectral irregularity, and spectral flux. When the attack portions are deleted from 12 sustained (aka continuant) tones (with attack time measured three different ways), the “remainder tones” are on average correctly identified almost at the same rate as the original sounds (85% vs 93% correct) and are better for identification than “attack-only tones.” Moreover, reverse playback of entire sustained tones does not affect their identification. These two results indicate the relative importance of steady-state and decay. Two different relational methods are (1) verbal attribute magnitude estimation, where timbres are rated on a scale from, say, “dull” to “sharp”; and (2) numerical ratings of timbre dissimilarity, which can be analyzed by MDS statistical algorithms to produce a “timbre space,” where each timbre occupies a point in the space and the distance between any two timbres represents their average perceptual dissimilarity. In the latter case, physical parameters such as attack time, spectral centroid, and spectral variance have been found to correlate well with

MDS dimensions. In one study, parameter salience was determined by testing how well listeners could detect various simplifications to time-varying spectral data after resynthesis, under the assumption that if a parameter is easily detected when a parameter is simplified, the parameter must have timbral saliency (McAdams et al., 1999). Another study with similar simplifications used a similarity rating method of testing subjects (Hajda, 1999). Both studies agreed that spectral flux, the amount of variation of the amplitude-normalized spectrum, is the most salient parameter of the sustained musical instrument sounds tested. The chapter closes with brief discussions of the effect of musical context on timbre and the perception of percussion (aka impulse) sounds.

Finally, in Chapter 8 Sophie Donnadieu considers a number of topics related to timbre perception. She begins by noting the difficulty of studying timbre due to the absence of a satisfactory definition, its multidimensional nature, and a diversity of notions about the types of sound sources that produce timbre, whether they be isolated tones, multiple pitches on a single instrument, combinations of different instruments, or unfamiliar sounds produced by sound synthesis. Next, the concept of perceptual dimensions is discussed, with an emphasis on MDS methods, and the results of several MDS experiments are described (e.g., Grey and Moorer, 1977; McAdams et al., 1995). Usually two or three dimensions can be resolved and correlated (either qualitatively or quantitatively) with spectrotemporal features such as “temporal envelope,” “spectral envelope,” and “spectral flux.” Next she introduces the concept of “specificities,” whereby different instruments have unique aspects of timbral quality, such as special types of attacks or special spectral or formant characteristics. The effect of listener musical experience is also explored, and musicianship is found to affect the precision and coherence of judgments. Furthermore, the predictive power of timbre spaces is discussed in terms of interpolating along dimensions using morphing techniques, perception of “timbral intervals,” auditory streaming, and the effect of context. Finally, attempts to evaluate the efficacy of verbal attributes such as “smooth” vs “rough” for describing timbre are discussed. In the next section Donnadieu looks at the idea of timbral categorization. According to categorization theory, timbre is mentally organized by clusters, rather than as a continuum, e.g., any sound with certain characteristics might be categorized as a “trumpet.” Or it is also plausible that timbres are strictly grouped by listeners according to physical sound-production characteristics (e.g., instrument size, shape, material, and manner of excitation) which are inferred from the corresponding sounds. Donnadieu describes her own experiment on categorization processes and finds that timbral categories correspond to perceptual reality while at the same time they are related to the physical functioning of musical instruments. She concludes by describing several studies, including one of her own, which use a physical parameter continuum (e.g., attack time) to test the relationship between “identification” and “discrimination.” While most studies seem to suggest that categorical perception is salient and is based on feature detection, her study on a rise-time continuum for struck and bowed vibraphones supported a theory of noncategorical perception. Therefore, the conditions under which categorical vs noncategorical perception of timbre occur is still an open question.

These eight chapters give eight different perspectives on the problem of understanding musical sounds from an analytical point of view. They hopefully will give the reader a broad insight into how sounds can be analyzed, illustrated, modified, synthesized, and perceived.

J.W.B.
Urbana, Illinois, U.S.A.
February, 2005

References

- Beauchamp, J. W. and Fornango, J. P. (1966). "Transient Analysis of Harmonic Musical Tones by Digital Computer," 31st Convention of the Audio Eng. Soc. Convention, Audio Engr. Soc. Preprint No. 479.
- Beauchamp, J. W. (1969). "A Computer System for Time-Variant Harmonic Analysis and Synthesis of Musical Tones," in *Music by Computers*, H. F. von Forester and J. W. Beauchamp, eds. (J. Wiley, New York), pp. 19–62.
- Beauchamp, J. W. (1974). "Time-variant spectra of violin tones," *J. Acoust. Soc. Am.* **56**(3), 995–1004.
- Beauchamp, J. W. (1975). "Analysis and Synthesis of Cornet Tones Using Nonlinear Inter-harmonic Relationships," *J. Audio Eng. Soc.* **23**(10), 778–795.
- Beauchamp, J. W. (1993). "Unix Workstation Software for Analysis, Graphics, Modification, and Synthesis of Musical Sounds," 94th Convention of the Audio Eng. Soc., Berlin, Audio Eng. Soc. Preprint No. 3479.
- Dolson, M. (1986). "The Phase Vocoder: A Tutorial," *Computer Music J.* **10**(4), 14–27.
- Flanagan, J. L. and Golden, R. M. (1966). "Phase Vocoder," *Bell System Technical J.* **45**, 1493–1509. Reprinted in *Speech Analysis*, R. W. Schafer and J. D. Markel, eds. (IEEE Press, New York), 1979, pp. 388–404.
- Freedman, M. D. (1967). "Analysis of Musical Instrument Tones," *J. Acoust. Soc. Am.*, **41**(4), 793–806.
- Freedman, M. D. (1968). "A Method for Analyzing Musical Tones," *J. Audio Eng. Soc.* **16**(4), 419–425.
- Grey, J. M. (1975). "An Exploration of Musical Timbre," unpublished doctoral dissertation, Stanford University, Stanford, CA. Also available as Stanford Dept. of Music Report STAN-M-2.
- Grey, J. M. (1977). "Multidimensional perceptual scaling of musical timbres," *J. Acoust. Soc. Am.* **61**(5), 1270–1277.
- Grey, J. M. and Moorer, J. A. (1977). "Perceptual evaluations of synthesized musical instrument tones," *J. Acoust. Soc. Am.* **62**(2), 454–462.
- Grey, J. M. and Gordon, J. W. (1978). "Perceptual effects of spectral modifications on musical timbres," *J. Acoust. Soc. Am.* **63**(5), 1493–1500.
- Hajda, J. M. (1999). "The Effect of Time-Variant Acoustical Properties on Orchestral Instrument Timbres," doctoral dissertation, University of California, Los Angeles. UMI number 9947018.
- Helmholtz, H. von ([1877] 1954). *On the Sensation of Tone as a Psychological Basis for the Study of Music*, 4th ed. Trans., A. J. Ellis., ed. (Dover, New York).
- Keeler, J. S. (1972). "Piecewise-Periodic Analysis of Almost-Periodic Sounds and Musical Transients," *IEEE Trans. on Audio and Electroacoustics* **AU-20**(5), 338–344.

- Luce, D. A. (1963). *Physical Correlates of Non-Percussive Musical Instruments*, PhD dissertation, Massachusetts Institute of Technology, Cambridge, MA.
- Luce, D. and Clark, M. (1967), "Physical Correlates of Brass-Instrument Tones," *J. Acoust. Soc. Am.* **42**(6), 1232–1243.
- Luce, D. A. (1975). "Dynamic Spectrum Changes of Orchestral Instruments," *J. Audio Eng. Soc.* **23**(7), 565–568.
- McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). "Perceptual scaling of synthesized musical timbres : Common dimensions, specificities, and latent subject classes," *Psychol. Res.* **58**, 177–192.
- McAdams, S., Beauchamp, J. W., and Meneguzzi, S. (1999). "Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters," *J. Acoust. Soc. Am.* **105**(2), 882–897.
- McAulay, R. J. and Quatieri, T. F. (1986). "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. on Acoust., Speech, and Signal Processing ASSP-34*(4), 744–754.
- Miller, J. R. and Carterette, E. C. (1975). "Perceptual space for musical structure," *J. Acoust. Soc. Am.* **58**(3), 711–720.
- Moorer, J. A. (1978). "The Use of the Phase Vocoder in Computer Music Applications," *J. Audio Eng. Soc.* **26**(1/2), 42–45.
- Portnoff, M. R. (1976). "Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform," *IEEE Trans. Acoust. Speech, and Signal Processing ASSP-24*, 243–248. Reprinted in *Speech Analysis*, R. W. Schafer and J. D. Markel, eds. (IEEE Press, New York), pp. 405–410.
- Risset, J.-C. and Mathews, M. V. (1969). "Analysis of Musical-Instrument Tones," *Physics Today* **22**(2), 23–30.
- Serra, X. and Smith, J. O. (1990). "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition," *Computer Music J.* **14**(4), 12–24.
- Smith, J. O. and Serra, X. (1987). "PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation," *Proc. 1987 Int. Computer Music Conf.*, Urbana, IL (Int. Computer Music Assn., San Francisco), pp. 290–297. Also available as Report No. STAN-M-43, Dept. of Music, Stanford Univ., 1987.
- Strong, W. and Clark, M. (1967a). "Synthesis of Wind-Instrument Tones," *J. Acoust. Soc. Am.* **41**(1), 39–52.
- Strong, W. and Clark, M. (1967b). "Perturbations of Synthetic Orchestral Wind-Instrument Tones," *J. Acoust. Soc. Am.* **41**(2), 277–285.
- von Forester, H. F. and Beauchamp, J. W., eds. (1969). *Music by Computers* (J. Wiley, New York).
- Wedin, L. and Goude, G. (1972). "Dimension analysis of the perception of instrumental timbre," *Scand. J. Psych.* **13**, 228–240.
- Wessel, D. L. (1973). "Psychoacoustics and Music: A Report From Michigan State University," *Page: Bulletin of the Computer Arts Society* **30** (London, U.K.).
- Wright, M., Beauchamp, J., Fitz, K., Rodet, X., Röbel, A., Serra, X., and Wakefield, G. (2001). "Analysis/synthesis comparison," *Organized Sound* **5**(3), 173–189.
- Zwicker, E. (1961). "Subdivision of the Audible Range into Critical Bands (Frequenzgruppen)," *J. Acoust. Soc. Am.* **33**(2), 248.

Acknowledgments

I wish to acknowledge the following people who made many valuable suggestions regarding the text: Stephen McAdams and John Hajda, for their work on the Donnadieu chapter, and Larry Heyl, who spent many hours deciphering all of the chapters. Special thanks go to my wonderful wife Karen Fuchs-Beauchamp for the enormous time she spent reconciling the references and the Index and, in general, for helping me surmount various hurdles in completing the book.

J.W.B.

Contents

<i>Preface</i>	vii
<i>Acknowledgments</i>	xv
1. Analysis and Synthesis of Musical Instrument Sounds	1
<i>James W. Beauchamp</i>	
1 Analysis/Synthesis Methods	2
1.1 Harmonic Filter Bank (Phase Vocoder) Analysis/Synthesis	3
1.1.1 Frequency Deviation and Inharmonicity	3
1.1.2 Heterodyne-Filter Analysis Method.....	5
1.1.2.1 Window Functions.....	5
1.1.2.2 Harmonic Analysis Limits	10
1.1.2.3 Synthesis from Harmonic Amplitudes and Frequency Deviations.....	12
1.1.3 Signal Reconstruction (Resynthesis) and the Band-Pass Filter Bank Equivalent.....	12
1.1.4 Sampled Signal Implementation.....	13
1.1.4.1 Analysis Step	14
1.1.4.2 Synthesis Step	17
1.1.4.2.1 Piecewise Constant Amplitudes and Frequencies	20
1.1.4.2.2 Piecewise Linear Amplitude and Frequency Interpolation	20
1.1.4.2.3 Piecewise Quadratic Interpolation of Phases.....	21
1.1.4.2.4 Piecewise Cubic Interpolation of Phases.....	23
1.2 Spectral Frequency-Tracking Method.....	26
1.2.1 Frequency-Tracking Analysis	27
1.2.2 Frequency-Tracking Algorithm	29
1.2.3 Fundamental Frequency (Pitch) Detection.....	33

1.2.4	Reduction of Frequency-Tracking Analysis to Harmonic Analysis	36
1.2.5	Frequency-Tracking Synthesis	37
1.2.5.1	Frequency-Tracking Additive Synthesis	37
1.2.5.2	Residual Noise Analysis/Synthesis	39
1.2.5.3	Frequency-Tracking Overlap-Add Synthesis.....	40
2	Analysis Results Using SNDAN	42
2.1	Analysis File Data Formats	43
2.2	Phase-Vocoder Analysis Examples for Fixed-Pitch Harmonic Musical Sounds.....	44
2.2.1	Spectral Centroid	45
2.2.2	Spectral Envelopes	50
2.2.3	Spectral Irregularity	55
2.3	Phase-Vocoder Analysis of Sounds with Inharmonic Partials	58
2.3.1	Inharmonicity of Slightly Inharmonic Sounds: The Piano.....	60
2.3.2	Measurement of Tones with Widely Spaced Partials: The Chime	62
2.3.3	Measurement of a Sound with Dense Partials: The Cymbal.....	66
2.3.4	Spectrotemporal Incoherence	67
2.3.5	Inverse Spectral Density: Cymbal, Chime, and Timpani.....	69
2.4	Frequency-Tracking Analysis of Harmonic Sounds	75
2.4.1	Frequency-Tracking Analysis of Steady Harmonic Sounds.....	75
2.4.2	Frequency-Tracking Analysis of Vibrato Sounds: The Singing Voice	75
2.4.3	Frequency-Tracking Analysis of Variable-Pitch Sounds	81
3	Summary.....	82
	References	86
2.	Fundamental Frequency Tracking and Applications to Musical Signal Analysis	90
	<i>Judith C. Brown</i>	
1	Introduction to Musical Signal Analysis in the Frequency Domain.....	90
2	Calculation of a Constant-Q Transform for Musical Analysis.....	93
2.1	Background	93
2.2	Calculations	93
2.3	Results	96

3	Musical Fundamental-Frequency Tracking Using a Pattern-Recognition Method	99
3.1	Background	99
3.2	Calculations	100
3.3	Results	101
4	High-Resolution Frequency Calculation Based on Phase Differences	103
4.1	Introduction	103
4.2	Results Using the High-Resolution Frequency Tracker	104
5	Applications of the High-Resolution Pitch Tracker	105
5.1	Frequency Ratios of Spectral Components of Musical Sounds	105
5.1.1	Background	106
5.1.2	Calculation	107
5.1.3	Results	107
5.1.3.1	Cello	108
5.1.3.2	Alto Flute	110
5.1.4	Discussion	110
5.2	Perceived Pitch Center of Bowed String Instrument Vibrato Tones	111
5.2.1	Background	111
5.2.2	Experimental Method	112
5.2.2.1	Sound Production and Manipulation	112
5.2.2.2	Listening Experiments	112
5.2.3	Results	113
5.2.3.1	Experiment 1: NonProfessional-Performer Listeners	113
5.2.3.2	Experiment 2: Graduate-Level and Professional Violinist Listeners	114
5.2.3.3	Experiment 3: Determination of JND for Pitch	114
6	Summary and Conclusions	116
Appendix A:	An Efficient Algorithm for the Calculation of a Constant-Q Transform	116
Appendix B:	Single-Frame Approximation—Calculation of Phase Change for a Hop Size of One Sample	117
References	119
3.	Beyond Traditional Sampling Synthesis: Real-Time Timbre Morphing Using Additive Synthesis	122
	<i>Lippold Haken, Kelly Fitz, and Paul Christensen</i>	
1	Introduction	122
2	Additive Synthesis Model	123
2.1	Real-Time Synthesis	124

2.2	Envelope Parameter Streams.....	125
2.3	Noise Envelopes.....	125
3	Additive Sound Analysis	125
3.1	Sinusoidal Analysis.....	125
3.2	Noise-Enhanced Sinusoidal Analysis	125
3.3	Spectral Reassignment	128
3.3.1	Time Reassignment	128
3.3.2	Frequency Reassignment.....	130
3.3.3	Spectral-Reassignment Summary.....	130
4	Navigating Source Timbres: Timbre Control Space.....	131
4.1	Creating a New Timbre Control Space.....	135
4.2	Timbre Control Space with More Control Dimensions	135
4.3	Producing Intermediate Timbres: Timbre Morphing	135
4.4	Weighting Functions for Real-Time Morphing.....	136
4.5	Time Dilation Using Time Envelopes.....	136
4.6	Morphed Envelopes.....	137
4.7	Low-Amplitude Partials	138
5	New Possibilities for the Performer: The Continuum Fingerboard.....	139
5.1	Previous Work	140
5.2	Mechanical Design of the Playing Surface.....	141
6	Final Summary	142
	References	142

4. A Compact and Malleable Sines+Transients+Noise Model for Sound

Scott N. Levine and Julius O. Smith III

1	Introduction	145
1.1	History of Sinusoidal Modeling.....	146
1.2	Audio Signal Models for Data Compression and Transformation	148
1.3	Chapter Overview.....	149
2	System Overview.....	150
2.1	Related Current Systems.....	150
2.2	Time-Frequency Segmentation.....	151
2.3	Reasons for the Different Models.....	151
3	Multiresolution Sinusoidal Modeling	152
3.1	Analysis Filter Bank.....	154
3.2	Sinusoidal Parameters.....	155
3.2.1	Sinusoidal Tracking	155
3.2.2	Masking	155
3.2.3	Sinusoidal Trajectory Elimination.....	157
3.2.4	Sinusoidal Trajectory Quantization	158
3.3	Switched Phase Reconstruction	158
3.3.1	Cubic-Polynomial Phase Reconstruction.....	160

3.3.2	Phaseless Reconstruction.....	160
3.3.3	Phase Switching.....	161
4	Transform-Coded Transients.....	161
4.1	Transient Detection.....	162
4.2	A Simplified Transform Coder.....	163
4.3	Time-Frequency Pruning	164
5	Noise Modeling	164
5.1	Bark-Band Quantization.....	165
5.2	Line-Segment Approximation.....	166
6	Applications.....	167
6.1	Sinusoidal Time-Scale Modification	170
6.2	Transient Time-Scale Modification.....	170
6.3	Noise Time-Scale Modification	170
7	Conclusions	170
8	Acknowledgment.....	171
	References	171
5.	Spectral Envelopes and Additive + Residual Analysis/Synthesis	175
	<i>Xavier Rodet and Diemo Schwarz</i>	
1	Introduction	175
2	Spectral Envelopes and Source–Filter Models.....	178
2.1	Source–Filter Models	178
2.2	Source–Filter Models Represented by Spectral Envelopes.....	181
2.3	Spectral Envelopes and Perception	184
2.4	Source and Spectrum Tilt.....	186
2.5	Properties of Spectral Envelopes.....	187
3	Spectral Envelope Estimation Methods.....	188
3.1	Requirements	190
3.2	Autoregression Spectral Envelope	190
3.2.1	Disadvantage of AR Spectral Envelope Estimation.....	193
3.3	Cepstrum Spectral Envelope	194
3.3.1	Disadvantages of the Cepstrum Method.....	196
3.4	Discrete Cepstrum Spectral Envelope.....	197
3.5	Improvements on the Discrete Cepstrum Method.....	200
3.5.1	Regularization	200
3.5.2	Stochastic Smoothing (the Cloud Method)	200
3.5.3	Nonlinear Frequency Scaling.....	202
3.6	Estimation of the Spectral Envelope of the Residual Signal.....	204
4	Representation of Spectral Envelopes	205
4.1	Requirements	205
4.2	Filter Parameters	206

4.3	Frequency Domain Sampled Representation	206
4.4	Geometric Representation.....	207
4.5	Formants.....	208
4.5.1	Formant Wave Functions.....	208
4.5.2	Basic Formants.....	209
4.5.3	Fuzzy Formants	209
4.5.4	Discussion of Formant Representation	210
4.6	Comparison of Representations	210
5	Transcoding and Manipulation of Spectral Envelopes	211
5.1	Transcodings.....	211
5.1.1	Converting Formants to AR-Filter Coefficients	211
5.1.2	Formant Estimation	211
5.2	Manipulations.....	212
5.3	Morphing	212
5.3.1	Shifting Formants.....	213
5.3.2	Shifting Fuzzy Formants	214
5.3.3	Morphing Between Well-Defined Formants	215
5.3.4	Summary of Formant Morphing	215
6	Synthesis with Spectral Envelopes.....	216
6.1	Filter Synthesis	216
6.2	Additive Synthesis	217
6.3	Additive Synthesis with the FFT^{-1} Method.....	217
7	Applications.....	218
7.1	Controlling Additive Synthesis	218
7.2	Synthesis and Transformation of the Singing Voice	219
8	Conclusions	220
9	Summary.....	220
	Appendix: List of Symbols	221
	References	222
6.	A Comparison of Wavetable and FM Data Reduction Methods for Resynthesis of Musical Sounds	228
	<i>Andrew Horner</i>	
1	Introduction	228
2	Evaluation of Wavetable and FM Methods.....	229
3	Comparison of Wavetable and FM Methods	231
3.1	Generalized Wavetable Matching	232
3.2	Wavetable-Index Matching.....	232
3.3	Wavetable-Interpolation Matching.....	234
3.4	Formant-FM Matching	236
3.5	Double-FM Matching	237
3.6	Nested-FM Matching	238
4	Results	240
4.1	The Trumpet	241

4.2	The Tenor Voice.....	243
4.3	The Pipa	245
5	Conclusions	245
	Acknowledgments	247
	References	247
7.	The Effect of Dynamic Acoustical Features on Musical Timbre	250
	<i>John M. Hajda</i>	
1	Introduction	250
2	Global Time-Envelope and Spectral Parameters	251
2.1	Salience of Partitioned Time Segments.....	251
2.2	Relational Timbre Studies.....	258
2.2.1	Temporal Envelope.....	260
2.2.2	Spectral Energy Distribution	261
2.2.3	Spectral Time Variance	262
3	The Experimental Control of Acoustical Variables.....	263
4	Conclusions and Directions for Future Research.....	267
	References	268
8.	Mental Representation of the Timbre of Complex Sounds	272
	<i>Sophie Donnadieu</i>	
1	Timbre: A Problematic Definition	272
2	The Notion of Timbre Space.....	274
2.1	Continuous Perceptual Dimensions.....	274
2.1.1	Spectral Attributes of Timbre.....	274
2.1.2	Temporal Attributes of Timbre	281
2.1.3	Spectrotemporal Attributes of Timbre	283
2.2	The Notion of Specificities.....	285
2.3	Individual and Group Listener Differences.....	286
2.4	Evaluating the Predictive Power of Timbre Spaces	290
2.4.1	Perceptual Effects of Sound Modifications	290
2.4.2	Perception of Timbral Intervals	290
2.4.3	The Role of Timbre in Auditory Streaming.....	292
2.4.4	Context Effects	294
2.5	Verbal Attributes of Timbre	296
2.5.1	Semantic Differential Analyses	296
2.5.2	Relations Between Verbal and Perceptual Attributes or Analyses of Verbal Protocols.....	296
3	Categories of Timbre	297
3.1	Studies of the Perception of Causality of Sound Events	299
3.2	Categorical Perception: A Speech-Specific Phenomenon	301

3.2.1	Definition of the Categorical Perception Phenomenon.....	301
3.2.2	Musical Categories: Plucking and Striking vs Bowing.....	302
3.2.2.1	Are the Same Feature Detectors Used for Speech and Nonspeech Sounds?.....	303
3.2.2.2	Categorical Perception in Young Infants	304
3.2.2.3	The McGurk Effect for Timbre.....	305
3.2.3	Is There a Perceptual Categorization of Timbre?.....	306
4	Conclusions	312
	References	313
	<i>Index</i>	320

Analysis and Synthesis of Musical Instrument Sounds

JAMES W. BEAUCHAMP

Introduction

For synthesizing a wide variety of musical sounds, it is important to understand which acoustic properties of musical instrument sounds are related to specific perceptual features. Some properties are obvious: Amplitude and fundamental frequency easily control loudness and pitch. Other perceptual features are related to sound spectra and how they vary with time. For example, tonal “brightness” is strongly connected to the centroid or tilt of a spectrum. “Attack impact” (sometimes called “bite” or “attack sharpness”) is strongly connected to spectral features during the first 20–100 ms of sound, as well as the rise time of the sound. Tonal “warmth” is connected to spectral features such as “incoherence” or “inharmonicity.”

Experienced musical listeners can usually identify which instruments are present in a music recording, although identification accuracy varies with the prominence of an instrument (in the music), familiarity, number of instruments, etc. Listeners can even track an individual instrument, by “pushing other instruments into the background,” as it moves up and down the pitch scale. Something about the integrity of an individual instrument’s scope of spectral possibilities makes experienced musical listeners able to consider a group of notes to be “from that instrument.” This may be aided by listeners’ ability to visualize the physical apparatus that produces a group of sounds previously heard. However, it is also probable that listeners can learn to hear these connections without ever having seen a physical instrument producing the sounds, simply by listening to recordings.

Despite the current lack of a comprehensive theory of timbre, it is highly probable that such a theory will eventually be based on data obtained from time-varying spectrum analysis. Section 1 of this chapter examines some useful methods for analysis and synthesis of musical sounds based on the short-time Fourier transform. Section 2 investigates various characteristics of instrumental sound spectra in an effort to gain an understanding of that which makes different musical sounds sound different, i.e., how they might evoke unique timbres. Throughout, the SNDAN analysis/synthesis software package (Beauchamp, 1993) is used to illustrate examples of musical sound spectral analysis.

1 Analysis/Synthesis Methods

While mathematical representations of musical instrument sounds are not unique, it is very useful to represent such sounds as a collection of sine waves (sinusoids) with time-varying amplitudes, frequencies, and phases and possibly also with an additive noise signal having certain time-varying spectral properties. With this model, it is assumed that a musical sound signal $s(t)$ can be expressed as

$$s(t) = \sum_{k=1}^{K(t)} A_k(t) \cos(\theta_k(t)) + n(t), \quad (1.1a)$$

where

$$\theta_k(t) = 2\pi \int_0^t f_k(\tau) d\tau + \theta_{k_0}. \quad (1.1b)$$

The various parameters are defined as follows:

t = time.

$A_k(t)$ = amplitude of the k th sine wave (frequency component or partial) at time t .

k = partial number.

$K(t)$ = number of sinusoidal partials, which may vary with time.

$\theta_k(t)$ = phase of partial k at time t .

$f_k(t)$ = frequency of partial k at time t .

$\theta_{k_0} = \theta_k(0)$ = initial phase of partial k (phase at time = 0).

$n(t)$ = additive noise signal, whose short-term spectrum varies with time.

The instantaneous phase of each partial is intrinsically bound to its initial phase and its instantaneous frequency. Given the starting phase and the frequency (the phase derivative), the phase is known, at least theoretically, at each instant of time. Note that if the time scale or frequencies are altered, the relative phases among the partials will change.

The noise term $n(t)$ can be omitted from the model if the noise is considered to be embedded in the individual partials. The decision about whether noise should be separate from the sinusoids or contained within them depends on the type of analysis used, the nature of the noise, and convenience when doing the synthesis, especially if modifications such as time-stretching are to be done. In most of the examples presented in this chapter, noise will be assumed to be embedded in the amplitude and frequency time functions for the individual partials. Therefore, with this assumption, a musical instrument signal can be represented strictly as

$$s(t) = \sum_{k=1}^{K(t)} A_k(t) \cos(2\pi \int_0^t f_k(\tau) d\tau + \theta_{k_0}). \quad (1.2)$$

What remains, given the representation of Eq. 1.2, is to estimate its various parameters, namely, $K(t)$, $A_k(t)$, $f_k(t)$, and θ_{k_0} for $1 \leq k \leq K$. In this chapter, two different methods of analysis, both of which are examples of short-time Fourier analysis, are presented. One is called the harmonic filter bank or phase vocoder method and the other the frequency-tracking or McAulay–Quatieri (MQ) method.

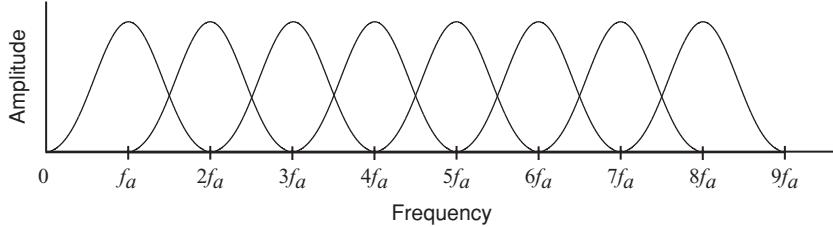


FIGURE 1.1. Overlapping band-pass analysis filter responses centered at harmonics of f_a .

1.1 Harmonic Filter Bank (Phase Vocoder) Analysis/Synthesis

Harmonic filter bank or phase vocoder analysis simulates a bank of overlapping band-pass filters each centered on an integer multiple of a base frequency f_a ; i.e., at harmonic frequencies $f_k = kf_a$, for $k = 1, \dots, K$, where f_a is referred to as the analysis frequency, and K is a constant number of harmonics. Each filter function $W_k(f - f_k)$ has a maximum value of unity at $f = kf_a$. Also, each filter function is zero or very small for $f \leq (k-1)f_a$ and $f \geq (k+1)f_a$. Such a filter bank, consisting of a series of overlapping bell-shaped curves, one for each band-pass filter, is depicted in Fig. 1.1. This filter bank has the special property that for a periodic signal with constant fundamental frequency exactly at f_a and fixed harmonic amplitudes A_k , each filter will produce a sine wave with frequency $f_k = kf_a$ and amplitude A_k , i.e.,

$$s_k(t) = A_k \cos(2\pi kf_a t + \theta_{k_0}). \quad (1.3)$$

1.1.1 Frequency Deviation and Inharmonicity

If, on the other hand, the amplitudes and frequencies are allowed to vary with time (but not too fast!) and each k th harmonic frequency is confined to a narrow range around kf_a , the filter outputs will closely—although not perfectly—replicate the terms in the summation of Eq. 1.2. In this case, it is useful to define

$$f_k(t) = kf_a + \Delta f_k(t), \quad (1.4a)$$

where $\Delta f_k(t)$ is a time-varying frequency deviation.

The frequency deviation can be written as

$$\Delta f_k(t) = f_k(t) - kf_a, \quad (1.4b)$$

and the relative frequency deviation as

$$\frac{\Delta f_k(t)}{k} = \frac{f_k(t)}{k} - f_a. \quad (1.4c)$$

Also useful is the normalized frequency deviation

$$\frac{\Delta f_k(t)}{kf_a} = \frac{f_k(t)}{kf_a} - 1, \quad (1.4d)$$

which gives the fractional deviation of a frequency with respect to its harmonic value. For example, if $\Delta f_k / kf_a$ varies by ± 0.06 (or 6%), the k th harmonic frequency varies upward and downward by approximately one semitone with respect to its center position, kf_a . A well-known measure of microtonal pitch is the logarithmic cents measure, where there are 100 cents per semitone. Normalized frequency deviation can be expressed in terms of cents deviation using the formula

$$\Delta \text{cents}(t) = 1200 \cdot \log_2 \left(\frac{\Delta f_k(t)}{kf_a} \right). \quad (1.4e)$$

A sound is instantaneously harmonic if all frequencies track one another such that

$$\Delta f_k(t) = k \Delta f_1(t), \quad (1.5a)$$

which leads to a definition of *inharmonicity*:

$$I_k(t) = \frac{\Delta f_k(t)}{k \Delta f_1(t)} - 1. \quad (1.5b)$$

In practice, if the amplitude of the first harmonic is too small, Δf_1 may be poorly defined, and Eq. (1.5b) may result in a poor estimate of inharmonicity. To circumvent this problem, a composite fundamental frequency deviation is defined as

$$\Delta f_{c1}(t) = \frac{\sum_{k=1}^5 A_k(t) \Delta f_k(t) / k}{\sum_{k=1}^5 A_k(t)}, \quad (1.5c)$$

which is the relative-amplitude-weighted sum of the harmonic-normalized first five harmonic frequency deviations. This is an ad hoc formula based on research on the relative dominance of low harmonics for determining pitch (e.g., Moore et al., 1985) and the observation that most musical instruments have their strongest harmonics within the first five. Note that if all the harmonic amplitudes are equal, the ordinary average of the relative frequency deviations results. But with unequal amplitudes, stronger amplitudes dominate the formula. Thus, for cases where A_1 is weak, Δf_{c1} should be substituted for Δf_1 in Eq. (1.5b).

Owing to analysis and signal imperfections, some small amount of inharmonicity will appear to be present in the analysis of the most harmonious of tones. However, Eq. (1.5b) is especially useful for cases when the signal has appreciable amounts of inharmonicity.

A problem arises when the frequencies of a sound to be analyzed have too much deviation from harmonic frequency values, whether it be due to frequency modulations or long-term inharmonicity. For the harmonic case, a fundamental frequency that deviates by Δf_1 from f_a translates into a change of $k\Delta f_1$ from kf_a , which is the center frequency of the k th harmonic analysis filter, also called the k th *bin*. When $k\Delta f_1 \geq 0.5 f_a$, the k th frequency component is reported from the $(k + 1)$ st bin with as much or greater amplitude than from the k th bin. Meanwhile, the k th harmonic bin's output will also include the effect of the $(k - 1)$ st harmonic.

Thus, while a moderate amount of fundamental frequency deviation typically does not cause appreciable analysis error in the lower harmonics, at a certain harmonic the analysis accuracy for the upper partials will be affected. This is a basic limitation of the harmonic filter bank approach.

1.1.2 Heterodyne-Filter Analysis Method

The filter-bank analyzer is implemented by a method known by various names (e.g., phase vocoder, short-time Fourier transform) including the heterodyne filter method (Beauchamp and Fornango, 1966; Beauchamp, 1969), which is derived from traditional Fourier series analysis. Accordingly, the complex amplitude of the k th harmonic of $s(t)$ is given by

$$\tilde{c}_k(t) = \int_{-\infty}^{\infty} w(t - \tau) e^{-j2\pi kf_a \tau} s(\tau) d\tau, \quad (1.6a)$$

where $w(t)$ is the impulse response of a low-pass filter. Equation (1.6a) can be interpreted as being the combination of two operations:

- (1) Heterodyne (i.e., multiplication) of the signal $s(t)$ by the complex exponential function $e^{-j2\pi kf_a t}$ [which can also be written as $\cos(2\pi kf_a t) - j \sin(2\pi kf_a t)$], where f_a is the analysis frequency.
- (2) Low-pass filtering of this product by convolution with a special “window” function $w(t)$, which in general is an even function of t .

The heterodyne operation shifts the frequency kf_a within $s(t)$ to $f = 0$ and frequencies in the vicinity of kf_a to the vicinity of zero. Then the low-pass filter attempts to remove all components except those whose frequencies are less than $f_a/2$. To illustrate, let's define

$$s'_k(t) = e^{-j2\pi kf_a t} s(t) \quad (1.6b)$$

as the heterodynized signal. Then the low-pass operation can be accomplished by

$$\tilde{c}_k(t) = w(t) * s'_k(t), \quad (1.6c)$$

where ‘ $*$ ’ indicates convolution. In terms of Fourier transforms Eq. (1.6c) becomes

$$\tilde{C}_k(f) = W(f) S'_k(f) = W(f) S(f + kf_a). \quad (1.6d)$$

The Fourier transform of $w(t)$, $W(f)$, is also known as the frequency response or the filter characteristic of $w(t)$, whereas $S(f)$ is the spectral characteristic, or simply the spectrum, of $s(t)$. Because the low-pass region of $W(f)$ corresponds to the frequency range $(-0.5 f_a, 0.5 f_a)$ and the frequency range $((k - 0.5) f_a, (k + 0.5) f_a)$ of $S(f)$ has been translated to this region by virtue of $S(f + kf_a)$, $\tilde{C}_k(f)$ ideally contains only the portion of $S(f)$ corresponding to a $\pm 0.5 f_a$ band around kf_a , and, consequently, is the equivalent of the output of a symmetric band-pass filter.

1.1.2.1 Window Functions

Window functions are particular versions of $w(t)$ that are time-limited and whose Fourier transforms have “nice” low-pass characteristics. These functions are

referred to as window functions or simply windows because they can be visualized as providing a “window” on a particular segment of the signal. These functions are therefore zero outside a time interval $-T \leq t \leq T$. They are also even functions [i.e., $w(t) = w(-t)$], with the result that their Fourier transforms are real and their phase responses are zero.

The simplest possible window is the rectangular window, which for our application is defined as

$$w(t) = \begin{cases} f_a, & |t| \leq 0.5/f_a \\ 0, & |t| > 0.5/f_a \end{cases}. \quad (1.7a)$$

Note that f_a , the analysis frequency, is associated with the height and width of the window. In this case, the window width is $1/f_a$, and, because the height is f_a , the area of the window is 1.0. [Other windows will be given in terms of $w(t)/f_a$ in order to simplify the formulas.] In comparison to other useful window functions, the rectangular window has a very inferior response for $f > f_a$. However, in a certain sense, it does afford the best time resolution.

A much better and very convenient window function is the hanning (aka Hann) window:

$$\frac{w(t)}{f_a} = \begin{cases} \cos^2(0.5\pi t f_a) = 0.5 + 0.5 \cos(\pi t f_a), & |t| \leq 1/f_a \\ 0, & |t| > 1/f_a \end{cases} \quad (1.7b)$$

The width of this window is $2/f_a$, its peak amplitude is again f_a , and its area is again 1.0.

A variation on this window function is the Hamming window:

$$\frac{w(t)}{f_a} = \begin{cases} 0.5 + 0.426 \cos(\pi t f_a), & |t| \leq 1/f_a \\ 0, & |t| > 1/f_a \end{cases} \quad (1.7c)$$

Like the hanning, the Hamming is a 2-term window function having a window width $2/f_a$, but with a peak amplitude of $0.926 f_a$. Note the discontinuity at $t = \pm 1/f_a$. Its area is again 1.0.

A more sophisticated window function is the 4-term Blackman–Harris window:

$$\frac{w(t)}{f_a} = \begin{cases} 0.25 + 0.3403 \cos(0.5\pi t f_a) + 0.0985 \cos(\pi t f_a), & |t| \leq 2/f_a \\ +0.0081 \cos(1.5\pi t f_a), & |t| > 2/f_a \\ 0, & |t| > 2/f_a \end{cases} \quad (1.7d)$$

The width of this window is $4/f_a$, and its peak amplitude is $0.6969 f_a$. Again the area is 1.0.

More details on the behavior of these window functions are given in Harris (1978) and Nuttal (1981). Figure 1.2a compares the four window functions given above (normalized by f_a). They can be generalized to the form:

$$\frac{w(t)}{f_a} = \begin{cases} \sum_{p=0}^{P-1} \alpha_p \cos(2\pi p f_a t / P), & |t| \leq \frac{P}{2f_a} \\ 0, & |t| > \frac{P}{2f_a} \end{cases}, \quad (1.8)$$

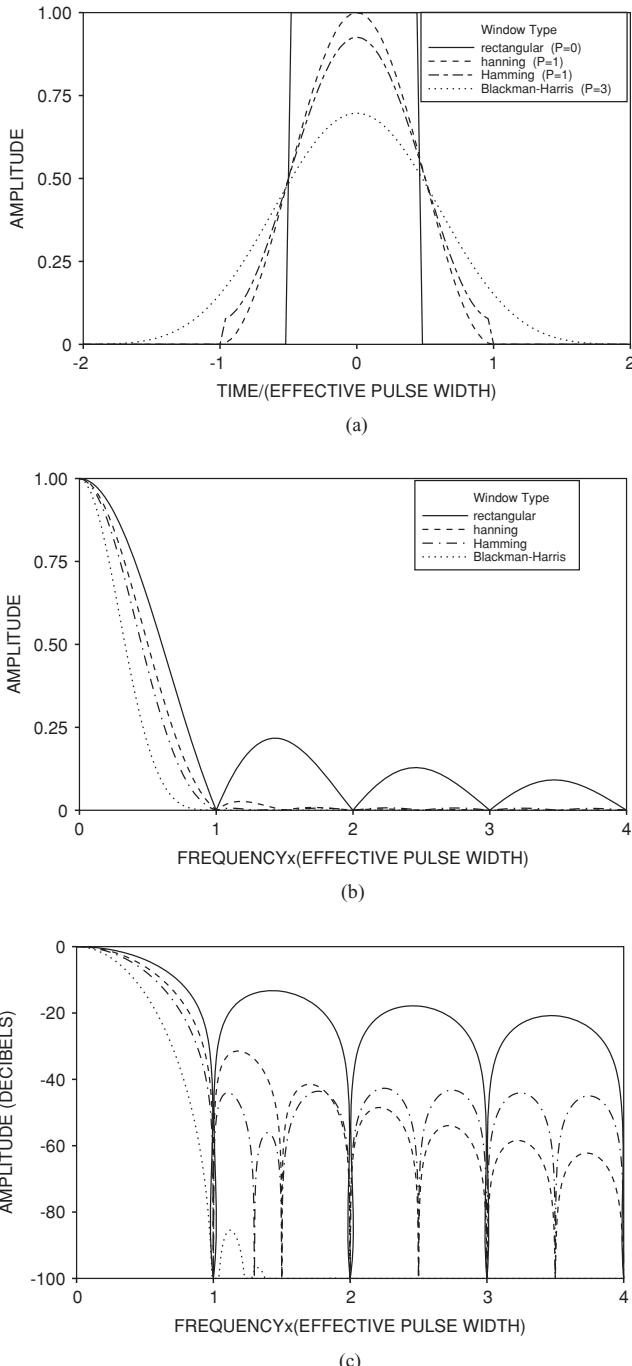


FIGURE 1.2. Comparison of four window types: rectangular, hanning, Hamming, and Blackman-Harris. (a) Normalized window functions, $w(f_a)/f_a$; (b) Normalized window frequency responses, $W(f/f_a)$; (c) Window responses in decibels, $20\log(W(f/f_a))$.

where P is the number of terms in the summation and $\alpha_0 = 1/P$. Then, Eq. (1.6a) for this class of window functions can be written

$$\tilde{c}_k(t) = f_a \int_{t-\frac{P}{2f_a}}^{t+\frac{P}{2f_a}} \frac{w(t-\tau)}{f_a} e^{-j2\pi kf_a\tau} s(\tau) d\tau \quad (1.9a)$$

$$= f_a \sum_{p=0}^{P-1} \alpha_p \int_{t-\frac{P}{2f_a}}^{t+\frac{P}{2f_a}} \cos(2\pi p f_a(t-\tau)/P) e^{-j2\pi kf_a\tau} s(\tau) d\tau. \quad (1.9b)$$

The frequency responses of these window functions can be calculated easily by taking their Fourier transforms according to

$$W(f) = \int_{-\infty}^{\infty} w(\tau) e^{-j2\pi f\tau} d\tau \quad (1.10a)$$

$$= f_a \sum_{p=0}^{P-1} \alpha_p \int_{t-\frac{P}{2f_a}}^{t+\frac{P}{2f_a}} \cos(2\pi p f_a \tau / P) e^{-j2\pi kf\tau} d\tau, \quad (1.10b)$$

where f is the frequency.

Knowing that

$$\begin{aligned} \int_{-T}^T \cos(\beta\tau) e^{-j\omega\tau} d\tau &= \frac{\sin((\omega + \beta)T)}{\omega + \beta} + \frac{\sin((\omega - \beta)T)}{\omega - \beta} \\ &= T [\text{sinc}((\omega + \beta)T) + \text{sinc}((\omega - \beta)T)], \end{aligned} \quad (1.11)$$

and taking $\omega = 2\pi f$, $T = P/(2f_a)$, and $\beta = 2\pi p f_a / P$, the general formula for the frequency response becomes

$$W(f) = \frac{P}{2} \sum_{p=0}^{P-1} \alpha_p \left(\text{sinc}\left(\pi \left(\frac{Pf}{f_a} + p\right)\right) + \text{sinc}\left(\pi \left(\frac{Pf}{f_a} - p\right)\right) \right). \quad (1.12a)$$

From Eq. (1.12a), considering that $\alpha_0 = 1/P$ and $f = 0$, it follows that $H(0) = P\alpha_0 = 1.0$, the maximum value of the response. Also, if the frequency is a harmonic of f_a , i.e., $f = kf_a$, $k = 1, 2, 3, \dots$, it can be seen that $W(kf_a) = 0$. The first zero, which occurs at $f = f_a$, defines the end of the low-frequency response. Because of the zero positions, this type of response is perfect for analysis of absolutely periodic signals having fundamental frequency f_a . Another interesting result is that for $f = qf_a/P$, $q = 1, 2, \dots, P-1$, $W(qf_a/P) = 0.5P\alpha_q$. This allows a quick calculation of some frequency-response values for $0 < f < f_a$ (the “pass band”) in terms of the window function coefficients. Note that the decibel equivalent of W , $W_{\text{db}} = 20 \log_{10}(W)$, is zero for $f = 0$ and less than zero for $f > 0$. The “half-way” pass-band values at $W_{\text{db}}(f_a/2)$ are, respectively, -3.9 , -6.0 , -7.4 , and -20.1 dB for the rectangular, hanning, Hamming, and 4-term Blackman–Harris windows.

Equation (1.12a) can also be written (Nuttall, 1981) as

$$W(f) = P \operatorname{sinc} \left(\frac{\pi Pf}{f_a} \right) \sum_{p=0}^{P-1} \frac{(-1)^p \alpha_p}{1 - \left(\frac{pf_a}{Pf} \right)^2}. \quad (1.12b)$$

The sinc function shows that for $f \geq f_a$ (the “stop band”) the response has zeros which are separated by f_a/P . (Zeros for $f < f_a$ are cancelled by singularities due to particular summation term denominators.) Of particular importance is the response for frequencies halfway between the zero frequencies above f_a , i.e., the “half-way” stop-band values at $W((q + .5)f_a/P)$ for $q = P, P + 1, P + 2, \dots$. These values, which are hopefully small, give an idea of how well the filter rejects unwanted frequencies. It turns out that these maximum stop-band responses (in terms of W_{db}) for the rectangular, hanning, Hamming, and 4-term Blackman–Harris windows are, respectively, $-13.5, -31.5, -43.2$, and -92.0 dB. The $W(f)$ and $W_{\text{db}}(f)$ responses are compared in Figs. 1.2b and 1.2c.

Another very useful window function is the Kaiser–Bessel window (Kaiser and Schafer, 1980; Harris, 1978; Nuttall, 1981), which is defined in the time domain by

$$w(t) = \frac{1}{T} \frac{\alpha}{\sinh(\alpha)} I_o \left(\alpha \sqrt{1 - (2t/T)^2} \right), \quad |t| < \frac{T}{2}, \quad (1.13a)$$

where I_o is the zeroth-order modified Bessel function of the first kind, α is a fixed parameter, and T is the window width. By varying α , different frequency responses can be achieved. The general formula for the frequency response is

$$W(f) = \frac{\sinh(\alpha \sqrt{1 - (\pi Tf/\alpha)^2})}{\sinh(\alpha) \sqrt{1 - (\pi Tf/\alpha)^2}}. \quad (1.13b)$$

When $\pi Tf/\alpha > 1$, the square roots of this rather peculiar function become imaginary and the numerator sinh function turns into a sin function. When $\pi Tf/\alpha = 1$, the roots are zero, and $W(f) = \alpha / \sinh(\alpha)$. The first zero occurs when the argument of the sin is π , and this leads to $f_o = \sqrt{1 + (\alpha/\pi)^2}/T$. For $f > f_o$ the function approximately follows a sinc function:

$$W(f) = \frac{\alpha}{\sinh(\alpha)} \operatorname{sinc} \left(\alpha \sqrt{(\pi Tf/\alpha)^2 - 1} \right) \approx \frac{\alpha}{\sinh(\alpha)} \operatorname{sinc}(\pi Tf). \quad (1.13c)$$

So for a given window width T , the first zero frequency and the amount of stop-band rejection depends on the value of the parameter α , and there is a trade-off between the two. For example, to mimic a Hamming window, taking $\alpha = 5.441$ forces $f_o = 2/T$. The minimum stop-band attenuation is approximately 40 dB, which is comparable to the Hamming. If the first zero is moved to $f_o = 4/T$, the stop-band attenuation becomes approximately 92 dB, like the 4-term Blackman–Harris. Two other things are obvious from Eq. (1.13c): (1) The sidelobe peak values of the Kaiser–Bessel window are spaced by $1/T$, with zero values half-way in between. (2) The peak values decrease in amplitude by -6 dB/octave. But,

most importantly, the Kaiser–Bessel window is a chameleon that can mimic other optimum windows depending on the value of α .

1.1.2.2 Harmonic Analysis Limits

An important problem occurs when the input fundamental frequency is detuned from f_a by an amount Δf . Then harmonic k is detuned by $f_{k1} = k\Delta f$, and this becomes the output frequency after heterodyning by kf_a , as opposed to zero which occurs when tuning is perfect. Meanwhile, the neighboring harmonics, which should be rejected, have frequencies at $(k - 1)(f_a + \Delta f)$ and $(k + 1)(f_a + \Delta f)$, and after heterodyning by kf_a these frequencies become $f_{k2} = -f_a + (k - 1)\Delta f$ and $f_{k3} = f_a + (k + 1)\Delta f$, respectively. Thus, analysis accuracy can be measured by taking the difference between the amplitude of the desired harmonic k and the amplitudes of the undesired harmonics $k - 1$ and $k + 1$, which may corrupt the k th harmonic amplitude measure. This is tantamount to comparing $W_{\text{db}}(f_{k1})$ with $W_{\text{db}}(f_{k2})$ and $W_{\text{db}}(f_{k3})$. (Note again, that if $\Delta f = 0$, there is no problem!) $W_{\text{db}}(f_{k1}) - W_{\text{db}}(f_{k2})$ and $W_{\text{db}}(f_{k1}) - W_{\text{db}}(f_{k3})$ give measures of the relative rejection of the unwanted components.

To take a concrete example, let $\Delta f = 0.03 f_a$ (approximately a half-semitone) and $k = 3$ (third harmonic). Then $f_{31} = 0.09 f_a$, $f_{32} = -0.94 f_a$, and $f_{33} = 1.12 f_a$. The rectangular, hanning, Hamming, and 4-term Blackman–Harris window responses (in decibels) are compared in the following table:

Window Type	$W_{\text{db}}(f_{31} = 0.09 f_a)$	$W_{\text{db}}(f_{32} = -0.94 f_a)$	$W_{\text{db}}(f_{33} = 1.12 f_a)$	$W_{\text{db}}(f_{31}) - W_{\text{db}}(f_{32})$	$W_{\text{db}}(f_{31}) - W_{\text{db}}(f_{33})$
rectangular	-0.1	-24.0	-19.6	23.9	19.5
hanning	-0.2	-32.2	-32.3	32.0	36.1
Hamming	-0.2	-38.6	-44.1	38.4	43.9
Blackman–Harris	-0.4	-70.5	-92.1	70.1	91.7

For another example, again let $\Delta f = 0.03 f_a$ and take $k = 10$ (tenth harmonic). Then $f_1 = 0.3 f_a$, $f_2 = 0.73 f_a$, and $f_3 = 1.33 f_a$. The four window responses are now:

Window Type	$W_{\text{db}}(f_1 = 0.3 f_a)$	$W_{\text{db}}(f_2 = 0.73 f_a)$	$W_{\text{db}}(f_3 = 1.33 f_a)$	$W_{\text{db}}(f_1) - W_{\text{db}}(f_2)$	$W_{\text{db}}(f_1) - W_{\text{db}}(f_3)$
rectangular	-1.3	-9.7	-13.7	8.4	12.4
hanning	-2.1	-14.4	-35.3	12.3	33.2
Hamming	-2.5	-17.7	-61.8	15.2	59.3
Blackman–Harris	-4.9	-33.4	-115.7	28.5	110.8

It should be clear from these numbers that it is more difficult to isolate a higher harmonic. For positive mistuning, harmonic $k - 1$ causes more corruption of harmonic k than harmonic $k + 1$ does. (However, for negative mistuning the opposite is true.) Also, isolation of a harmonic improves with the sophistication of the window

type. For example, the 4-term Blackman–Harris is better than the Hamming, the Hamming is better than the hanning, and the hanning is better than the rectangular. However, there are at least a couple more issues to consider in determining the best window.

For one thing, the hanning-vs-Hamming tradeoff comes out differently if the corruption caused by several harmonics surrounding the one being analyzed are considered. That is, harmonic k with frequency $k(f_a + \Delta f)$ can be corrupted by harmonics $\dots, k - 3, k - 2, k - 1, k + 1, k + 2, k + 3, \dots$ having frequencies $\dots, (k - 3)(f_a + \Delta f), (k - 2)(f_a + \Delta f), (k - 3)(f_a + \Delta f), (k + 1)(f_a + \Delta f), (k + 2)(f_a + \Delta f), (k + 3)(f_a + \Delta f), \dots$, not just the immediate neighbors of harmonic k . From Fig. 1.2c, it is evident that the hanning response function provides better rejection than the Hamming for $f/f_a > 2$, which should reduce the corruption of non-immediate-neighbor harmonics. Thus, the best window to use depends on the nature of the signal’s spectrum and the particular harmonic to be analyzed.

Another concern is the narrowness of the 4-term Blackman–Harris response for the pass-band region $0 \leq f/f_a \leq 1$ and the corresponding opulent time-domain width ($4/f_a$) of its window function [see Eq. (1.7d)]. Even though its response side lobes are lower, its main lobe is more sensitive to frequency detuning than the other window functions. Also, the Blackman–Harris’s relatively wide time-window can cause time-resolution problems with attendant loss of some detail. Thus, the hanning and Hamming window functions, in addition to being somewhat cheaper to compute, have some possible accuracy advantages over the 4-term Blackman–Harris.

In connection with filter response functions, one might ask “Why not use an ideal rectangular filter response?” Such a response, in its low-pass form, is defined as

$$W(f) = \begin{cases} 1.0, & |f| \leq 0.5f_a \\ 0, & |f| > 0.5f_a \end{cases} \quad (1.14a)$$

and gives ideal results in the frequency domain in that it perfectly separates harmonics of f_a , is relatively impervious to frequency changes, and still yields a summed response of 1.0. However, this $W(f)$ corresponds to the time-window function

$$\frac{w(t)}{f_a} = \frac{\sin(\pi f_a t)}{\pi t f_a}, \quad (1.14b)$$

which is not time-limited and converges very slowly to zero as time increases. While this window performs very precisely in the frequency domain, it would give rise to much time-domain distortion.

1.1.2.3 Synthesis from Harmonic Amplitudes and Frequency Deviations

According to Fourier series theory, an analyzed signal can be synthesized using

$$\hat{s}(t) = \sum_{k=-\infty}^{\infty} \tilde{s}_k(t) = \sum_{k=-\infty}^{\infty} \tilde{c}_k(t)e^{j2\pi kf_a t} \quad (1.15a)$$

$$= c_0(t) + \sum_{k=1}^{\infty} (\tilde{c}_k(t)e^{j2\pi kf_a t} + \tilde{c}_{-k}(t)e^{-j2\pi kf_a t}) \quad (1.15b)$$

$$= A_0(t) + \sum_{k=1}^{\infty} A_k(t) \cos(2\pi kf_a t + \theta_k(t)) \quad (1.15c)$$

$$= A_0(t) + \sum_{k=1}^{\infty} A_k(t) \cos(2\pi(kf_a t + \int_0^t \Delta f_k(t) dt) + \theta_{k0}), \quad (1.15d)$$

where for $k \geq 1$,

$$A_k(t) = 2 |\tilde{c}_k(t)| = 2\sqrt{(\text{Re}(\tilde{c}_k(t)))^2 + (\text{Im}(\tilde{c}_k(t)))^2}, \quad (1.15e)$$

$$\theta_k(t) = \text{atan2}(\text{Im}(\tilde{c}_k(t)), \text{Re}(\tilde{c}_k(t))), \quad (1.15f)$$

$$\theta_{k0} = \theta_k(0), \quad (1.15g)$$

$$\Delta f_k(t) = \frac{1}{2\pi} \frac{d\theta_k(t)}{dt}, \quad (1.15h)$$

and where $\tilde{c}_k(t)$ is defined in Eq. (1.6a).

Note: $\text{atan2}(y,x)$ is a function that is available in C and other computer languages, and unlike $\text{atan}(y/x)$, it correctly computes the angle (in radians) of the phasor $x + jy$, even when x is nonpositive.

Eq. (1.15d) gives a general equation for time-varying synthesis, and $A_k(t)$, $\Delta f_k(t)$, and θ_{k0} are the parameters that must be known in order for synthesis to proceed. However, for real audio signals, only a finite number of harmonics (K) are needed. This is usually given by

$$K = \text{floor}(0.5 f_s / f_a), \quad (1.15i)$$

where $0.5 f_s$ is the Nyquist or half-sample frequency (see Section 1.1.4).

1.1.3 Signal Reconstruction (Resynthesis) and the Band-Pass Filter Bank Equivalent

Theoretically, if all band-pass filter outputs are combined, the original signal can be accurately reconstructed, regardless of whether the signal's frequency components line up with the filter center frequencies, using

$$\hat{s}(t) = \sum_{k=-\infty}^{\infty} \tilde{s}_k(t) = s_0(t) + \sum_{k=1}^{\infty} (\tilde{s}_k(t) + \tilde{s}_{-k}(t)) \approx s(t). \quad (1.16a)$$

The near identity of Eq. 1.16a becomes a true or approximate identity if the equivalent band-pass filter transfer functions add up to 1.0 or close to it—assuming that, in practice, one is careful to retain the proper phases of the components in the

synthesis process, or else improper phase cancellations will occur. These filters can be derived as follows: Substituting the definition for $\tilde{c}_k(t)$ given in Eq. (1.6a) into Eq. (1.15a) gives

$$\hat{s}(t) = \sum_{k=-\infty}^{\infty} e^{j2\pi kf_a t} \sum_{-\infty}^{\infty} w(t-\tau) e^{-j2\pi kf_a \tau} s(\tau) d\tau \quad (1.16b)$$

$$= \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} w(t-\tau) e^{j2\pi kf_a(t-\tau)} s(\tau) d\tau \quad (1.16c)$$

$$= \sum_{k=-\infty}^{\infty} (w(t)e^{j2\pi f_a t}) * s(t). \quad (1.16d)$$

Taking the Fourier transform yields:

$$\hat{S}(f) = \sum_{k=-\infty}^{\infty} W(f - kf_a) S(f) = S(f) \sum_{k=-\infty}^{\infty} W(f - kf_a) = S(f) W_{\text{sum}}(f). \quad (1.16e)$$

Note that $W(f - kf_a)$ is the band-pass transformation of the low-pass window response function, i.e., the $W(f)$ response function has just been shifted to the right by amount kf_a . Thus, whether the synthesized signal $\hat{s}(t)$ equals the original signal $s(t)$ hinges on whether the sum of the shifted window functions, which form a harmonic filter bank (see Fig. 1.1), adds up to unity [i.e., $W_{\text{sum}}(f) \equiv 1.0$]. Figure 1.3 shows such sums for four window functions discussed above, where frequency is normalized by f_a . Each sum consists of 25 individual band-pass filter responses, ranging from $k = -10$ to $k = 15$, but only $0 \leq k \leq 5$ is shown. For an infinite number of filters, the rectangular-window summed response is theoretically 1.0, independent of frequency, but it is slow to converge and shows some variation for a finite number. The Hanning window response converges rapidly to 1.0. The Hamming window response exhibits a 1.4 dB ripple, which is probably difficult to detect aurally. However, the 4-term Blackman–Harris window response varies by 8.1 dB between band centers and half-band centers. Thus, as discussed above, the Blackman–Harris window, which gives perfect results for a perfectly periodic signal of frequency f_a , does not perform well for the harmonic filter bank on periodic signals whose frequencies vary substantially from harmonics of f_a . However, it must be said that if the Blackman–Harris window were narrowed somewhat, thus widening its low-pass frequency response, a better overall effect could be obtained, even though the responses at the harmonics of f_a would no longer be zero.

1.1.4 Sampled Signal Implementation

Although an analog implementation of the continuous-time analyzer described above is a possibility, a sampled signal implementation on a computer is much more practical. This requires that the signal be stored as a series of samples $s(n/f_s)$, $n = 0, 1, 2, \dots$, where f_s is the sample frequency. Of course, input of samples from an analog source requires an analog-to-digital converter (ADC), and playback

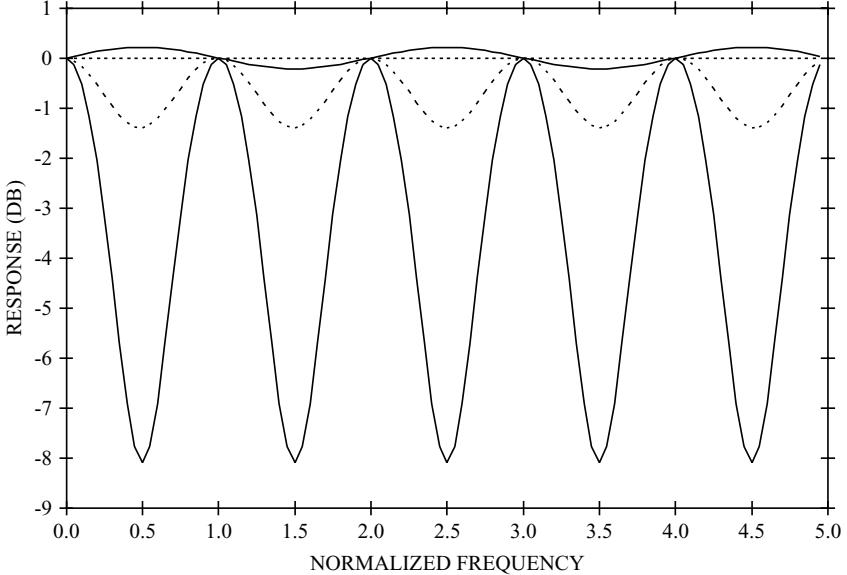


FIGURE 1.3. Sums of overlapping band-pass analysis filter responses $W((f - kf_a)/f_a)$ for $k = -10, -9, \dots, 15$: rectangular (upper solid curve), hanning (dotted curve), Hamming (dashed curve), and Blackman–Harris (lower solid curve). These give the overall resynthesis frequency responses to an arbitrary input signal.

from the computer requires a digital-to-analog converter (DAC). A typical sample frequency, which is frequently used in computer applications and for compact discs (CDs), is 44,100 Hz. This is high enough that signal frequencies up to 20,000 Hz, a frequency roughly corresponding to the upper limit of human hearing, are well resolved.

1.1.4.1 Analysis Step

The objective of the analysis step is to compute the starting phases, amplitudes, and frequency deviations of K harmonics of the input signal at a series of time frames i , which occur at a rate considerably lower than the sample rate. For our phase-vocoder method the frame rate is equal to $f_a/2$. Computation corresponds to the sampled equivalent of Eq. (1.9). First, sample numbers n and m are defined to be series of integers that define times at $t_n = n/f_s$ and $\tau_m = m/f_s$. Then, substituting $t \leftarrow t_n$ and $\tau \leftarrow \tau_m$ in Eq. (1.9) gives:

$$\tilde{c}_k(n/f_s) = f_a \sum_{m=-N/2}^{n+N/2-1} w'((n-m)/f_s) e^{-j2\pi kf_a m/f_s} s(m/f_s)/f_s, \quad (1.17a)$$

where $N \cong Pf_s/f_a$ is the length of the window function w' in samples and w' is the normalized version of w , i.e., $w'() = w()/f_a$. Eq. (1.17a) gives the complex amplitude of the k th harmonic. Note that all of the time functions in this definition

are sampled at intervals of $1/f_s$. For the window functions discussed above, recall that $P = 1$ for the rectangular window, $P = 2$ for the Hamming and hanning window functions, and $P = 4$ for the fourth-order Blackman–Harris window, so that N corresponds to 1, 2, or 4 periods of the frequency f_a . The center of the window function occurs when $n = m$. For convenience, $\tilde{c}_k(n/f_s)$ can be replaced by $\tilde{c}_k(n)$, $w'((n - m)/f_s)$ by $w'(n - m)$, and $s(m/f_s)$ by $s(m)$, so that Eq. (1.17a) now reads as:

$$\tilde{c}_k(n) = \frac{f_a}{f_s} \sum_{m=n-N/2}^{n+N/2-1} w'(n - m) e^{-j2\pi km f_a / f_s} s(m) \quad (1.17b)$$

$$= \frac{P}{N} \sum_{m=n-N/2}^{n+N/2-1} w'(n - m) e^{-j2\pi Pkm/N} s(m). \quad (1.17c)$$

Equation (1.17c) can be thought of as a discrete approximation to Eq. (1.9) in terms of the sum of N values, where N is an even number. As can be made obvious by using a small value of N (e.g., $N = 4$), this formula represents an assymetrical sampling of $w'()$, with $N/2$ points to the left of the middle and $N/2 - 1$ points to the right. This can be easily fixed with a slight shift of the $w'()$ function, by 0.5 point. Also, the fast Fourier transform (FFT) is usually used for computation, which for most algorithms means that (1) Eq. (1.17c) should be in the form of the discrete Fourier transform (DFT) and (2) N should be a power of 2.

For the requirement that N be a power-of-two, the signal $s(n)$ must be resampled in order to produce exactly $N = 2^M$ points, where M is an integer. In order to avoid undersampling, the signal must be resampled at a higher sample rate $f'_s \geq f_s$. Therefore, let

$$N = 2^M = 2^{\text{ceil}(\log_2(Pf_s/f_a))}, \quad (1.18a)$$

and the new sample rate becomes

$$f'_s = \frac{Nf_a}{P}. \quad (1.18b)$$

For example, if a 261.6 Hz (middle C) tone is digitized at a 44,100 Hz sample rate and analyzed using a Hamming window ($P = 2$) of width $2/f_a$, then $Pf_s/f_a = 337.16$, $N = 512$, and the new sample rate is 66,969.6 Hz. Several methods for changing the sample rate are available [e.g., see Smith and Gossett (1984)]. The method programmed by Maher (1989) for use in the SNDAN analysis/synthesis package (Beauchamp, 1993) convolves a Hamming-windowed sinc function with the input signal, and the upsampled result is linearly interpolated.

For the DFT requirement, the substitution $m \leftarrow m + n - N/2$ is made in Eq. (1.17c), resulting in

$$\tilde{c}_k(n) = e^{j\pi k P(1-2n/N)} \frac{P}{N} \sum_{m=0}^{N-1} w'(N/2 - m) s(m + n - N/2) e^{-j2\pi k Pm/N}. \quad (1.19a)$$

With $P = 1$, the summation of Eq. (1.19a) is in the correct form for the DFT. However, with $P > 1$, Eq. (1.19a) indicates analysis only at frequencies P/N ,

$2P/N, 3P/N, \dots$, whereas the DFT is defined for all frequencies $0, 1/N, 2/N, 3/N, \dots$. This problem is solved by taking the FFT for all frequencies (replacing kP by k' and letting $k' = 0, 1, 2, \dots$) and then retaining only the components needed (i.e., $k' = P, 2P, 3P, \dots$). For example, for $P = 2$, all of the harmonics of $0.5f_a$ are first computed, and then the odd-numbered components are thrown away while keeping the even-numbered ones, which correspond to harmonics of f_a . Thus, if the DFT is defined as:

$$X(n, k') = \sum_{m=0}^{N-1} w'(N/2 - m)s(m + n - N/2)e^{-j2\pi k'm/N}, \quad k' = 0, \dots, N-1, \quad (1.19b)$$

Eq. (1.19a) becomes

$$\tilde{c}_k(n) = e^{j\pi k P(1-2n/N)} \frac{P}{N} X(n, Pk), \quad k = 1, \dots, K. \quad (1.19c)$$

Another implication of Eq. (1.19a) is that $\tilde{c}_k(n)$ needs to be computed for all integer values of n . However, it turns out that $\tilde{c}_k(n)$ can be accurately represented by considerably fewer samples due to the inherent low bandwidth of this function. Assuming that the bandwidth of $\tilde{c}_k(n)$ is confined to f_a , which is approximately true for each of the window functions discussed above (except the rectangular window), $\tilde{c}_k(n)$ can be minimally sampled at a frequency of $2f_a$, which corresponds to two points per period of the input signal or $2P$ points sampled evenly within the N -point window. The analysis-sample spacing or hop size (in signal samples) is then $H = 0.5N/P$, i.e., $0.25N$ for the Hamming or hanning or $0.125N$ for the 4-term Blackman–Harris window. Therefore, values only need to be computed for $n = Hi$, where i is the frame number, so that Eq. (1.19c) then becomes

$$\tilde{c}_k(Hi) = e^{j\pi k(P-i)} \frac{P}{N} X(Hi, Pk), \quad k = 1, \dots, K. \quad (1.19d)$$

This is a DFT with a constant multiplier (P/N) and an extra phase shift of $\pi k(P - i)$. Because the phase shift is always an integer multiple of π , it is equivalent to a shift of either 0° or 180° .

Computation of amplitude (magnitude), phase, and frequency follows from Eqs. (1.15e)–(1.15h). First, the real and imaginary parts of $\tilde{c}_k(Hi)$, which naturally result from an FFT or DFT, are taken to be $a_k(i)$ and $b_k(i)$, respectively. Then, for $k \geq 1$,

$$A_k(Hi) = 2\sqrt{a_k^2(i) + b_k^2(i)}, \quad (1.20a)$$

$$\theta_k(Hi) = \text{atan } 2(b_k(i), a_k(i)), \quad (1.20b)$$

$$\theta_{k0} = \theta_k(0). \quad (1.20c)$$

Calculating the frequency deviation from one frame to the next requires some care. It is essentially a matter of computing the difference between the phase of each frame and that of the preceding frame and multiplying the result by a suitable scale factor. However, if the phase is advancing and crosses the $+\pi$ boundary, it will immediately jump negative to be slightly greater than $-\pi$. This does not mean that the frequency is suddenly negative. Conversely, if the phase is receding

and crosses the $-\pi$ boundary, it will suddenly jump positive to be slightly less than $+\pi$, which falsely implies a positive frequency. It is better to imagine that the phase is progressing around a circle and choose the phase difference that is smallest in that angular regime. One way to handle this problem is by using a modulo function:

$$\Delta\theta_k(Hi) = \text{mod}(\theta_k(H(i+1)) - \theta_k(Hi); -\pi, \pi), \quad (1.20d)$$

which automatically confines the phase to the range $[-\pi, \pi]$.

Another method is to use the identity formula

$$\begin{aligned} \Delta\theta_k(Hi) &= \text{atan} 2(b_k(i+1)a_k(i) - a_k(i+1)b_k(i), a_k(i)a_k(i+1) \\ &\quad + b_k(i)b_k(i+1)), \end{aligned} \quad (1.20e)$$

which obviates having to calculate the individual phases according to Eq. (1.20b).

Because the time between frames is always $0.5/f_a$, the slope of the phase with respect to time, which gives the frequency deviation of harmonic k , becomes

$$\Delta f_k(Hi) = \frac{\Delta\theta_k(Hi)}{\pi} f_a, \quad (1.20f)$$

and the total estimated frequency of harmonic k is given by

$$f_k(Hi) = kf_a + \Delta f_k(Hi) = \left(1 + \frac{\Delta\theta_k(Hi)}{k\pi}\right) kf_a. \quad (1.20g)$$

Finally, given $\theta_k(Hi)$ and $\Delta f_k(Hi)$ for frame i , the phase for the next frame $i + 1$ can be recovered using

$$\theta_k(H(i+1)) = \theta_k(Hi) + \Delta\theta_k(Hi) = \theta_k(Hi) + \pi \frac{\Delta f_k(Hi)}{f_a}. \quad (1.20h)$$

In this way, except for roundoff error, the initial phases $\{\theta_k(0)\}$ and frequency deviations $\{\Delta f_k(Hi)\}$ of the harmonics at each frame are sufficient to recover the phases $\{\theta_k(Hi)\}$ of all of the harmonics at each frame. The frequency deviations are preferred for file storage over the phases because they are intuitively more useful for sound data examination and manipulation. Thus, according to this design, an analysis file contains the initial harmonic phases and for each harmonic and frame number the amplitude/frequency-deviation pairs $\{\{A_k(Hi), \Delta f_k(Hi)\}, k = 1, \dots, K\}, i = 0, \dots, I - 1\}$, where I is the total number of frames.

Figure 1.4 shows the fixed filter bank analysis of an F₄ trumpet tone played *ff* with analysis frequency $f_a = 350$ Hz in terms of harmonic amplitude vs frequency and time. Fig. 1.4a is a 3D display with amplitude being the vertical dimension (frequency deviations are not shown) and Fig. 1.4b is a 2D display showing harmonic frequencies vs time with harmonic amplitudes depicted in terms of darkness.

1.1.4.2 Synthesis Step

Synthesis can be accomplished either by using inverse FFTs and overlap-adds of adjacent windows or by straightforward sinusoidal (oscillator bank) additive synthesis.

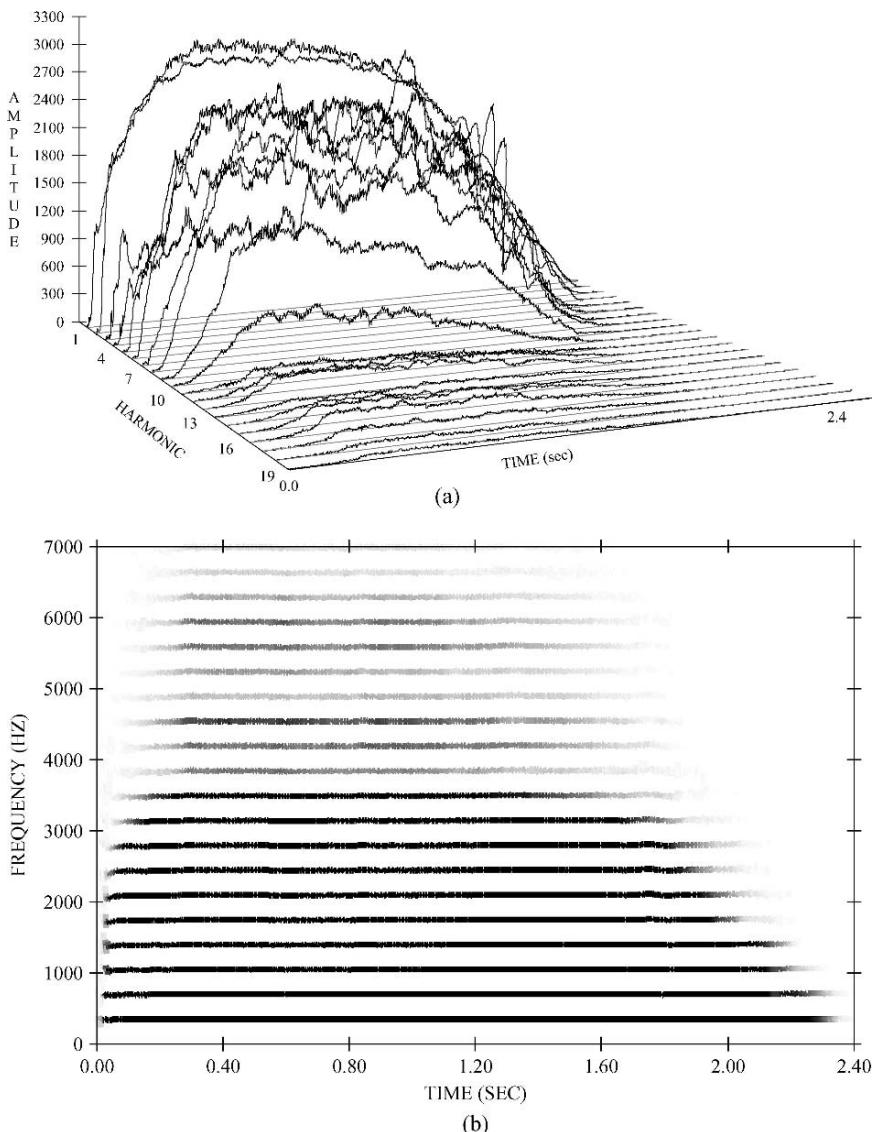


FIGURE 1.4. Time-variant analysis of an F_4 (350 Hz) trumpet tone played *ff*: (a) amplitude-vs-harmonic number-vs-time (3D) graphs of the harmonic envelopes. (b) Frequency-vs-time (2D) graphs of the harmonics. Amplitude is indicated by darkness in the 2D graph.

With overlap-add one must take care that the effect of the analysis window function disappears. This can be shown to be true for the cosine-term windows considered above when they are overlapped with window centers spaced by $0.5/f_a$. Assuming no time-scale modification is needed, spectrum data manipulations can be made on the harmonic amplitude and frequency data, but these data must be

converted to real and imaginary parts for use with the inverse FFT. Time-scale modification with this method is not as straightforward as with the oscillator bank, but it can be done by resampling in the frequency domain (Rodet and DePalle, 1992; George and Smith, 1992). The principal advantage of the method over additive synthesis of sinusoids is the increased speed of synthesis when a large number of harmonics is used.

With time-varying sinusoidal additive synthesis, computation is very direct and is based on the following formula, derived from Eq. (1.15d) [with the $A'_0(t)$ term omitted]:

$$\hat{s}(n) = \sum_{k=1}^K A'_k(n) \cos \left[\frac{2\pi}{f_s''} \left(k f_a n + \sum_{m=0}^{n-1} \Delta f'_k(m) \right) + \theta_{k_0} \right], \quad (1.21a)$$

where f_s'' is the synthesis sample frequency, $A'_k(n)$ and $\Delta f'_k(n)$ are the synthesis harmonic amplitudes and frequency deviations for harmonic k at sample n , respectively.

With additive synthesis there are a few issues to consider. First, as the sample counter n advances beyond zero, the cosine argument advances by $2\pi(k f_a + \Delta f'_k(n))/f_s''$ on each sample, giving

$$\hat{s}(n) = \sum_{k=1}^K A'_k(n) \cos(\Theta_k(n)), \quad (1.21b)$$

where $\Theta_k(n)$, the “total synthesis phase,” is computed recursively using

$$\Theta_k(n+1) = \text{mod} \left(\Theta_k(n) + \frac{2\pi}{f_s''} (k f_a + \Delta f'_k(n)); -\pi, \pi \right). \quad (1.21c)$$

Second, because in the analysis step A_k and Δf_k are only computed for the frame boundaries, i.e., every H samples at the analysis sample rate, there is the issue of how to interpolate A_k and Δf_k between these boundaries. In the next four sections methods for phase reconstruction using zeroth-(constant), first-(linear), second-(quadratic), and third-(cubic) order phase interpolations are examined.

Third, it is desirable to match the analysis frame boundary phases with the “synthesis offset phase,” which can be defined from Eq. (1.21a) as

$$\theta'_k(n) = \frac{2\pi}{f_s''} \sum_{m=0}^{n-1} \Delta f'_k(m) + \theta_{k_0}, \quad (1.21d)$$

or recursively using

$$\theta'_k(n+1) = \theta'_k(n) + \frac{2\pi}{f_s''} \Delta f'_k(n). \quad (1.21e)$$

Recall that at each frame boundary θ_k can be computed recursively from the analyzed phase and frequency deviation of the previous frame using Eq. (1.20h). One would hope, then, that these could be matched using Eq. (1.21d) or (1.21e).

1.1.4.2.1 Piecewise Constant Amplitudes and Frequencies. With piecewise constant amplitudes and frequencies, the waveform recreated by “identity resynthesis” (resynthesis without any spectral modifications) is the original signal amplitude-modulated by the analysis window function. In the Hamming or hanning case, four frames per window are used, so the portion of the window used in synthesis varies in amplitude between 0.8536 and 1.0, a variation of about 1.4 dB. This could be compensated by multiplying by an inverse function. Merely changing the amplitudes of the spectrum should produce similar results, but when frequencies or the time scale are altered, unpredictable results can take place.

What happens to the synthesis phase offset for the constant frequency synthesis case can be seen by looking at the first frame. With the synthesis sample rate of f_s'' , the number of samples between frames is $H' = 0.5f_s''/f_a$. After H' samples the synthesis offset phase becomes

$$\theta'_k(H') = \theta_{k_0} + H' \frac{2\pi}{f_s''} \Delta f_k(0) \quad (1.22a)$$

$$= \theta_{k_0} + \frac{f_s''}{2f_a} \frac{2\pi}{f_s''} \left(\frac{\theta_k(H) - \theta_k(0)}{\pi} \right) f_a \quad (1.22b)$$

$$= \theta_k(H). \quad (1.22c)$$

Therefore, the analysis offset phase is exactly matched. However, the primary objection to this method is the discontinuity of frequency at the boundaries. This is not a problem under identity conditions but may produce perceptible artifacts when the time or frequency scale is changed.

1.1.4.2.2 Piecewise Linear Amplitude and Frequency Interpolation. This is the most commonly used method. Both amplitudes and frequencies are linearly cross-faded between frame boundaries. Thus,

$$A'_k(n) = \frac{H'(i+1) - n}{H'} A_k(Hi) + \frac{n - H'i}{H'} A_k(H(i+1)), \quad H'i \leq n < H'(i+1), \quad (1.23a)$$

and

$$\begin{aligned} \Delta f'_k(n) &= \frac{H'(i+1) - n}{H'} \Delta f_k(Hi) + \frac{n - H'i}{H'} \Delta f_k(H(i+1)), \\ H'i &\leq n < H'(i+1). \end{aligned} \quad (1.23b)$$

While this works very well in general (mainly because human ears are relatively insensitive to slowly changing phase errors), it can be shown that unless the frequency is constant, Eq. (1.23b) will result in the wrong phase values at the frame boundaries. In fact, for $i = 0$, it can be shown that successive application of Eqs. (1.23b) and (1.21c) will result in

$$\theta'_k(H') \cong \theta_k(H) + \frac{1}{2} \Delta^2 \theta_k(0), \quad (1.23c)$$

where

$$\Delta^2\theta_k(0) = \theta_k(2H) - 2\theta_k(H) + \theta_k(0). \quad (1.23d)$$

$\Delta^2\theta_k(Hi)$ may be thought of as the phase acceleration between frames. Unless the frequency is fixed, this error will accumulate from frame to frame.

1.1.4.2.3 Piecewise Quadratic Interpolation of Phases. With this method, the harmonic phases are matched at the frame boundaries while the frequencies are matched halfway between the boundaries (Ding and Qian, 1997). As in the previous section, frequency varies linearly with time although it is not directly matched to the phase-difference frequency. The offset phase is reconstructed by using a series of parabolas each of which extend from the midpoint of one frame to the midpoint of the next. Recalling that H' is the number of samples in a synthesis frame, let parabola 01 extend from $n = -H'/2$ to $n = H'/2$ with value at $n = 0$ matching the original initial phase $\theta_k(0)$. Similarly, let parabola 12 extend from $n = H'/2$ to $n = 3H'/2$ with value at $n = 1$ matching the original phase $\theta_k(H)$. These two phase parabolas can then be written as

$$\theta_{k'01}(n) = \theta_k(0) + \frac{A_0}{2}n + \frac{B_0}{2H'}n^2, \quad -\frac{H'}{2} \leq n \leq \frac{H'}{2} \quad (1.24a)$$

$$\theta_{k'12}(n) = \theta_k(H) + \frac{A_1}{2}(n - H') + \frac{B_1}{2H'}(n - H')^2, \quad \frac{H'}{2} \leq n \leq \frac{3H'}{2}, \quad (1.24b)$$

where A_0 , B_0 , A_1 , and B_1 are constants to be determined. Note that these parabolas give the correct phases when $n = 0$ and $n = H'$.

Then derivatives are taken with respect to n to reveal the corresponding frequency deviation functions for segments 01 and 12:

$$\Delta f_{k'01}(n) = \frac{A_0}{2} + \frac{B_0}{H'}n, \quad -\frac{H'}{2} \leq n \leq \frac{H'}{2}, \quad (1.24c)$$

$$\Delta f_{k'12}(n) = \frac{A_0}{2} + \frac{B_0}{H'}(n - H'), \quad \frac{H'}{2} \leq n \leq \frac{3H'}{2}, \quad (1.24d)$$

where ‘‘real frequency’’ in Hz is related to these by a factor of $f_s/2\pi$.

Continuity of phase requires that $\theta'_{k'01}(H'/2) = \theta'_{k'12}(H'/2)$, i.e., setting Eqs. (1.24a) and (1.24b) equal at their connecting point gives

$$\theta_k(0) + \frac{A_0}{2}\frac{H'}{2} + \frac{B_0}{2H'}\left(\frac{H'}{2}\right)^2 = \theta_k(H) + \frac{A_1}{2}\left(\frac{H'}{2} - H'\right) + \frac{B_1}{2H'}\left(\frac{H'}{2} - H'\right)^2, \quad (1.25a)$$

resulting in

$$2A_0 + B_0 + 2A_1 - B_1 = \frac{8}{H'}(\theta_k(H) - \theta_k(0)) = \frac{8}{H'}\Delta\theta_k(0), \quad (1.25b)$$

where the phase difference, $\Delta\theta_k(0) = \theta_k(H) - \theta_k(0)$, should be taken in the sense of $\text{mod}(\Delta\theta_k(0); -\pi, \pi)$.

Also, continuity of frequency requires that $\Delta f'_{k01}(H'/2) = \Delta f'_{k12}(H'/2) = \Delta f'_{k1}$; i.e., setting Eqs. (1.24a) and (1.24b) equal at their connecting point gives

$$\frac{A_0}{2} + \frac{B_0}{H'} \left(\frac{H'}{2} \right) = \frac{A_1}{2} + \frac{B_1}{H'} \left(\frac{H'}{2} - H' \right) = \Delta f'_{k1} \quad (1.26a)$$

or

$$\frac{A_0 + B_0}{2} = \frac{A_1 - B_1}{2} = \Delta f'_{k1}, \quad (1.26b)$$

resulting in

$$B_1 = A_1 - (A_0 + B_0), \quad (1.26c)$$

so that Eq. (1.25b) becomes

$$3A_0 + 2B_0 + A_1 = \frac{8}{H'} \Delta \theta_k(0). \quad (1.27)$$

The other end points of the 01 and 12 straight lines can be defined as

$$\Delta f'_{k01}(-H'/2) = \frac{A_0}{2} + \frac{B_0}{H'} \left(\frac{-H'}{2} \right) = \frac{A_0 - B_0}{2} = \Delta f'_{k0}, \quad (1.28a)$$

$$\Delta f'_{k12}(3H'/2) = \frac{A_1}{2} + \frac{B_1}{H'} \left(\frac{3H'}{2} - H' \right) = \frac{A_1 + B_1}{2} = \Delta f'_{k2}. \quad (1.28b)$$

Then, Eqs. (1.26b), (1.28a), and (1.28b) combine to yield

$$\begin{aligned} A_0 &= \Delta f'_{k0} + \Delta f'_{k1}, \quad B_0 = \Delta f'_{k1} - \Delta f'_{k0}, \quad A_1 = \Delta f'_{k1} + \Delta f'_{k2}, \\ B_1 &= \Delta f'_{k2} - \Delta f'_{k1}, \end{aligned} \quad (1.28c)$$

so that Eq. (1.27) becomes

$$\Delta f'_{k0} + 6\Delta f'_{k1} + \Delta f'_{k2} = \frac{8}{H'} \Delta \theta_k(0). \quad (1.29a)$$

This equation can be generalized to a sequence of equations

$$\Delta f'_{ki} + 6\Delta f'_{ki+1} + \Delta f'_{ki+2} = \frac{8}{H'} \Delta \theta_k(iH), \quad (1.29b)$$

where i is the frame number, running from $i = 0$ to $i = I - 1$ (there are I frames). This results in I equations with $I + 2$ unknowns. However, because $\Delta f'_{k0}$ and $\Delta f'_{kI+1}$ are really outside of the legitimate region for frame analysis, values for them must be extrapolated by extending the 12 and $I-1, I$ straight lines, giving

$$\Delta f'_{k0} = 2\Delta f'_{k1} - \Delta f'_{k2} \text{ and } \Delta f'_{kI+1} = 2\Delta f'_{kI} - \Delta f'_{kI-1}, \quad (1.29c)$$

so that the equation for $i = 0$ becomes

$$8\Delta f'_{k1} = \frac{8}{H'} (\theta_k(H) - \theta_k(0)), \quad (1.29d)$$

and a similar equation obtains for $i = I - 1$. Therefore, there are now I equations with I unknowns, and these can be readily solved by a process of elimination.

The same process can be applied to interpolation of the amplitude values.

The quadratic interpolation method works very well. However, its one big disadvantage is that it must be applied to the signal as a whole rather than recursively as the signal progresses. Therefore its use is restricted to non-real-time applications on fairly short sound files. It works fine on single musical instrument tones.

1.1.4.2.4 Piecewise Cubic Interpolation of Phases. In this case, it is assumed that both the phases and the frequencies are known at the beginning and end of each frame. By asserting these, continuity of frequency and phase can be guaranteed, and unlike the quadratic case, each frame segment is computed independently (McAulay and Quatieri, 1986). To begin, the equation for phase is postulated in terms of a cubic function of n :

$$\theta_k(n) = \theta_a + \gamma(n - H'i) + \alpha(n - H'i)^2 + \beta(n - H'i)^3, \quad H'i \leq n \leq H'(i+1). \quad (1.30a)$$

At $n = H'i$ the phase is obviously

$$\theta_k(H'i) \equiv \theta_a. \quad (1.30b)$$

At $n = H'(i+1)$, the phase becomes

$$\theta_k(H'(i+1)) \equiv \theta_b + 2\pi M = \theta_a + \gamma H' + \alpha H'^2 + \beta H'^3. \quad (1.30c)$$

The term $2\pi M$, where M is an integer, is added to the computed phase because it can only be computed in terms of its principal value in the range $(-\pi, \pi)$. It turns out that the best M to use is the one which produces the smoothest phase function, where smoothness corresponds to the minimum mean-squared average of the phase function's second derivative. This is a method of "phase unwrapping."

The derivative of Eq. (1.30a) with respect to n gives the frequency function

$$\Delta f_k(n) = \gamma + 2\alpha(n - H'i) + 3\beta(n - H'i)^2, \quad (1.30d)$$

which at the two frame boundaries become

$$\Delta f_k(H'i) \equiv \Delta f_a = \gamma \quad (1.30e)$$

and

$$\Delta f_k(H'(i+1)) \equiv \Delta f_b = \Delta f_a + 2\alpha H' + 3\beta H'^2. \quad (1.30f)$$

Using Eqs. (1.30b), (1.30c), (1.30e), and (1.30f), α and β can be solved to produce

$$\alpha = \frac{3}{H'^2} \Delta \theta - \frac{1}{H'} \Delta^2 f \quad (1.30g)$$

and

$$\beta = \frac{-2}{H'^3} \Delta \theta + \frac{1}{H'^2} \Delta^2 f, \quad (1.30h)$$

where

$$\Delta \theta \equiv \theta_b - \theta_a - \Delta f_a H' + 2\pi M \quad (1.30i)$$

and

$$\Delta^2 f \equiv \Delta f_b - \Delta f_a. \quad (1.30j)$$

Note that the only right-side unknown is M . This is evaluated by considering it to be a continuous variable and finding the value of it which minimizes

$$g(M) = \frac{1}{4H'} \int_{H'i}^{H'(i+1)} \left(\frac{\partial^2 \theta(n)}{\partial n^2} \right)^2 dn = \frac{1}{4H'} \int_0^{H'} (2\alpha + 6\beta n)^2 dn \quad (1.31a)$$

$$= \frac{1}{H'} \int_0^{H'} (\alpha^2 + 6\alpha\beta n + 9\beta^2 n^2) dn = \alpha^2 + 3\alpha\beta H' + 3\beta^2 H'^2. \quad (1.31b)$$

Substituting Eqs. (1.30g) and (1.30h) into (1.31b) gives

$$g(M) = \frac{3(\Delta\theta)^2 - 3H' \cdot \Delta^2 f \cdot \Delta\theta + H'^2(\Delta^2 f)^2}{H'^4}. \quad (1.31c)$$

Noting from Eq. (1.30i) that $\Delta\theta$ can be written as $\Delta\theta = \Delta\theta_1 + 2\pi M$, the derivative of Eq. (1.31c) with respect to M can be taken and set to zero. This gives

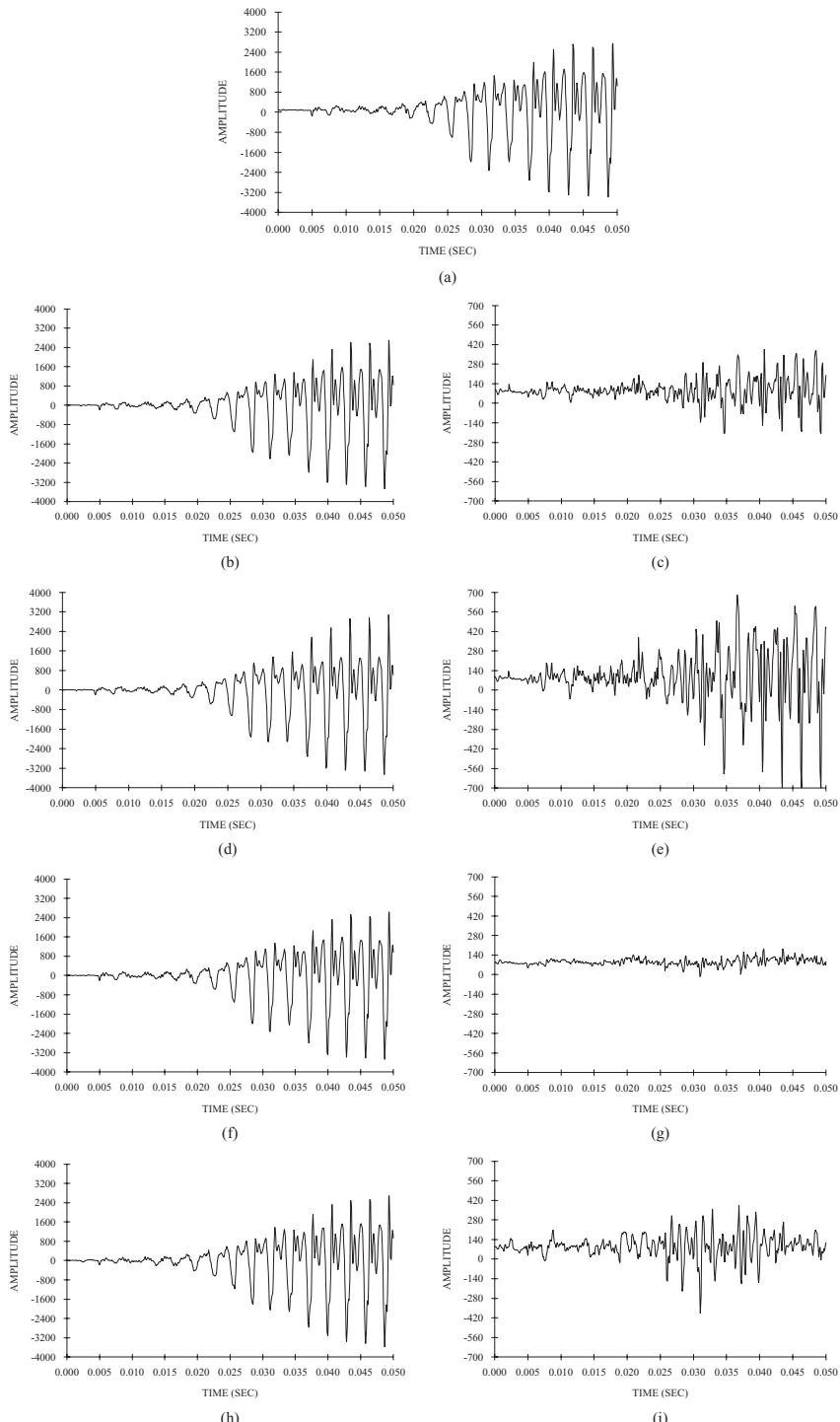
$$M = \frac{1}{2\pi} \left(\frac{H'}{2} (\Delta f_a + \Delta f_b) - (\theta_b - \theta_a) \right). \quad (1.31d)$$

This continuous result should be rounded to the nearest integer before substituting into Eqs. (1.30g) and (1.30h) for computation of the cubic polynomial coefficients α and β . These coefficients and the fact that $\gamma = \Delta f_a$ are then used for the phase formula of Eq. (1.30a).

Compared to the quadratic method, a primary advantage of the cubic interpolation method is that each frame is handled separately, so there is no problem with real-time applications other than the time involved in computing each frame. Moreover, because phase as well as frequency is accounted for on each frame boundary, large relative phase errors between frames should not occur. Thus, the cubic method ordinarily works well for time-stretching applications. However, Ding and Qian (1997) have shown that the quadratic method is more stable in the face of random initial phases and small amounts of noise added to frame-boundary phase measurements for the case of a fixed-frequency offset.

Figure 1.5 compares the first 50 ms of the original F_4 *ff* trumpet tone signal with resyntheses of the tone using piecewise-linear frequency, piecewise-constant frequency, quadratic phase, and cubic phase interpolation. It also gives the differences between the original and the four cases. From the differences, it appears that piecewise quadratic is much superior to the other methods. However, in listening to the difference signals, it is apparent that the piecewise cubic produces the

FIGURE 1.5. First 50 ms of original (middle top), resynthesized (left), and difference signals (right) for the F_4 trumpet tone using various methods of phase interpolation: (a) original; (b) and (c) piecewise-linear frequency; (d) and (e) piecewise-constant frequency; (f) and (g) piecewise-quadratic phase, (h) and (i) piecewise-cubic phase.



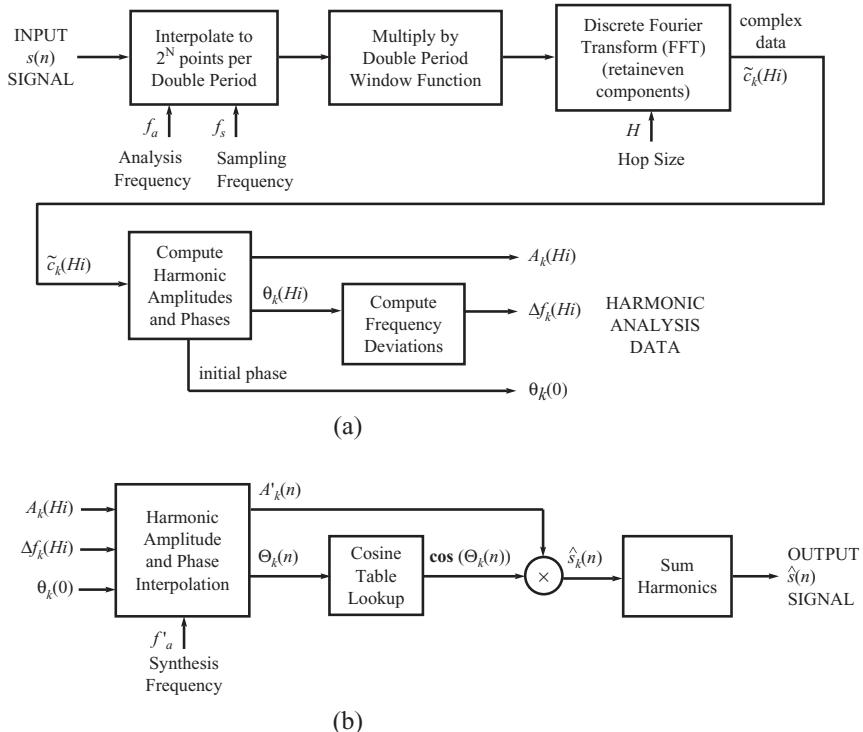


FIGURE 1.6. Block diagram of the fixed filter bank phase-vocoder analysis/resynthesis method, where n = sample number, k = harmonic number, and i = frame number: (a) analysis method; (b) additive resynthesis method based on harmonic analysis data.

perceptually smaller result because, unlike the other methods that still retain the pitch of the original, at least in this case, it produces a pitchless broadband residual.

In summary, Fig. 1.6 gives a block diagram of an analysis/synthesis system based on the fixed filter-bank (phase vocoder) approach.

1.2 Spectral Frequency-Tracking Method

When the input signal is more complex than a single quasiperiodic sound such as a fixed-pitch trumpet tone, the harmonic analysis/synthesis method may not be sufficient. This is true of sounds that contain inharmonic partials or significant noise, but it is especially true of signals having large pitch variations or those containing several instrument tones at different pitches. The frequency-tracking or MQ method, which was introduced by McAulay and Quatieri (1986) for speech and extended by Smith and Serra (1987) for music applications, takes the position that a sound signal is composed of collections of sinusoids having arbitrary frequencies (i.e., with no particular ratios between frequencies) and that each frequency

component does not necessarily exist during the entire duration of the signal. In fact, some frequency components may have extremely short lives, typically ones that form clusters to imitate bursts of noise.

1.2.1 Frequency-Tracking Analysis

It is well known that individual frequency components, i.e., sinusoids or partials, can be observed and measured as peaks in a discrete Fourier transform if the component frequencies are spaced substantially farther apart than the bin frequencies of the analysis and the bandwidth of the analysis window function. A useful criterion (Smith and Serra, 1987) is that for adjacent component frequencies, e.g., f_1 and f_2 , to be resolved, they must be separated by at least the window function bandwidth given by

$$\Delta f_w = B_w \frac{f_s}{N} = B_w \Delta f_b, \quad (1.32)$$

where B_w is the window bandwidth in bins, f_s is the sample frequency, N is the number of samples in the window function, and Δf_b is the bin separation frequency. Based on the window transform's first zero frequency, $B_w = 2$ for the rectangular window, $B_w = 4$ for the hanning and Hamming windows, and $B_w = 8$ for the 4-term Blackman–Harris window.

Figure 1.7 illustrates the magnitude spectrum analysis of two superimposed sinusoids of different frequency for various window types and frequency separations. Zero fill, whereby the FFT length is artificially increased by added zeros to the left and to the right of the window function, can be used to reveal the true nature of the window transform functions. In this case, the magnitude transforms of two window transform functions, one for each sinusoid frequency, are superimposed. With f_s taken to be equal to N , the bin frequencies have integer values. For the rectangular window, if f_1 and f_2 are set to integers separated by at least 2, the peaks are clearly discernable. However, if with the same separation they are set to frequencies half-way between the integers (worst case), they are less distinct unless zero fill is used. Also, a very significant amount of sidelobe behavior is visible. With a hanning or Hamming window and the same frequency component spacing, the components cannot be separated even when zero fill is used. However, when the frequency spacing is increased to 3, separation is very clear. For the 4-term Blackman–Harris window, a spacing of 3 also works well. So a separation of 3 bin frequencies is adequate for the three window types. Also, the hanning and Hamming sidelobe amplitudes are very small, and the Blackman–Harris sidelobes are not visible.

Therefore, in order to accurately resolve peaks in a magnitude spectrum their frequencies must be separated by at least three bin frequency units. Otherwise, components will appear to be merged and cannot be separated easily. A typical situation is

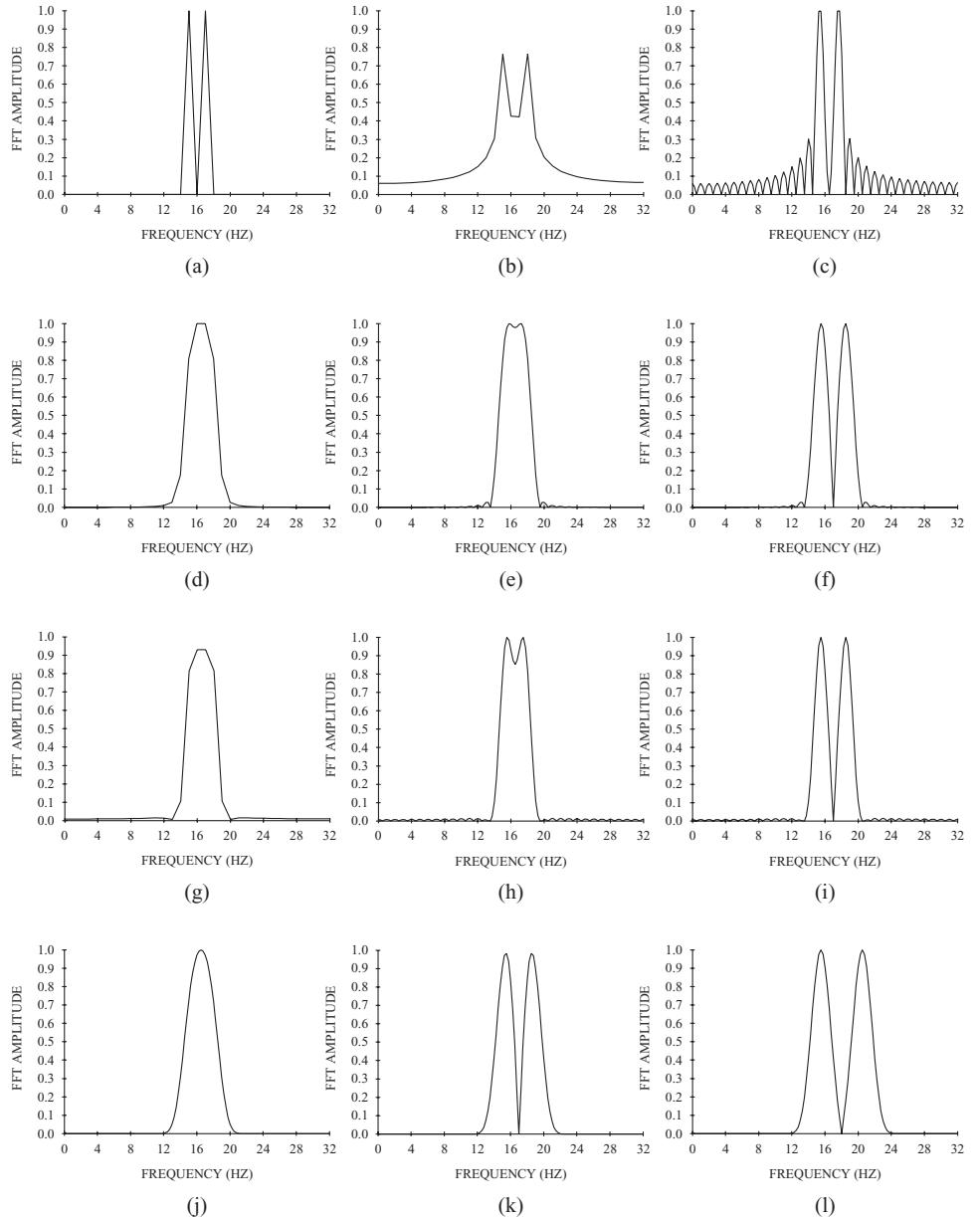


FIGURE 1.7. Discrete Fourier transform of a windowed signal consisting of two unit-amplitude sinusoids with frequencies f_1 and f_2 for various window functions, with and without zero-fill. In all cases the sampling frequency is 64. (a)–(c) rectangular window responses: (a) $f_1 = 15$, $f_2 = 17$, no zero-fill; (b) $f_1 = 15.5$, $f_2 = 17.5$, no zero-fill; (c) $f_1 = 15.5$, $f_2 = 17.5$, with zero-fill. (d)–(f) Hanning window responses: (d) $f_1 = 15.5$, $f_2 = 17.5$, no zero-fill; (e) $f_1 = 15.5$, $f_2 = 17.5$, with zero-fill; (f) $f_1 = 15.5$, $f_2 = 18.5$, with zero-fill. (g)–(i) Hamming window responses: (g) $f_1 = 15.5$, $f_2 = 17.5$, no zero-fill; (h) $f_1 = 15.5$, $f_2 = 17.5$, with zero-fill; (i) $f_1 = 15.5$, $f_2 = 18.5$, with zero-fill. (j)–(l) Blackman–Harris window responses (all with zero-fill): (j) $f_1 = 15.5$, $f_2 = 17.5$; (k) $f_1 = 15.5$, $f_2 = 18.5$; (l) $f_1 = 15.5$, $f_2 = 20.5$.

where $f_s = 44,100$ Hz, $N = 1024$, and the bin separation frequency is $\Delta f_b = 43$ Hz. It is clear that under these circumstances the lowest fundamental frequency of a harmonic tone that can be analyzed properly is about 130 Hz. The time resolution in terms of the window width is $N/f_s = 23$ ms. Going to lower frequencies by a certain factor compromises the time resolution by the inverse factor.

Because the detected peaks must be at least three window bins apart, for each frame i the maximum number of peaks K_i that can be detected in a spectrum is $N/6$. Because the frequency resolution is $\Delta f_b = f_s/N$, the maximum number of peaks (or partials) is equal to $f_s/(6\Delta f_b)$. For example, if a minimum peak separation of 40 Hz were required at a sample rate of 44,100 Hz, a power-of-2 value of N must first be chosen to yield a value at least that small. The maximum useable bin separation frequency would be $40/3 = 13.3$ Hz. A bin separation of 10.8 Hz is given by $N = 4096$, so the corresponding minimum peak separation would be 32.4 Hz, and the maximum number of peaks that could be resolved would be $4096/6 \approx 683$. It is assumed that each peak corresponds to a sinusoid in the signal.

1.2.2 Frequency-Tracking Algorithm

Assuming that frequency components can be resolved, the frequency-tracking method consists of the following four steps:

1. Successive FFTs (corresponding to frames) of overlapped windowed segments of the input signal are computed. A window function such as the Kaiser with $\alpha = 6.3$ can be used for good peak separation. Usually a zero-fill factor of at least 1.0 is used, but larger (integer) factors might be useful. This yields an FFT window size of N' , as opposed to N , the width of the window function. The real and imaginary parts of the FFT are retained and the magnitude values are computed. [See Eqs. (1.13e) and (1.19a.)] The FFTs are overlapped by a hop size of H samples or $\Delta t_{\text{frame}} = H/f_s$ s. Also, the FFT bin frequency spacing (after zero-padding) is $\Delta f_{\text{FFT}} = f_s/M'$.

2. For each frame i , K_i spectrum peaks are identified from the magnitude spectrum. Each peak is determined by three consecutive FFT magnitude values $A_{\xi-1}$, A_ξ , and $A_{\xi+1}$ (ξ is the FFT bin number variable), where A_ξ is the largest of the three. The estimated frequency and true maximum value are found by parabolic interpolation. Zero-fill helps the interpolation, because more points are automatically inserted between the window-function bins, and the interpolation is band-limited. However, direct interpolation can be implemented by fitting a smooth curve to the three points. Ideally, a best-fit shifted version of the transformed window function should be used. In practice, it has been found that fitting a quadratic to the log of the magnitude function yields adequate results with much less computation (Smith and Serra, 1987). Thus, the peak frequency is given by

$$f_k = (\xi + p)\Delta f_{\text{FFT}}, \quad (1.33a)$$

where

$$p = 0.5 \frac{\log(A_{\xi-1}A_{\xi+1})}{\log(A_{\xi-1}A_{\xi+1}/A_{\xi}^2)}. \quad (1.33b)$$

Next, the amplitude and phase of each peak is calculated. The peak amplitude is computed using

$$A_k = \frac{A_{\xi}}{(A_{\xi-1}/A_{\xi+1})^{p/4}}. \quad (1.33c)$$

The corresponding phase is computed by first interpolating the values of the real and imaginary parts, a_{ξ} and b_{ξ} , at the peak frequency using a formula analogous to Eq. (1.33c), and then computing the phase according to Eq. (1.20b). The frequency, amplitude, and phase of each peak (A_k , f_k , θ_k) are thus computed and retained. Other FFT bin information is discarded.

Usually not every local maximum is chosen to be a peak. For example, peaks may be ignored that are not above a predefined threshold. The threshold can vary with frequency. E.g., a threshold that lowers as frequency increases may be desirable because even though the higher-frequency components of most musical sounds are generally weaker than the lower-frequency components, they are still very audible. This threshold variation can be accomplished by preprocessing the signal with a simple first-order digital filter and then applying a fixed threshold (Beauchamp, 1993). Serra (1989) discusses the use of two thresholds, one absolute and one relative to the highest peak of the spectrum of the current frame. He also uses the log equivalent of a peak-to-valley ratio defined by

$$pvr = \frac{A_{\xi}}{\sqrt{A_{\xi-\delta_1}A_{\xi+\delta_2}}}, \quad (1.33d)$$

where $\xi - \delta_1$ and $\xi + \delta_2$ are the bins corresponding to the first minimum below and above the peak bin ξ . Unless pvr is above a designated threshold, the peak would be rejected. Fitz et al. (1992) discuss using logarithmic bands within which weak peaks may be masked by a strong peak and are therefore discarded.

Figure 1.8 shows a typical magnitude spectrum with peaks above an amplitude threshold of 100 identified.

3. Frequency-vs-time tracks are formed by connecting peaks of consecutive frames. This turns out to be the most crucial aspect of the analysis method, and there is probably no perfect way to do it. Rules for forming tracks form a heuristic method, which is not guaranteed to be optimal in any sense. The basic procedure is to find the best match between the peaks of frame i with frame $i + 1$. Matches are attempted between corresponding frequencies that are close together. If the number of peaks in frames i and $i + 1$ are K_0 and K_1 , respectively, and $K_0 > K_1$, some of the tracks will have to end (“death”). On the other hand, if $K_0 < K_1$, some new tracks will begin (“birth”). Tracks could also begin or end because the only available potential matches have excessive frequency differences. Detailed procedures for peak-tracking are given by McAulay and Quatieri (1986), Smith and Serra (1987), Serra (1989), and Maher (1989). Fitz

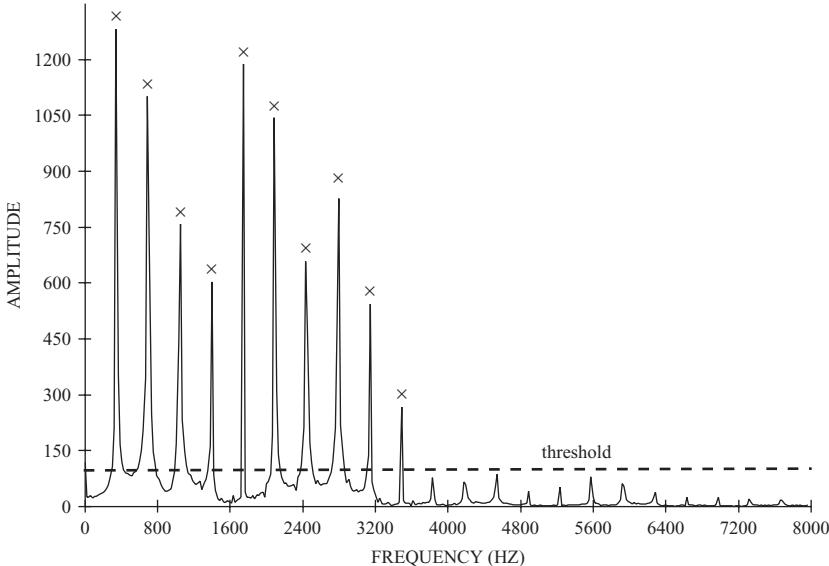


FIGURE 1.8. Magnitude spectrum of the F₄ trumpet tone with peaks whose amplitudes are above a threshold of 100 marked.

et al. (1992) discuss the possibility of hysteresis in tracking, where the end of a track is not counted as a real end until it persists for several frames. However, it is obvious that the best frequency tracker should look at the signal as a whole, or at least in large chunks, rather than just a few frames. Depalle et al. (1993a, 1993b) developed a method based on a hidden Markov model (HMM) that uses computed probabilities of peak trajectories to determine improved overall tracking.

Figure 1.9a shows a set of tracks for a tenor voice sound (sung at G₃) with vibrato. Note that with the fixed-filter-bank (phase-vocoder) method it would be difficult to isolate the harmonics because when harmonic frequency deviations exceed $0.5 f_a$ —in this case when the harmonic number is greater than about 8—the filter responses begin to seriously overlap. Frequency tracking alleviates that problem. Fig. 1.9b shows the same data plotted in three dimensions, which provides a view of the amplitude of the various tracks, as well as their frequencies, as functions of time.

4. Peak data for each track are written to a file. For each frame i , the number of peaks K_i is given and each peak k is represented by amplitude $A_{k,i}$, frequency $f_{k,i}$, phase $\theta_{k,i}$, and a “link” $\kappa_{k,i}$ giving the peak number of the next frame to which it is connected. An alternative to giving the next-frame peak number is to number the tracks and give the track number of each peak, but a problem with this method is that the number of tracks usually changes continually throughout the sound, so that the track numbers soon get out of frequency order.

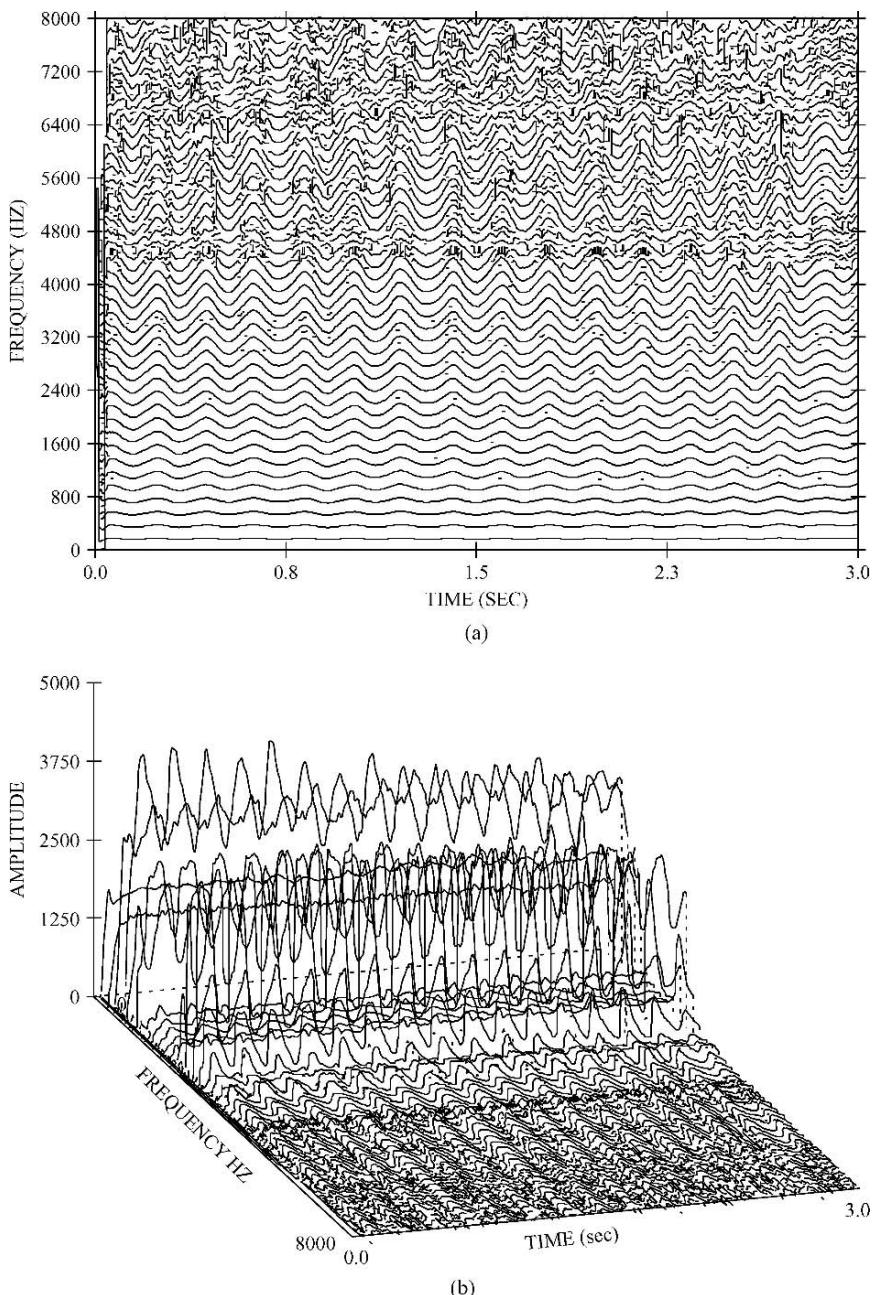


FIGURE 1.9. Frequency-tracking analysis of a G₃ tenor voice sound: (a) frequency-vs-time 2D display showing positions of tracks; (b) amplitude-vs-frequency-vs-time 3D display which shows the strengths and frequencies of the various harmonics.

1.2.3 Fundamental Frequency (Pitch) Detection

There is some controversy about the use of the term *pitch* to mean fundamental frequency (or f_0). Most speech communication researchers [e.g., Hess (1983) and Schroeder (1999)] use the two terms interchangeably, whereas most auditory science researchers insist on a clear distinction between them. According to the latter, pitch is strictly a percept and should not be confused with frequency. However, the pitch of a sound corresponds to the frequency of a sine tone that is judged to have the same pitch. For periodic signals, perceived pitch generally corresponds to its fundamental frequency, although pathological cases are easy to construct. For example, if only a few upper harmonics are resynthesized, the pitch may be associated either with the center of the harmonic band or with the greatest common divisor of the harmonic frequencies. Pitch ambiguity can also arise when only odd harmonics are present and the fundamental component is missing. Moreover, in typical music performance, pitch is highly variable, and not all sounds are equally harmonic. Some short sounds may be very noise-like and yet will be perceived (by musically experienced listeners) as particular musical pitches. Unlike speech, where pitch tends to change smoothly over time and is restricted in range, in typical solo musical passages pitches continually change from one relatively constant value to another, large leaps often occur, and spans of two octaves or more are possible.

What are the requirements of a good musical pitch detector? First, the detector should yield frequency-vs-time data for recordings of solo acoustical musical instruments. Recordings with reverberation present difficulties because echoes overlaying the intended sounds tend to confuse detectors. (Humans seem to have the uncanny ability to ignore these echoes.) Second, the detector should yield a pitch-vs-time graph that music experts agree corresponds to what they hear. If the input signal is a recording of a written score performance, it is reasonably easy to assess the accuracy of the pitch detector. On the other hand, if the performance is an improvisation, a transcription of some sort must be produced before evaluation can be done. A useful form for the transcription is a series of “events,” which gives the start-time, end-time, and average pitch (in log frequency units) of each note. Another useful parameter is an estimate of the definiteness of the pitch for each event. Such a transcription can be produced by using a sound file editor to play back isolated segments of the file and determining the pitch by comparison to a tone generator. The pitch-vs-time graph can be compared to these data visually or by using a computer to tally the errors on a note-by-note or frame-by-frame basis. Another way to assess the quality of pitch detection is to resynthesize the input signal using the detected pitch information. If the amplitude is very low when some errors occur, the pitch errors may be inaudible and can also be easily gated out in the pitch-vs-time graph. Musical pitch detectors should be evaluated with a large corpus of material, i.e., recordings of solo passages with little or no reverberation and corresponding transcriptions. Unfortunately, such a corpus does not currently exist in the public domain.

Pitch detectors can either work directly with time-domain samples or with frequency-domain spectra. In the time domain, autocorrelation period detectors

have proven useful (Moorer, 1974, 1975; Boersma, 1993; de Cheveigné and Kawahara, 2002). In essence, a signal is compared with a time-delayed version of itself, either by multiplying the two signals together and averaging or by subtracting the two and averaging the magnitudes of the differences, both over a certain time window. In the former case, the first significant maximum indicates the period, whereas in the latter case the first significant minimum indicates the period. Assuming that a sufficiently wide window is chosen, frequency-domain spectra show the positions of the harmonics, and the positions of the harmonics can be used to predict the fundamental frequency. Frequency-domain harmonic-matching methods for musical pitch detection have been developed by Piszczałski and Galler (1979), Doval and Rodet (1991), Brown (1992), and Maher and Beauchamp (1994). Another distinct method uses the cepstrum, which seeks to determine the periodicity of the DFT magnitude by taking its log and then applying a second FFT. This has been used extensively in speech applications (Noll, 1967) but seldom for music [for an exception, see Chen (2001)]. All of these methods rely on the selection of a minimum or maximum of a function. Determining which of several maxima or minima correspond to the correct fundamental frequency turns out to be the biggest problem in making these methods reliable. Roads (1996) gives an extensive overview of several music pitch-detection methods.

Maher and Beauchamp discuss a pitch detector based on spectral peaks called a Two-Way Mismatch Algorithm (Beauchamp et al., 1993; Maher and Beauchamp, 1994). This method was further developed by Cano (1998). The algorithm compares the frequencies of the peaks with the frequencies of the harmonics of a series of hypothetical fundamental frequencies. The name of the algorithm comes from its method of comparing the “measured” peak frequencies $\{f_k, k = 1, \dots, K\}$ with the nearest “predicted” harmonic frequencies $\{nf_o, n = 1, \dots, N\}$ [(where $N = \text{ceil}(\max(f_k/f_o))$)] and, in reverse, comparing the harmonic frequencies with the nearest peak frequencies. For example, suppose there are peak frequencies of 90, 180, 270, 360, and 450 Hz and predicted harmonic frequencies of 50, 100, 150, 200, 250, 300, 350, 400, and 450 Hz. Then there are peak-to-harmonic frequencies $90 \rightarrow 100, 180 \rightarrow 200, 270 \rightarrow 250, 360 \rightarrow 350$, and $450 \rightarrow 450$. However, in reverse there are harmonic-to-peak frequencies $50 \rightarrow 90, 100 \rightarrow 90, 150 \rightarrow 180, 200 \rightarrow 180, 250 \rightarrow 270, 300 \rightarrow 270, 350 \rightarrow 360, 400 \rightarrow 360$, and $450 \rightarrow 450$. While in the first case the absolute frequency differences are 10, 20, 20, 10, and 0, in the second case they are 40, 10, 30, 20, 20, 30, 10, 40, and 0. The “measured-to-predicted error” depends on the first set of numbers, whereas the “predicted-to-measured error” depends on the second set, and the total error depends on the weighted sum of the two. However, the amplitudes of the peak components also matter, as weak amplitude components (which might be spurious) are not as important as strong amplitudes. Moreover, the actual frequencies of the components may be important. A general formula for the frequency mismatch error is

$$\text{Err}_{\text{total}} = \text{Err}_{p \rightarrow m} + \rho \text{Err}_{m \rightarrow p} \quad (1.34a)$$

$$= \frac{1}{N} \sum_{n=1}^N E_W (\Delta f_n, nf_o, a_{k \leftarrow n}) + \frac{\rho}{K} \sum_{k=1}^K E_W (\Delta f_k, f_k, a_k). \quad (1.34b)$$

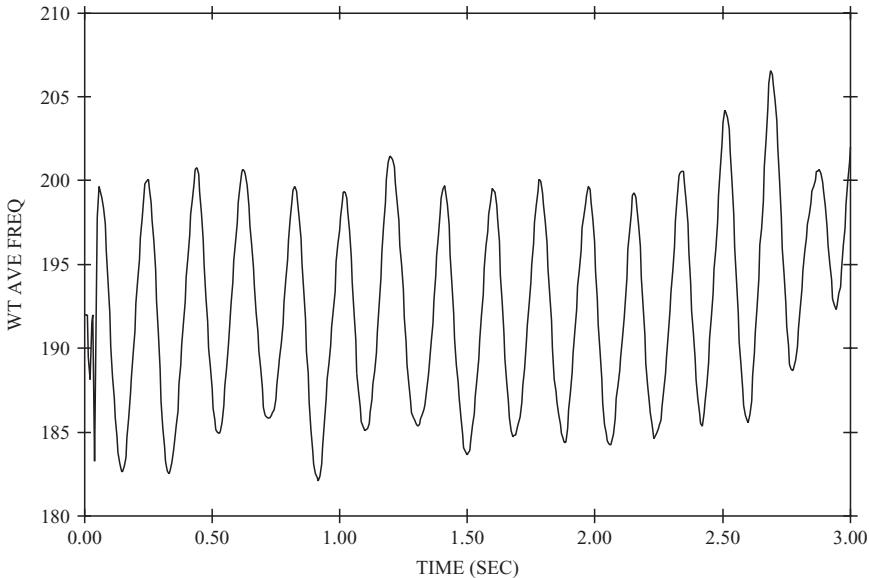


FIGURE 1.10. Fundamental frequency f_0 vs time for the G3 tenor voice sound.

Note that the forms of the two error terms are the same and that ρ is a factor used to weight the relative importance of the two terms in calculating the total error. Δf_n is the magnitude of the difference between the frequency of the n th harmonic and the k th peak frequency closest to it, i.e., $\Delta f_n = |f_n - f_k|$. a_k is the normalized amplitude of the k th peak component, i.e., $a_k = A_k / \max(A_k)$ so that the maximum value of a_k is 1.0. For the first summation, $a_{k \leftarrow n}$ refers to the normalized amplitude of the peak component k that is closest to the harmonic n . Δf_k is the magnitude of the difference between the frequency of the k th peak frequency and the n th harmonic closest to it, i.e., $\Delta f_k = |f_n - f_k|$. A simplified version of the error function used by Maher and Beauchamp (1994) is

$$\text{Err}_{\text{total}} = \frac{1}{N} \sum_{n=1}^N (1 + q \cdot a_n) \frac{\Delta f_n}{f_n^p} + \frac{\rho}{K} \sum_{k=1}^K (1 + q \cdot a_k) \frac{\Delta f_k}{f_k^p}, \quad (1.34c)$$

where q , p , and ρ were taken to be 1.4, 0.5, and 0.33, respectively.

For each frame of the analysis, the fundamental frequency f_0 is varied over a designated frequency range and the “true” fundamental frequency is deemed to be the one that yields the lowest value of $\text{Err}_{\text{total}}$.

A graph of f_0 vs time for the tenor voice whose spectral data are shown in Fig. 1.9 is given in Fig. 1.10. A graph of f_0 vs time for a clarinet solo passage translated into equal-tempered pitch units is shown in Fig. 1.11a. For comparison, Fig. 1.11b shows the corresponding musical score.

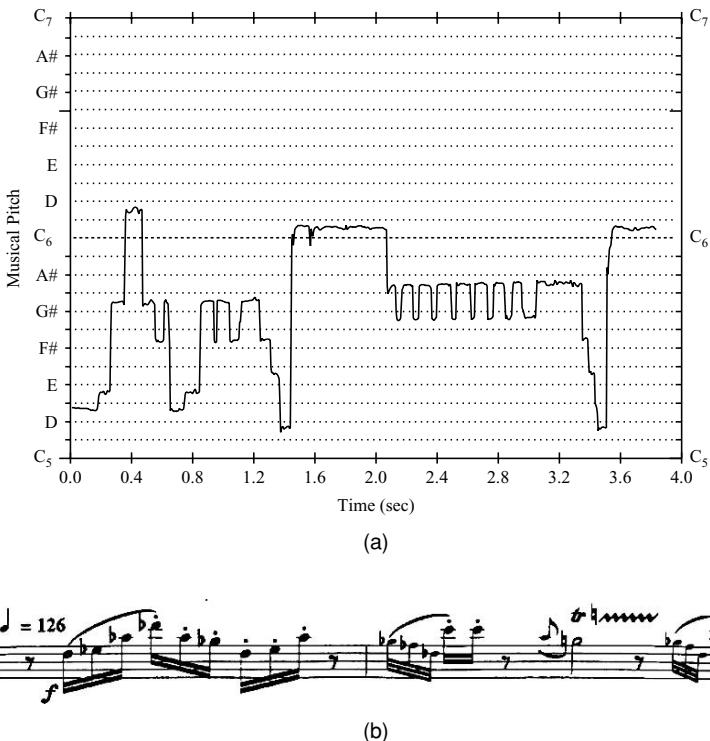


FIGURE 1.11. Pitch detection of a solo clarinet passage: (a) fundamental frequency vs time (from Beauchamp et al., 1993, Fig. 9, reproduced by permission of the Audio Engineering Society); (b) equivalent musical score. (from Messiaen, 1941)

1.2.4 Reduction of Frequency-Tracking Analysis to Harmonic Analysis

In 1989 Robert Maher wrote a program [described in Beauchamp (1993)] that reduces frequency-tracking data to harmonic data in a format almost identical to that used for the harmonic filter bank (aka phase vocoder) analysis/synthesis described in Section 1.1. This format is known as “mq.an,” as opposed to “mq,” the format used for frequency-tracking analysis/synthesis, and “pv.an,” used for the phase vocoder. (These formats are discussed in detail in Section 2.1.)

The principal difference between the “mq” and “pv.an” formats is that with the latter the amplitude and associated deviation from fixed harmonic frequency are stored for each harmonic at each frame, whereas with the “mq” format a set of amplitudes and corresponding absolute frequencies are given for each frame. Also, the “pv.an” format assumes that the same number of spectral components are used throughout the sound, whereas with the “mq” format the number of components varies with time. Another difference is that with the “pv.an” format only initial phases of the harmonics are given, whereas the “mq” format includes the starting phases for each time frame.

Frequency-tracking (“mq”) data can be converted to the harmonic format (“mq.an”) albeit with some loss of spectral information. First, a fundamental-frequency-vs-time track must be produced (see Section 1.2.3). Then that track is used to guide the extraction of each harmonic amplitude based on “mq” components whose frequencies are within a certain designated neighborhood of the expected harmonic frequency. The neighborhood is specified in terms of a “harmonic acceptance interval” (Δk), which is the fraction of departure from the expected harmonic frequency for a peak to be considered as a valid harmonic. Given an expected time-varying fundamental frequency f_0 and harmonic k , the amplitude and frequency of the harmonic are taken from the strongest amplitude peak component whose frequency lies between $(1 - \Delta k)kf_0$ and $(1 + \Delta k)kf_0$. In this way, any deviations from perfect harmonicity are preserved as long as they do not exceed the harmonic acceptance interval. Δk is typically taken to be 0.03 or 3%, corresponding to a semitone acceptance interval.

For each harmonic k , the frequency deviation that is actually stored is the difference between the extracted frequency and kf_a , where f_a is the expected average fundamental given by the user. This is appropriate for signals that have only one pitch. However, if the signal’s pitch variation exceeds a semitone, it is usually more appropriate to set f_a to zero and store absolute frequencies instead of frequency deviations.

Figure 1.12a, b shows block diagrams of a frequency-tracking analysis system with “.mq” (tracks) and “.mq.an” (harmonics) outputs.

1.2.5 Frequency-Tracking Synthesis

As with the phase vocoder, frequency-tracking synthesis can proceed either by sinusoidal additive synthesis (McAulay and Quatieri, 1986; Smith and Serra, 1987) or overlap-add synthesis (Rodet and Depalle, 1992).

1.2.5.1 Frequency-Tracking Additive Synthesis

Frequency-tracking additive synthesis uses a model similar to Eq. (1.2). Note that, unlike the phase vocoder, there is not a fixed number of sinusoidal components for the entire duration of a sound. As Eq. (1.2) only holds for the duration of each frame, it can be rewritten as

$$s(t_i + \Delta t) = \sum_{k=1}^{K_i} A_{k,i}(t_i + \Delta t) \cos(2\pi \int_{t_i}^{t_i + \Delta t} f_{k,i}(\tau) d\tau + \theta_{k,i}). \quad (1.35a)$$

where $t_i = it_{\text{frame}}$, is the beginning time of each frame, $0 \leq \Delta t < t_{\text{frame}}$ is the time between frames, K_i is the number of sinusoid partials or tracks, which varies from one frame to the next, $A_{k,i}(t)$ and $f_{k,i}(t)$ are functions that describe the change of the k th track amplitude and frequency during the i th frame, and $\theta_{k,i}$ is the k th track starting phase for the i th frame.

The discrete implementation follows from Eq. (1.21b):

$$\hat{s}_i(n) = \sum_{k=1}^{K_i} A'_{k,i}(n) \cos(\Theta_{k,i}(n)), \quad H'i \leq n < H'(i+1) \quad (1.35b)$$

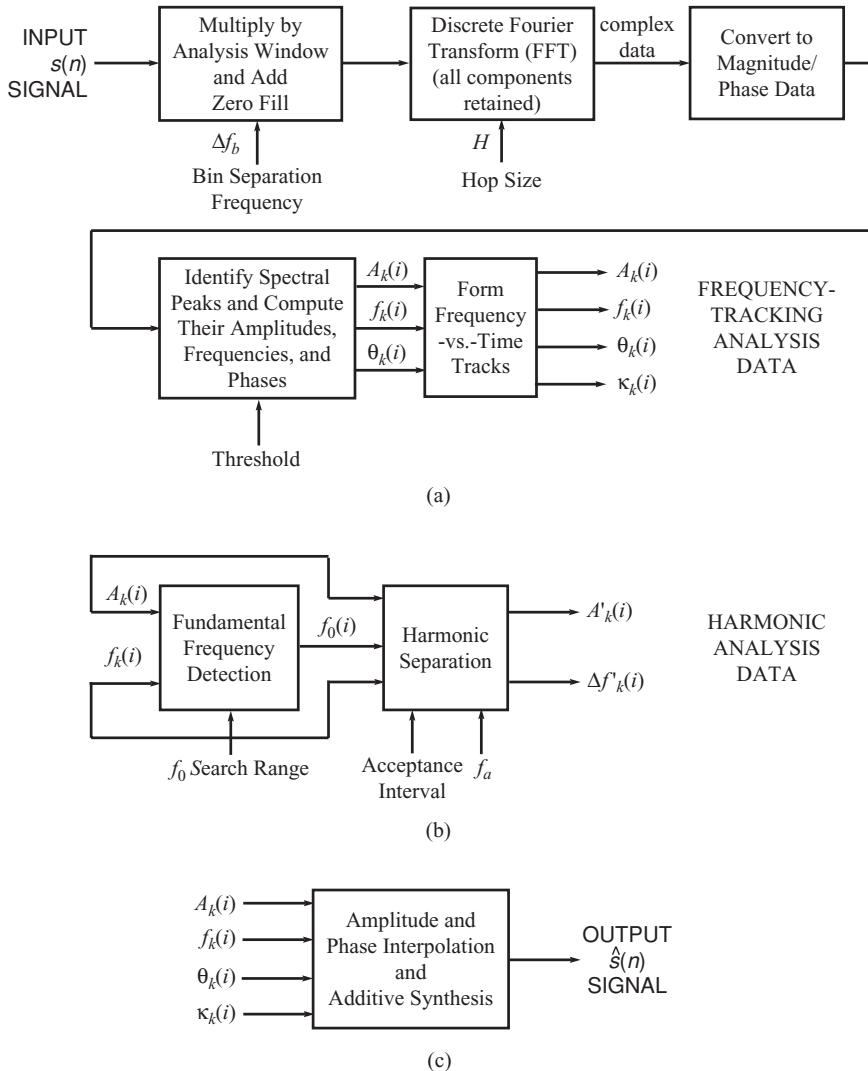


FIGURE 1.12. Block diagram of frequency-tracking analysis/resynthesis system: (a) analysis method; (b) fundamental frequency detection and reduction to harmonic analysis data; (c) resynthesis method based on frequency-tracking data.

n is now the sample counter, i is the frame counter, and H is the number of samples per frame. Amplitudes are generally linearly interpolated between frame boundaries whereas phases are interpolated using cubic interpolation as described in Section 1.1.4.2.4. H' can be different than the original H in order to allow time-scaling or synthesis at a different sample rate. It is desirable that frequencies,

amplitudes, and phases be matched at the frame boundaries. However, if frequency shifting is to be done, it is nearly impossible to specify what the most appropriate phase values should be, so a less complex phase interpolation algorithm may be sufficient.

Equations (1.35a) and (1.35b) are actually incomplete because they do not show that the k th partial may actually connect to a different partial number whenever a track birth or death occurs. In practice this is accomplished by associating a “link” value κ_k with the amplitude, frequency, and phase data for each frame and gives the partial number of the next frame the current partial is connected to.

Figure 1.12c shows a block diagram for frequency-tracking additive synthesis.

1.2.5.2 Residual Noise Analysis/Synthesis

Many musical sounds seem to have appreciable noise embedded in them. Wind instruments have varying amounts of attack noise and breath noise. Bowed strings have scrape noise. Percussion instruments have impact noise and perhaps damping noise at the ends of sounds. The problem is to make a clear distinction between noise (the random or stochastic part) and tone (the pitched or deterministic part). The main idea is to do the best possible analysis of the pitched part of a signal, $s_p(t)$, and then subtract the pitched resynthesis from the original signal, $s(t)$, to produce a residual noise, $n(t) = s(t) - s_p(t)$. Hopefully, the pitched components in $s(t)$ will be cancelled by this subtraction, leaving only a noise component with no audible pitch. It requires that pure sinusoidal components be properly identified with proper amplitudes, frequencies, and phases. This separation process was first implemented by Serra and Smith (1990) and is extended in Chapter 4 by Levine and Smith to include separate coding of transients.

Another viewpoint is that noise can be considered to be a modulation (amplitude and/or frequency) of the sinusoidal components. Indeed, if one views the graphs of amplitude and frequency vs time for various instruments, it is easy to see that the basically smooth functions are colored by a certain amount of what appears to be random variation. If these curves are smoothed and the sounds resynthesized, one can easily hear the reduction in noise (McAdams et al., 1999). Fitz et al. (2000) and Fitz and Haken (2002) have implemented an analysis/synthesis method that includes random modulation of sinusoid amplitudes that is described in Chapter 3.

In order to ensure that the deterministic part consists of truly sinusoidal partials, Serra (1989) and Serra and Smith (1990) eliminate peaks by (1) use of a minimum-peak-height parameter, (2) specification of ranges where peaks are expected, and (3) spectral peak continuation (SPC). SPC requires some knowledge of the signal to be analyzed. Not all peaks are considered to be equally important, and the goal is to form tracks only from the most important ones. The concept of frequency guides, which are similar to tracks, is used to assist in the selection of tracks. Because an individual sound is generally more stable in the middle or end of its time-span, track formation can start at these locations instead of at the attack, where sound spectra are frequently much more complex and more difficult to track reliably.

Once the deterministic tracks are determined, the sound is resynthesized from these tracks and then subtracted from the original signal to form the stochastic part of the signal. If the sinusoid phases are not computed accurately, the subtraction of the deterministic part from the original can be done in the frequency domain on the magnitude spectrum data.

The stochastic or noise residual signal is then approximated for synthesis. First the magnitude spectrum of the residual is approximated as a spectral envelope using a piecewise linear fit. If the residual was formed in the frequency domain, its magnitude spectrum is readily available; otherwise, it must be computed from the time-domain version. Synthesis proceeds by overlap-add using the computed magnitude spectra and random phase which changes on every frame. A hanning window is used for the overlap-add synthesis. An alternative approach is to use a linear predictive coding (LPC) approximation to the noise residual spectral envelope which is driven by white noise.

The main virtue of the sinusoid-plus-noise model is that, if separated properly, the noise part stays noiselike even when the sound is stretched, unlike the situation when sinusoidal components are time-stretched. In the latter case, the originally rapid variations of the sinusoids, which may imitate noise well at the original rate, become audible time variations, significantly changing the character of the sound.

Serra (1997) discusses using the two-way-mismatch pitch detection algorithm (discussed in Section 1.2.3) to enhance the pitched part separation in the case of quasiperiodic input signals. Fundamental frequency estimates are used to refine positions of the peaks and are also used to adjust the sizes of the analysis windows in an effort to improve the time-frequency tradeoff of the analysis.

1.2.5.3 Frequency-Tracking Overlap-Add Synthesis

Rodet and Depalle (1992) developed a method of overlap-add synthesis called inverse-FFT synthesis, which operates from arbitrary amplitude/frequency/phase-vs-time track data. Another way of thinking about the data is that they consist of a series of spectral envelopes which can be used to represent both sinusoidal and noise-residual data. The authors claim that with their FFT^{-1} approach computation is typically reduced by a factor of 10–30 over that required for the straightforward additive synthesis approach.

A description of the method begins by looking at the formula for a single sinusoid or partial of arbitrary amplitude A_k , frequency f_k , and zero-time phase θ_{k0} over a limited time interval. Such a sinusoid $s_k(t)$ would be defined by

$$s_k(t) = A_k \cos(2\pi f_k t + \theta_{k0}), -T/2 \leq t \leq T/2, \quad (1.36a)$$

where t = time, k = partial (or track) number, and T = window size.

Then the question is, what does this signal look like in the frequency domain? This can be computed by first multiplying $s_k(t)$ by a window function $w(t)$, such as the hanning window,

$$w(t) = \begin{cases} (2/T) \cos^2(\pi t/T), & |t| \leq T/2 \\ 0, & |t| > T/2 \end{cases}, \quad (1.36b)$$

and then applying the short-time Fourier transform,

$$\begin{aligned}
 S_k(f) &= \int_{-T/2}^{T/2} w(t)s_k(t)e^{-j2\pi ft}dt \\
 &= (A_k/2)e^{j\theta_k} \int_{-T/2}^{T/2} w(t)e^{-j(2\pi(f-f_k))}dt \\
 &\quad + (A_k/2)e^{-j\theta_k} \int_{-T/2}^{T/2} w(t)e^{-j2\pi(f+f_k)}dt \\
 &= (A_k/2)e^{j\theta_k} W(f - f_k) + (A_k/2)e^{-j\theta_k} W(f + f_k),
 \end{aligned} \tag{1.36c}$$

where $W(f)$, the Fourier transform of $w(t)$, is given by

$$W(f) = \text{sinc}(\pi f T) + 0.5[\text{sinc}(\pi(fT - 1)) + \text{sinc}(\pi(fT + 1))]. \tag{1.36d}$$

The bandwidth of this hanning response in the frequency domain is $4/T$, based on the response up to the first zero ($f = \pm 2/T$). However, if the side-lobe response is required to be below 0.01 (-40 dB), the width should be extended to $6/T$. It follows that if the frequency f_k of the sine-wave signal in Eq. (1.36a) is greater than $3/T$, for the purposes of calculating the $f > 0$ response, the second term of Eq. (1.36c) can be ignored and rewritten as the following approximation:

$$\hat{S}_k(f) \cong \begin{cases} A_k e^{j\theta_k} W(f - f_k), & |f - f_k| \leq 3/T \\ 0, & |f - f_k| > 3/T \end{cases} \tag{1.36e}$$

Similarly, for the $f < 0$ response, the second term would be used exclusively, resulting in an equation similar to Eq. (1.36e).

It follows that if the window size T is greater than 3 divided by the lowest frequency in the signal, Eq. (1.36e) will be adequate. Keep in mind, however, that even though the hanning window's transform is real, unless θ_k is 0 or π , a rare occurrence, $\hat{S}_k(f)$ will be complex.

For discrete signal synthesis, a sinusoid partial is sampled at frequency f_s and partitioned into 50% overlapped windows of size M with $M = Tf_s$. Assuming no zero fill, the frequencies of the DFT are spaced by $\Delta f = f_s/M = 1/T$. This means that $W(f - f_k)$ is sampled at six points within its principal nonzero region defined by Eq. (1.36e) (four points within its main lobe). Therefore, to synthesize a windowed sine wave at amplitude A_k , frequency f_k , and phase θ_k , the sampled frequency domain function $\hat{S}_k(m\Delta f)$, where m is the transform bin number in the range $(f_k T - 3, f_k T + 3)$, must be constructed first. Next, the complex response functions for all sinusoids k in the current frame must be summed, making sure that negative frequency as well as the positive frequency responses are included:

$$\hat{S}(m\Delta f) = \sum_{k=-K}^K \hat{S}_k(m\Delta f). \tag{1.36f}$$

Then, the inverse DFT of $\hat{S}(m\Delta f)$ is taken to obtain the samples $w(n/f_s)\hat{s}(n/f_s)$. Finally, successive frames, which are separated by $M/2$ samples, are overlap-added.

This is tantamount to forming

$$\begin{aligned} & (T/2)w(n/f_s)\hat{s}(n/f_s) + (T/2)w((n - M/2)/f_s)\hat{s}(n/f_s) \\ &= ((T/2)w(n/f_s) + (T/2)w((n - M/2)/f_s))\hat{s}(n/f_s) \quad (1.36g) \\ &= \hat{s}(n/f_s). \end{aligned}$$

The last step of Eq. (1.36g) is true if the window and its shifted version add to unity. With the definition of Eq. (1.36b) and noting that $n = tf_s$ and $M = Tf_s$, we have

$$\begin{aligned} & (T/2)w(n/f_s) + (T/2)w((n - M/2)/f_s) \\ &= \cos^2(\pi t/T) + \cos^2(\pi(t - T/2)/T) \\ &= \frac{1}{2} + \frac{1}{2}\cos(2\pi t/T) + \frac{1}{2} + \frac{1}{2}\cos(2\pi t/T - \pi) \quad (1.36h) \\ &= 1 + \frac{1}{2}\cos(2\pi t/T) - \frac{1}{2}\cos(2\pi t/T) = 1. \end{aligned}$$

At least this is what would happen if the transforms were taken from actual DFTs of the signal. However, the assumption implicit in Eq. (1.36a) is that frequencies are constant during each frame, whereas, in general, frequencies are actually changing during the frames. Therefore, Eq. (1.36g) is only an approximation. In fact, the second signal term in the first step of Eq. (1.36g) is generally slightly different than the first. This can cause phase cancellations between corresponding frequency components of the two frames, which would cause undesirable amplitude modulation. Rodet and Depalle (1992) discuss a method of phase adjustment to minimize this problem. They also discuss implementing linear interpolation between frames by multiplying the windowed signals of Eq. (1.35g) by ratios of triangular windows divided by the $w()$ windows. As an extension of the basic technique, Goodwin and Rodet (1994) discuss changing the basic assumption of Eq. (1.36a) to one where frequency changes linearly in time over the window. Depending on the amount of frequency change (it is usually small), the constructed frequency-domain windows $W(f - f_k)$ will be warped compared to their zero-frequency-change versions. The payoff is that phases of adjacent frames should now line up and phase cancellation should no longer be a problem.

Figure 1.13 shows a block diagram of the inverse-FFT system.

2 Analysis Results Using SNDAN

The SNDAN software package, developed for Unix at the University of Illinois at Urbana-Champaign and ported to DOS by Richard Dobson in the United Kingdom, can be used to perform spectral analysis, graph the results, and perform spectrotemporal modifications and resynthesis from the spectral data. Two time-varying spectrum analyzers are provided, a phase vocoder analyzer (pvan) and a frequency-tracking (McAulay and Quatieri, 1986) analyzer (mqan). As described in Section 1.1, the phase vocoder can be tuned for the fundamental frequency of

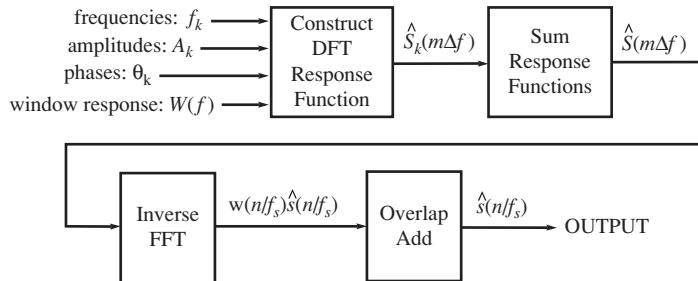


FIGURE 1.13. Block diagram of an inverse-FFT synthesis system for resynthesis from a single frame of amplitudes, frequencies, and phases. Responses for all components k are summed before the inverse FFT occurs. The output signal is achieved by overlap-add of adjacent frames' signal segments.

a fixed-pitch input signal so that the bins line up with the harmonics of the input signal. On the other hand, the frequency-tracking analyzer saves data in the form of spectral peaks for each frame, where each peak is represented by four numbers: amplitude, frequency (in Hz), phase (in radians), and a “link” that gives the track number of the next frame’s track to which the current peak is linked. If it is zero, it is assumed that the track ends (“dies”) at that frame.

2.1 Analysis File Data Formats

Data saved in an analysis file (by either pvan or mqan) are comprised of three parts. Parts 1 and 2 comprise the file header. Part 3 comprises the analysis data, which can be one of three forms: ‘pv.an’ (phase vocoder output), ‘mq’ (frequency-tracker output), or ‘mq.an’ (‘mq’ converted to the ‘an’ format).

- 1 Musicological data. This consists of the following text information:
 - (a) performer name
 - (b) instrument played
 - (c) date of recording
 - (d) pitch played (e.g., C₄)
 - (e) performed dynamic (e.g., *ff*)
 - (f) vibrato (yes or no)
 - (g) portion of original sound (e.g., “all” or “all but attack”)
 - (h) date of analysis
 - (i) additional comments
- 2 File-critical data, consisting of
 - (a) data type (e.g., “simple” for ‘an’ analysis or “MQ” for ‘mq’ analysis)
 - (b) sample rate of signal analyzed (in samples/second)
 - (c) sound duration (in seconds)
 - (d) maximum amplitude of input signal

- (e) analysis frequency (in Hz)
 - (f) time between analysis frames (in seconds)
 - (g) analysis block size (FFT length) (in samples)
 - (h) number of harmonics ('an' data only)
 - (i) number of channels (e.g., 1 for monaural, 2 for stereo)
 - (j) number of frames in analysis file
 - (k) analysis reinterpolation factor (seldom used)
- 3 The analysis data. This depends on whether the data are 'an' or 'mq':
- (a) 'an' data (usually produced by pvan):
 - (i) initial phase (in radians) for all harmonics.
 - (ii) for each frame: {amplitude, frequency deviation or absolute frequency (in Hz)} for all harmonics.
 - (b) 'mq' data (usually produced by mqan):

for each frame: number of spectral peaks; peak data consisting of {amplitude, frequency (in Hz), phase (in radians), and link} for all peaks.

As discussed in Section 1.2.4, by using frequency detection (Beauchamp et al., 1993; Maher and Beauchamp, 1994) and harmonic separation (Beauchamp, 1993), it is possible to convert an 'mq' data file to an 'mq.an' data file. When frequency deviates considerably, f_a is usually set to zero and the "frequency deviations" become the actual partial frequencies. The principal reason for going through this procedure is that the 'an' harmonic format is much simpler than the 'mq' frequency-tracking format, and more software has been developed for it. A side benefit is that the procedure performs a certain amount of noise reduction on the signal.

2.2 Phase-Vocoder Analysis Examples for Fixed-Pitch Harmonic Musical Sounds

Once a signal has been analyzed by pvan and the analysis data are placed in a data file, the program monan can display data in a number of different ways. Figure 1.14 shows individual amplitude-vs-time graphs for the first six harmonics of a long (8 s) trumpet sound. Fig. 1.15 shows the corresponding normalized-frequency-deviation-vs-time graphs, where the deviations are given in terms of $\Delta f_k / (kf_a)$ and k is the harmonic number. A problem with computing the frequency deviation using the phase-vocoder method is that, when an amplitude is low, the corresponding computed frequency becomes very noisy. While the sound can be resynthesized with high quality using these data, if the amplitude spectrum is altered such that weak harmonics are amplified, the frequency noise can create noticeable synthesis artifacts. This noise problem can often be cured by selectively filtering or zeroing the frequency-deviation data.

The harmonic amplitude and frequency graphs can also be shown in composite fashion. A three-dimensional harmonic amplitude-vs-time graph for a shorter trumpet tone at this pitch is shown in Fig. 1.4a. A composite frequency-vs-time

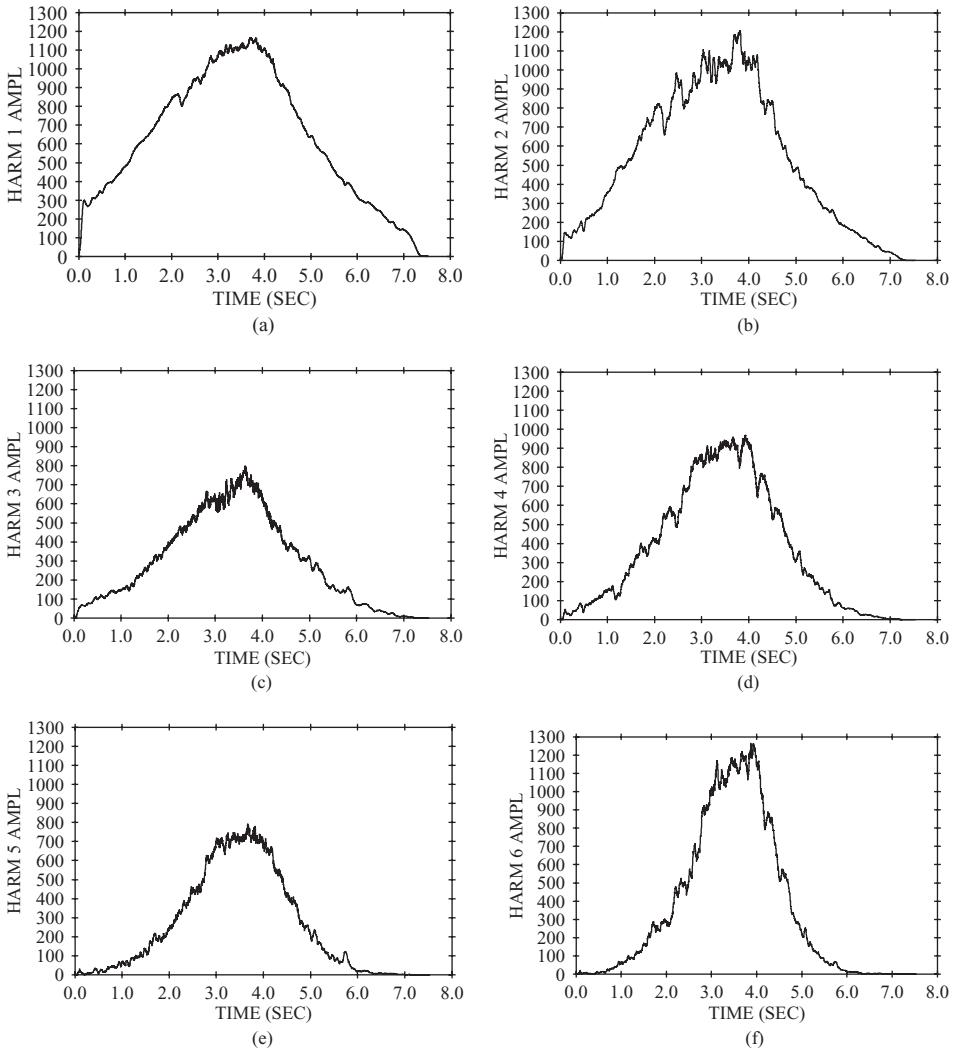


FIGURE 1.14. Phase vocoder analysis of an F4 (350 Hz) trumpet tone played *pp* < *ff* > *pp*: (a)–(f) harmonic amplitudes A_k vs time for harmonics $1 \leq k \leq 6$.

graph for the same tone, where darkness indicates amplitude, is shown in Fig. 1.4b. Because lower amplitude harmonics register very light on this scale, the attendant frequency noise is hardly visible.

2.2.1 Spectral Centroid

Many other forms of data display are provided by monan. For example, a well-known and popular measure of the spectrum is the spectral centroid, which is

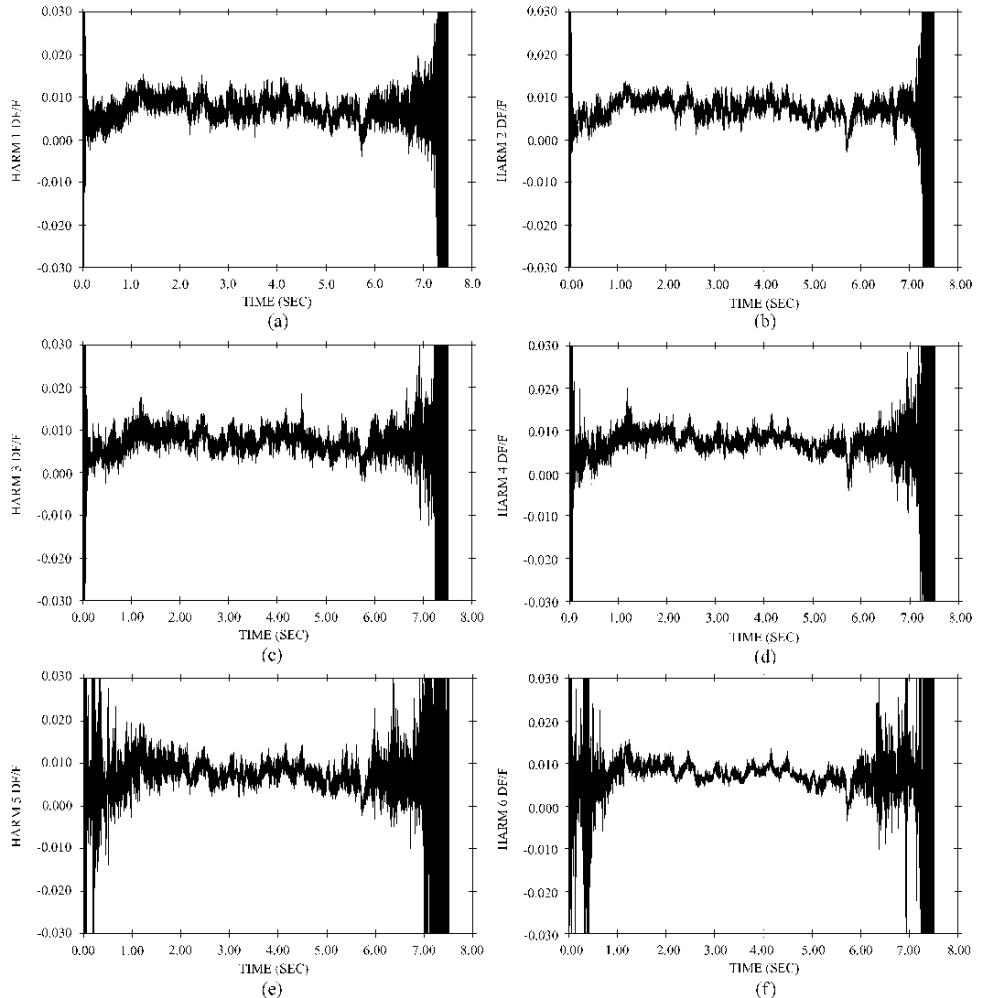


FIGURE 1.15. Phase vocoder analysis of the F_4 trumpet tone: (a)–(f) relative frequency deviations $\Delta f_k / (k f_a)$ vs time for harmonics $1 \leq k \leq 6$. A relative deviation of 0.03 (i.e., 3%) corresponds to approximately a half-semitone change of frequency.

closely related to perceptual brightness. For many musical instrument sounds, it is important that this varies with time. When normalized by the fundamental frequency, the centroid is defined by

$$BR(t) = \frac{\sum_{k=1}^K k A_k(t)}{\sum_{k=1}^K A_k(t)}. \quad (1.37a)$$

Note that $BR(t)$ can be thought of as a time-variant spectral-amplitude-averaged harmonic number, so that Eq. (1.37a) can be rewritten as

$$BR(t) = \sum_{k=1}^K \alpha_k(t) k, \quad (1.37b)$$

where

$$\alpha_k(t) = \frac{A_k(t)}{\sum_{k=1}^K A_k(t)}. \quad (1.37c)$$

The time-varying weight $\alpha_k(t)$ gives the fraction of the total amplitude applied to the harmonic number variable k , providing an amplitude-average value of the harmonic number in Eq. (1.37b). (Of course, an ordinary average of k would be amplitude-independent and would not be useful.) As explained later in this chapter, amplitude-averaging is a very useful operation, and can be applied to a variety of situations.

A version of the unnormalized spectral centroid is also useful:

$$f_c(t) = (BR(t) - 1) f_a. \quad (1.37d)$$

The spectral centroid can be thought of as a measure or indicator of the richness or breadth of the spectrum. Because it is independent of the actual amplitude scale of a signal, if the amplitudes of Eq. (1.37) were multiplied by a constant factor, BR would not change. However, BR (or f_c) often has a close relationship to the overall RMS amplitude of the signal, defined as

$$A_{\text{rms}}(t) = \sqrt{\sum_{k=1}^K A_k^2(t)}. \quad (1.38)$$

Figures 1.16a and 1.16b show $BR(t)$ and $A_{\text{rms}}(t)$ for the long trumpet sound, and Fig. 1.16c shows $BR(t)$ plotted against $A_{\text{rms}}(t)$. Note that there is a burst of high BR at the beginning of the sound (occurring at low amplitude) followed by a dip and then a slower increase, which indicates that upper harmonics become stronger as time progresses. This can be verified by looking at Figs. 1.4a and 1.4b. Also, BR decays to a low value at the end of the sound. To disguise the effect of background noise or breath noise at the end of a sound, a small constant can be added to the numerator and denominator of Eq. (1.37a). The strong relationship between BR and A_{rms} is evident from Fig. 1.16c.

Spectral centroid can be used to quantize the fact that some instruments' sounds (at the same pitch) are brighter than others (e.g., trumpet is generally brighter than French horn). Also, many instrument sounds exhibit significant changes of centroid during sounds that are very noticeable by listeners (McAdams et al., 1999). Table 1.1 gives maximum and average BR values for a number of instruments all playing E₄^b (311.1 Hz).

Spectral centroid can be modified by replacing the harmonic amplitudes by those which are multiplied by a monotonic increasing or decreasing function of

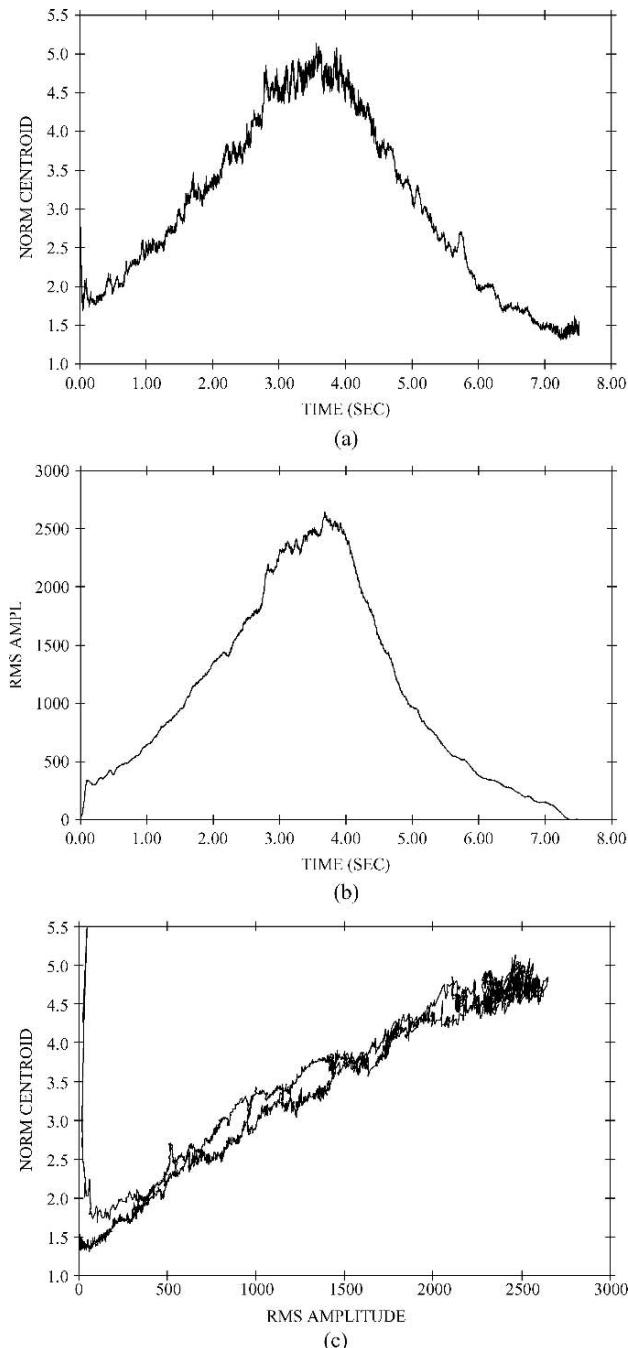


FIGURE 1.16. For the F₄ trumpet tone: (a) normalized spectral centroid *BR* vs time; (b) RMS amplitude vs time; (c) *BR* vs RMS amplitude (31 harmonics used in calculations).

TABLE 1.1 Average and Maximum Normalized Centroids for 14 Instrument Sounds

Instrument	Average centroid	Maximum centroid
Bassoon	3.2	8.6
Cello	4.6	14.6
Clarinet	6.4	11.1
Flute	3.4	11.2
Harp	1.6	15.2
Harpsichord	7.9	31.0
Marimba	1.4	6.7
Oboe	4.5	6.3
Recorder	2.0	6.5
Alto saxophone	4.1	9.8
Trumpet	4.9	5.8
Horn	2.5	5.4
Vibraphone	1.3	6.7
Violin	4.6	7.5

harmonic, such as

$$A_k \leftarrow k^p A_k. \quad (1.39)$$

When $p > 0$, the centroid is increased, and when $p < 0$, the centroid is decreased. Because the relation between centroid and p is monotonically increasing, virtually any centroid can be matched using a straightforward optimization technique such as the Newton method.

For brass tones there are strong nonlinear relationships between harmonic amplitudes (Beauchamp, 1975; Benade, 1976). Fig. 1.17 shows graphs of $A_k(t)$ vs $A_{\text{rms}}(t)$ for the long trumpet sound for the first six harmonics. Note that as the harmonic number increases, the curves are pushed to the right. Sampling the spectrum at high RMS amplitude automatically yields a spectrum with more relative energy in the upper partials than at low RMS amplitude.

For the flute, the situation is similar but more complex. When listening to a long swell tone one can easily hear harmonics popping in and out. Figs. 1.18 and 1.19 and 1.21 and 1.22 show graphs for a flute sound comparable to Figs. 1.14–1.17 for the trumpet. The more jagged nature of the flute's harmonic envelopes is evident from Fig. 1.18. For example, $A_5(t)$ for the trumpet and the flute can be seen in Figs. 1.14e and 1.18e, respectively. Note how the trumpet's harmonic 5 amplitude rises and falls smoothly, whereas the flute's amplitude suddenly rises to a peak level at 3.5 s and then falls again at 5.2 s. Also, as seen from Figs. 1.17 and 1.22, a much "tighter" relationship is evident between the individual harmonic and corresponding RMS amplitude for the trumpet than for the flute. Nevertheless, while the flute's normalized spectral centroid and RMS amplitude (see Figs. 1.21a and 1.21b) do not vary with time as smoothly as the trumpet's (Figs. 1.17a and 1.17b), both RMS-vs-centroid curves follow definite trends upward.

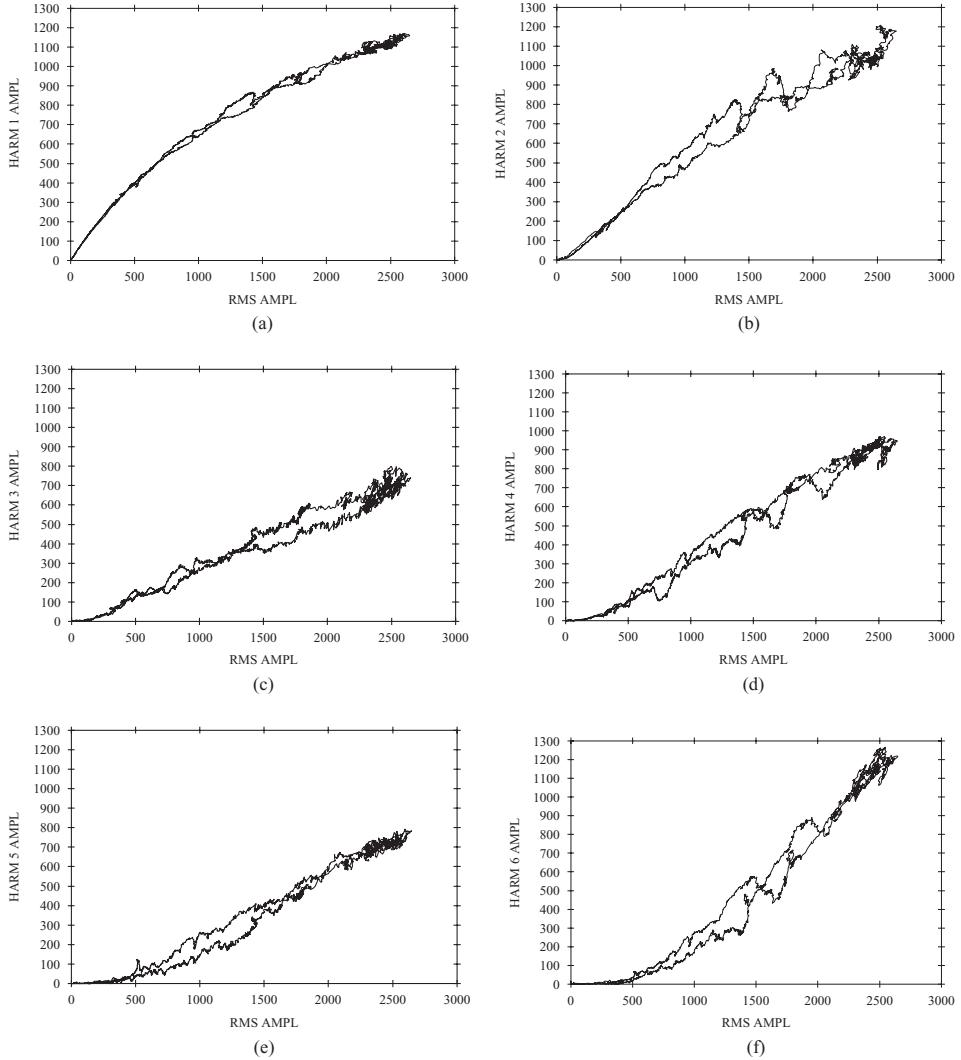


FIGURE 1.17. For the F_4 trumpet tone: (a)–(f) harmonic amplitudes A_k vs RMS amplitude for harmonics $1 \leq k \leq 6$.

For completeness, Figs. 1.19 and 1.20 show the flute's original and smoothed normalized frequency deviations for harmonics 1–6.

2.2.2 Spectral Envelopes

Another important feature of musical sounds is the spectral envelope. Spectra can be shown as vertical line graphs, which emphasize the individual harmonics,

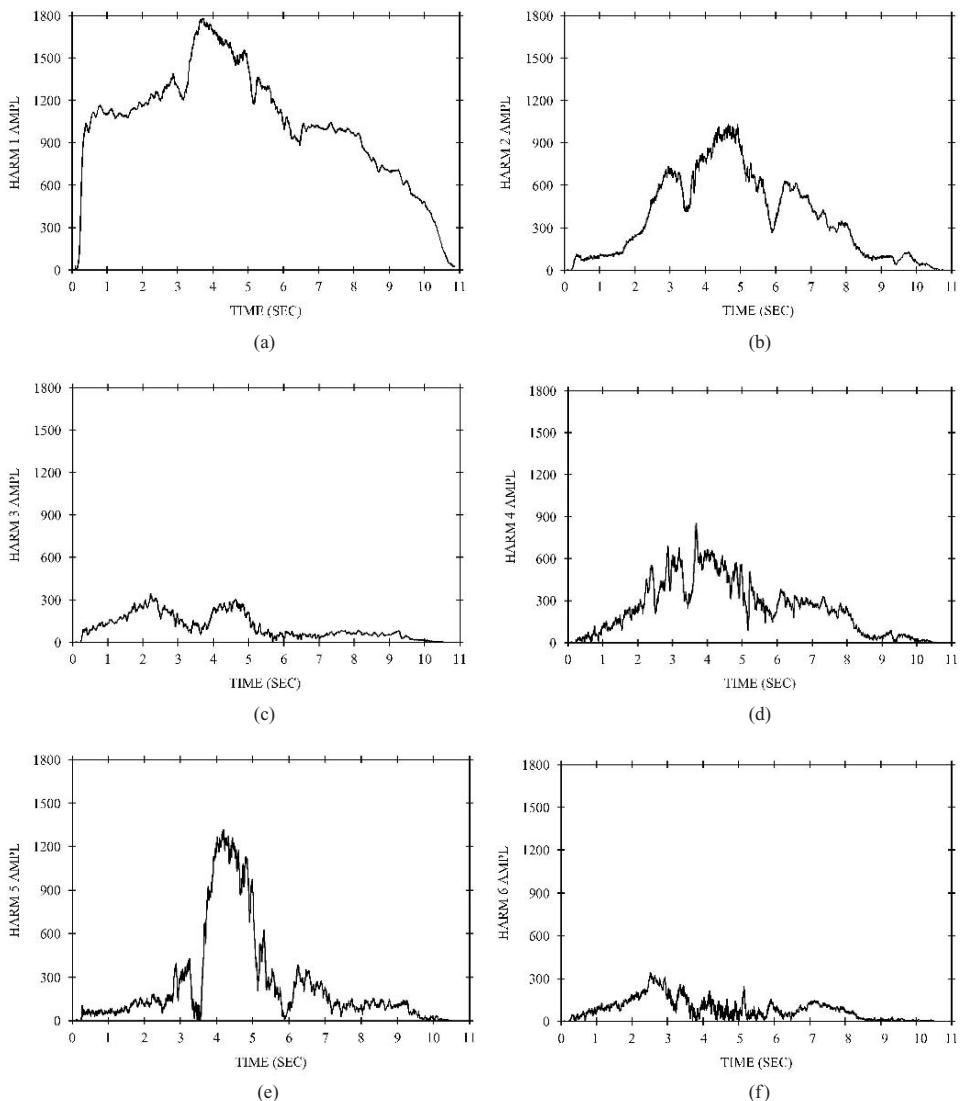


FIGURE 1.18. Phase vocoder analysis of an E₄ (330 Hz) flute tone played *pp* < *ff* < *pp* : (a)-(f) harmonic amplitudes A_k vs time for harmonics $1 \leq k \leq 6$.

or as connected line graphs, which show the overall shape of the spectrum. For the latter, there is a choice of whether to use linear or log frequency for the horizontal axis. In any case, there is also a choice between linear or decibel amplitude scale for the vertical axis. Fig. 1.23 shows spectral envelopes for the long trumpet and flute tones, both as vertical lines (linear amplitude vs linear frequency) and as connected graphs (linear amplitude vs linear frequency)

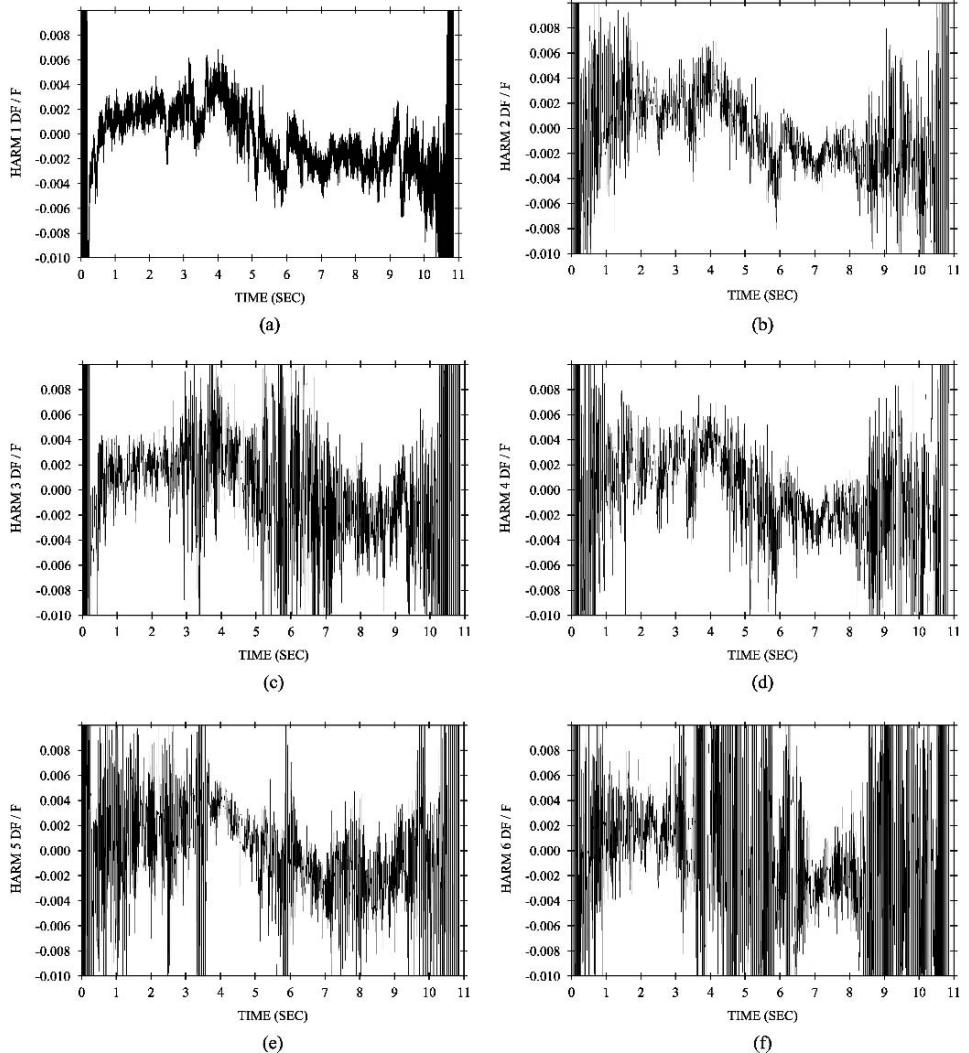


FIGURE 1.19. Phase vocoder analysis of the E₄ flute tone: (a)–(f) relative frequency deviations $\Delta f_k/(k f_a)$ vs time for harmonics 1 to 6. A relative deviation of 0.01 (i.e., 1%) corresponds to approximately a one-sixth-semitone change of frequency.

and decibel amplitude vs log frequency). These were computed at the apexes of the sounds' RMS amplitudes. It is apparent, especially from the dB-vs-log-f plots, that the flute spectrum is much more jagged and that it rolls off more quickly (has a smaller bandwidth) than the trumpet. Methods for representing spectral envelopes are discussed in detail in Chapter 5 by Rodet and Schwarz.

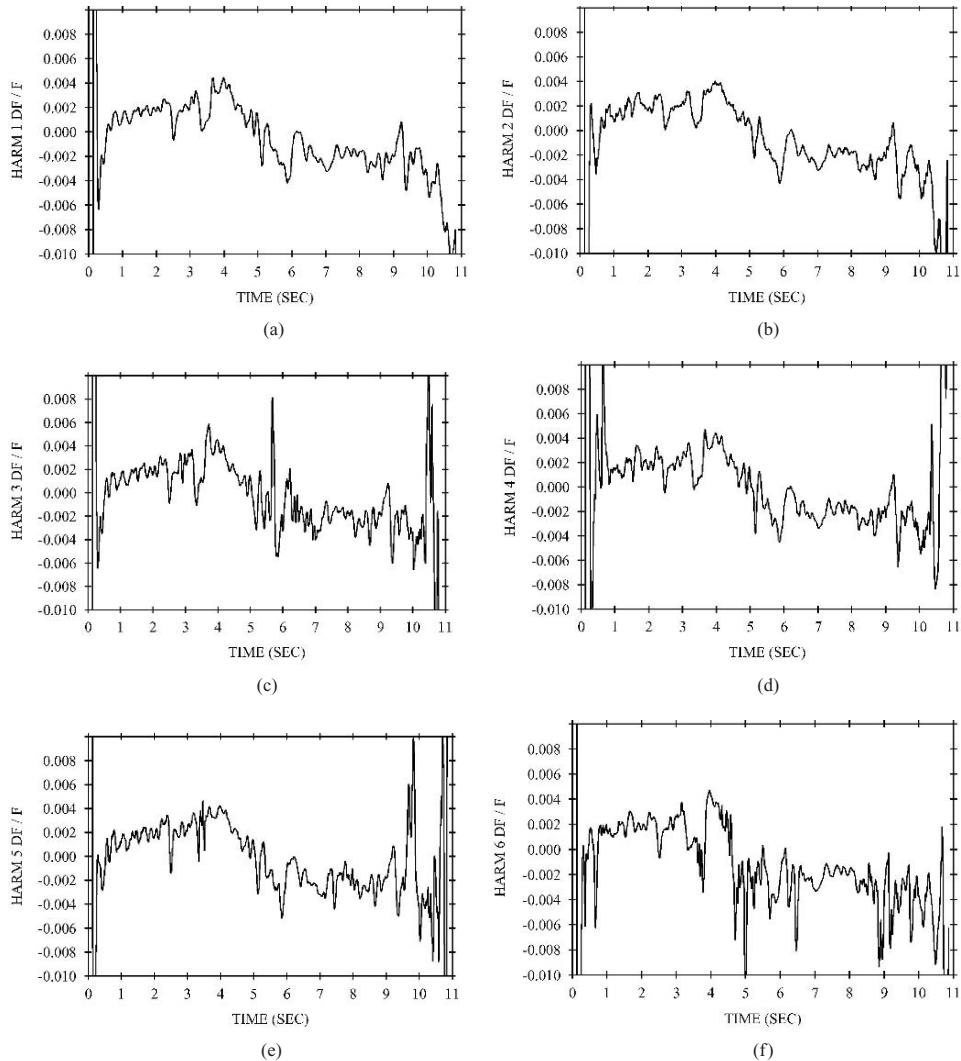
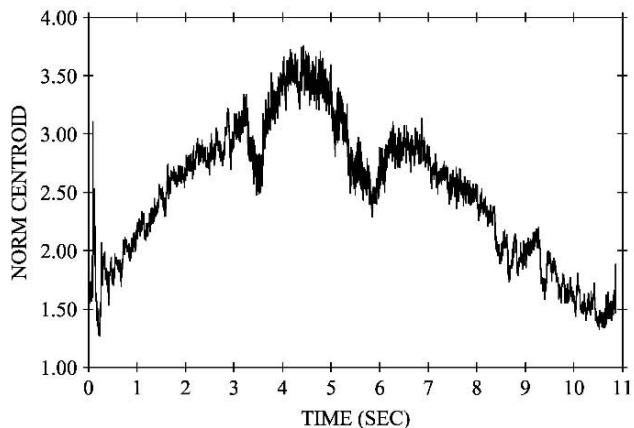
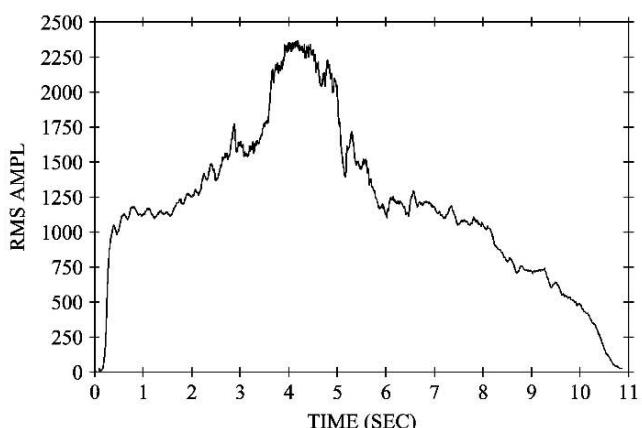


FIGURE 1.20. For the E₄ flute tone: (a)–(f) relative frequency deviations $\Delta f_k/(kf_a)$ vs time smoothed by a 5-Hz-cutoff low-pass filter for harmonics 1 to 6.

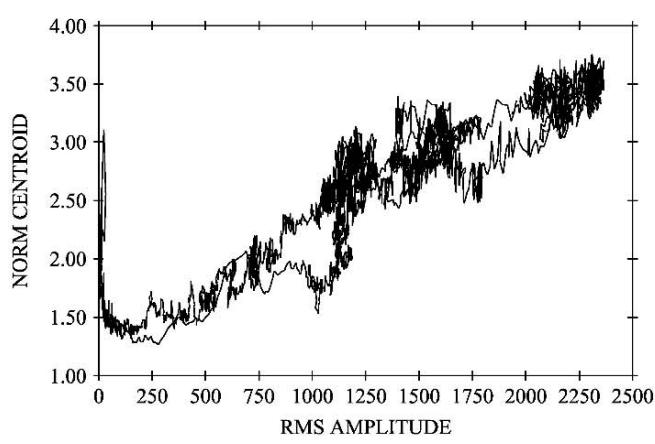
Average spectral envelopes can be computed for groups of tones for the same instrument. This was first done by Luce (1963), Strong and Clark (1967a,b), Luce and Clark (1967), and Luce (1975) for tones played at specific dynamic levels (***pp***, ***mf***, and ***ff***). Rather than using dynamics to segregate the spectral envelopes, the spectral envelopes of Fig. 1.24 were clustered and averaged based on the spectral centroids [as defined by Eq. (1.37d)]. Details on this type of computation are given by Beauchamp and Horner (1995).



(a)



(b)



(c)

FIGURE 1.21. For the E_4 flute tone: (a) normalized spectral centroid BR vs time; (b) RMS amplitude vs time; (c) BR vs RMS amplitude (31 harmonics used in calculations).

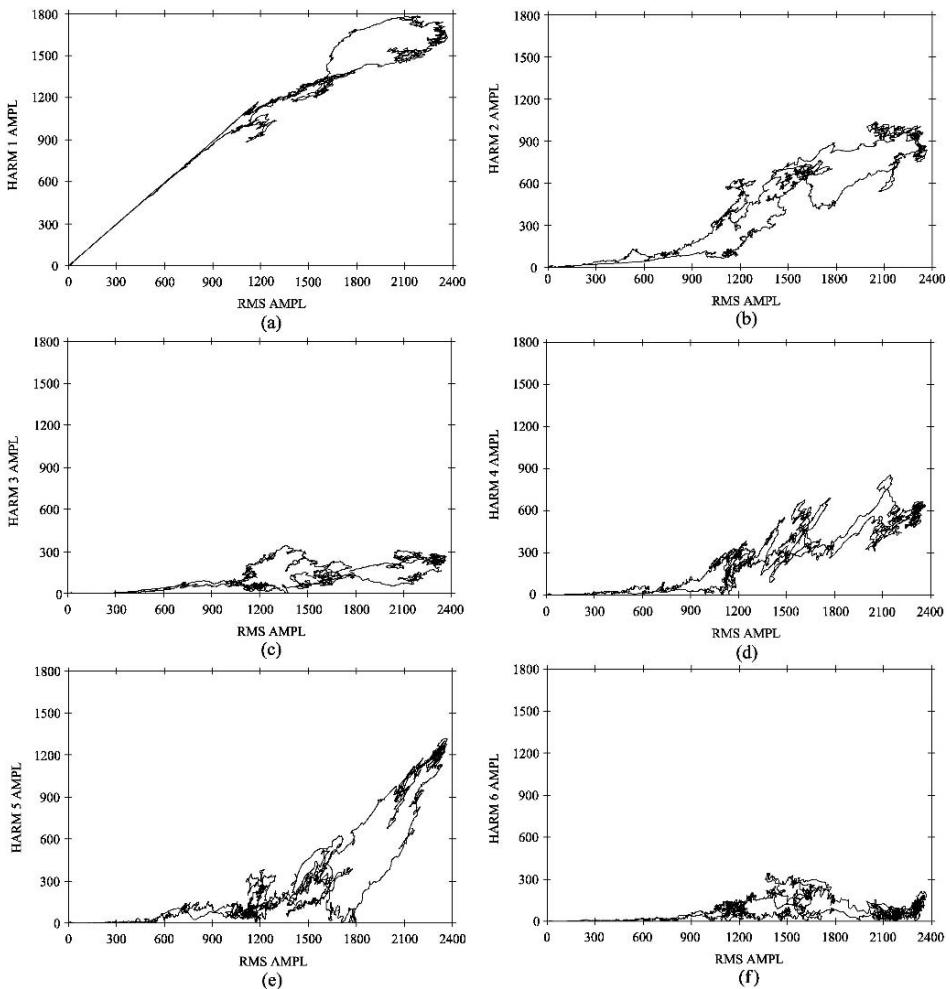


FIGURE 1.22. For the E₄ flute tone: (a)–(f) Harmonic amplitudes A_k vs RMS amplitude for harmonics 1 to 6.

2.2.3 Spectral Irregularity

A measure of “jaggedness” or spectral irregularity compares the spectrum to a smoothed version of itself:

$$SIR(i) = \frac{\sum_{k=2}^{K-1} \vec{A}_k(i) \| A_k - \vec{A}_k(i) \|}{A_{\text{rms}}(i) \sum_{k=2}^{K-1} \dot{\vec{A}}_k(i)}, \quad (1.40a)$$

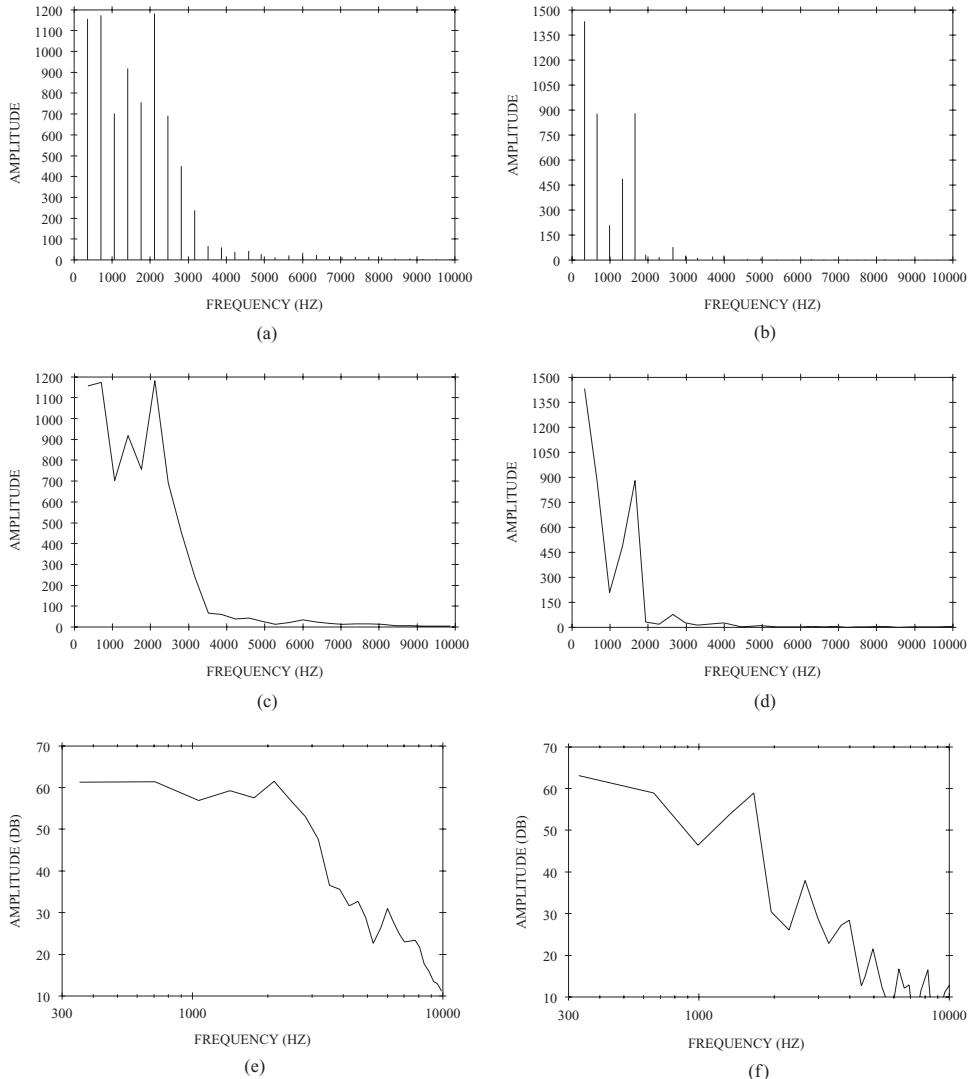


FIGURE 1.23. Spectral envelopes for the F_4 trumpet (left, taken at $t = 3.7$ s.) and the E_4 flute (right, taken at $t = 4.2$ s.) tones: (a), (b) linear amplitude-vs-frequency bar graphs; (c), (d) linear amplitude-vs-frequency line graphs; (e), (f) decibel amplitude-vs-log frequency line graphs.

where i is the analysis frame number and

$$\ddot{A}_k(i) = (A_{k-1}(i) + A_k(i) + A_{k+1}(i))/3 \quad (1.40b)$$

is the spectrally smoothed harmonic amplitude. Note that with Eq. (1.40a) the magnitude of the difference between the original harmonic amplitude and its smoothed

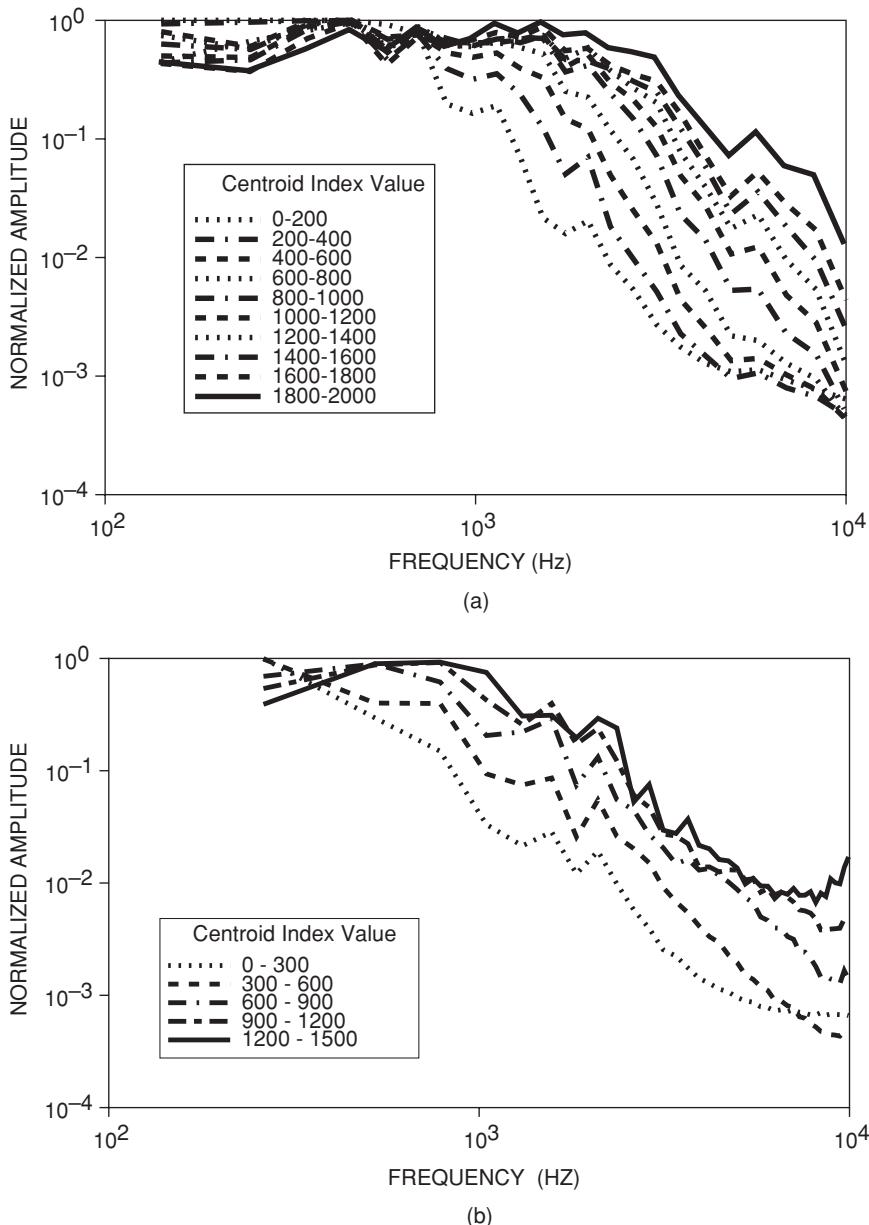


FIGURE 1.24. Spectral envelope families for the trumpet and flute where each envelope corresponds to a range of unnormalized spectral centroid. (a) Trumpet spectral envelopes are averaged for tones over a range of pitches (from F_3 to F_5) and dynamics (*pp* to *ff*). (Horner and Beauchamp, 1995, Fig. 6, reproduced by permission of the Audio Eng. Soc.) (b) Flute spectral envelopes are averaged for tones at a single pitch (C_4) and a range of dynamics (*pp* to *ff*).

version is amplitude-averaged over the harmonics and then is normalized by the RMS amplitude so that the value of *SIR* is independent of any amplitude scaling and will lie between 0 and 1. Fig. 1.25 shows *SIR*-vs-time measurements for short trumpet and flute sounds. In general, the flute spectrum has a greater irregularity than that of the trumpet. Amplitude-averaging over the entire sounds gives *SIR* = 0.11 for the flute and 0.06 for the trumpet.

Spectral irregularity can be modified. For example, a spectrum can be smoothed by using the method of Eq. (1.40b). This is tantamount to a simple low-pass filter operating in the frequency domain. A high-pass filter could be used to increase the irregularity. Another method is to process the spectrum in a nonlinear fashion. For example, a harmonic amplitude can be replaced by an exponentiated version:

$$A_k \leftarrow A_k \left(\frac{A_k}{\max(A_k)} \right)^p. \quad (1.41)$$

When $p > 0$, peak amplitudes are increased relative to weaker amplitudes. The opposite is true for $p < 0$. The result is to accentuate or deaccentuate spectral irregularity. A feature of this formulation is that the maximum spectrum amplitude is not changed.

Spectral irregularity appears to have a profound effect on a sound's timbre. McAdams et al. (1999), found that on average, 96% of the time, subjects could correctly distinguish between sounds synthesized with full data and those with spectrally smoothed data. (The lowest figure, 82%, was for the trumpet, which had a relatively smooth spectrum to begin with.) One might argue that the ability to distinguish was due to the smoothing operation altering the spectral centroid. However, except for a marimba sound (which had few partials), the change of centroid due to this operation was less than about 4%. In another study, Horner et al. (2004) found that listeners could discriminate random spectral changes, which increase spectral irregularity, with 78–90% accuracy if the average random spectral error was 24%. Still, despite its obvious importance, no particular perceptual attribute has been found to correspond with spectral irregularity.

2.3 Phase-Vocoder Analysis of Sounds with Inharmonic Partials

Sounds with inharmonic partials can be divided into three categories:

1. Sounds with nearly harmonic partials. Most plucked (e.g., guitar) or struck string (e.g., piano) sounds fall into this category. Equations frequently given [e.g., Fletcher (1964); Lattard (1993)] for the partial frequencies (aka mode frequencies) of a plucked or struck string are

$$f_k = k f_o \sqrt{1 + B k^2} \approx k f_1 [1 + (B/2)(k^2 - 1)], \quad (1.42a)$$

where B is the inharmonicity constant, f_o is frequency for a string with no stiffness, and f_1 is the actual fundamental frequency. For piano sounds B is normally in the range of 0.0001 to 0.001 for f_1 below 1000 Hz and 0.001 to 0.01 above that

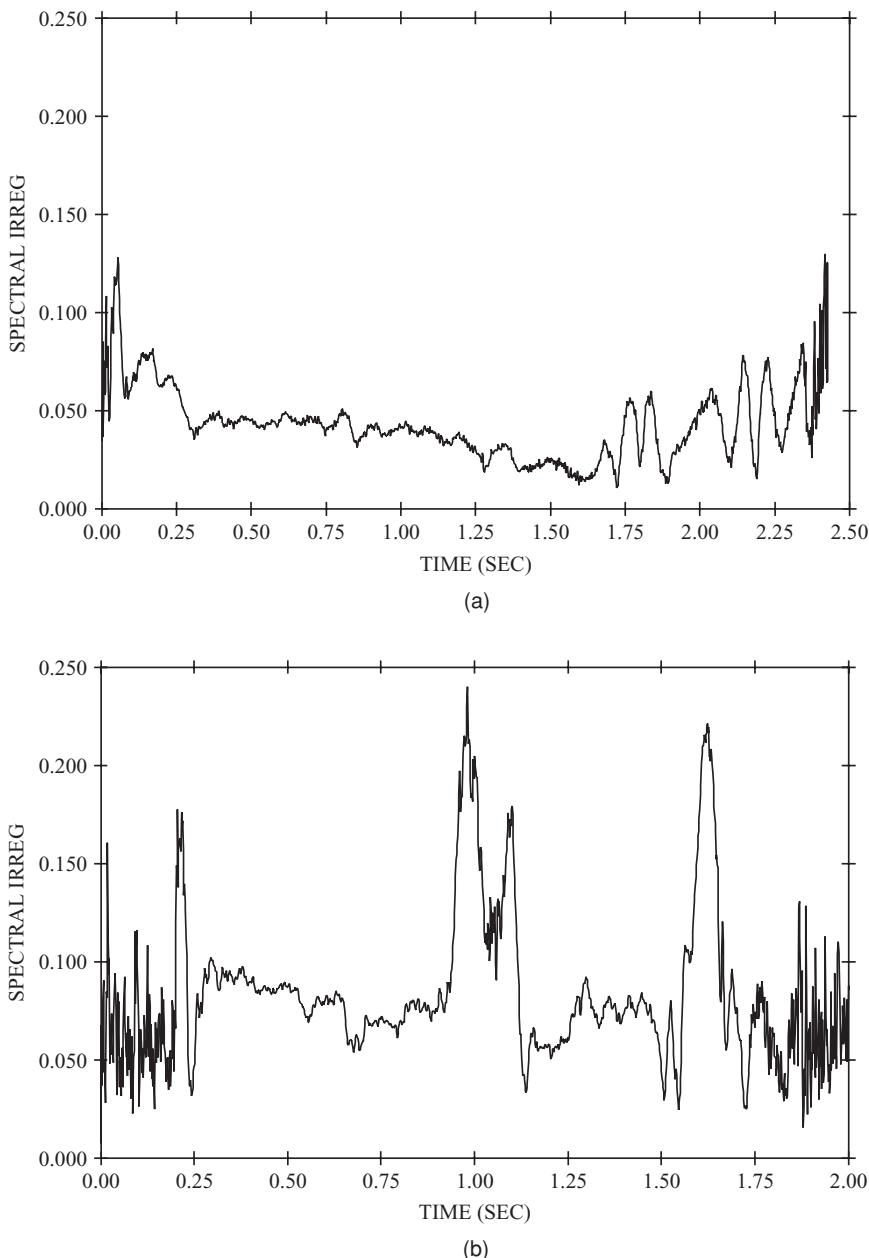


FIGURE 1.25. Spectral irregularity vs time for (a) the F₄ *ff* trumpet tone and (b) an E₄ *ff* flute tone.

frequency. Solving for B in terms of the amount of stretching of the k th partial $\Delta f_k = f_k - kf_1$ gives

$$B = \frac{2\Delta f_k}{(k^2 - 1)kf_1}, k > 1. \quad (1.42b)$$

2. Sounds with widely spaced (sparse) partials. Wooden and metal bar instruments, such as xylophone, marimba, vibraphone, and chimes, fall into this category. An approximate formula for their partial frequencies (Morse, 1976, p. 162) is

$$f_k \cong f_o(2k + 1)^2, \quad (1.43)$$

where f_o is a constant that depends on the geometry and type of material used. Bells, which can be thought of as “deformed thick metal plates,” also fall into this category, but the formula for their mode frequencies is much more complex (Morse, 1976).

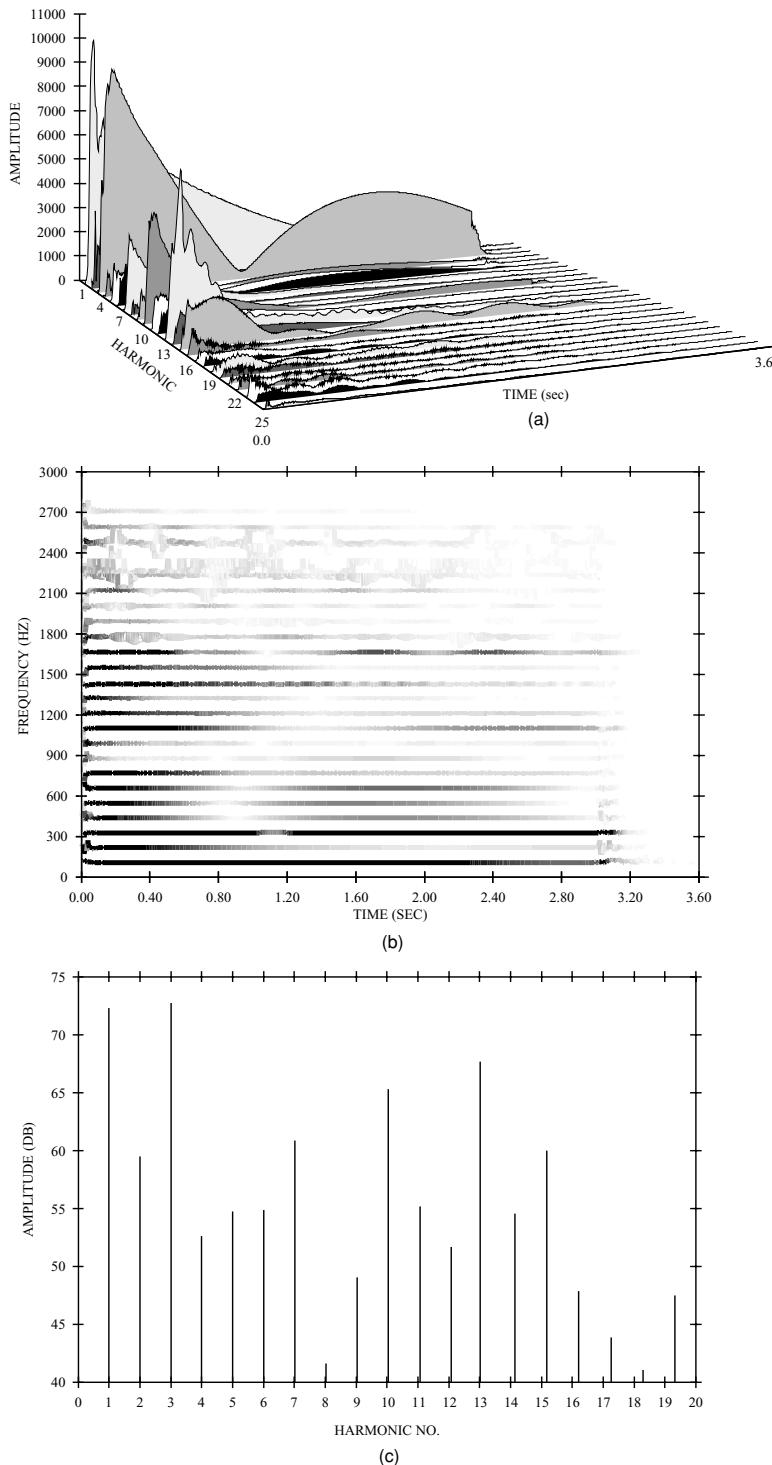
3. Sounds with closely spaced (dense) partials. Stretched membranes and thin metal plates, such as drums, cymbals, and gongs (tam-tams), fall into this category. While simplified formulas for their mode frequencies exist (Morse, 1976), they do not correspond to actual measurements.

At first glance, it might seem that frequency-tracking analysis would be the best approach for analysis of sounds with inharmonic partials because that method is capable of handling arbitrary frequencies. However, there is a problem with tracking partial frequencies as they continually rise and fall in amplitude. For sounds with constant mode frequencies, the phase vocoder seems to be more robust for analysis, even though choosing the best analysis frequency is a non-trivial problem. For sounds that are nearly harmonic, an analysis frequency close to the fundamental usually is sufficient. For widely spaced frequencies it is possible to choose a frequency that is close to an integer divisor of the partial frequencies. For closely spaced frequencies it seems best to choose a very low fundamental, in the neighborhood of 10–20 Hz. This gives a quite good frequency resolution, but any lower value would seriously compromise time resolution.

2.3.1 Inharmonicity of Slightly Inharmonic Sounds: The Piano

Figure 1.26a shows amplitude-vs-harmonic-vs-time (3D) phase vocoder data for an A₂ piano tone, where the phase-vocoder analysis frequency f_a was chosen to correspond to the tone’s fundamental frequency. Note that the partials have different decay rates, and some partials exhibit marked undulatory behavior. This is further shown in the composite frequency-vs-time (2D) graph (i.e., a spectrogram) of Fig. 1.26b, where darkness indicates amplitude. This does not illustrate the

FIGURE 1.26. Phase-vocoder spectral analysis of an A₂ (110 Hz) piano tone. (a) 3D spectrogram showing the amplitude-vs-time behaviors of the partials. (b) 2D spectrogram, where darkness indicates amplitude. (c) Time-averaged spectrum of the piano tone in terms of amplitude (in dB) vs average normalized frequency (f_k/f_a).



inharmonicity phenomenon very well, however, because the amount by which the partial frequencies exceed their corresponding harmonics slowly migrates upward as the partial numbers increase. The inharmonicity constant B can be measured by first estimating $\Delta f_k = \bar{f}_k - k\bar{f}_1$ from these frequency data (using amplitude-averaging over time) and then computing B using Eq. (1.42b). Unfortunately, the results are not as consistent as one might expect, as the lower partials yield values lower than the higher ones and some individual partials yield values that are much higher or lower than expected.

Reasons for compromised accuracy in computing B include:

First, when a partial is weak or contains a strong oscillation—probably due to it consisting of two or more closely tuned components—the resulting frequency-vs-time graph is noisy or may contain a sudden jump due to phase interaction of the components.

Second, when a partial number reaches a certain value, the corresponding frequency component begins to appear in two adjacent analysis bin outputs, so that the frequencies appearing in those outputs are actually the same. This is not a problem with the combined frequency plots, but it is a problem with the B computation, because the subtraction of $k\bar{f}_1$ is now incorrect. The worst case is when $f_k = (k + 0.5)f_a$, which, if the frequencies follow Eq. (1.42a), occurs when $k_{\max} \cong \sqrt[3]{1/B}$. For example, when $B = 0.0001$, $k_{\max} = 21$ yields this situation. After this point, each harmonic bin output is actually the sum of two components, which further obscures the analysis. This lack-of-separation problem can be alleviated by analyzing with $f_a = 0.5\bar{f}_1$, thus halving the bin filter bandwidths, so that each frequency component will appear in only one harmonic output. However, with the phase vocoder it then becomes cumbersome to determine which partials belong to which harmonic bin. Moreover, until the k_{\max} point is reached, this method yields the same results as analysis with $f_a = \bar{f}_1$.

Figure 1.26c shows the phase-vocoder-measured time-averaged amplitudes and frequencies of the first 19 partials of a piano tone. Here, the gradually increasing shift of the partial frequencies with respect to the harmonic positions is very apparent. Calculated B values [using Eq. (1.42b)] for individual partials of several piano tones at different pitches are shown in Table 1.2. It appears that B measurements often start out low (even negative) with notable exceptions for partials 2–4 and then settle into fairly fixed values as partial number increases. At a certain point the B calculation decreases and then turns negative (not shown) due to the partial moving into the next analysis bin. By increasing the analysis frequency slightly, proper B calculation can be extended to higher partials at the expense of accuracy in computing the lower partial frequencies.

2.3.2 Measurement of Tones with Widely Spaced Partial: The Chime

According to Eq. (1.43), the frequencies of a bar are $9f_o, 25f_o, 49f_o, 81f_o, 121f_o, 169f_o, 225f_o, 289f_o, 361f_o, \dots$ for modes $k = 1, 2, 3, 4, 5, 6, 7, 8, 9, \dots$. Dividing by 9 gives $f_1, 2.78f_1, 5.45f_1, 9.00f_1, 13.44f_1, 18.78f_1, 25.00f_1, 32.11f_1,$

TABLE 1.2 Calculated Inharmonicity Values B for Individual Partials of Several Piano Tones at Different Pitches.^a

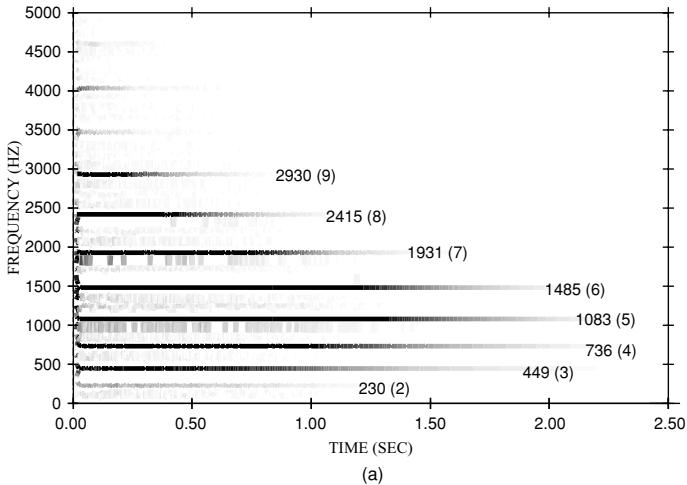
Pitch	f_a	$k:$	2	3	4	5	6	7	8	9	10	11	12	13	14
A ₀	27.1	$B \times 10^3:$	-3.001	-0.017	0.017	0.164	0.230	0.210	0.282	0.292	0.275	0.281	0.273	0.265	0.270
A ₁	54.7	$B \times 10^3:$	-0.241	0.291	0.316	0.023	0.080	0.076	0.076	0.080	0.093	0.086	0.083	0.094	0.093
A ₂	109.6	$B \times 10^3:$	0.225	0.012	0.032	0.050	0.051	0.076	0.075	0.089	0.089	0.090	0.097	0.013	0.098
A ₃	219.1	$B \times 10^3:$	0.513	0.389	0.355	0.369	0.275	0.301	0.285	0.300	0.274	0.271	0.265	0.262	
A ₄	440.1	$B \times 10^3:$	-3.245	-0.237	0.198	0.391	0.537	0.583	0.623	0.579	0.643	0.478			
A ₅	875.5	$B \times 10^3:$	2.393	2.731	2.681	2.516	2.747	2.332							
A ₆	1754.7	$B \times 10^3:$	7.955	8.424	7.404	3.647									
A ₇	3532.6	$B \times 10^3:$	4.230	9.101	0.242										

^a f_a is the fundamental analysis frequency corresponding to the pitch of each tone. k is the series of partial numbers for each tone.

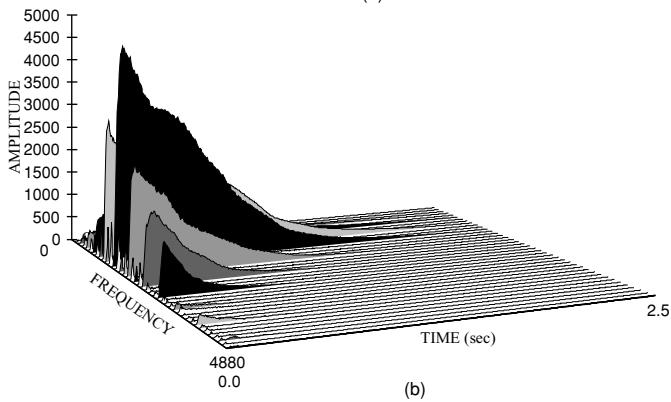
$40.11f_1, \dots$, so these frequencies are clearly more separated than a harmonic series. A chime (or tubular bell) only approximately follows this series, and the author's measurements of an F₄[#] chime tone yielded ratios of 2.82, 5.49, 9.00, 13.23, 18.16, 23.62, 29.70, and 35.82 for modes 2 through 9. (Mode 1 was too weak to measure. The actual frequencies measured were 230, 449, 736, 1083, 1485, 1931, 2415, and 2930 Hz.) Analyzing at $f_a = f_4/6 = 122$ Hz, these frequencies roughly corresponded to harmonics 2, 4, 6, 9, 12, 16, 20, and 24. Note the gradual spreading of the position of these partials. The strike-tone pitch of a chime tone is generally attributed to $f_4/2$ (Rossing, 1976); in this case, it is $736/2 = 368$ Hz, which is close to the standard frequency for F₄[#] (370.0 Hz). The virtual pitch explanation for this pitch calculation is based on the fact that modes 4, 5, and 6, which in our analysis correspond to harmonics 6, 9, 12, can be reduced to the ratios 2:3:4, which form harmonics 2–4 of a harmonic series. The missing fundamental of that series is then $f_4/2$. However, it could also be $f_5/3$ (361 Hz) or $f_6/4$ (371.3 Hz). The average of these three numbers is 366.8 Hz.

Figure 1.27a shows a 2D spectrogram of the F₄[#] chime tone, where intensity is indicated by darkness, and Fig. 1.27b shows the amplitude-vs-frequency-vs-time 3D graph. Fig. 1.27c shows the time-averaged spectrum of the sound. Average frequencies and mode numbers are labeled on Fig. 1.27a. It is evident that the strongest modes are modes 5 and 6.

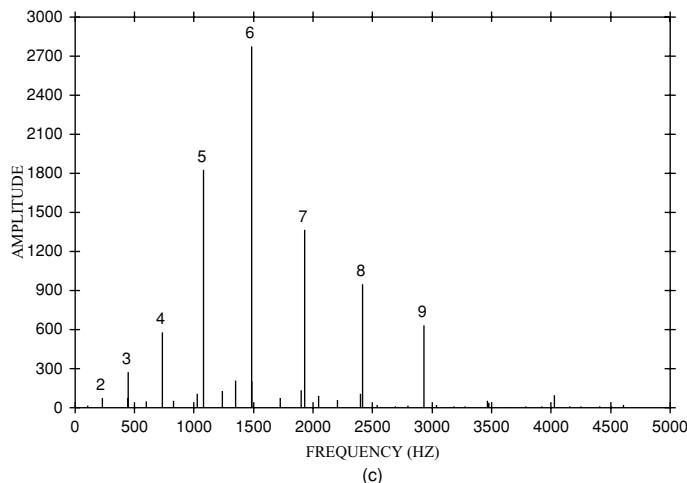
Different partials have different rates of decay, as shown in Fig. 1.28a. This means that the proportion of each harmonic relative to the RMS amplitude changes over time, as demonstrated by Fig. 1.28b. Modes 7, 8, and 9 have the most rapid decays—in the range of -35 to -60 dB/s, whereas the remaining modes decay much more slowly so that at the end of the sound, only modes 3 through 6 are left. Relative to the RMS amplitude, at the beginning of the sound the strongest mode is 6, followed by 5, 7, 8, 9, 4, 3, and 2 in order of strength. In the middle of the sound (at 1.25 s) the strength order is 5, 6, 4, 3, 7, 2, and 8, whereas at the end of the sound the order is 4, 5, 3, and 6. Therefore, the quality of the sound changes dramatically from the beginning to the end of the sound.



(a)



(b)



(c)

FIGURE 1.27. Phase vocoder spectral analysis of an $F_4^{\#}$ chime tone using an analysis frequency of 122 Hz. (a) 2D spectrogram, where darkness indicates amplitude; the frequency and number (in parentheses) of each mode is given. (b) 3D spectrogram showing the amplitude-vs-time behaviors of the modes. (c) Time-averaged spectrum of the chime tone with mode numbers labeled.

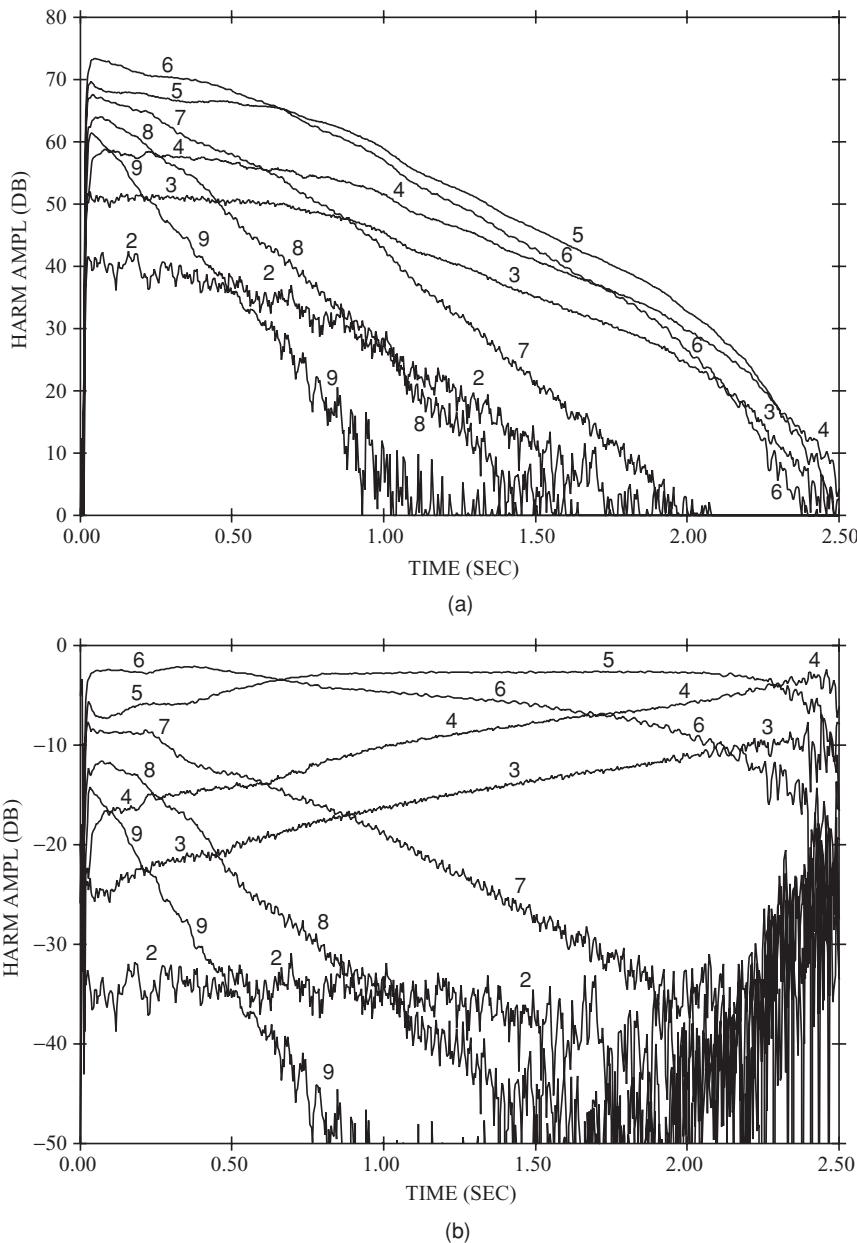


FIGURE 1.28. Amplitude-vs-time envelopes of modes 2–9 of the $F_4^{\#}$ chime tone superimposed. (a) Amplitude-vs-time envelopes superimposed. (b) Amplitude-vs-time envelopes normalized by the instantaneous RMS amplitude, showing the dominance of modes 6, 5, and 7 at the beginning of the tone and modes 4, 5, and 3 at the end of the tone.

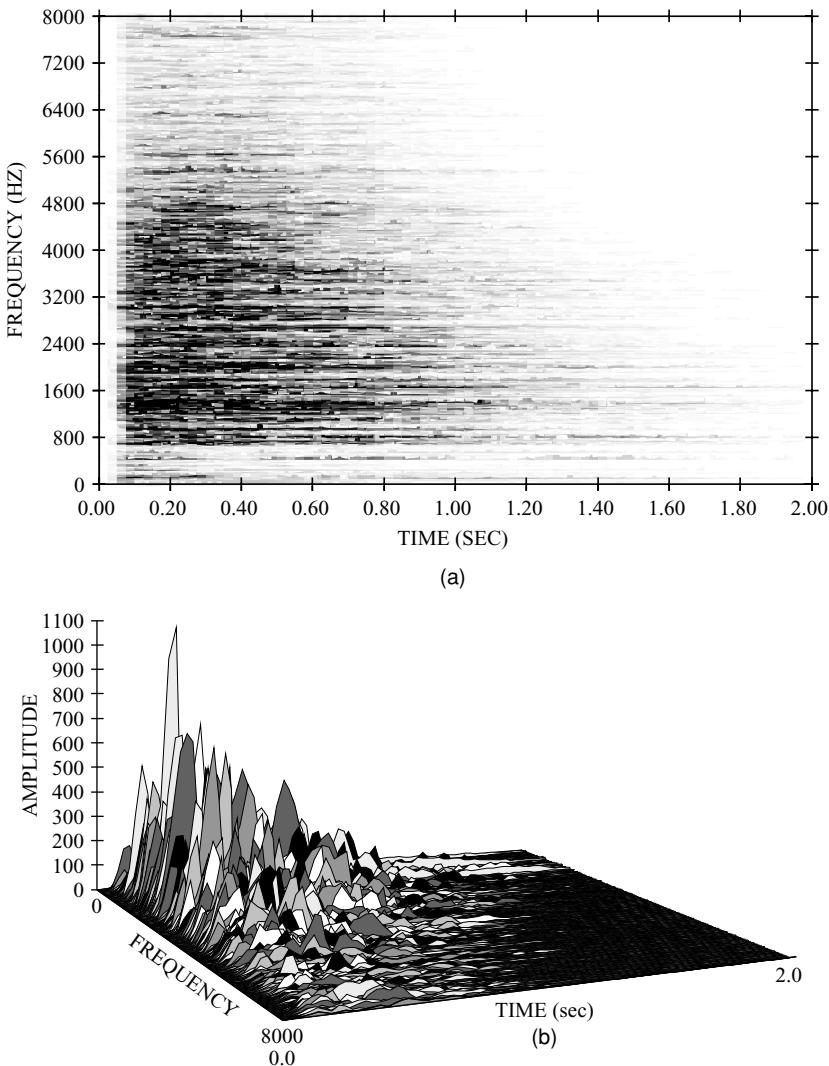


FIGURE 1.29. Phase vocoder spectral analysis of a cymbal sound using an analysis frequency of 20 Hz. (a) 2D spectrogram. (b) 3D spectrogram, showing the complexity of this sound in terms of the extreme amplitude variation of the individual analysis bins.

2.3.3 Measurement of a Sound with Dense Partials: The Cymbal

The spectra of thin plates and stretched membranes can be very dense. Figs. 1.29a and 1.29b show 2D and 3D time-varying spectra of a cymbal sound. In this case, the analysis frequency was taken to be 20 Hz, a compromise that provides adequate resolution of both time and frequency. The time-varying spectrum graphs indicate that the cymbal's modes have a high density. The 3D graph also shows that the amplitudes of the individual modes are extremely variable.

As shown in Fig. 1.30a, the average spectrum of the cymbal sound is very irregular. The greatest concentration of energy is between 800 and 8000 Hz, and it is particularly strong in the region 800 to 4800 Hz. The strongest vibration modes are at 1340 and 2200 Hz, corresponding to harmonics 67 and 110 of the analysis. Figs. 1.30b and 1.30c show that the amplitude-vs-time behavior of these modes is characterized by an amplitude decay with a superimposed seemingly random perturbation. The correlation between these curves is weak. However, the RMS amplitude-vs-time curve of the cymbal sound (shown in Fig. 1.30d), which takes into account all of the analysis bins, is quite smooth and decays according to an approximately exponential curve, so that its dB-vs-time curve is almost linear. The best-fit decay rate of this curve is -22.4 dB/sec.

2.3.4 Spectrotemporal Incoherence

For the cymbal sound, it is quite obvious from Figs. 1.29a and 1.29b and also Figs. 1.30b and 1.30c that the amplitude-vs-time behaviors of the various harmonic bins are not well correlated. Measuring standard correlations between all bin combinations would require $K^2 I$ multiplies and adds, where K is the number of bins and I is the number of frames, and it is not obvious how the results of the different bins should be combined. Spectrotemporal incoherence (SI) can be calculated simply by measuring how well a time-varying spectrum compares with a totally coherent version of itself. A totally coherent spectrum can be obtained by setting each bin amplitude to be proportional to the sound's original average amplitude and to vary in time according to its RMS amplitude:

$$\hat{A}_k(i) = \frac{\bar{A}_k A_{\text{rms}}(i)}{\sqrt{\sum_{k=1}^K \bar{A}_k^2}}, \quad (1.44a)$$

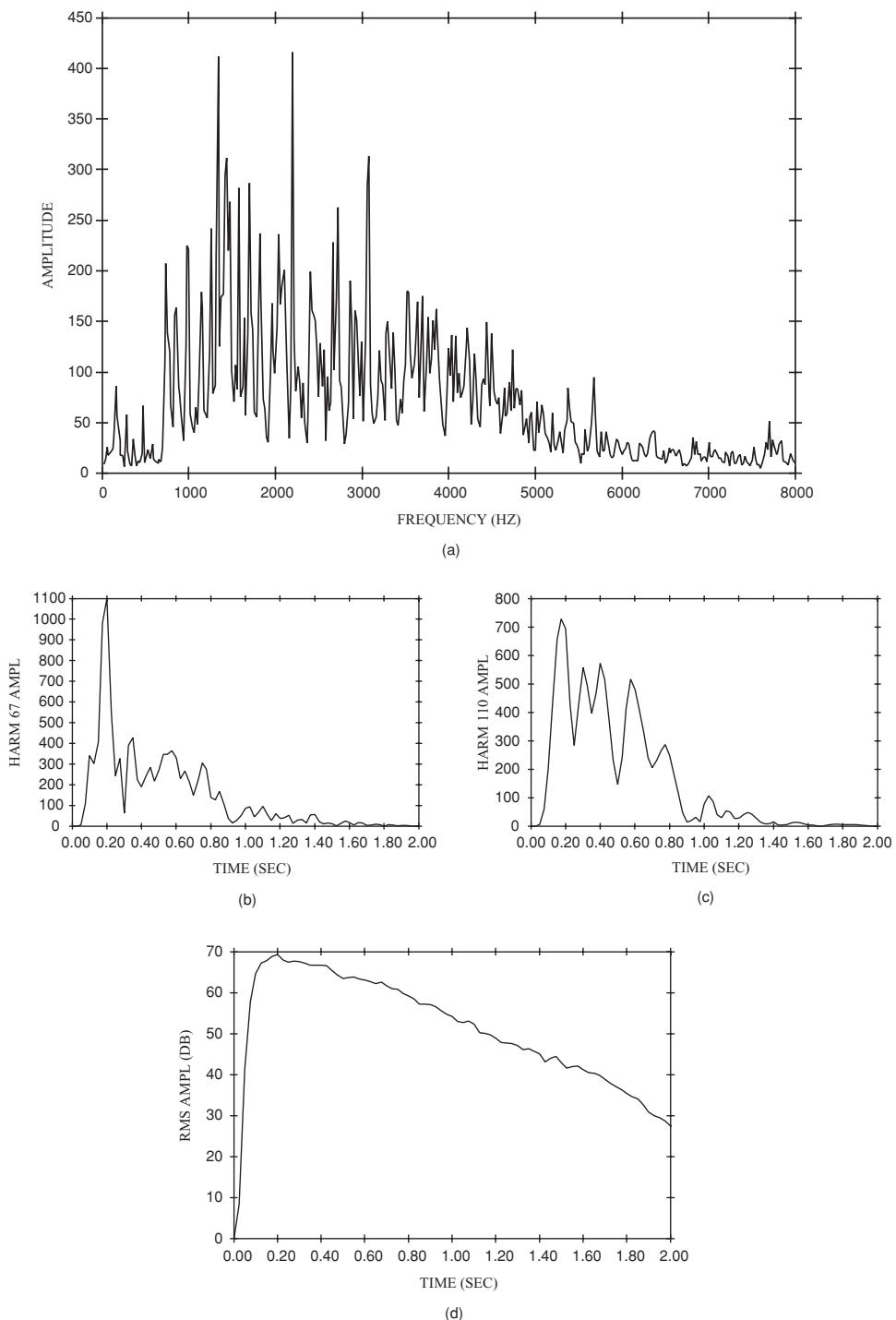
where i is the frame number, and

$$\bar{A}_k = \frac{\sum_{i=0}^{I-1} (A_k(i))^2}{\sum_{i=0}^{I-1} A_k(i)} \quad (1.44b)$$

is the k th bin amplitude amplitude-averaged over all frames. Note that according to this formulation all of the coherent bin amplitude-vs-time functions are proportional to the RMS amplitude-vs-time function.

The spectrotemporal incoherence SI is then defined as

$$SI = \left(\frac{\sum_{i=0}^{I-1} \sum_{k=1}^K (A_k(i) - \hat{A}_k(i))^2}{\sum_{i=0}^{I-1} \sum_{k=1}^K (A_k(i))^2} \right)^{\frac{1}{2}}. \quad (1.44c)$$



By design, SI is zero if all harmonic amplitudes are proportional to the RMS amplitude and to each other. For the cymbal sound illustrated in Figs. 1.29 and 1.30, $SI = 0.46$. By contrast, SI values for the trumpet, flute, piano, and chime sounds discussed above were measured to be 0.19, 0.26, 0.30, and 0.17, respectively. So, according to the definition of Eq. (1.44c), the cymbal sound is a comparatively incoherent—i.e., complex—sound.

If the cymbal sound is resynthesized after all harmonic amplitude-vs-time functions are replaced by their RMS equivalents [as defined by Eq. (1.44a)], the result sounds similar to exponentially decaying white noise, although some of the cymbal quality remains. Note that with this type of sound, there is no change in spectral centroid or, in fact, no change in the normalized average spectrum. The noise-like quality is largely due to the bin frequencies, which vary quite randomly over the individual bin ranges. If these frequencies are set to their corresponding fixed bin frequencies (i.e., integer multiples of 20 Hz), the resynthesized sound has a much more “coherent” character.

2.3.5 Inverse Spectral Density: Cymbal, Chime, and Timpani Compared

Spectral density can be measured by counting the number of spectral peaks above a certain threshold and dividing by the frequency range of these peaks. A more convenient measure is inverse spectral density (ISD), which, as it implies, is just the inverse of the spectral density, or, in other words, the average frequency difference between adjacent spectral peaks. The problem is to adequately define a “spectral peak.” The simple solution is to define peaks as local maxima of the amplitude spectrum that exceed a given threshold. The assumption is that each peak corresponds to a sinusoid, corresponding to a modal frequency, in the sound. As long as the threshold is well below the maximum magnitude of the spectrum, the ISD is quite stable as the threshold is varied, as illustrated by a graph of ISD -vs-threshold (in decibels) for the cymbal sound shown in Fig. 1.31a. It is also fairly constant as a function of time for a fixed threshold, as shown in Fig. 1.31b.

For the cymbal sound, the ISD , appears to be in the range of 60–100 Hz. It would be larger if minor peaks were ignored. Minor peaks are those that are too close to major peaks, particularly if they are in valleys in between peaks. Serra (1989) discusses a way to exclude some peaks based on “peak height,” but this definition is difficult to implement. Theoretically, peaks below a threshold can be excluded, but practically, this method fails unless the spectral envelope corresponding to the desired peaks is relatively flat. Indeed, a spectrum-flattening algorithm might improve the ISD measurement.

FIGURE 1.30. Further analysis of the cymbal sound. (a) Time-averaged spectrum showing dominant modes at bins 67 and 110 (frequencies 1340 and 2200 Hz). (b, c) Amplitude-vs time envelopes of bins 67 and 100. (d) RMS amplitude (in dB) vs time, indicating an exponential decay rate of -22.4 dB/s.

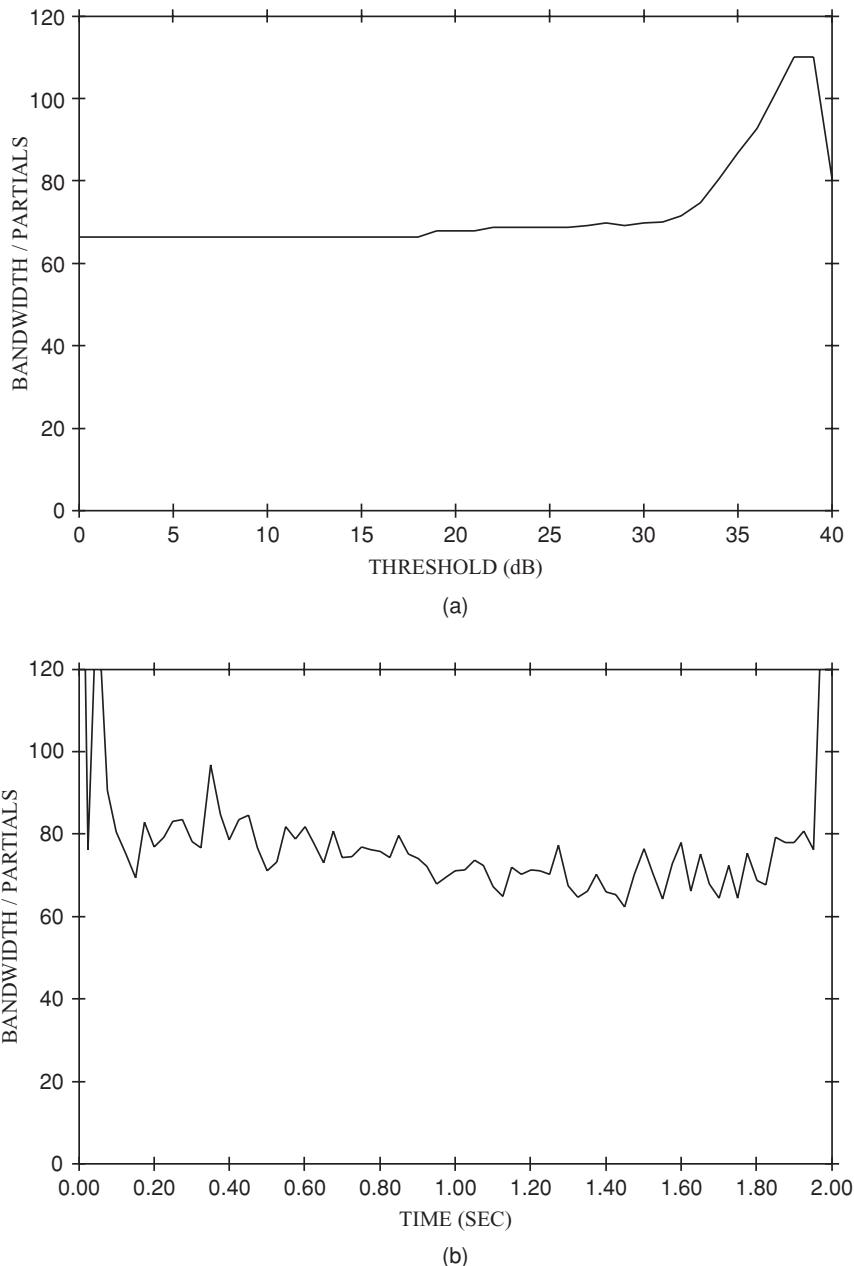


FIGURE 1.31. Inverse spectral density (*ISD*) of the cymbal sound in terms of bandwidth divided by number of spectral peaks (partials) in the band for an RMS-amplitude-normalized spectrum. (a) *ISD*-vs-dB threshold for $t = 0.1$ s. (b) *ISD*-vs-time for a threshold of 34 dB, which is 30 dB below the peak normalized spectrum level of 64 dB, indicating an approximately 80 Hz average value. ($t = 0.1$ s on the *ISD*-vs-time graph corresponds to 34 dB on the *ISD*-vs-dB graph.)

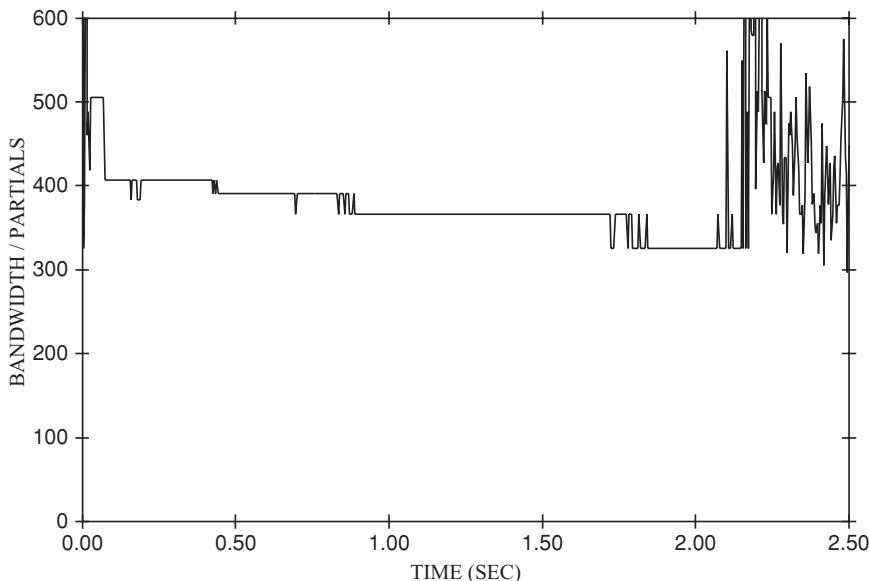


FIGURE 1.32. Chime tone inverse spectral density (*ISD*) vs time, indicating an approximate 400 Hz average value. The threshold is 44 dB, 30 dB below the maximum 74 dB maximum spectrum level.

By contrast, the *ISD* for the chime tone is much higher, starting out about 400 Hz and decreasing to a somewhat lower value during most of its duration, as shown in Fig. 1.32. This value results from the average difference between its distinctive mode frequencies.

An E₃ (165 Hz) timpani sound was analyzed by the phase vocoder, again with $f_a = 20$ Hz, and the resulting 2D and 3D spectrum-vs-time graphs are shown in Figs. 1.33a and 1.33b. The spectrum is again initially very dense, but not as broadband as that of the cymbal, and certain distinct modes emerge as the sound progresses. A 3D spectrum graph with amplitudes normalized by the RMS amplitude, as shown in Fig. 33c, makes this even more obvious. In fact, this emergence of certain prominent frequencies that are more-or-less equally spaced gives rise to the distinct pitch that is characteristic of the timpani sound.

As indicated by Fig. 1.34a, the decay rate of this timpani sound is faster than that of the cymbal (see Fig. 1.30d), and a best straight-line fit to the RMS amplitude-vs-time curve gives -25.3 dB/s. Also, the *ISD* varies mainly between 80 and 120 Hz, as shown in Fig. 1.34b. Thus, its value is somewhat larger than the cymbal's. Inspection of snapshot spectra of the cymbal and timpani (see Figs. 1.35a and 1.35b) indicates that their spectral peaks are similarly spaced. So why do the two instruments sound so different? For one thing, for the most prominent modes the timpani's peak bands are narrower than those of the cymbal, implying that they actually represent distinct sinusoidal modes. Also, the 2D and 3D spectra of

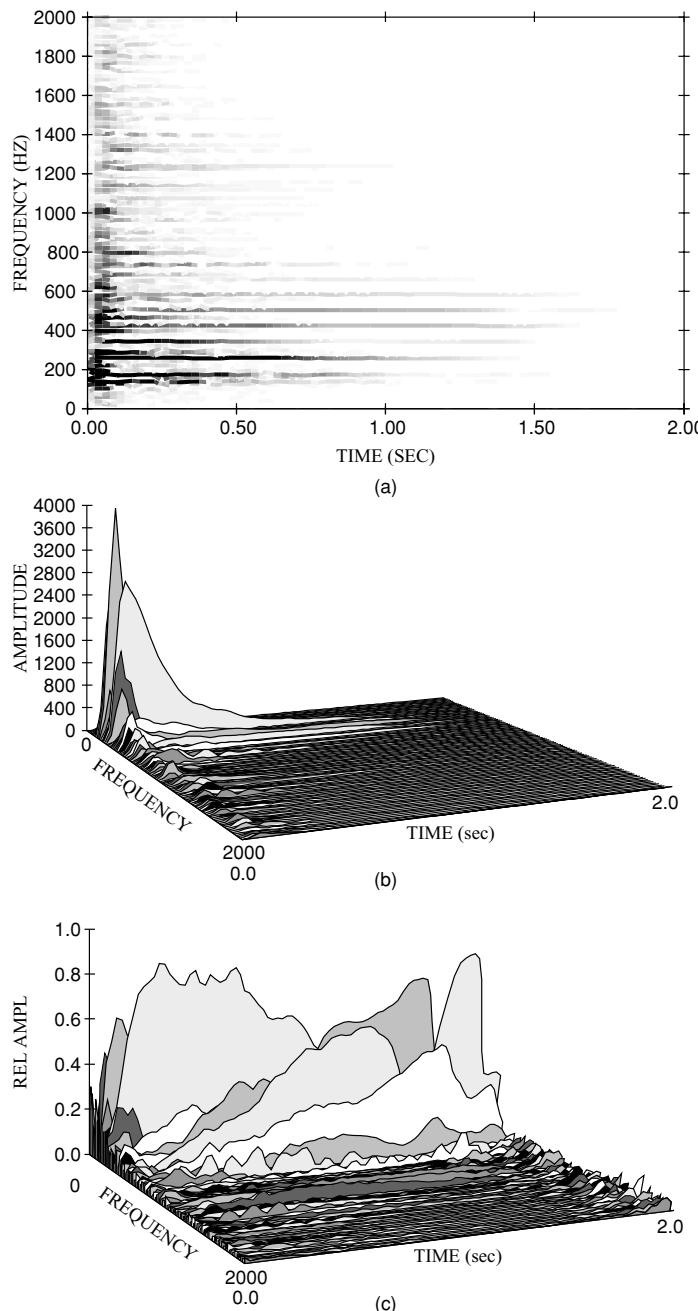


FIGURE 1.33. Phase vocoder spectral analysis of a timpani sound using an analysis frequency of 20 Hz. (a) 2D spectrogram. (b) 3D spectrogram showing individual mode amplitude-vs-time envelopes. (c) 3D spectrogram with mode amplitudes normalized by the instantaneous RMS amplitude.

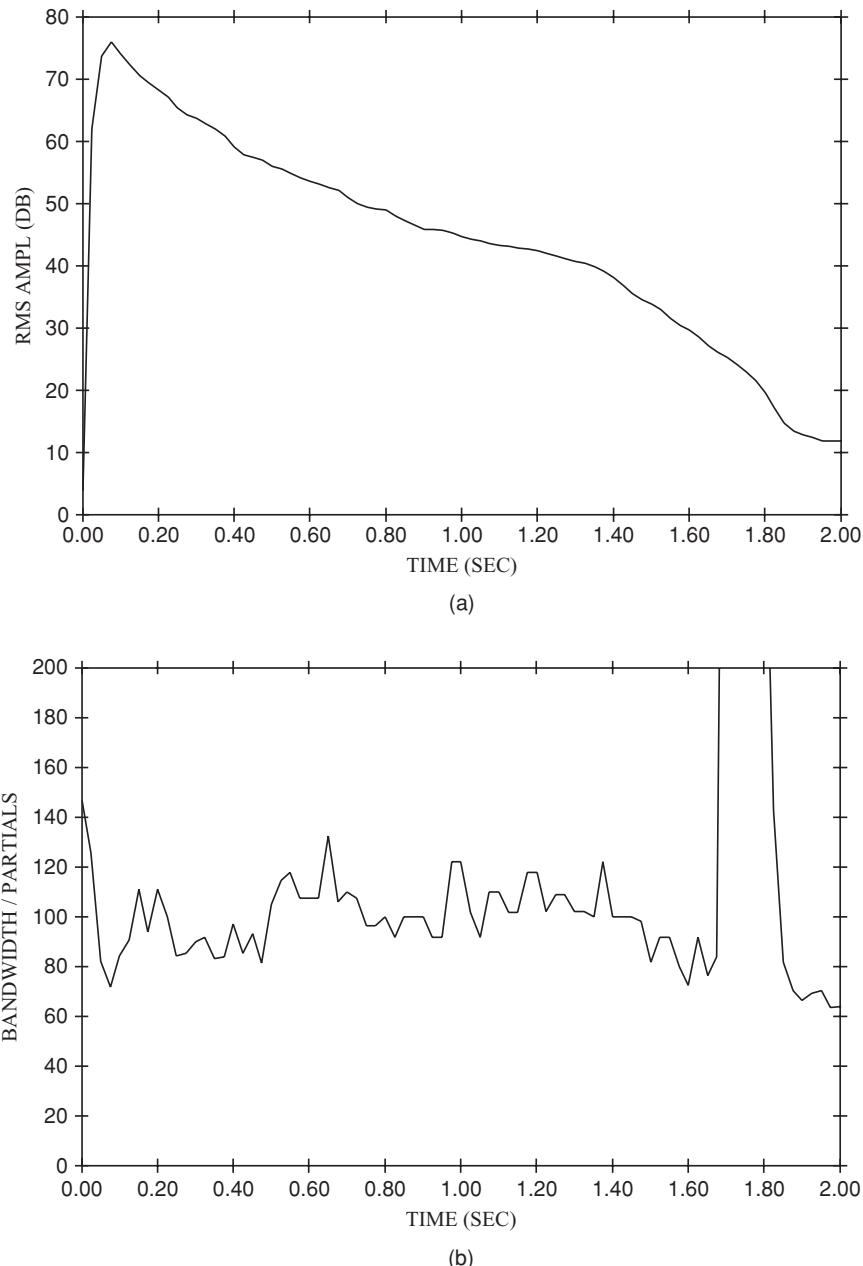


FIGURE 1.34. Further analysis of the timpani sound. (a) RMS amplitude (in dB) vs time, indicating a best fit exponential decay rate of -25.3 dB/s . (b) ISD-vs-time for a threshold of 45 dB, which is 30 dB below the peak normalized spectrum level of 75 dB, indicating an approximately 100 Hz average value.

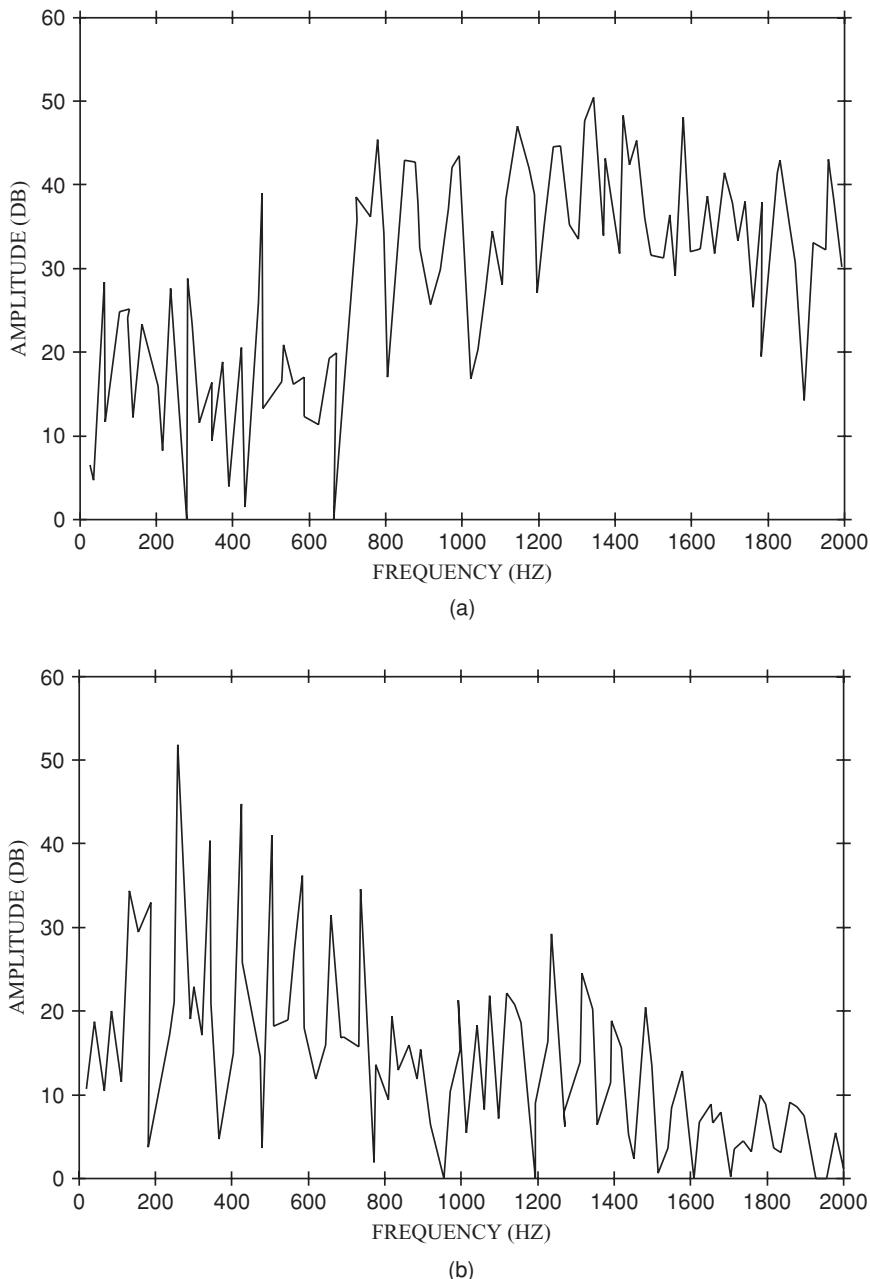


FIGURE 1.35. “Snapshot” spectra of the sounds of the (a) cymbal and (b) timpani in terms amplitude in dB vs frequency for $20 \leq f \leq 2000$ Hz with 20 Hz resolution. Both are taken at $t = 0.6$ s.

the cymbal and the timpani (Figs. 1.29 and 1.33), show that the timpani's peak amplitudes decay very smoothly whereas the cymbal's amplitudes are highly variable. Moreover, as mentioned above, the frequencies of the strong cymbal modes are also highly variable, while those of the timpani are quite stable. Figs. 1.36a and 1.36b show a comparison between frequency-vs-time curves for strong modes of the cymbal and the timpani. The timpani mode obviously displays much less frequency variation than that of the cymbal.

Based on frequencies up to 2000 Hz and a threshold amplitude of 100, Fig. 1.37 compares the spectral centroids of the cymbal and timpani tones. The timpani's centroid is substantially lower than the cymbal's, and both decline somewhat over time.

2.4 Frequency-Tracking Analysis of Harmonic Sounds

Virtually any kind of sound can be analyzed using the frequency-tracking (mq) method. However, success depends on whether frequencies are tracked correctly or not. The frequency-tracking method works best if partials are not too numerous or too close together, especially if the partials' amplitudes fluctuate. When using SNDAN's mqan program for frequency-tracking analysis, the user must specify the lowest frequency to be resolved and a threshold value (in dB) below which spectral peaks are ignored. The programs fdetect and harmformat can be used to detect the (time-varying) fundamental frequency of the data and separate it into harmonics so that it can be displayed using SNDAN's monan program. The user must specify the range of fundamental frequencies expected and a harmonic acceptance interval for separation into harmonics.

2.4.1 Frequency-Tracking Analysis of Steady Harmonic Sounds

To demonstrate how this method works for a typical harmonic sound with little frequency change, Figs. 1.38 and 1.39 show individual harmonic amplitude and normalized frequency deviation graphs for the first six harmonics of the F₄ trumpet tone after the frequency-tracking analysis is reduced to harmonics. Comparable graphs given in Figs. 1.14 and 1.15 obtained by phase-vocoder analysis look very similar; the only difference is the amount of fine detail noise that occurs in the phase-vocoder and is missing in the mq graphs. This can be attributed to two features of the frequency-tracking analysis: thresholding, which effectively gates out some of the noise, and the relatively wide window, which narrows the FFT filters.

2.4.2 Frequency-Tracking Analysis of Vibrato Sounds: The Singing Voice

While small amounts of vibrato can be handled well by the phase vocoder, large amounts cause problems. Fig. 1.40 compares the 3D spectrum of a G₃ tenor voice sound (singing the vowel “ah”) obtained by harmonic phase-vocoder analysis with analysis frequency 192 Hz with result of the harmonic-reduced frequency-tracking

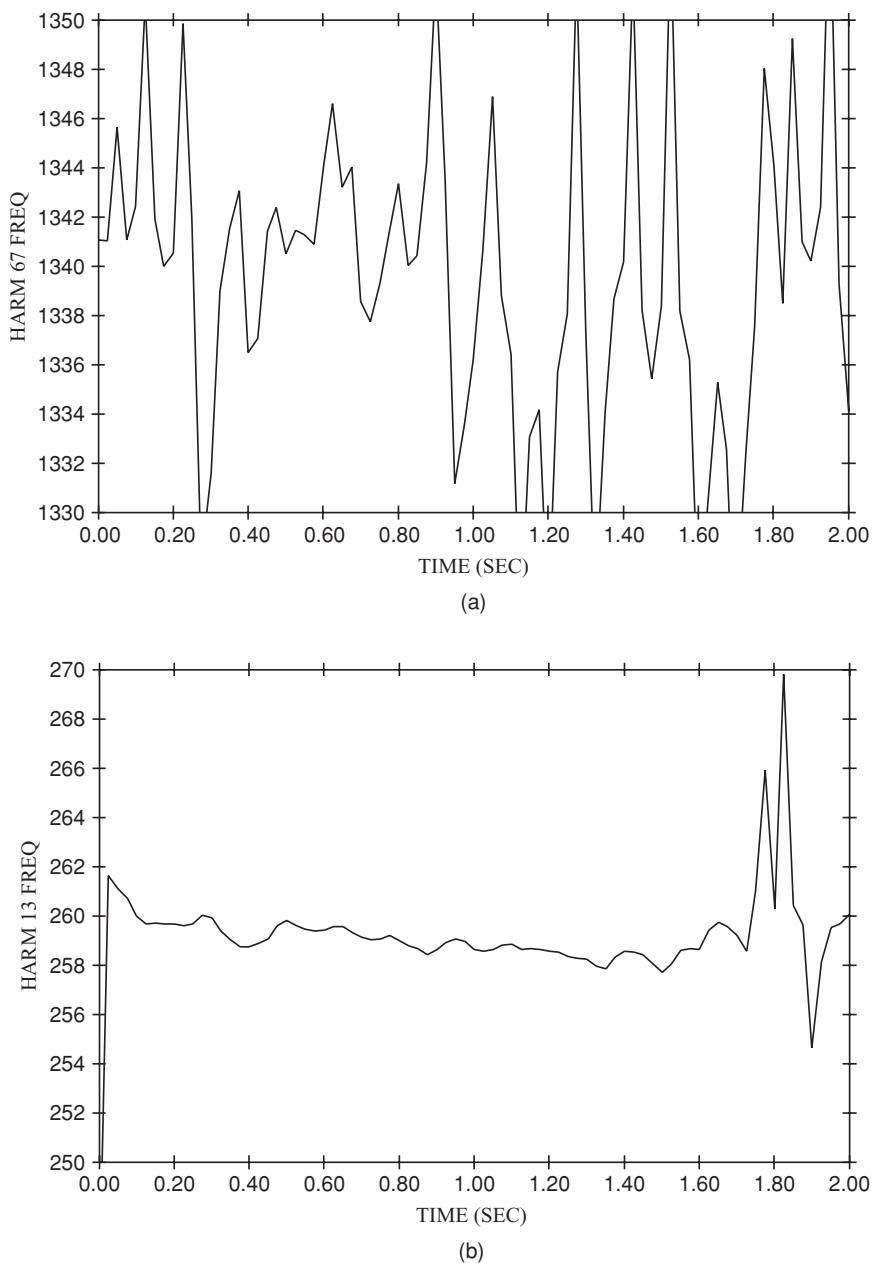


FIGURE 1.36. Frequency-vs-time of a prominent mode for (a) the cymbal (bin 67, 1340 Hz) and (b) the timpani (bin 13, 260 Hz).

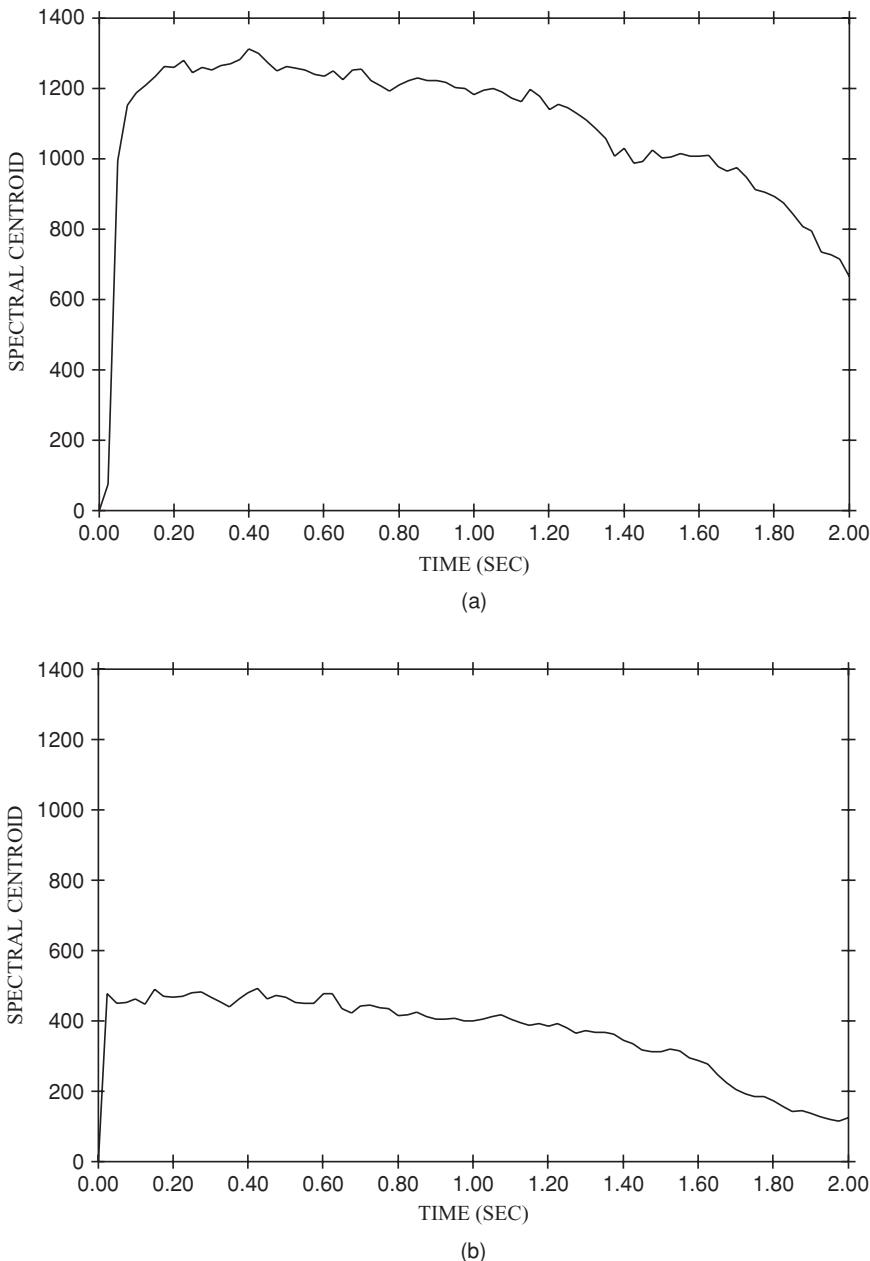


FIGURE 1.37. Spectral centroid vs time for (a) cymbal, (b) timpani, based on frequencies below 2000 Hz with amp = 100 threshold.

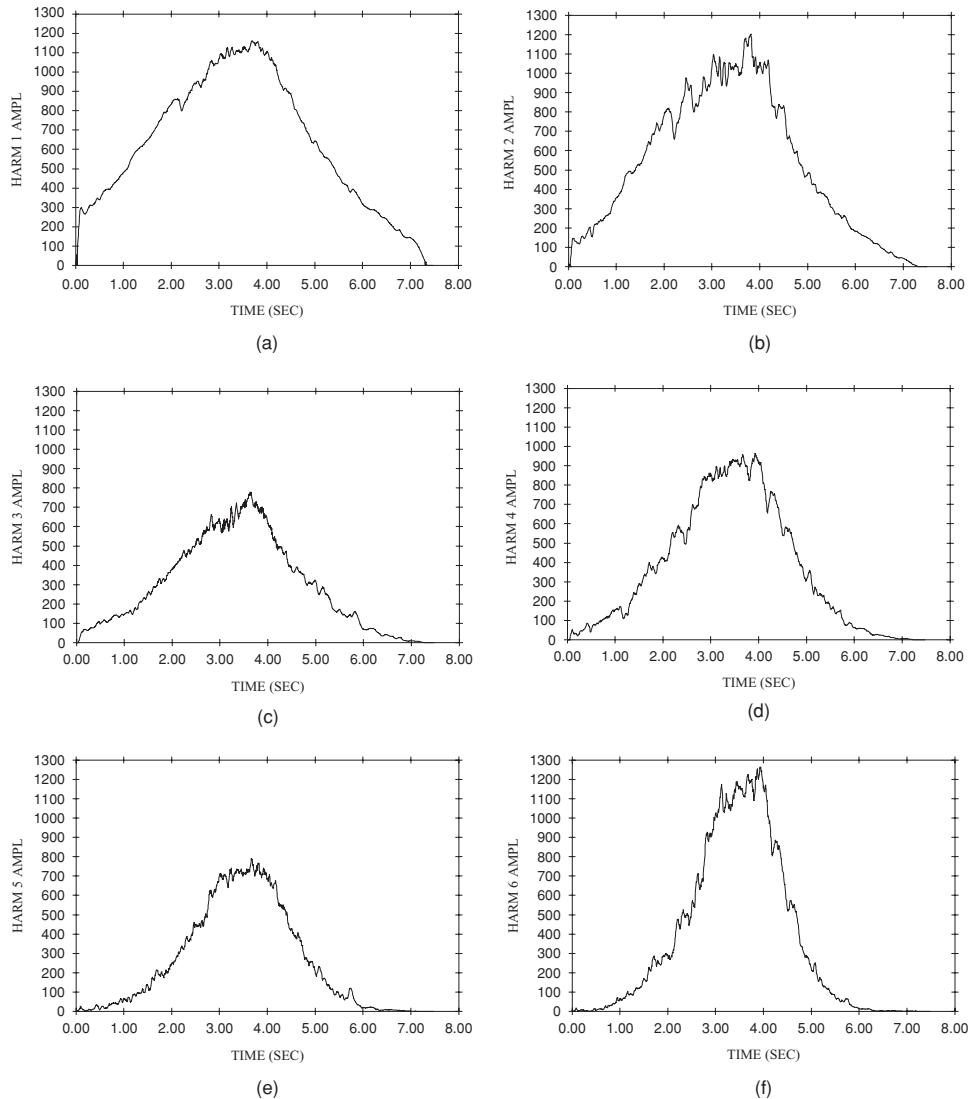


FIGURE 1.38. Frequency-tracking analysis of the F_4 , $pp <ff> pp$ trumpet tone reduced to harmonics of 350 Hz: (a)–(f) amplitude-vs-time curves for harmonics 1–6. Comparison to Fig. 1.14 shows very close correspondence of the frequency-tracking and phase-vocoder methods for this case.

method. It is obvious that frequency-tracking achieves by far the clearer analysis in this case. For this sound, harmonics 13 and 14 are very strong and coincide with the “singer’s formant” (Sundberg, 1974), with harmonic 13 operating on the low side and harmonic 14 operating on the high side of the formant resonance

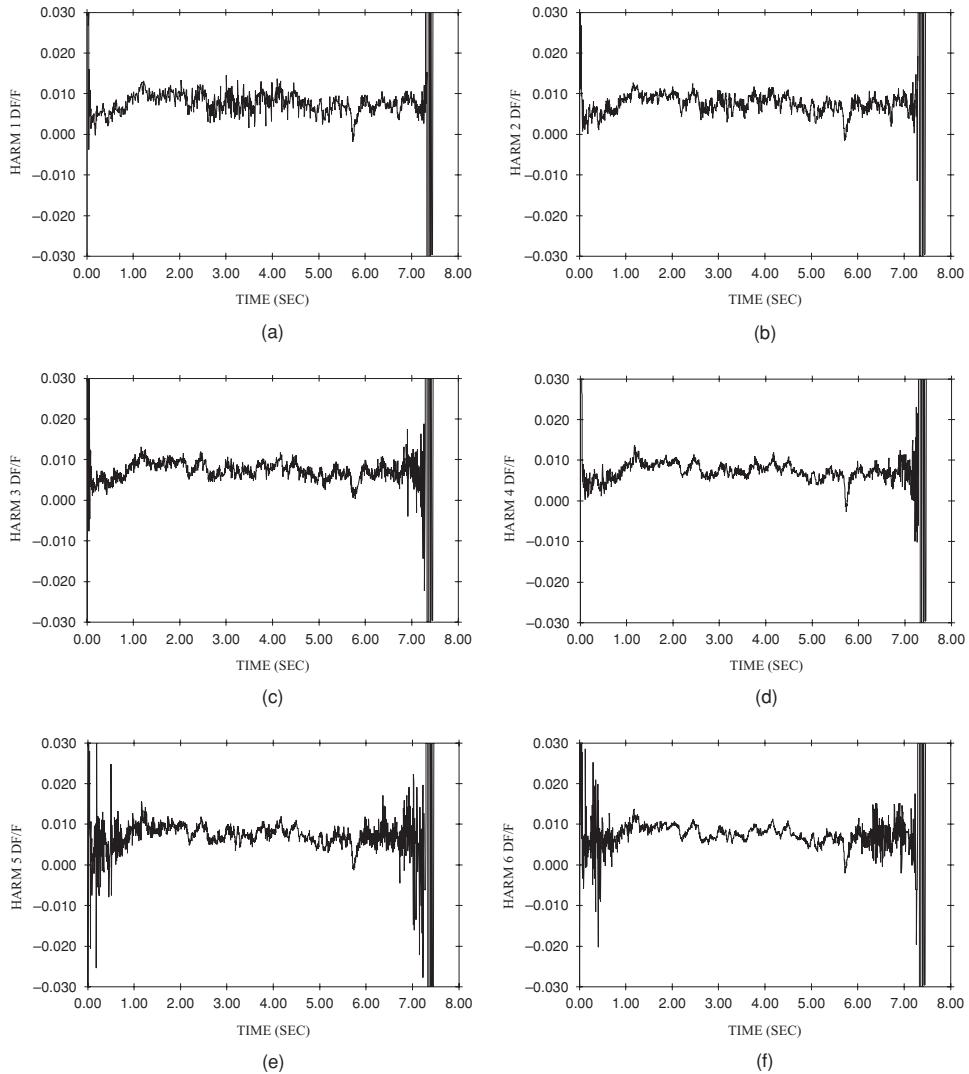


FIGURE 1.39. Same analysis as Fig. 1.38: (a)–(f) Normalized frequency deviations for harmonics 1–6. Comparison to Fig. 1.15 shows very close correspondence of the frequency-tracking and phase-vocoder methods, except that the data obtained from frequency-tracking analysis exhibit less noise.

curve, respectively. Because the frequency of the fundamental, as shown in the pitch detection result of Fig. 1.41a, swings from roughly 183 to 207 Hz, it is expected that the 14th harmonic will swing over 14 times those numbers, i.e., from 2562 to 2898, and this result is confirmed by the frequency-tracking result of Fig. 1.41b. But these numbers exceed the bandwidth of the phase-vocoder bin

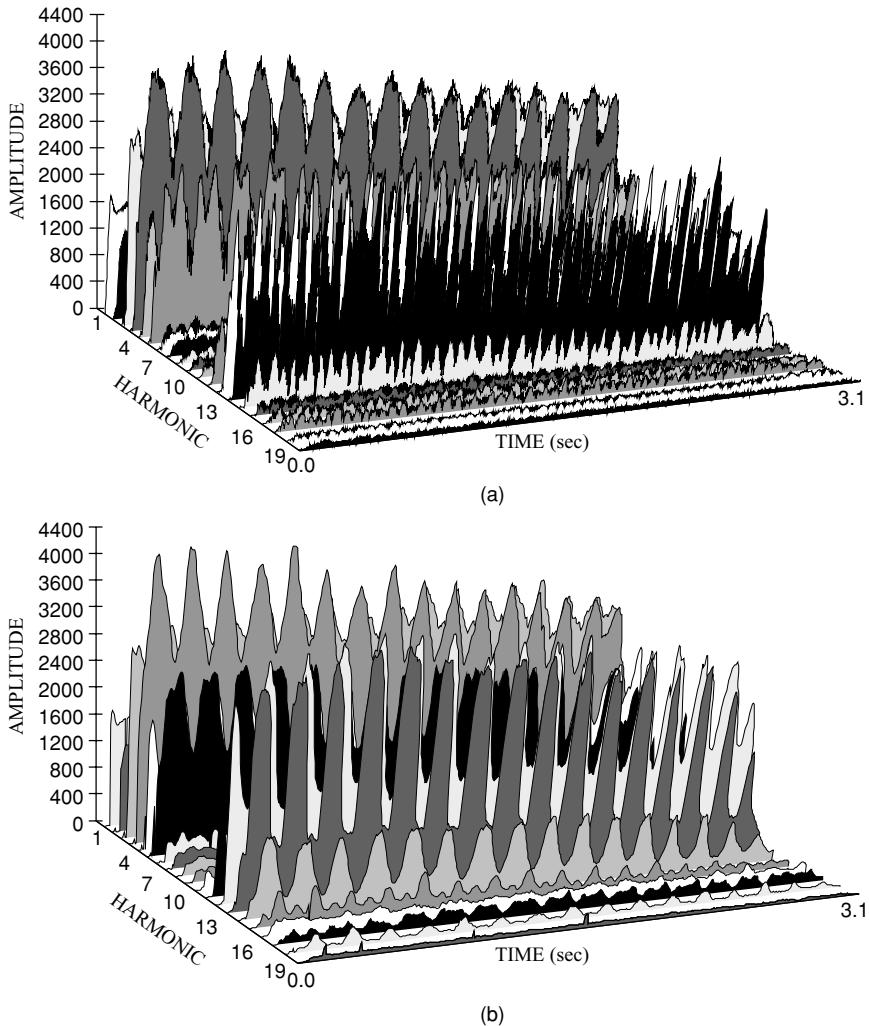


FIGURE 1.40. Analysis of the G_3 (192 Hz) tenor voice tone: (a) 3D spectrum obtained from phase-vocoder analysis at $f_a = 192$ Hz. (b) 3D spectrum obtained from frequency-tracking analysis reduced to harmonics.

centered on the 14th harmonic, which essentially extends from 2592 to 2754 Hz. So while the problem with phase-vocoder analysis on this type of sound is small for the fundamental, shown in Fig. 1.41c, it becomes enormous for upper partials such as the 14th harmonic, as seen in Fig. 1.41d, which is seriously corrupted by neighboring harmonic bins.

The frequency-tracking amplitude- and frequency-vs-time data of Figs. 1.40 and 1.41 can be combined with superimposed graphs of $A_k(t)$ vs $f_k(t)$ as shown in

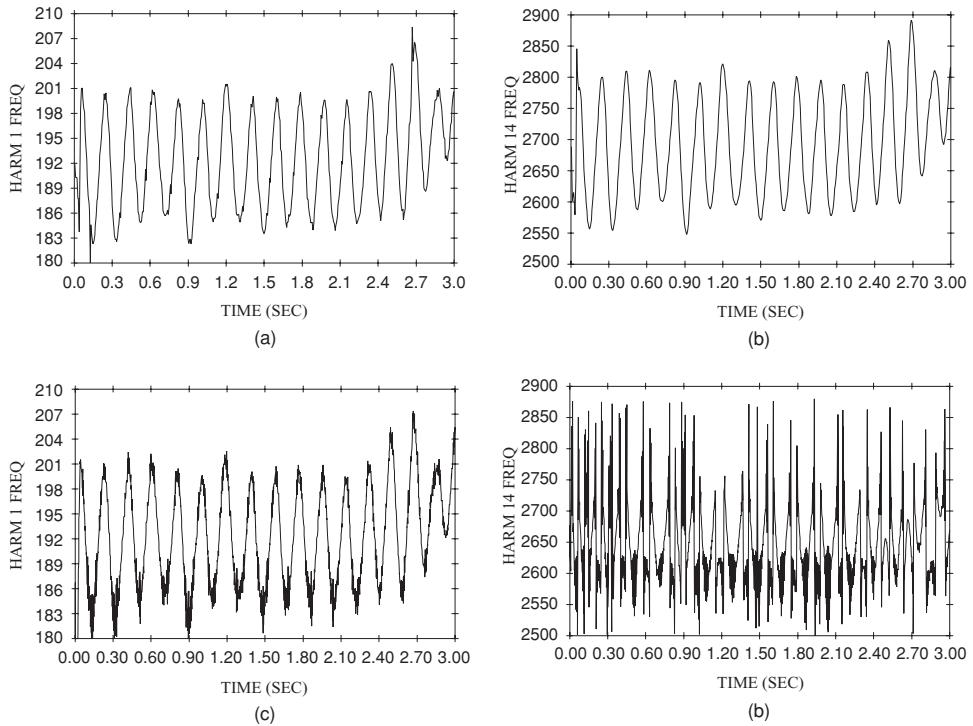


FIGURE 1.41. Analysis of the G_3 (192 Hz) tenor voice tone: (a, b) f_1 , f_{14} vs time for frequency-tracking analysis. (c, d) f_1 , f_{14} vs time for phase-vocoder analysis at $f_a = 192$ Hz.

Fig. 1.42. This composite graph illustrates the formant nature of the vocal sound, where the first and second formants and the singer's formant stand out at about 700, 1100, and 2800 Hz, respectively. The overall curve may also be described as a “low-pass filter” with a cutoff at about 700 Hz followed by a rolloff of about -12 dB/octave. Maher and Beauchamp (1990) give other similar voice analysis examples.

2.4.3 Frequency-Tracking Analysis of Variable-Pitch Sounds

The frequency-tracking method is especially useful when pitch varies by a semi-tone (approximately 6% change) or more. The two-way-mismatch pitch-detection algorithm is described in Section 1.2.3, and results for the tenor voice and a clarinet passage are given in Figs. 1.10 and 1.11. Fig. 1.43a shows the 2D display of the frequency-tracking analysis of a solo saxophone passage. Note that harmonic tracks corresponding to a changing fundamental dominate the display. The pitch-vs-time graph for this sample is shown in Fig. 1.43b. These data can be used in conjunction with the frequency-tracking analysis data to separate the harmonics from other spectral information and write a file in the same format as that produced

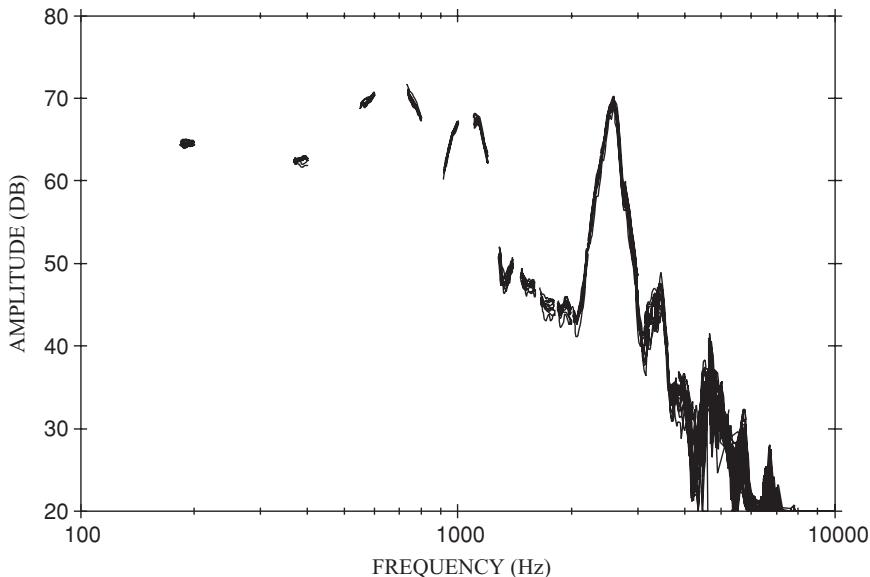


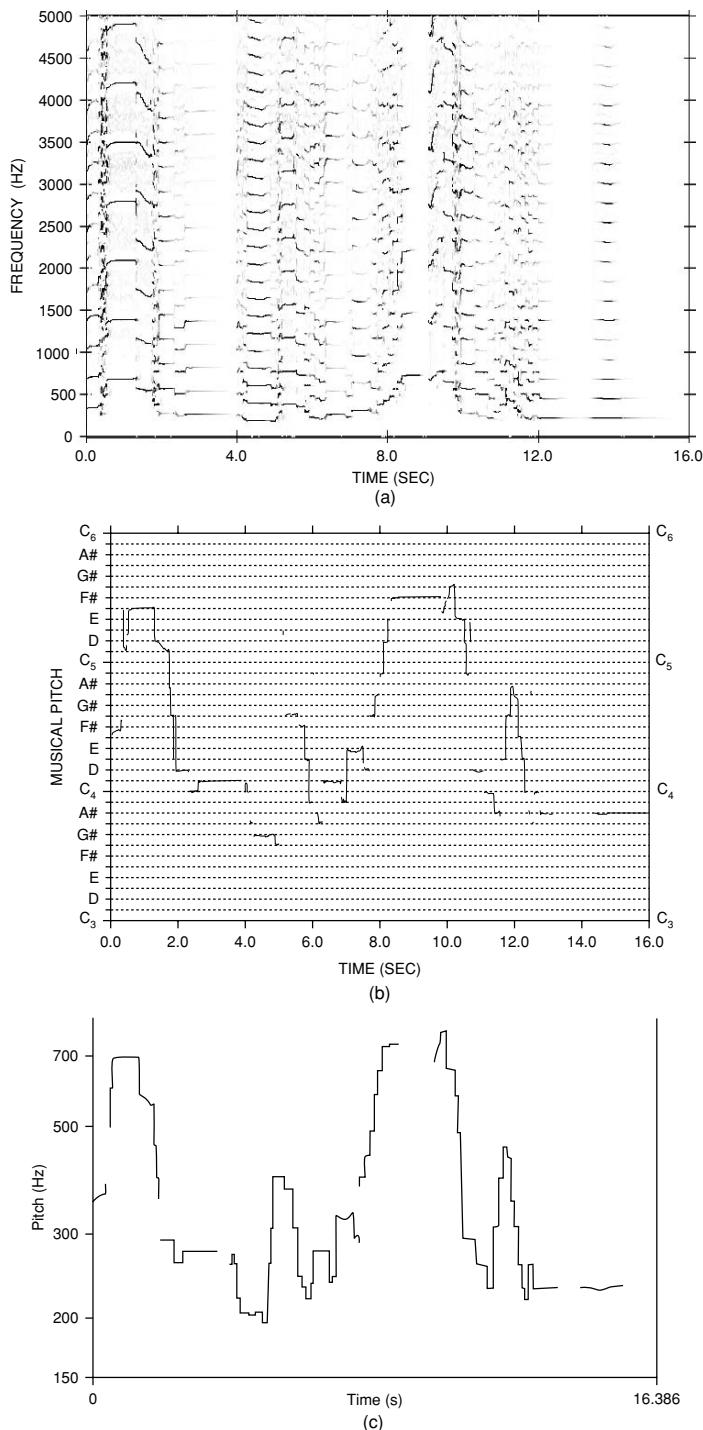
FIGURE 1.42. Further analysis of the G_3 (192 Hz) tenor voice tone: Superimposed $A_k(t)$ -vs- $f_k(t)$ patterns (obtained from frequency-tracking analysis) which trace out a “formant characteristic” of the voice spectrum.

by the phase-vocoder program. The separated harmonics can then be resynthesized to produce a sound that consists solely of harmonic partials and excludes any incidental inharmonic frequencies or noises, including the effects of reverberation. Besides producing a “cleaned-up” version of the sound track, resynthesis is a good test of the pitch detection algorithm, because inaccuracies in the fundamental-frequency data in the resynthesized sound are immediately obvious to the ear. For comparison to Fig. 1.43b, Fig. 1.43c shows the result of pitch detection using the Praat program written by Paul Boersma and David Weenink (Boersma, 1993), based on the time-domain autocorrelation method.

3 Summary

The methods for analysis and synthesis of monophonic sound signals described in this chapter are based on the sum-of-sinusoids model and rely heavily on the

FIGURE 1.43. Frequency-tracking analysis of an alto saxophone solo: (a) 2D time-varying spectrum; (b) musical-pitch vs. time obtained by two-way-mismatch analysis of time-varying spectrum data; (c) log-frequency vs time obtained by the time-domain autocorrelation method.



short-time Fourier transform as implemented in digital form on a computer. The two principal methods are the harmonic filter bank (phase-vocoder) and the frequency-tracking (mq) method.

The phase vocoder relies on the intrinsic harmonic filtering characteristic of the short-time Fourier transform. While the Fourier transform algorithm basically converts a signal waveform into a collection of complex terms, one for each harmonic, these are easily converted into amplitude and phase form. Then, using an approximation to the phase derivative, frequency deviation can be computed from the phases of adjacent time frames. The window size for Fourier analysis is chosen to be an integer multiple of the expected period of the input signal. For the harmonics to be cleanly separated, a multiple of 2 is appropriate for the hanning or Hamming window functions while 4 is appropriate for the 4-term Blackman–Harris window. The method is akin to classic Fourier series, except that, unlike the perfectly periodic case, it is expected that the harmonic amplitudes (and to a much lesser extent, the frequencies) vary with time. The method can be visualized as a bank of band-pass filters each of which is centered on a harmonic of the fixed analysis frequency. The frequencies of the input signal's harmonics can only vary a small amount from the band centers; otherwise the separation of the harmonics will be compromised.

The phase-vocoder analysis data, consisting of fixed analysis frequency, initial harmonic phases, followed by amplitude and frequency-deviation values for each harmonic on each frame, can be used to resynthesize the input signal. Generally, the output sound is difficult to discriminate from that of the input signal.

The phase vocoder is not restricted to the analysis of harmonic signals. This method can be used beneficially with any sound consisting of nearly constant frequency partials. The three cases are (1) sounds with nearly harmonic partials, (2) sounds with widely spaced partials, and (3) sounds with closely spaced partials. For sounds with nearly harmonic but progressively stretched partials, normal harmonic analysis will usually suffice, although it is true that the inharmonic upper partials will begin to line up with bins greater than those harmonically related to the fundamental. For sounds with widely spaced partials, choosing a sufficiently low fundamental or one which approximately divides the significant partial frequencies evenly yields an accurate analysis. For sounds with closely spaced partials, choosing a low fundamental around 20 Hz will result in a series of time-varying harmonic bands that capture the sound's essential spectral features.

The frequency-tracking method also relies on the STFT, but there is no requirement that the input signal consists of harmonically related sinusoids or that the signal has fixed frequencies. The window size, which is fixed, is set to be approximately three times the largest expected period, corresponding to the lowest spectral frequency, of the input signal. Provided that frequencies are not too close together, this method provides good separation between components. For each frame, amplitudes, frequencies, and phases of the components are estimated from the magnitude spectrum by quadratic interpolation of three points in the vicinity of each spectral peak. However, only peaks above a designated threshold are accepted. Component phases are estimated by interpolation of the phase-vs-frequency data.

Components are tracked from frame to frame using a heuristic algorithm based on connecting peaks whose frequencies are similar while attempting to maximize the lengths of the resulting tracks.

For pitch extraction of monophonic sounds consisting solely of harmonic partials, spectral peak data can be used to estimate a fundamental-frequency-vs-time function using the two-way-mismatch algorithm whereby the component peak frequencies are compared to the harmonics of a trial fundamental varied over a designated frequency range. For each frame, the fundamental frequency is chosen which minimizes an error function based on component-frequency/harmonic-frequency differences as well as the component amplitudes.

The fundamental-frequency-vs-time data in conjunction with the frequency-tracking analysis data can be used to produce a harmonic data set that is identical in form to that produced by the phase vocoder. However, the frequencies, while harmonically related, are not confined to small neighborhoods around fixed harmonic values. The same program used for phase-vocoder additive synthesis can be used to resynthesize sounds from these data. The resulting signal may sound “cleaned up” compared to the original.

For frequency-tracking resynthesis, component sinusoids are reconstructed using the amplitude, frequency, and phase data for each spectral track on each pair of frames. Between frames the sinusoid phases are computed using a cubic method that matches the frequencies (phase slopes) and phases at the end points. Although resynthesis quality is generally good to excellent, quality can be compromised by poor tracking. Also, some additive noise is usually lost due to associated spectral peaks below the designated threshold being ignored. With careful resynthesis, it is possible to subtract the resynthesized signal from the original to produce a noise residual. The residual, which can be separately modeled, may be added to the sinusoidal resynthesis to produce a more natural-sounding output signal.

It has been found that the phase-vocoder method is most robust for sounds with fixed frequencies, even if they are inharmonically related. On the other hand, for sounds with widely varying frequencies, the frequency-tracking method is clearly superior. For example, vocal sounds with vibrato are served well by the frequency-tracking method as are sounds with extremely variable pitch. In the latter case, a separate pitch detection step can be done which may be used for music transcription and to steer reduction to a harmonic-only format. However, the frequency-tracking method may suffer from momentary drop-outs, which if they occur in the higher frequencies are particularly audible, and poor partial tracking, which produces artifacts when time-stretching is employed. For fixed-pitch harmonic sounds, the frequency-tracking and phase-vocoder analysis/synthesis methods have been found to give comparable results.

Various measures of the time-varying spectrum are useful and are correlated with listener ability to distinguish among musical sounds. Two that have proven useful are the spectral centroid and spectral irregularity, both of which can be measured as functions of time. Spectral centroid is a measure of an instrument’s bandwidth and is strongly associated with the perception of “brightness.” Some instrument sounds tend to be brighter than others (e.g., trumpet is generally brighter

than French horn), and many exhibit significant changes of centroid during sounds that are very noticeable by listeners. While reduction in spectral irregularity (by spectral smoothing) is highly discriminable for sounds that exhibit high values of it, no particular timbral percept has been associated with it. However, brass instrument sounds generally have smoother spectra than those of woodwinds and strings.

Spectra can be altered or simplified in various ways. Variation of the normalized spectrum can be eliminated by replacing the original time-varying spectrum with one which is proportional to the product of the RMS amplitude-vs-time envelope and the average spectrum. Spectral centroid can be altered by multiplying the spectral amplitudes by a function which increases or decreases with respect to frequency, with a positive or negative trending exponential function being a logical choice. Using an optimization method, the average or instantaneous centroid can be matched to an arbitrary value, provided there is sufficient harmonic energy to begin with. Spectral irregularity can be decreased by spectral smoothing, but it can also be increased or decreased by replacing the spectral amplitudes by the same ones raised to a positive power, where powers less than 1 will diminish and powers greater than 1 will accentuate spectrum peaks.

Two other sound spectrum measurements on sound which may prove important for timbre perception are spectro-temporal incoherence (*SI*) and inverse spectral density (*ISD*). *SI* is a measurement of how well individual partial amplitudes track the RMS amplitude envelope and can be applied to both harmonic and inharmonic sounds. *ISD* is a measurement of the average distance between significant partials and is most applicable to sounds with inharmonic partials. For meaningful *ISD* measurement, problems remain in determining the best way to determine the threshold above which components are considered to be significant and how to deal with extreme variations in the spectral envelope.

References

- Beauchamp, J. W. and Fornango, J. P. (1966). "Transient Analysis of Harmonic Musical Tones by Digital Computer," 31st Convention of the Audio Eng. Soc., New York, Audio Eng. Soc. Preprint No. 479.
- Beauchamp, J. W. (1969). "A Computer System for Time-Variant Harmonic Analysis and Synthesis of Musical Tones," in *Music by Computers*, H. von Foerster and J. W. Beauchamp, eds. (J. Wiley & Sons, New York), pp. 19–62.
- Beauchamp, J. W. (1975). "Analysis and Synthesis of Cornet Tones Using Nonlinear Inter-harmonic Relationships," J. Audio Eng. Soc., **23**(10), 778–795.
- Beauchamp, J. W. (1993). "Unix Workstation Software for Analysis, Graphics, Modification, and Synthesis of Musical Sounds," 94th Convention of the Audio Eng. Soc., Berlin, Audio Eng. Soc. Preprint No. 3479.
- Beauchamp, J. W., Maher, R. C., and Brown, R. (1993). "Detection of Musical Pitch from Recorded Solo Performances," 94th Convention of the Audio Eng. Soc., Berlin, Audio Eng. Soc. Preprint No. 3541.
- Beauchamp, J. W. and Horner, A. (1995). "Wavetable Interpolation Synthesis Based on Time-Variant Spectral Analysis of Musical Sounds," 98th Convention of the Audio Eng. Soc., Paris, Audio Eng. Soc. Preprint No. 3960.

- Benade, A. (1976). *Fundamentals of Musical Acoustics* (Oxford University Press, New York).
- Boersma, P. (1993). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proc. Institute of Phonetic Sciences* **17** (Amsterdam), 97–110.
- Brown, J. C. (1992). "Musical fundamental frequency tracking using a pattern-recognition method," *J. Acoust. Soc. Am.* **92**(3), 1394–1402.
- Cano, P. (1998). "Fundamental frequency estimation in the SMS analysis," *Proc. 1998 Digital Audio Effects Workshop (DAFX98)*.
- Chen, K. (2001). "Pitch-Synchronous Overlap-Add Musical Resynthesis with Variable Time-Scaling Set by a Human Conductor," unpublished masters thesis, Univ. of Illinois at Urbana-Champaign, Urbana, IL.
- de Cheveigné, A. and Kawahara, H. (2002). "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.* **111**(4), 1917–1930.
- Depalle, P., Garcia, G., and Rodet, X. (1993a). "Analysis of sound for additive synthesis: Tracking of partials using hidden Markov models," *Proc. 1993 Int. Computer Music Conf.*, Tokyo (Int. Computer Music Assoc., San Francisco), pp. 94–97.
- Depalle, P., Garcia, G., and Rodet, X. (1993b). "Tracking of partials for additive sound synthesis using hidden Markov models," *Proc. 1993 IEEE Conf. on Acoustics, Speech, and Signal Processing (ICASSP-93)*, Minneapolis (IEEE, New York), pp. I-225–228.
- Ding, Y., and Qian, X. (1997). "Processing of musical tones using a combined quadratic polynomial-phase sinusoid and residual (QUASAR) signal model," *J. Audio Eng. Soc.* **45**(7/8), 571–584.
- Doval, B. and Rodet, X. (1991). "Estimation of fundamental frequency of musical sound signals," *Proc. 1991 IEEE Conf. on Acoustics, Speech, and Signal Processing (ICASSP-91)*, Toronto (IEEE, New York), pp. 3657–3660.
- Fitz, K., Walker, W., and Haken, L. (1992). "Extending the McAulay-Quatieri analysis for synthesis with a limited number of oscillators," *Proc. 1992 Int. Computer Music Conf.*, San Jose, CA (Int. Computer Music Assoc., San Francisco), pp. 381–382.
- Fitz, K., Haken, L., and Christensen, P. (2000). "A new algorithm for bandwidth association in bandwidth-enhanced additive sound modeling," *Proc. 2000 Int. Computer Music Conf.*, Berlin (Int. Computer Music Assoc., San Francisco), pp. 384–387.
- Fitz, K., and Haken, L. (2002). "On the use of time-frequency reassignment in additive sound modeling," *J. Audio Eng. Soc.* **50**(11), 879–893.
- Fletcher, H. (1964). "Normal vibration frequencies of a stiff piano string," *J. Acoust. Soc. Am.* **36**(1), 203–209.
- George, E. B., and Smith, M. J. T. (1992). "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones," *J. Audio Eng. Soc.* **40**(8), 497–516.
- Goodwin, M., and Rodet, X. (1994). "Efficient Fourier synthesis of nonstationary sinusoids," *Proc. 1994 Int. Computer Music Conf.*, Aarhus, Denmark (Int. Computer Music Assoc. San Francisco), pp. 333–334.
- Harris, F. J. (1978). "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE* **66**(1), 51–83.
- Hess, W. (1983). *Pitch Determination of Speech Signals* (Springer-Verlag, New York).
- Horner, A. and Beauchamp, J. (1995). "Synthesis of Trumpet Tones Using a Wavetable and a Dynamic Filter," *J. Audio Eng. Soc.* **43**(10), 799–812.
- Horner, A., Beauchamp, J., and So, R. (2004). "Detection of Random Alterations to Time-Varying Musical Instrument Spectra", *J. Acoust. Soc. Am.* **166**(3), 1800–1810.

- Kaiser, J. F., and Schafer, R. W. (1980). "On the use of the I_0 -sinh window for spectrum analysis," IEEE Trans. on Acoustics, Speech and Signal Processing **ASSP-28**(1), 105–106.
- Lattard, J. (1993). "Influence of inharmonicity on the tuning of a piano—Measurements and mathematical simulation," J. Acoust. Soc. Am. **94**(1), 46–53.
- Luce, D. A. (1963). *Physical Correlates of Non-Percussive Musical Instruments*, PhD dissertation, Massachusetts Institute of Technology, Cambridge, M.A.
- Luce, D. and Clark, M. (1967). "Physical Correlates of Brass-Instrument Tones," J. Acoust. Soc. Am. **42**(6), 1232–1243.
- Luce, D. A. (1975). "Dynamic Spectrum Changes of Orchestral Instruments," J. Audio Eng. Soc. **23**(7), 565–568.
- Maher, R. C. (1989). "An approach for the separation of voices in composite musical signals," unpublished Ph.D. dissertation, University of Illinois at Urbana-Champaign, Urbana, IL.
- Maher, R. C., and Beauchamp, J. W. (1990). "An investigation of vocal vibrato for synthesis," Applied Acoustics **30**(2&3), 219–245.
- Maher, R. C. and Beauchamp, J. W. (1994). "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," J. Acoust. Soc. Am. **95**(4), 2254–2263.
- McAdams, S., Beauchamp, J. W., and Meneguzzi, S. (1999). "Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters," J. Acoust. Soc. Am. **105**(2), 882–897.
- McAulay, R. J., and Quatieri, T. F. (1986). "Speech analysis/synthesis based on a sinusoidal representation," IEEE Trans. on Acoustics, Speech, and Signal Processing **34**(4), 744–754.
- Messiaen, O. (1941). *Quatuor pour la fin du temp.*
- Moore, B. C. J., Glasberg, B. R., and Peters, R. W. (1985). "Relative dominance of individual partials in determining the pitch of complex tones," J. Acoust. Soc. Am. **77**(5), 1853–1860.
- Moorer, J. A. (1974). "The optimum comb method of pitch period analysis of continuous digitized speech," IEEE Trans. on Acoustics, Speech and Signal Processing **ASSP-22**(5), 330–338.
- Moorer, J. A. (1975). *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer*, Report No. STAN-M-3, Center for Computer Research in Music and Acoustics (CCRMA), Dept. of Music, Stanford, CA.
- Morse, P. M. (1976). *Vibration and Sound* (American Institute of Physics for the Acoustical Society of America, Melville, NY).
- Noll, A. M. (1967). "Cepstrum pitch determination," J. Acoust. Soc. Am. **41**(2), 293–309.
- Nuttall, A. H. (1981). "Some windows with very good sidelobe behavior," IEEE Trans. on Acoustics, Speech, and Signal Processing **ASSP-29**(1), 84–91.
- Piszczalski, M., and Galler, B. A. (1979). "Predicting musical pitch from component frequency ratios," J. Acoust. Soc. Am. **66**(3), 710–720.
- Roads, C. (1996). *The Computer Music Tutorial* (MIT Press, Cambridge, MA).
- Rodet, X., and Depalle, P. (1992). "Spectral envelopes and inverse FFT Synthesis," *93rd Convention of the Audio Engineering Society*, San Francisco, Audio Eng. Soc, Preprint 3393.
- Rossing, T. D. (1976). "Acoustics of percussion instruments—Part I," The Physics Teacher **14**(9), 546–556.
- Schroeder, M. R. (1999). *Computer Speech: Recognition, Compression, Synthesis* (Springer, New York).

- Serra, X. (1989). *A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition*, Report No. STAN-M-58, Center for Computer Research in Music and Acoustics (CCRMA), Dept. of Music, Stanford, CA.
- Serra, X., and Smith, J. O. (1990). "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music J.* **14**(4), 12–24.
- Serra, X. (1997). "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, C. Roads, S. T. Pope, A. Piccialli, and G. De Poli, eds. (Swets and Zeitlinger, Exton, PA), pp. 91–122.
- Smith, J. O., and Gossett, P. (1984). "A flexible sampling-rate conversion method," *Proc. 1984 IEEE Conf. on Acoustics Speech, and Signal Processing* (ICASSP-84), San Diego (IEEE, New York), pp. 19.4.1–19.4.2.
- Smith, J. O., and Serra, X. (1987). "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," *Proc. 1987 Int. Computer Music Conf.*, Urbana, IL (Int. Computer Music Assoc., San Francisco), pp. 290–297.
- Strong, W. and Clark, M. (1967a). "Synthesis of Wind-Instrument Tones," *J. Acoust. Soc. Am.* **41**(1), 39–52.
- Strong, W. and Clark, M. (1967b). "Perturbations of Synthetic Orchestral Wind-Instrument Tones," *J. Acoust. Soc. Am.* **41**(2), 277–285.
- Sundberg, J. (1974). "Articulatory interpretation of the 'singing formant,'" *J. Acoust. Soc. Am.* **55**(4), 838–844.

Fundamental Frequency Tracking and Applications to Musical Signal Analysis

JUDITH C. BROWN

1 Introduction to Musical Signal Analysis in the Frequency Domain

The constant-Q spectral transform (Brown, 1991) can be used to analyze musical signals and can be effectively employed as a front end for measurements of fundamental frequency. This transform also has advantages for the analysis of musical signals over the conventional discrete Fourier transform, or FFT in its fast-Fourier-transform implementation. Because the FFT computes frequency components on a linear scale with a particular fixed resolution or bandwidth (frequency spacing between components), it frequently results in too little resolution for low musical frequencies and better resolution than needed at high frequencies.

For example, if we consider a sampling frequency (f_s) equal to 22050 Hz and a window size (N) of 512 samples, then the frequency resolution is $f_s/N = 43$ Hz for the entire range of frequencies from 0 to 11025 Hz. For the fundamental of the lowest note of the violin, G_3 at 196 Hz, this is 22% of its frequency, whereas a musical semitone corresponds to a 6 % spacing. Therefore, all the information about three to four adjacent musical notes is contained in one frequency bin. At the upper end of the piano the frequency of the note C_8 is 4186 Hz, and the next lowest note would have a 6% frequency separation equal to 251 Hz. With the same 43 Hz frequency resolution, roughly six bins correspond to this single note difference, giving excess and unneeded information.

Thus, it is clear that the conventional discrete Fourier transform (DFT) is inefficient for musical applications. What is needed is information about the spectral components contained within a musical instrument's full frequency range. For example, piano notes are tuned approximately in equal temperament with frequencies $f_k \cong f_{\min} \cdot (2^{(1/12)k})$, where f_{\min} is the frequency of the lowest note and f_k is the frequency of the note k semitones above f_{\min} . For our calculations, we can choose analysis frequencies $f_j = f_{\min} \cdot (2^{(1/24)j})$, giving two frequency bins per musical note (quartertone spacing). Because the frequency resolution in Hz is equal to the frequency difference between bins, the resolution is given by $\Delta f_j = f_{j+1} - f_j = 2^{1/24} \cdot f_j - f_j$, and the ratio of frequency to resolution or Q ,

TABLE 2.1. Frequencies of
Musical Notes of the Octave
Beginning on Middle C on a Piano

Note	Frequency (Hz)
C ₄	261.63
C ₄ [#]	277.18
D ₄	293.66
D ₄ [#]	311.13
E ₄	329.63
F ₄	349.23
F ₄ [#]	369.99
G ₄	392.00
G ₄ [#]	415.30
A ₄	440.00
A ₄ [#]	466.16
B ₄	493.88
C ₅	523.25

defined as $f_j/\Delta f_j = 1/(2^{1/24} - 1) \cong 34$, is a constant. Thus, the transform is equivalent to a 1/24th octave constant-Q filter bank.

For a musical example, see Table 2.1, where the note frequencies for the octave beginning on middle C (aka C₄) are shown. The frequencies are given by

$$f_k = (2^{1/12})^k 261.63, \quad (2.1)$$

where $k = 0$ to 12 for this octave. It is also clear that

$$\log(f_k) = \frac{\log(2)}{12} k + \log(261.63). \quad (2.2)$$

Thus, the log-frequencies of the notes are linearly related to note number k .

An extremely important property of the constant-Q transform which follows is that for sounds made up of harmonic frequency components, the non-uniform spacing of the components shown as a function of bin number is independent of fundamental frequency. The spacing pattern is shown in Fig. 2.1, which is the plot of a hypothetical spectrum with equal amplitude frequency components at 100 Hz, 200 Hz, 300 Hz, ..., 1000 Hz. The positions on the horizontal axis corresponding to $\log_{10}(\text{frequency})$ are spaced the same for any set of harmonically related components. For example, the spacing between the first two harmonics is $\log_{10}(200) - \log_{10}(100) = \log_{10}(2)$, that between the second and third harmonics is $\log_{10}(3/2)$, and so forth.

Although this was shown for the example of a fundamental frequency of 100 Hz and log base 10, it holds for any fundamental. That is, the absolute positions on the log-frequency axis depend on the frequency of the fundamental, but the relative positions of the harmonics with respect to each other are invariant. Thus, these spectral components form an invariant “pattern” in the log-frequency domain, and

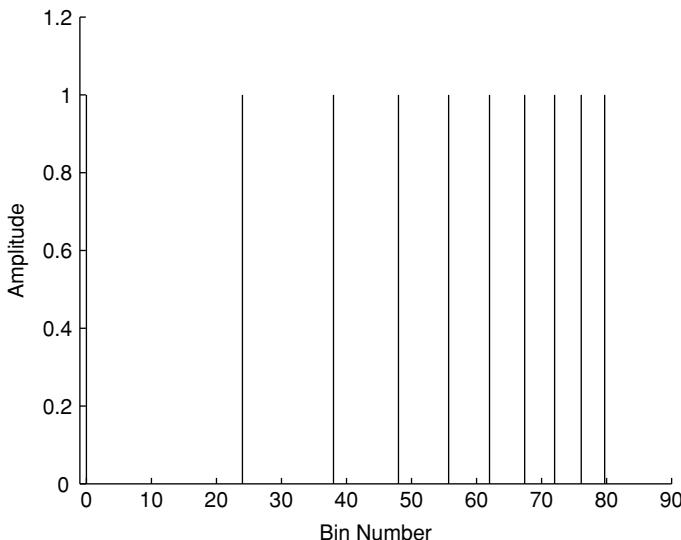


FIGURE 2.1. A Fourier transform pattern of 10 equal-amplitude harmonic frequency components plotted against log frequency (constant-Q bin number) for a bin spacing of 24 bins per octave.

this pattern is the same for all sounds with harmonic frequency components. Of course, the amplitudes of the components may vary from one harmonic to the next, reflecting differences in the timbres of the sounds analyzed.

By comparison, the conventional DFT in a linear plot against frequency exhibits a constant separation between component frequencies for harmonic musical sounds. This is the dominant feature of the spectral patterns produced, and both the component separations and the overall positions of the patterns vary with fundamental frequency. The result is that it is difficult to pick out differences in other features of the sound, such as spectral shape, attack, decay, and absolute position in the frequency domain, which identifies the fundamental frequency.

The log-frequency representation, on the other hand, gives a unique spacing pattern for harmonic spectral components, and thus, the problem of fundamental frequency tracking becomes a problem of recognizing this pattern. In addition to its practical advantages, this idea has theoretical appeal for its similarity to modern theories of pitch perception based on pattern recognition (Gerson and Goldstein, 1978). In one of these theories, the perception of the pitch of a sound with a missing fundamental is explained by the “pattern” formed by the remaining harmonics on the basilar membrane. In Section 3 a computer algorithm that recognizes the pattern made by these harmonics in the log-frequency domain will be discussed. We will see that it can correctly identify the fundamental frequency even in those cases where there is no spectral energy at the frequency of the fundamental.

Next, we will discuss how the constant-Q transform can be implemented in a straightforward calculation. Then, we will apply pattern-matching techniques to

the pitch detector problem. After that, we will describe how a phase-vocoder-based method can be used for extremely accurate fundamental frequency measurements. Finally, this method will be applied to answer two questions: First, how precisely harmonic are musical sounds, i.e., how close to exact integers are the ratios of the frequency components? Second, what exact pitch is perceived when sounds are played with vibrato or frequency modulation?

2 Calculation of a Constant-Q Transform for Musical Analysis

2.1 Background

The constant-Q filter bank and its similarity to the auditory system has been explored in several theses (Petersen, 1980; Seneff, 1985; Stautner, 1983), each of which reference the literature extensively. For those who wish to review techniques of digital signal processing, an article by Higgins (1976) is recommended as a background discussion of sampling effects in the calculation of the discrete Fourier Transform. The theory of the short-time Fourier transform was originally developed by Schroeder and Atal (1962) and was extensively reviewed by Nawob et al. (1983).

Various schemes for implementing constant-Q spectral analysis outside a musical context have been published (Braccini and Oppenheim, 1974; Gambardella, 1971, 1979; Harris, 1976; Helms, 1976; Oppenheim, et al., 1971; Youngberg and Boll, 1978). Music researchers at the Center for Computer Research in Music and Acoustics (CCRMA) at Stanford have used a “bounded Q” Transform (Kashima and Mont-Reynaud, 1985, cited in Chafe et al., 1985) similar to that of Harris (1976). Kronland-Martinet (1988) and his group at Marseilles have employed a wavelet transform for musical analysis and synthesis. This is a constant-Q method similar to the Fourier transform and to the method described in this chapter, but it is based on the use of wavelets as generalized basis functions.

The method described below has two advantages over previous methods: The first is its simplicity; and the second is that it is calculated for frequencies that are exponentially spaced with two frequency components per musical half-step, giving exactly the information that is needed for musical analysis with sufficient resolution to distinguish adjacent musical notes. Furthermore, a sound with harmonic frequency components results in a constant pattern in the log-frequency domain.

2.2 Calculations

As mentioned in the introduction, the optimum spacing of frequency components for musical analysis corresponds to quarter-tone spacing of the equal-tempered scale. The frequency of the k th spectral component or *bin number* is thus given by

$$f_k = (2^{1/24})^k f_{\min}, \quad (2.3)$$

where f_k varies from f_{\min} to an upper frequency chosen to be below the Nyquist frequency. The minimum frequency f_{\min} is an adjustable parameter and can be chosen to be the lowest frequency about which information is desired, for example, a frequency conveniently below that of the open G string for calculations on sound produced by a violin. If desired, the bin number can be calculated from the frequency using

$$k = 24 \log_2(f_k/f_{\min}) \cong 79.73 \log_{10}(f_k/f_{\min}). \quad (2.4)$$

The resolution or bandwidth Δf_k is defined as the difference between consecutive bin frequencies, and Q is the ratio of frequency to bandwidth. Then for quarter tone ($\cong 3\%$) resolution, we have

$$Q = f_k/\Delta f_k = f_k/((2^{1/24} - 1)f_k) \cong 34. \quad (2.5)$$

However, for the discrete Fourier transform, the bandwidth is equal to the sample rate divided by the window size (the number of samples analyzed in the time domain). Thus, the window size $N[k]$ is equal to the sample rate divided by the bandwidth. If the ratio of frequency to bandwidth is a constant (constant-Q), the window size varies inversely with the bin frequency:

$$N[k] = f_s/\Delta f_k = f_s/(f_k/Q) = f_s Q/f_k, \quad (2.6)$$

where f_s is the sampling rate.

From Eq. (2.6) we see that for the constant-Q case that

$$f_k = Q f_s / N[k]. \quad (2.7)$$

This corresponds to a digital frequency of $2\pi Q/N[k]$.

Because the period in samples for frequency f_k is f_s/f_k , it follows from Eq. (2.6) that a window of length $N[k]$ contains Q complete cycles for each frequency f_k . This makes physical sense, because in order to distinguish between f_{k+1} and f_k when their ratio is $2^{1/24} \cong 34/33$, we must examine at least 33 cycles.

For comparison, it is interesting to consider the conventional discrete Fourier transform in terms of the quality factor $Q = f_k/\Delta f_k$. Here, because $f_k = kf_s/N$, where N is fixed, and $\Delta f_k = f_s/N = \Delta f$, a fixed quantity, we see that $f_k/\Delta f$ is equal to the bin number k , and this is also the number of periods of frequency f_k which occur in the fixed window.

An expression for the k th spectral component for the constant-Q transform can be derived by considering the corresponding component for the short-time DFT (Oppenheim and Schafer, 1975):

$$X[k] = \sum_{n=0}^{N-1} w[n]x[n]e^{-j2\pi kn/N} \quad (2.8)$$

Here $x[n]$ is the n th sample of the digitized temporal function being analyzed. The digital frequency is $2\pi k/N$. For each k the period in samples is N/k , and the number of cycles analyzed is equal to k . $w[n]$ gives the shape of the window function, which is discussed below.

According to Eq. (2.7), the digital frequency of the constant-Q k th component is $2\pi Q/N[k]$. The window function has the same shape for each component, but its length is determined by $N[k]$ so it is a function of k as well as n . We must also normalize by dividing the sum by $N[k]$ since the number of terms varies with k . Equation (2.8) thus can be rewritten as

$$X^{cq}[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} w[n, k]x[n]e^{-j2\pi kn/N[k]}, \quad (2.9)$$

which is the constant-Q transform of the signal $x[n]$ (see Appendix A).

For each bin frequency the period in samples is $N[k]/Q$, so we always analyze Q cycles. A comparison of variables used in the calculation of the constant-Q and the conventional Fourier transforms is given in Table 2.2.

TABLE 2.2. Comparison of the discrete Fourier transform (DFT) and the constant-Q transform

Parameter	DFT	Constant-Q transform
Frequency f_k	$k\Delta f$ (Linear in k)	$(2^{1/24})^k \cdot f_{\min}$ (Exponential in k)
Window size	Constant = N	Variable = $N[k] = f_s Q/f_k$
Resolution Δf	Constant = f_s/N	Variable = f_k/Q
$\frac{f_k}{\Delta f_k}$	Variable = k	Constant = Q
Cycles in window	Variable = k	Constant = Q

In practice, Eq. (2.9) is used as the basis for our calculations with $N[k] = N_{\max}/(2^{1/24})^k$. N_{\max} is Q times the period of the lowest analysis frequency in samples. The Nyquist condition becomes $2\pi Q/N[k] < \pi$, which means $N[k] > 2Q$. This is identical to the usual statement that there must be at least two samples per period to avoid aliasing.

The simple choice of a window function $w[n,k]$ equal to unity over the interval $(0, N[k] - 1)$, results in the rectangular window, which can be shown to have maximum “spillover” into adjacent frequency bins (Harris, 1978). Instead, we use a Hamming window

$$w[n, k] = \alpha + (1 - \alpha)\cos(2\pi n/N[k]), \quad (2.10)$$

where $\alpha = 25/46$ and $0 \leq n \leq N[k] - 1$. This choice results in a worst case spillover of -42 dB.

The calculation of the constant-Q transform described in this section is very straightforward both computationally and conceptually. It does not, however, take advantage of the computational efficiency of the fast Fourier transform. It is possible to transform a DFT into a constant-Q transform as described by Brown and Puckette (1992), thus taking advantage of the speed of the FFT calculation. This method involves the calculation of kernels which are applied to each subsequent FFT. Only a few multiplications are required for the calculation of each component of the constant-Q transform, so this transformation adds a small amount to the computation time. Details are given in Appendix A.

2.3 Results

Most examples were taken from sounds of musical instruments digitized from live performances. Other examples were generated using Barry Vercoe's Csound software (Boulanger, 2000; Vercoe, 1986). Calculations were carried out every 500 samples, corresponding to about 15 ms at a sample frequency of 32000 samples per second.

Figures 2.2 and 2.3 are graphs of the constant-Q transform amplitude on the left vertical axis plotted against bin number on the horizontal axis and time on the right vertical axis. Fig. 2.2 is for a violin performance of a G major diatonic scale starting at G_3 and ending at G_5 . Note that the maxima for the fundamental frequencies start at 196 Hz, corresponding to bin 13 (a very weak fundamental) and end at 784 Hz, corresponding to bin 61. This follows from Eq. (2.4) when $f_{\min} = 134.65$ Hz. An important consideration is the percentage difference of nearby frequencies that can be resolved. Note that frequencies are resolved up to the 20th harmonic, where the frequencies differ by about 5%. In Fig. 2.3 the violin starts at D_5 , corresponding to 587 Hz at bin 51, and glissandos to A_5 , corresponding to 880 Hz at bin 65. Associated spectral changes are also apparent. This example of continuous fundamental frequency change will be referred to again in the next section on fundamental frequency tracking.

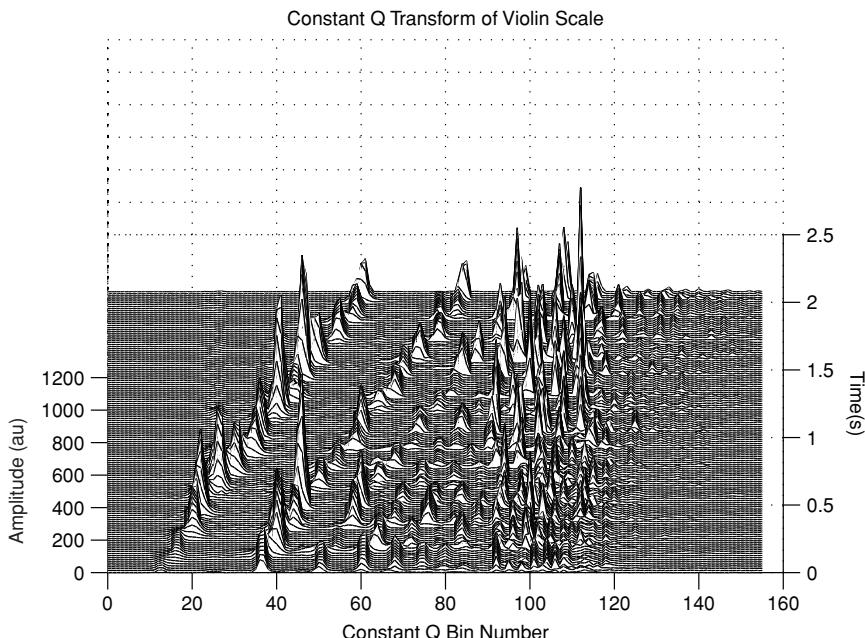


FIGURE 2.2. Constant-Q transform of violin playing a G major diatonic scale from G_3 (196 Hz) to G_5 (784 Hz).

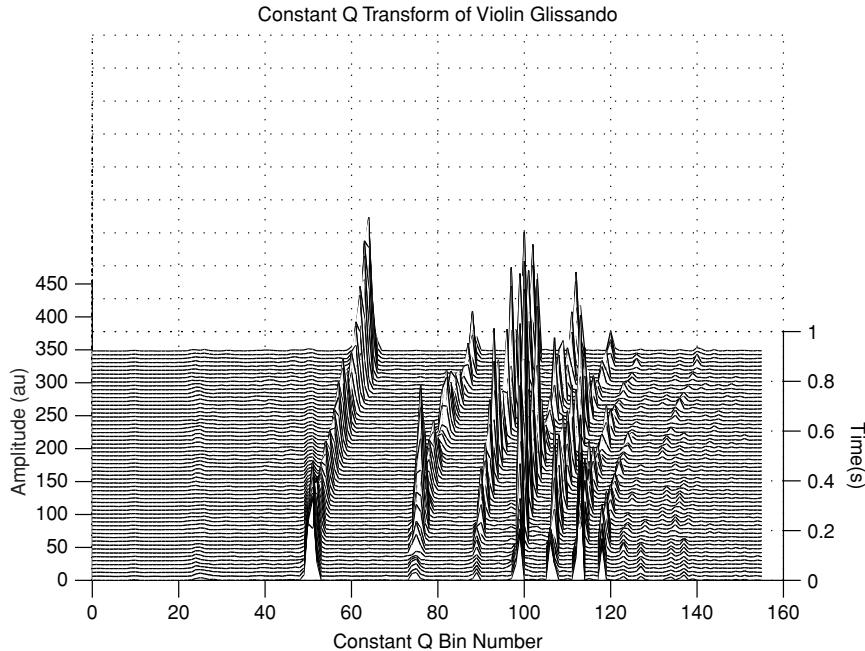


FIGURE 2.3. Constant-Q transform of violin glissando from D₅ (587 Hz) to A₅ (880 Hz).

Figure 2.4 shows the constant-Q transform of three octave-related notes (G₃ = 196 Hz, G₄ = 392 Hz, and G₅ = 784 Hz) generated by software. Each sound contains 20 harmonics of equal amplitude. Note that, as shown in the figure, the amplitudes differ slightly due to the positions of the corresponding frequencies relative to the center frequencies of the bins into which they fall. However, it is clear that the patterns, in terms of frequency spacings, are identical; only the relative positions on the frequency axis indicate that the notes are different. Fig. 2.5 represents a 512-point traditional DFT of this same sound for comparison. Here the harmonics are equally spaced, and this is the major feature that stands out. The resolution for this case is 62.5 Hz, allowing the harmonics of even the lowest note to be resolved.

Figures 2.6 and 2.2 offer a comparison of the traditional (Fig. 2.6) and constant-Q (Fig. 2.2) transforms for the sound of a violin. Each shows the transform magnitude for the G major diatonic scale played from G₃ to G₅. It is very difficult to say anything at all about spectral content for the conventional plot of Fig. 2.6; it is even difficult to determine note changes for the low-frequency notes. On the other hand, Fig. 2.2 very clearly indicates not only the note changes but also the spectral content; for example, G₃ and A₃ have almost undetectable fundamentals. Most striking of the spectral features is the formant in the region of 3000 Hz.

Figure 2.7 shows the constant-Q transform for the violin playing the note D₅ = 587 Hz (bin 51) with vibrato. The second harmonic is considerably weaker for the

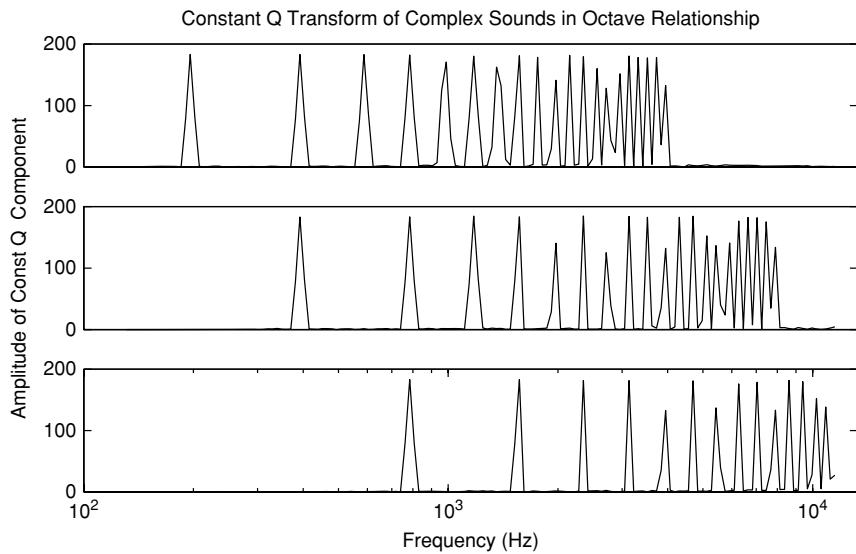


FIGURE 2.4. Constant-Q transforms of three complex sounds with fundamentals G_3 (196 Hz), G_4 (392 Hz), and G_5 (784 Hz), each having 20 harmonics of equal amplitude.

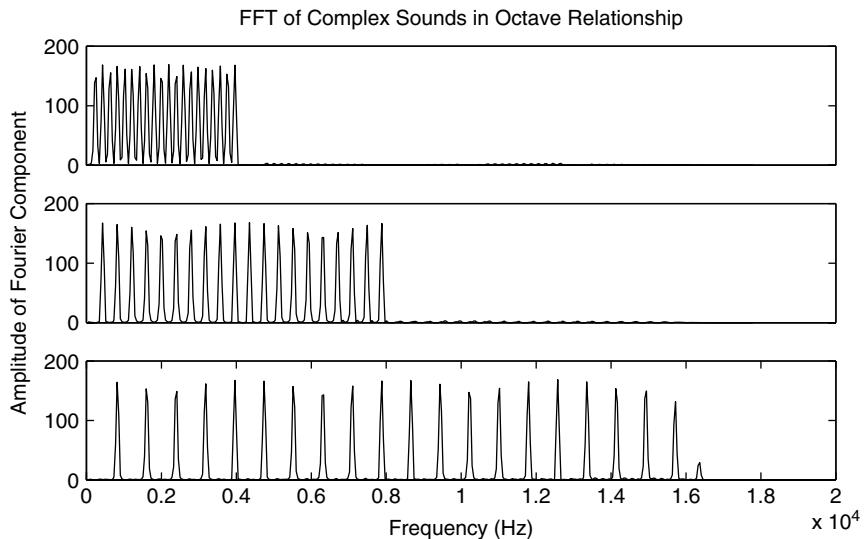


FIGURE 2.5. Discrete Fourier transforms of three complex sounds with fundamentals G_3 (196 Hz), G_4 (392 Hz), and G_5 (784 Hz), each having 20 harmonics of equal amplitude.

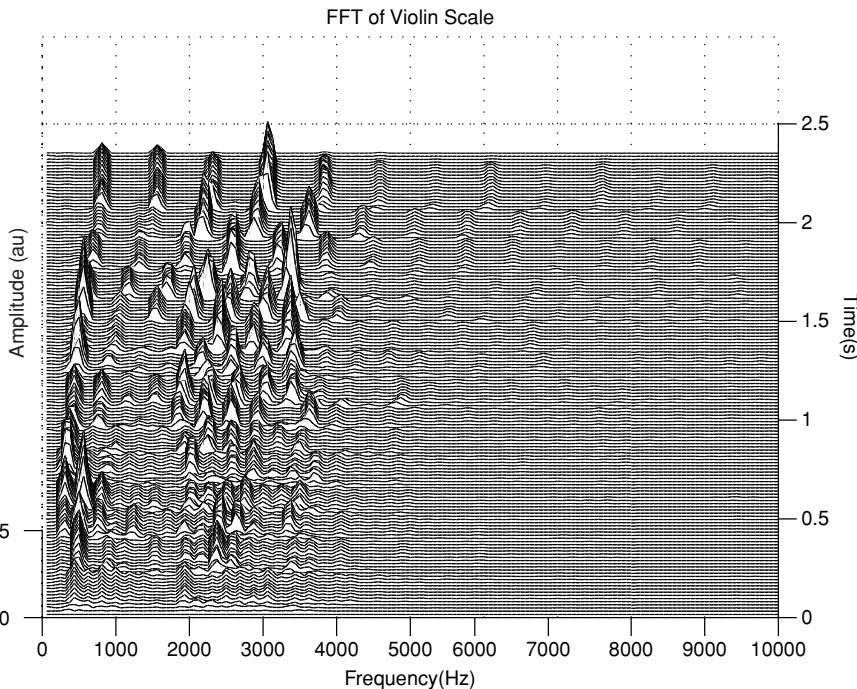


FIGURE 2.6. Discrete Fourier transform of violin playing a G major diatonic scale from G₃ (196 Hz) to G₅ (784 Hz).

higher region of the vibrato while the 7th and 9th harmonics (bins 118 and 126) are weaker for the lower-frequency region. Most remarkable in the spectrum is the extremely strong 6th harmonic (bin 113). This harmonic falls right in the 3000 Hz formant region mentioned above and is amplified by a violin body resonance which occurs in this region.

3 Musical Fundamental-Frequency Tracking Using a Pattern-Recognition Method

3.1 Background

The problem of musical pitch tracking has received relatively little attention in comparison to the massive efforts carried out by the speech community for use with various speech encoders for communications purposes. Musical applications have, for the most part, been in the area of intelligent systems, where an accurate pitch tracker is a necessity at the front end. For a more complete review of previous work in the field of musical pitch tracking see Brown and Zhang (1991).

Most efforts at musical pitch tracking have taken place in the frequency domain (Amuedo, 1985; Chafe and Jaffe, 1986; Terhardt, 1979; Terhardt et al., 1982) and

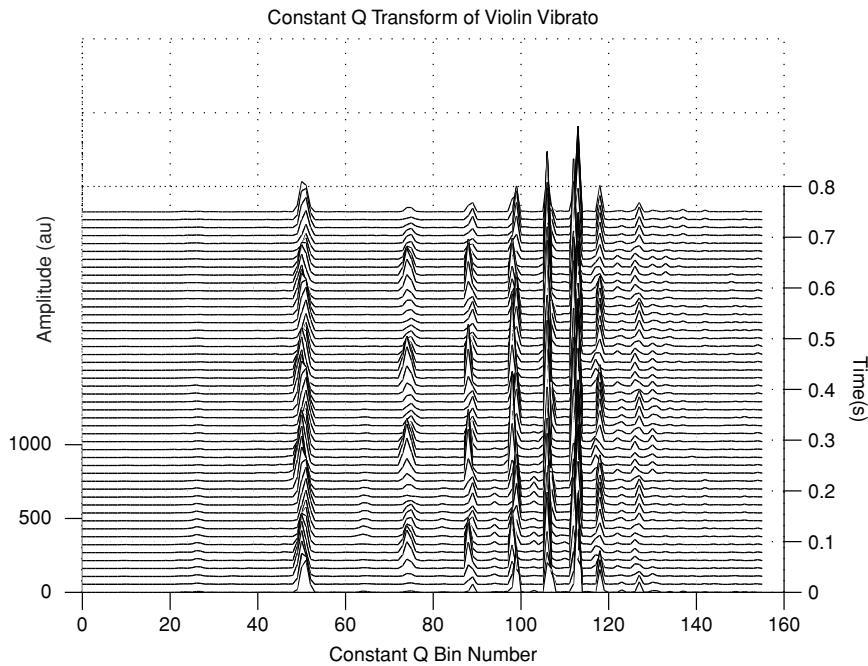


FIGURE 2.7. Constant-Q transform of violin playing D₅ (587 Hz) with vibrato.

have used a method similar to that of the Schroeder (1968) histogram method. After the calculation of a fast Fourier transform, a hypothesis is asserted for each frequency component of all possible fundamental frequencies for which it could be a harmonic, e.g., each frequency component is divided by integers and the results are entered in a table. The entries are weighted, and a decision is made based on criteria involving the number of components and their weights. The frequency is chosen that most closely meets previously determined criteria.

A similar pitch tracker (Piszczalski and Galler, 1979) took ratios of pairs of components to form their hypotheses for the fundamental and then proceeded as above. Duifhuis et al., (1982) studied speech segments using a method which most closely approaches that of this article. Following an FFT they kept a maximum of six peaks and then used a “harmonic sieve” to determine which of these peaks best fit the logarithmic spacing obtained with harmonic frequency components. This method was later refined by Scheffers (1983). Another method of this type was described by Maher and Beauchamp (1994), where pitch was chosen to minimize an error function based on the differences between spectrum peaks and corresponding harmonics of candidate fundamental frequencies.

3.2 Calculations

This method is based on the property summarized in Fig. 2.1 that a sound with harmonic frequency components has Fourier components with spacings in the

log-frequency domain that are independent of the fundamental frequency. If we assume that the frequencies in question are integer multiples of f_{\min} , the corresponding constant-Q bin numbers (for 10 harmonics) given by Eq. (2.4) are 0, 24, 38, 48, 55.7, 62, 67.4, 72, 76, and 79.7. Because this “pattern of 1’s” is constant for harmonic frequency components, it can be cross-correlated with the constant-Q transform of a sound, and a maximum should occur at the position of the fundamental. This is tantamount to shifting the template pattern of Fig. 2.1 to the right while multiplying it with the constant-Q magnitude spectrum X^{cq} of the signal and adding the results. The fundamental is detected when the maximum value of the sum occurs. In other words, we wish to find k_{\max} as the value of the bin number k which maximizes the expression

$$\sum_{h=1}^{h_{\max}} |X^{cq}[k + 24 \log_2(h)]|, \quad (2.11a)$$

where h is the harmonic number and h_{\max} is the maximum harmonic number. The actual fundamental frequency would be calculated using

$$f_0 = (2^{1/24})^{k_{\max}} f_{\min}. \quad (2.11b)$$

We see that as the convolution is computed, the first component of the harmonic pattern at some point coincides with the fundamental of the analyzed sound. This “asserts” the fundamental frequency corresponding to that position of the template as a hypothesis. As the template slides across the magnitude spectrum, the convolution obtains a number for each frequency corresponding to the sum of all the frequency components of the sound that are at harmonics of the test fundamental. Thus, in a very elegant and complete way we obtain results that previous researchers approached with the histogram method. There is a computational advantage as well in that we simply add components with appropriate spacing.

Note that this pitch tracker solves the problem of the “missing fundamental” in much the same manner as that hypothesized for humans. It essentially compares the harmonics present in the signal to a template and finds the best match. This is consistent with the pattern-matching theory (Gerson and Goldstein, 1978) of human pitch perception.

3.3 Results

As with any pattern-matching method, cross-correlation most unambiguously establishes the position of the pattern when it is close in shape to the “ideal pattern” (Duda and Hart, 1973). Thus, it is best if the number of components in the ideal pattern matches the average number of nonzero Fourier components for the particular instrument analyzed. This number therefore becomes an adjustable parameter to be optimized for each instrument.

Figures 2.8 and 2.9 show the spectrum and cross-correlation functions for two instruments with very different spectra. In Fig. 2.8 six components of the ideal pattern were used, while in Fig. 2.9 ten components were needed. For the effect

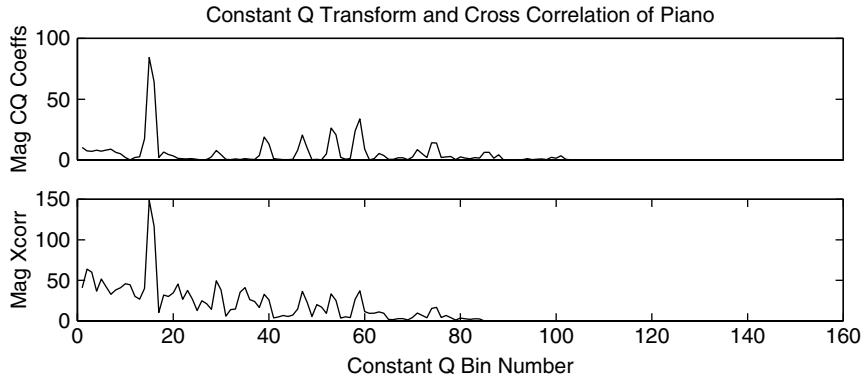


FIGURE 2.8. Constant-Q magnitude spectrum vs bin number for a 15 ms portion of a C₄ note produced by a piano (above) and the result of cross-correlation (below) of this spectrum with the function shown in Fig. 2.1.

of varying the number of components in the cross-correlation template on the frequency-tracking results for a particular instrument, see Brown (1992).

Figure 2.10 gives the pitch-tracking result in terms of fundamental frequency vs time for the violin scale spectrum shown in Fig. 2.2. Each point in this graph represents the peak of a cross-correlation calculation on an analysis frame similar to that of Fig. 2.9, corresponding to approximately 15 ms of sound. Because this is a diatonic scale, perfect results would consist of a sequential set of horizontal lines rising by one or two semitones corresponding to a half or a whole step in the scale. Thus, errors made by the pitch tracker are easily distinguished as points

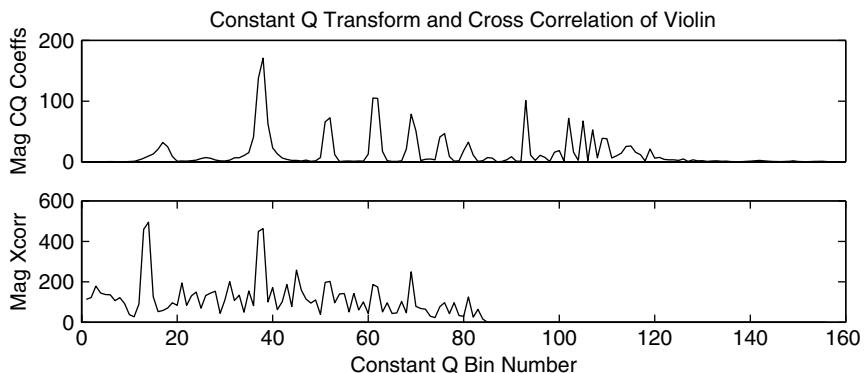


FIGURE 2.9. Constant-Q magnitude spectrum vs bin number for a 15 ms portion of a C₄ note produced by a violin (above) and the result of cross-correlation (below) of this spectrum with the function shown in Fig. 2.1.

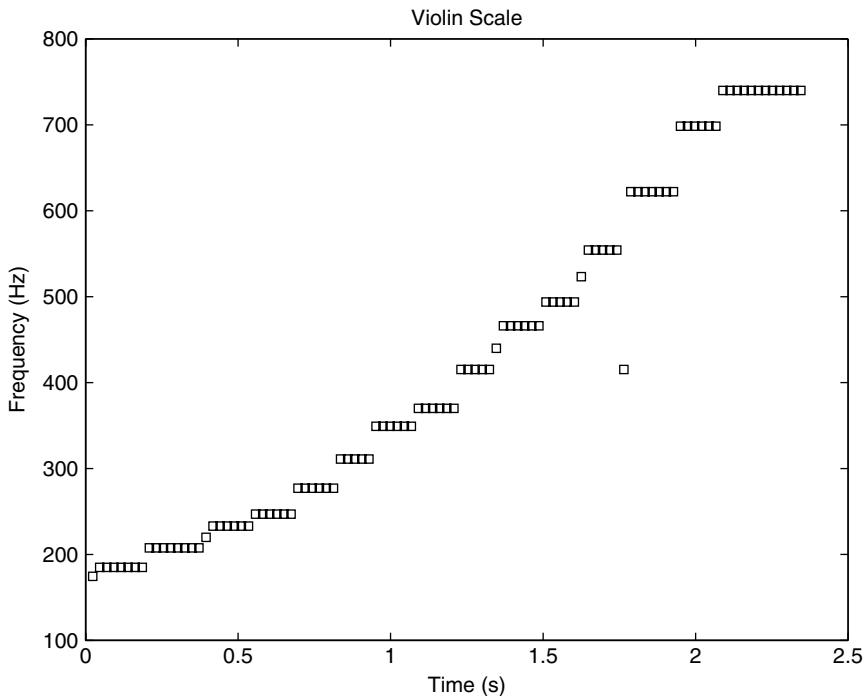


FIGURE 2.10. Pitch tracking results for a violin G major scale from G_3 to G_5 using cross-correlation with a pattern consisting of seven harmonics.

off the appropriate horizontal line. It can be seen that few errors occur, and these occur at note transitions, when more than one tone may be present.

The cross-correlation pattern recognition method has produced excellent pitch-tracking results for a variety of musical sounds whose spectra varied from a simple spectrum consisting of a strong fundamental with a few higher harmonics to an extremely complex spectrum where the fundamental was often weak and most energy was concentrated in higher harmonics over 2000 Hz. This success indicates that the algorithm has an ability to deal with a wide variety of musical sounds.

4 High-Resolution Frequency Calculation Based on Phase Differences

4.1 Introduction

The method of frequency determination described in the previous section [see also Brown (1992)] works extremely well for instruments playing discrete notes belonging to the equal-tempered scale. In that case, the smallest difference

between frequencies is approximately 6%, and the results can be reported as notes of the equal-tempered scale. However, a very different situation can arise in passages played by stringed or wind instruments. Unlike keyboard instruments, these instruments are not constrained to play discrete frequencies. Thus, frequency can vary continuously as in, for example, a glissando or vibrato. (See Figs. 2.3 and 2.7.) Moreover, even keyboard instruments can be tuned to temperaments other than equal-tempered. For all of these cases, in order to track the fundamental frequency accurately, the frequency determination must be much more accurate than a half-semitone or 3%.

The frequency of a particular Fourier component as obtained from the bin into which it falls in the magnitude spectrum is only as accurate as the resolution or frequency difference between bins, in our case 3%. This estimate can be improved by using the maximum of a quadratic fit to a maximum-amplitude bin and its two adjacent bins to estimate the amplitude and frequency of the underlying sinusoid (Smith and Serra, 1987). Even more accurate is a method we have developed which approximates frequency in terms of the phase change of a Fourier component. In our case, this is the component that our frequency tracker has selected as the correct fundamental frequency.

It has been long known that frequency can be determined much more accurately from phase change than by interpolation of the magnitude spectrum (Flanagan and Golden, 1966). However, there is a problem with determining the frequency from the phase difference over a reasonable hop size (samples between frames). This problem, called phase unwrapping, is caused by the fact that the phase change is only known modulo 2π . However, the problem does not arise with a hop size of one sample, because the highest digital frequency is π radians/sample. The only drawback is that this method requires the computation of an additional FFT.

Appendix B describes a method of obtaining an extremely precise value for the frequency of a particular bin. In our case, it is chosen to correspond to the fundamental frequency of the sound analyzed based on the phase difference corresponding to a hop of one sample. This is done *without* the calculation of an additional FFT by using an approximation based on periodicity. With this method we can accurately follow continuous frequency changes with low computational cost.

4.2 Results Using the High-Resolution Frequency Tracker

Precise frequency determination as described in Appendix B increases the total computation time by a negligible amount, because it is only carried out for three of the constant-Q bins used in the calculation. Once this bin is selected, a calculation is made to determine the corresponding bin number for the FFT. The real and imaginary parts of the FFT for this bin and those on either side of it were previously calculated, and only these three complex numbers are needed for the evaluation of the transform.

The power of this method is apparent when it is applied to the acoustic sounds for which it is intended. The circles shown in Figs. 2.11 and 2.12 indicate frequencies

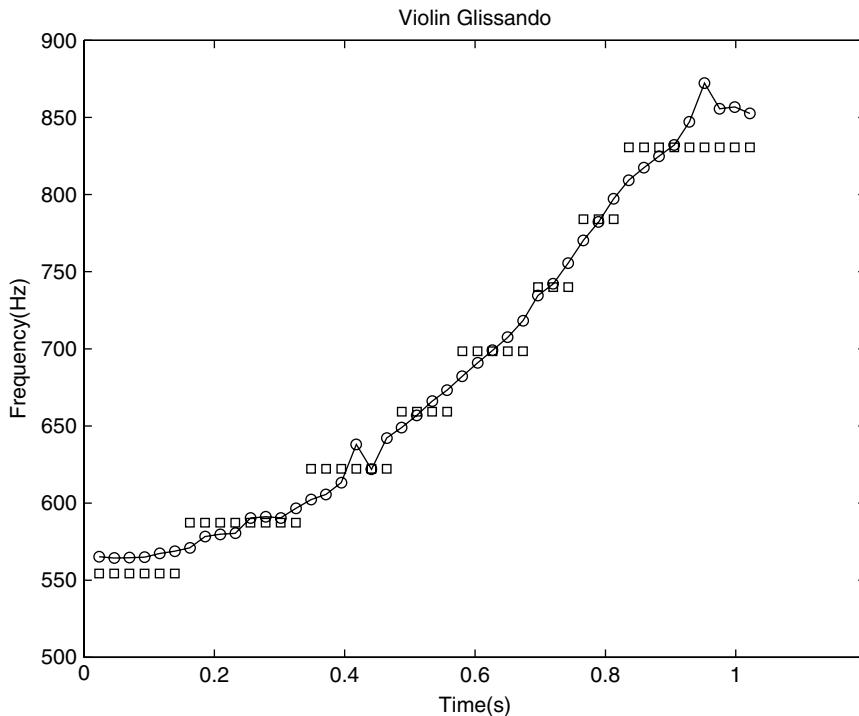


FIGURE 2.11. High-resolution frequency plotted against time for a violin executing a glissando. Squares represent the results of the fundamental-frequency tracker, and circles give the high-resolution results.

refined using the method of Appendix B, based on the frequency tracker output as described in Section 2, which is shown by the square symbols. The spectra of these sounds are shown in Fig. 2.3 (violin glissando) and Fig. 2.7 (violin vibrato).

5 Applications of the High-Resolution Pitch Tracker

Two of the applications that our high-resolution method has made possible are the measurement of the frequency ratios of musical sounds and a perception experiment with natural acoustic (as opposed to synthetic) sounds.

5.1 Frequency Ratios of Spectral Components of Musical Sounds

A knowledge of the exact ratios of the frequencies of the partials of sounds produced by musical instruments is important for an understanding of the underlying physics for producing these sounds. Also important is the application to the production of

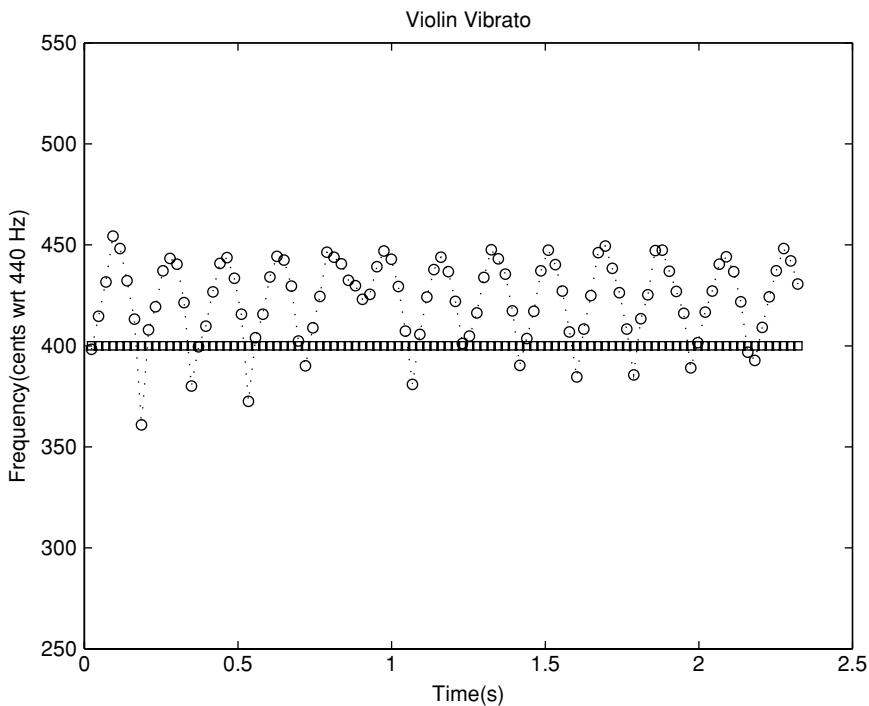


FIGURE 2.12. High-resolution frequency plotted against time for a violin executing vibrato. Symbols have the same meanings as in Fig. 2.11.

synthetic sounds that may be used in musical compositions for computers (Fletcher et al., 1962, 1965; Fletcher and Sanders, 1967).

5.1.1 Background

Early work by Fletcher et al., (1965) reported that the frequencies of the partials in steady tones produced by members of the string family “were found to be harmonic—that is, integral multiples of the fundamental frequency.” Almost a decade later, Beauchamp (1974) used “relative phase curves” to show that violin tone harmonics are locked together in the steady-state portions of open string tones but not during attack and decay transients or during vibrato. Soon after, in his textbook on musical acoustics, Benade (1976) stated that there is a wide class of instruments whose frequency ratios are related by precisely whole numbers, and these are the instruments producing sustained sounds. More recently Ando and Yamaguchi (1993) measured the statistical fluctuations of the note C₅ produced by a number of instruments both with and without vibrato. They found that, for a given sound, the standard deviations of the frequencies of all of its harmonics are nearly equal and conjecture that the reason for this is that they vary synchronously with the fundamental. However, Schumacher (1992) stated that sounds produced

by stringed instruments are aperiodic with the origin lying in the fundamental mechanisms of sound production, such as bow hair inhomogeneity for the bowed instruments.

Thus, while the existence of harmonic ratios in instrument tones had been both asserted or disclaimed widely in the past, there was little systematic effort to actually measure the frequencies of a variety of instruments purported to produce sounds with harmonics in integer or near-integer ratios and report them along with an assessment of the accuracy of the measurements. It was therefore of great interest for this author to measure the fluctuations in frequency for the fundamental and at the same time determine experimentally whether the higher harmonics exhibit identical fluctuations at integer ratios for various sustained-tone instruments.

5.1.2 Calculation

The single-frame approximation (Appendix B) described for determination of a precise value for the fundamental frequency can be equally well applied to determine the frequency of a component in any other FFT bin. In this calculation, the first step is a calculation of the “high-resolution” fundamental frequency. This frequency is then converted to a fractional FFT bin, and the original FFT is then tested for maxima at integer multiples of the fundamental. If a maximum is found, the frequency of that component is determined using the single-frame approximation. If no maximum is found, the two adjacent bins on either side are checked for maxima, and, if one was found, the frequency of that bin is recorded. If no maximum is found, a large negative frequency value is returned that goes off scale in the graphs.

Frequency measurements were made with a Hanning window of 25–100 ms, depending on the frequency range of the instrument, and a time advance or hop size of about 6 ms. Roughly 175 frequency measurements per second were made for each harmonic. Frequencies of the harmonics were then plotted in cents after first dividing each harmonic by its harmonic number. Thus, if all curves coincide, exact integer ratios must obtain to within 0.1%, which is the visual resolution of the curves. Results are presented graphically rather than in a table of averages with standard deviations because important information on frequency fluctuations is preserved in the graphs, which would be lost by taking numerical averages.

Calculations were carried out on digitized sounds produced by a clarinet, alto flute, voice, piano, violin, viola, and cello. The sounds produced by the stringed instruments included examples played pizzicato and bowed both with and without vibrato.

5.1.3 Results

Measured ratios were exactly equal to integers for all instruments except for the piano and string instruments played pizzicato. Anomalous behavior was observed in some regions for the fundamental frequency for vibrato sounds played by stringed instruments with the frequency deviation exceeding the extrema of the other harmonics divided by their harmonic number by about 1% on average.

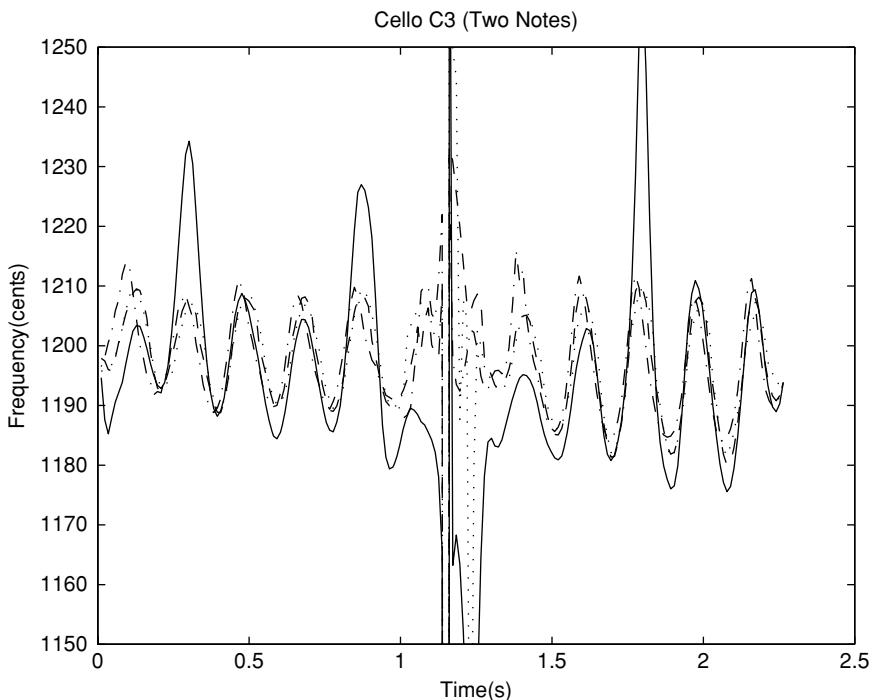


FIGURE 2.13. Measured frequency (in cents with respect to $C_2 = 65.41$ Hz) of the fundamental and harmonics 2–5 normalized by their harmonic numbers and plotted against time, for a cello executing the note C_3 with vibrato. The fundamental deviates from the normalized harmonics mainly at the extrema.

Graphical results for the cello are presented in Figs. 2.13 and 2.14. Figure 2.13 is rather typical of the results with vibrato where the fundamental is higher than the frequency of the other harmonics divided by their harmonic number at the position of some of the frequency maxima. For complete results see Brown (1996).

5.1.3.1 Cello

Vibrato: The graph of Fig. 2.13 shows harmonic frequency detection results from two successive C_3 notes performed with vibrato on a cello. The first note is a C_3 at the top of an ascending scale, immediately followed by a C_3 at the beginning of a descending scale. The region of the bow change between notes (occurring at $t = 1.2$ s) was very noisy as are the frequency measurements in this region.

Note that with the exception of the fundamental, all of the harmonics are in exact integer ratios, within the accuracy of the measurement method. The frequencies for this vibrato note show the same behavior seen for vibrato executed by the other stringed instruments in that the excursions of the fundamental exceed those of the other components. Here, in addition, there is highly anomalous behavior

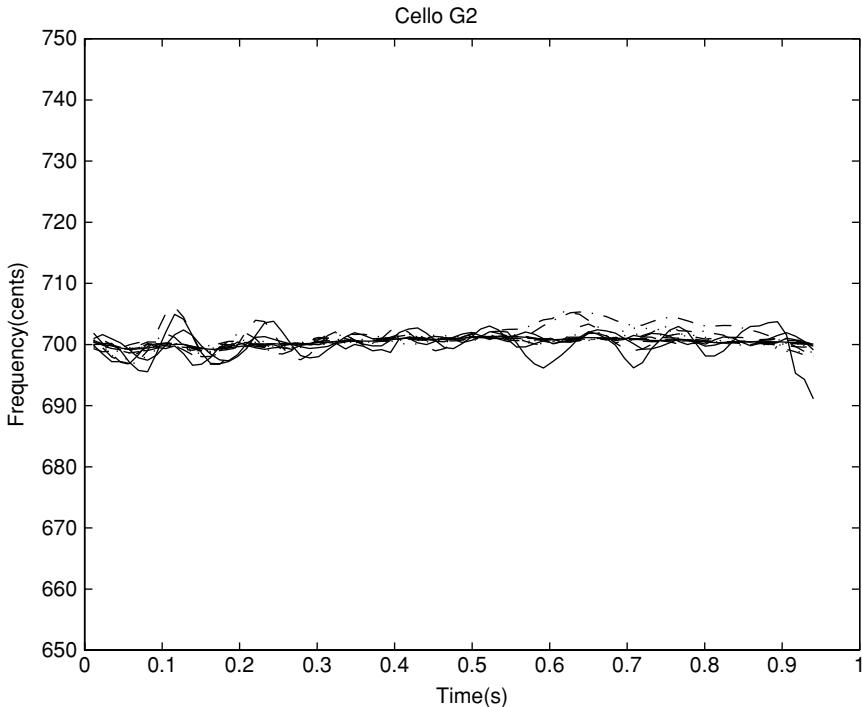


FIGURE 2.14. Measured frequency (in cents with respect to $C_2 = 65.41$ Hz) of the fundamental and harmonics 2 to 15 normalized by harmonic number and plotted against time, for a cello playing the note G_2 without vibrato.

at the position of the amplitude minima of the fundamental. [For corresponding amplitude plots see Brown (1996)]. At these points, the frequency of the fundamental can rise to as much as 40 cents or more above the frequencies of the other harmonics divided by their harmonic numbers.

The sounds with vibrato played on stringed instruments were the most difficult for the fundamental-frequency tracker for several reasons. First, the frequency is constantly changing due to the frequency modulation, and only an average can be measured due to the finite number of samples in the FFT window. Second, it is the motion of the performer's finger on the string which is causing this change in effective length, and there will always be unwanted fluctuations in bow pressure because humans are not mechanically perfect. Third, there may be more bow noise due to the varying conditions.

Open String: The result shown in Fig. 2.14 is a real tour-de-force. This sound was recorded at MIT with Yo-Yo Ma playing a G_2 , an open string tone with minimal frequency fluctuations. There are no errors in the frequency determinations out of over 1800 values. Measurements indicating exact harmonicity within ± 3 cents were possible up to the 25th harmonic.

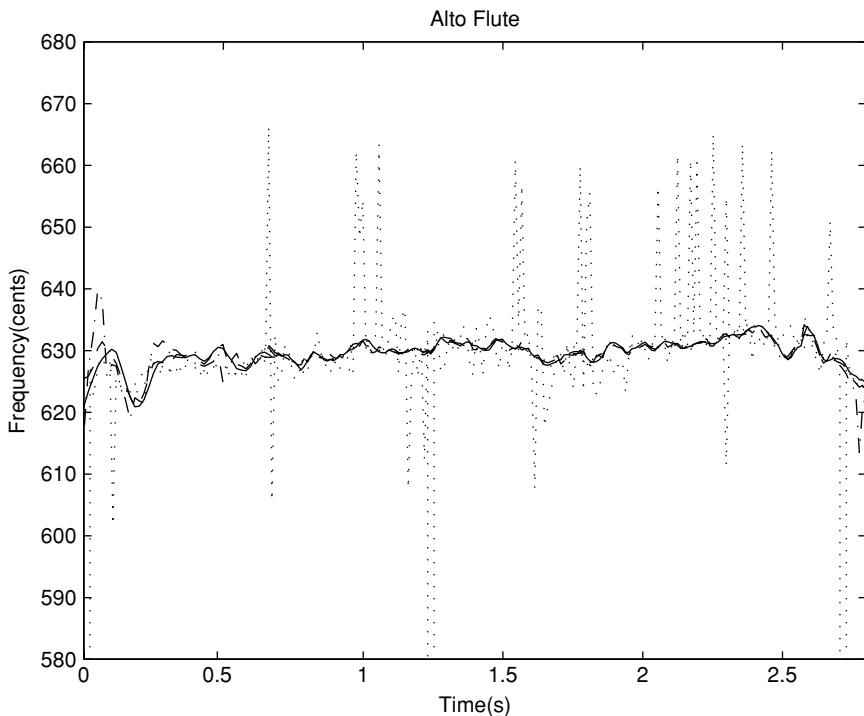


FIGURE 2.15. Measured frequency (in cents with respect to $A_4 = 440$ Hz) of the fundamental and harmonics 2 through 5 normalized by harmonic number and plotted against time, for an alto flute playing the note $D_5^{\#}$ without vibrato. The fourth and fifth harmonics are represented by dotted lines.

5.1.3.2 Alto Flute

The alto flute is a member of the woodwind family of instruments. A sound was played by this instrument without vibrato, and Fig. 2.15 clearly shows the small fluctuations that are characteristic of any musical note generated by a human performer. Exact integer ratios are made evident by the coincidence of solid curves in Fig. 2.15 for the first three harmonics. A number of strong deviations appear in harmonics 4 and 5 (dotted curves), especially in the 4th harmonic, which has the lowest amplitude, but these are probably due to breath noise in the sound. These results indicate near-perfect harmonicity and are a consequence of the flute's periodic tone production mechanism, often called an "air-reed."

5.1.4 Discussion

Continuously driven instruments such as bowed strings, winds, and the voice have phase-locked frequency components whose frequencies can be expressed as ratios of integers to within the currently achievable measurement accuracy of about 0.2%.

Because frequency fluctuations greater than the measurement accuracy are inherent in any sound produced by a human performer, improvement of the measurements is unnecessary. In fact, when we compare these results with measurements on synthetic sounds, where deviations from perfect harmonicity are an order of magnitude or more smaller, we see that frequency fluctuations are the limiting factor in this study rather than the accuracy of the frequency tracker.

On the other hand, sounds of impulsively driven instruments, such as the piano and pizzicato strings, have partials that deviate from integer ratios. These deviations (i.e., inharmonicities) are predicted by vibration theory and can easily be confirmed by measurement. In these cases, a brief excitation is followed by an independent decay for each component. The mechanism causing frequency deviations is the stiffness of the strings, and measurements were shown in detail to be in agreement with vibration theory by Fletcher (1964), where piano inharmonicity was found to be proportional to partial number squared.

It is generally believed that machine perception is inferior to that of the human perceptual system. In the case of pitch perception, a human perceives a sound with a complex spectrum as having a single pitch corresponding to the frequency of the fundamental. We have demonstrated that a computer can do this as well, and in addition is capable of extracting frequencies of the higher harmonics with as high precision as that of the fundamental.

5.2 *Perceived Pitch Center of Bowed String Instrument Vibrato Tones*

In this application, the precise measurement of fundamental frequency is not the final goal but an essential first step in determining the properties of the input signal, which is then used in a perception experiment.

The determination of the pitch center of frequency-modulated sounds has been the focus of a number of previous studies. The sources have usually been pure tones or synthetic complex sounds with well-defined spectral compositions. These synthetic sounds differ in temporal and spectral properties from sounds produced by musical instruments, and it is the latter acoustic sounds that performers are trained to produce and perceive in order to make intonation choices. Thus, sound samples played by a virtuoso violist were recorded specifically for this study and analyzed using our high-resolution-frequency method.

5.2.1 Background

The problem of determining the pitch center or the perceived pitch of frequency-modulated sounds has been studied over a long period of time by a number of researchers. The problem is of interest to psychoacousticians for giving them insight into the mechanisms of pitch perception. An understanding is also necessary for the study of intonation choices by string performers because most of their notes are played with vibrato. In fact, for a meaningful study of intonation, the following questions must be answered:

What pitch is perceived by experienced musical performers and listeners when a musical sound with vibrato is presented?

How do the accuracy and standard deviation of the responses of these experts compare for modulated and unmodulated sounds?

Results of three different experiments in recent decades (Iwamiya et al., 1983; Shonle and Horan, 1980; Sundberg, 1978) indicate that there is some question as to whether the mean pitch corresponds to the geometric or arithmetic mean frequency of a vibrato tone. All three experiments were conducted using the method of adjustment, which has certain problems in its rational underpinnings (Hake and Rodwan, 1966).

A study with a musical emphasis is a more meaningful way to address the question of the just noticeable difference (JND) of natural sounds because it is on these sounds that experts are trained. The study described here differs from the previous ones in that it was conducted with actual musical sounds. All subjects had experience performing on musical instruments whose tunings are continually adjusted during performance (in contrast to playing keyboard instruments, where the performer is not responsible for intonation). The two-interval/two-alternative forced choice experimental method was used, which has a distinct advantage over the method of adjustment.

Results are reported for two groups of subjects, divided according to their musical experience. The first group consisted of non-professional performers from the MIT Media Lab (MIT group). The second group consisted of advanced string players, graduate students studying violin at New England Conservatory (NEC), and a professional violinist from the Boston area (NEC group).

5.2.2 Experimental Method

5.2.2.1 Sound Production and Manipulation

All of the sounds used in this study were recorded at MIT with the professional violist (MT) playing a number of notes both with and without vibrato on a viola. They were analyzed using the high resolution fundamental frequency tracker described in Section 4. For the notes without vibrato, sound segments were chosen that had frequencies constant within a standard deviation of 2 cents or less.

5.2.2.2 Listening Experiments

Each listener was presented with the notes D₄, C₅[#], A₅, and G₆ with or without vibrato followed by the same note without vibrato at 10 pitch levels in a randomized order. The pitch levels in cents consisted of -15, -9, -6, -3, 0, +3, +6, +9, +15, and +21 relative to standard equal-tempered pitch. The mean (in cents) of the vibrato note corresponded to the 0 level of the non-vibrato note. The subject's task was to respond whether the second note was higher or lower in pitch than the first. Randomly mixed with these trials were an equal number of similar trials where the first note was replaced by a non-vibrato note. The purpose of these trials was to see how well listeners can distinguish between pitches without vibrato in

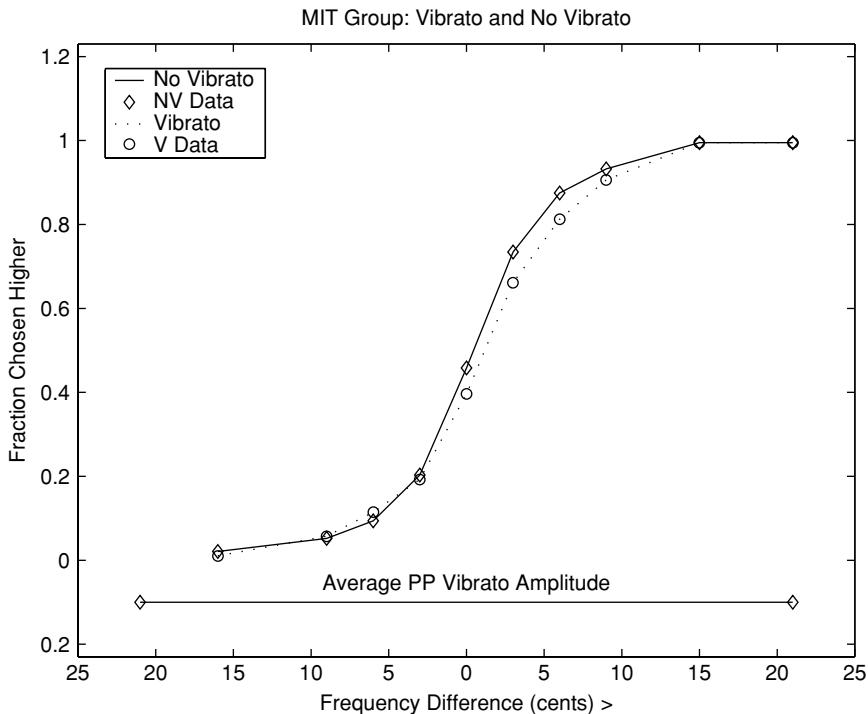


FIGURE 2.16. Fraction of subject responses (Experiment 1) that the target (non-vibrato) tone has a “higher” pitch than the vibrato tone plotted against target pitch level where 0 cents corresponds to the mean of the vibrato. The diamond points (solid curve) represent responses to control stimuli pairs both with no vibrato, whereas the circle points (dotted curve) represent the vibrato/non-vibrato case. The average peak-to-peak vibrato amplitude (frequency deviation) is included below the curves for comparison.

order to ascertain whether comparisons to the vibrato trials are meaningful. In all, each trial occurred eight times for a total of 640 trials for each subject.

5.2.3 Results

5.2.3.1 Experiment 1: Nonprofessional-Performer Listeners

The average psychometric curve for this group of listeners is presented in Fig. 2.16, where “fraction of ‘higher’ responses” are plotted vs the “pitch level” of the second (non-vibrato) sound whose pitch is the “target pitch.”

Note that if the frequency of the fixed target (second) sound is in fact higher than the mean frequency of the vibrato (first) sound (positive half of the abscissa) and all eight subjects’ responses were “higher,” this would correspond to a perfect score with an ordinate of 1. Similarly if the frequency of the target sound were lower (negative half of the abscissa), all subjects responding “lower” would correspond

to an ordinate 0. The abscissa value corresponding to an ordinate of 0.5 (meaning 50% responses higher) represents the pitch level judged to be the same as that of the vibrato sound. Recall that with our notation, a vibrato sound is labeled by its mean frequency. So an ordinate of 0.5 occurring at a value of 0 signifies that the pitch center of the vibrato was at its mean for that listener. For the non-vibrato curve, identical sounds were compared for the pitch level 0 position, and any deviation from ordinate 0.5 was statistical or indicated a bias on the part of the listener.

The similarity of the vibrato and non-vibrato pitch judgment curves in Fig. 2.16 is very striking. The two curves are almost identical. This is all the more impressive in view of the fact that the actual peak-to-peak vibrato deviation ranges up to twice the total extent of the pitch level axis. (For comparison, the average peak-to-peak vibrato deviation of the vibrato is indicated on the figure.) Yet the similarity of the curves implies that the average frequency of the vibrato sounds are perceived in exactly the same manner as the fixed-frequency sounds. There is not even a difference between the shapes of the judgment curves, which would have indicated more uncertainty in identifying the pitch of the sound with vibrato.

5.2.3.2 Experiment 2: Graduate-Level and Professional Violinist Listeners

This second group was chosen to determine whether string players perceive vibrato produced by stringed instruments in the same way that other musicians do. The average psychometric curve for this group is not pictured, but along with the first group of listeners, the data supported the conclusion that the pitch center of vibrato is at its mean. This psychometric curve is steeper around pitch level 0 than that of Fig. 2.16 indicating that these subjects are a little better at pitch discrimination than the first group. Alternatively, this could be due to the fact that the stimuli were produced by a stringed instrument, and these listeners had far more experience judging intonation of string sounds than those of the first group.

5.2.3.3 Experiment 3: Determination of JND for Pitch

Although the principal goal of this study was to determine the pitch center of vibrato musical tones, the simultaneous control experiment comparing frequency-modulated with unmodulated sounds provided an estimate of the JND for pitch for these subjects. This was estimated from the 76% correct point on the psychometric curve for the non-vibrato case resulting in JNDs of 2.8 cents for the MIT group and 2.5 cents for the NEC group with an upper bound on the error of ± 1 cent. The error was estimated from differences in the values at 24% and 76%, which represent the same sounds heard in reverse order. Therefore, the difference in the pitch center judgments by the moderately trained and highly trained groups seems inconsequential.

These JND values are slightly lower than values of 3.5–4 cents previously found for pure tones, summarized by Moore (1989), as would be expected for complex sounds (Spiegel and Watson, 1984). They fall within the range 1.7–7.5 cents reported by Spiegel and Watson (1984) for musicians discriminating square-wave stimuli, although they are smaller than their average values of 4.5 and 5.0

TABLE 2.3. Summary of Frequency Tracking Results on Viola Sounds^a

Note	Freq (cents)	ETD	Vnv Diff	Vib PP	Note	Freq (cents)	ETD	Vnv Diff	Vib PP
A ₄ open	-2				D ₄ open	-702			
A ₄ stdp	8	8			D ₄ stdp	-705	-5		
A ₄ vib	3	3	-5		D ₄ vib	-716	-16	-11	30
A ₄ stdp (2)	5.5				D ₄ stdp (2)	-706			
G ₃	-1400	0			G ₃ (2)	-1414	-14		
G ₃ vib	-1414	-14	-14	34	G ₃ vib (2)	-1422	-22	-8	38
G ₄	-200	0			G ₄ (2)	-208	-8		
G ₄ vib	-200	0			G ₄ vib (2)	-208	-8	0	
G ₅	998	-2			G ₅ (2)	998	-2		
G ₅ vib	998	-2	0	76	G ₅ vib (2)	998	-12	-10	36
G ₆	2200	0			G ₆ (2)	2213	13		
G ₆ vib	2208	8	8	37	G ₆ vib (2)	2214	14	1	43
E ₃ ^b	-1816	-16			E ₃ ^b (2)	-1815	-15		
E ₃ ^b vib	-1836	-36	-20	43	E ₃ ^b vib(2)	-1827	-27	-12	33
E ₄ ^b	-618	-18			E ₄ ^b (2)	-614	-14		
E ₄ ^b vib	-617	-17	1	100	E ₄ ^b vib (2)	-611	-11	3	80
E ₅ ^b	592	-8			E ₅ ^b (2)	593	-7		
E ₅ ^b vib	588	-12	-4	65	E ₅ ^b vib (2)	583	-17	-10	57
E ₆ ^b	1783	-17			E ₆ ^b (2)	1790	-10		
E ₆ ^b vib	1794	-6	11	40	E ₆ ^b vib (2)	1798	-2	8	40

Abbreviations: ETD = Difference from equal temperament; Vnv Diff = Vibrato–non-vibrato; Vib PP = peak-to-peak amplitude of the vibrato

^aFrequencies are given in cents relative to 440 Hz (A₄).

cents for frequencies 430–910 Hz. Moore and Glasberg (1990) report a JND of roughly 3 cents for complex tones containing the first six harmonics.

Although these results are in agreement with previous studies, it should be recalled that our experiments involve stimuli with a small frequency variation inherent in the use of actual musical sounds. In fact, these JNDs are only slightly greater than the standard deviations of the sounds being compared.

It is interesting to compare the JND to the control of a performer in repeating notes with the same frequency. The average of standard deviations of notes in Table 2.3 with respect to the same note (unmodulated) is 4.2 ± 3.9 cents. Thus, we can speculate that limits on intonation control are due in part to limits of motor control as well as pitch perception. There is also an inherent uncertainty of about 2 cents in the frequency produced by a bowed instrument due to the bowing mechanism (inhomogeneity of the bow hair, etc.) (McIntyre, Schumacher, and Woodhouse, 1981; McIntyre and Woodhouse, 1978).

The data reported here support the hypothesis that the pitch corresponding to the mean frequency of a frequency-modulated sound is the one which best matches that of an unmodulated sound. Furthermore, this modulated sound gives rise to nearly the same psychometric curve as that which results when a fixed-frequency sound is substituted for it. That is, for purposes of comparisons with a second sound, the vibrato tone is equivalent to a fixed-frequency sound having its mean

frequency. Equivalently, it can be stated that a human functioning as a frequency meter performs identically on an unmodulated sound and the mean of a frequency-modulated sound.

6 Summary and Conclusions

Calculation of a constant-Q transform (CQT) for musical analysis is described in Section 2 and Appendix A and compared to the discrete Fourier transform in detail. The resolution and flexibility in choice of analysis frequencies of the CQT make it advantageous as a tool for studying musical signals. Several graphical examples of violin sounds show clearly its advantage in visualizing timbral features as well. In addition to being accurate, it is esthetically attractive for its similarity to one of the prominent theories of human pitch perception. Another attractive feature is its versatility in choice of template: Different numbers of harmonics with varying amplitudes make it applicable for tracking any musical instrument.

The constant-Q transform lends itself to an elegant, as well as accurate, method of pitch tracking using pattern matching with an ideal template. As described in Section 3, the pattern-recognition pitch-tracking method for musical passages has been found to be accurate to the nearest quarter tone. Then a high-resolution-frequency determination method, described in Section 4 and Appendix B, based on the phase difference of adjacent frames, can be used as a post-processor where high precision is desired. Applications range from analysis of sounds with continuous frequency variation to the determination of temperament for performance studies in cognitive psychology.

Further applications of the high-resolution pitch tracker are described in Section 5. Exploitation of the accuracy of a phase-based method makes possible the calculation of an extremely accurate value of the frequency chosen by the template method. This accuracy makes it possible to carry out valuable experiments in many fields. Two examples are described: First, the direct determination of the harmonicity of higher harmonics for continuously excited and impulsively excited instruments. Second, a pitch perception experiment using acoustic sounds, rather than synthesized ones, to determine the pitch center of frequency-modulated sounds and to compare their JND with that of unmodulated sounds. Many future applications are possible based on these methods.

Appendix A: An Efficient Algorithm for the Calculation of a Constant-Q Transform

The calculation is based on a form of Parseval's equation (Oppenheim and Schafer, 1975), which states that for any two discrete functions of time $x[n]$ and $y[n]$:

$$\sum_{n=0}^{N-1} x[n]y^*[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k]Y^*[k], \quad (2.12)$$

where $X[k]$ and $Y[k]$ are the discrete Fourier transforms of $x[n]$ and $y[n]$, and $Y^*[k]$ and $y^*[n]$ are the complex conjugates of $y[n]$ and $Y[k]$ respectively.

Equation (2.9) can be rewritten as

$$X^{cq}[k_{cq}] = \sum_{n=0}^{N[k_{cq}]-1} w[n, k_{cq}] x[n] e^{-j\omega_{k_{cq}} n}, \quad (2.13)$$

where $X^{cq}[k_{cq}]$ is the k_{cq} component of the constant-Q transform. As before, the exponential has the effect of a filter for center frequency $\omega_{k_{cq}}$.

Using Eq. (2.12), Eq. (2.13) can be evaluated as follows: First, let

$$w[n, k_{cq}] e^{-j\omega_{k_{cq}} n} = \kappa^*[n, k_{cq}]. \quad (2.14)$$

Then, Eq. (2.12) can be written as

$$X^{cq}[k_{cq}] = \sum_{n=0}^{N-1} x[n] \kappa^*[n, k_{cq}] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] K^*[k, k_{cq}], \quad (2.15)$$

where $X^{cq}[k_{cq}]$ is the k_{cq} th constant-Q coefficient and $K[k, k_{cq}]$ is the discrete Fourier transform of $\kappa[n, k_{cq}]$, i.e.,

$$K[k, k_{cq}] = \sum_{n=0}^{N-1} \kappa[n, k_{cq}] e^{-j2\pi kn/N} = \sum_{n=0}^{N-1} w[n, k_{cq}] e^{j\omega_{k_{cq}} n} e^{-j2\pi kn/N}. \quad (2.16)$$

We will refer to $\{K[k, k_{cq}]\}$ in the frequency domain as the set of spectral kernels of the transformation and to the $\{\kappa[n, k_{cq}]\}$ as the set of temporal kernels. We have used a Hamming window as discussed in Section 2.

The kernels can be evaluated initially and do not contribute further to computation time. Furthermore, their values are close to zero outside a limited range and can be dropped, leading to only a few multiplications for each constant-Q coefficient $X^{cq}[k_{cq}]$. For further details see Brown and Puckette (1992).

Appendix B: Single-Frame Approximation—Calculation of Phase Change for a Hop Size of One Sample

If we assume that the signal $x[n]$ is periodic, the phase change for a hop size of one sample can be obtained from the following identity (Charpentier, 1986; Oppenheim and Schafer, 1975). If $T\{x[n]\} = X[k]$ is the k th component of the discrete Fourier transform of $x[n]$, then

$$T\{x[n+m]\} \cong e^{j2\pi km/N} X[k] \quad (2.17)$$

is the DFT after m samples.

The above equation applies to an unwindowed DFT. It is possible to use this result to obtain a hanning-windowed transform, because the effect of windowing can be calculated in the frequency domain for this window. We will use the notation

$X^H[k, n_o]$ to denote the hanning-windowed Fourier transform evaluated for a window beginning on sample n_o , that is,

$$X^H[k, n_o] = \sum_{n=0}^{N-1} x[n + n_o]w[n]e^{-j2\pi kn/N}, \quad (2.18a)$$

where

$$w[n] = 1/2 - 1/2 \cos(2\pi n/N) = 1/2[1 - (1/2)e^{j2\pi n/N} - (1/2)e^{-j2\pi n/N}]. \quad (2.18b)$$

Substituting this expression for the window into the preceding equation leads to

$$X^H[k, n_o] = (1/2)\{X[k] - (1/2)X[k+1] - (1/2)X[k-1]\}. \quad (2.19)$$

Substituting Eq. (2.17) with $m = 1$ into Eq. (2.19), the approximation for the hanning-windowed DFT after one sample is

$$\begin{aligned} X^H[k, n_o + 1] &= (1/2)\{e^{j2\pi k/N}X[k] - (1/2)e^{j2\pi(k+1)/N}X[k+1] \\ &\quad - (1/2)e^{j2\pi(k-1)/N}X[k-1]\} \end{aligned} \quad (2.20a)$$

$$\begin{aligned} &= (1/2)e^{j2\pi k/N}\{X[k] - (1/2)e^{j2\pi/N}X[k+1] \\ &\quad - (1/2)e^{-j2\pi/N}X[k-1]\}. \end{aligned} \quad (2.20b)$$

The digital frequency in radians per sample for the k th bin corresponding to the phase difference for a time advance of one sample is

$$\omega(k, n_o) = \text{mod}[\phi(k, n_o + 1) - \phi(k, n_o); 0, 2\pi], \quad (2.21a)$$

where

$$\phi(k, n_o + 1) = \text{atan}\{\text{Im}(X^H[k, n_o + 1])/\text{Re}(X^H[k, n_o + 1])\} \quad (2.21b)$$

and

$$\phi(k, n_o) = \text{atan}\{\text{Im}(X^H[k, n_o])/\text{Re}(X^H[k, n_o])\}. \quad (2.21c)$$

This expression for the phase difference holds for any DFT bin with the bin indicated by k . For use with a fundamental-frequency tracker, the calculation would only be used on the bin selected as winner by the tracker. This method is referred to as the single-frame approximation (SFA). Note that the frequency in Hz is given by

$$f(k, n_o) = [\omega(k, n_o)/(2\pi)]f_s, \quad (2.22)$$

where f_s is the sample frequency.

References

- Amuedo, J. (1985). "Periodicity estimation by hypothesis-directed search," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-85), Tampa, FL (IEEE, New York), pp. 395–398.
- Ando, S., and Yamaguchi, K. (1993). "Statistical study of spectral parameters in musical instrument tones," *J. Acoust. Soc. Am.* **94**(1), 37–45.
- Beauchamp, J. W. (1974). "Time-variant spectra of violin tones," *J. Acoust. Soc. Am.* **56**(3), 995–1004.
- Benade, A. H. (1976). *Fundamentals of Musical Acoustics* (Oxford University Press, New York).
- Boulanger, R. C. (2000). *The Csound Book: Perspectives in Software Synthesis, Sound Design, Signal Processing, and Programming* (MIT Press, Cambridge, MA).
- Braccini, C., and Oppenheim, A. V. (1974). "Unequal bandwidth spectral analysis using digital frequency warping," *IEEE Trans. on Acoustics, Speech, and Signal Processing ASSP-22*, 236–244.
- Brown, J. C. (1991). "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Am.* **89**(1), 425–434.
- Brown, J. C., and Zhang, B. (1991). "Musical frequency tracking using the methods of conventional and 'narrowed' autocorrelation," *J. Acoust. Soc. Am.* **89**(5), 2346–2354.
- Brown, J. C. (1992). "Musical fundamental frequency tracking using a pattern recognition method," *J. Acoust. Soc. Am.* **92**(3), 1394–1402.
- Brown, J. C., and Puckette, M. S. (1992). "An efficient algorithm for the calculation of a constant Q transform," *J. Acoust. Soc. Am.* **92**(5), 2698–2701.
- Brown, J. C. (1996). "Frequency ratios of spectral components of musical sounds," *J. Acoust. Soc. Am.* **99**(2), 1210–1218.
- Chafe, C., and Jaffe, D. (1986). "Source separation and note identification in polyphonic music," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-86), Tokyo (IEEE, New York), pp. 1289–1292.
- Chafe, C., Jaffe, D., Kashima, K., Mont-Reynaud, B., and Smith, J. (1985). "Techniques for note identification in polyphonic music," *Proc. 1985 Int. Computer Music Conf.*, Burnaby, B. C., Canada (Computer Music Assoc., San Francisco), pp. 399–405.
- Charpentier, F. J. (1986). "Pitch detection using the short-term phase spectrum," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-86), Tokyo (IEEE, New York), pp. 113–116.
- Duda, R. O., and Hart, P. E. (1973). *Pattern Classification and Scene Analysis* (John Wiley & Sons, NY).
- Duifhuis, H., Willemse, L. F., and Sluyter, R. J. (1982). "Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception," *J. Acoust. Soc. Am.* **71**, 1568–1580.
- Flanagan, J. L., and Golden, R. M. (1966). "Phase vocoder," *Bell Syst. Tech. J.* **45**, 1493–1509.
- Fletcher, H. (1964). "Normal vibration frequencies of a stiff piano string," *J. Acoust. Soc. Am.* **36**, 203–209.
- Fletcher, H., Blackham, E. D., and Geertsen, O. N. (1965). "Quality of violin, viola, cello and bass-viol tones," *J. Acoust. Soc. Am.* **37**, 851–863.
- Fletcher, H., Blackham, E. D., and Stratton, R. (1962). "Quality of piano tones," *J. Acoust. Soc. Am.* **34**, 749–761.

- Fletcher, H., and Sanders, L. C. (1967). "Quality of violin vibrato tones," *J. Acoust. Soc. Am.* **41**, 1534–1544.
- Gambardella, G. (1971). "A contribution to the theory of short-time spectral analysis with nonuniform bandwidth filters," *IEEE Transactions on Circuit Theory* **18**, 455–460.
- Gambardella, G. (1979). "The Mellin transforms and constant-Q spectral analysis," *J. Acoust. Soc. Am.* **66**, 913–915.
- Gerson, A., and Goldstein, J. L. (1978). "Evidence for a general template in central optimal processing for pitch of complex tones," *J. Acoust. Soc. Am.* **63**, 498–510.
- Hake, H. W., and Rodwan, A. S. (1966). "Perception and recognition," in *Experimental Methods and Instrumentation in Psychology* edited by J. B. Sidowski (McGraw Hill, New York), pp. 331–381.
- Harris, F. J. (1976). "High-resolution spectral analysis with arbitrary spectral centers and arbitrary spectral resolutions," *Computer Electr. Eng.* **3**, 171–191.
- Harris, F. J. (1978). "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE* **66**(1), 51–83.
- Helms, H. D. (1976). "Power spectra obtained from exponentially increasing spacings of sampling positions and frequencies," *IEEE Trans. on Acoustics, Speech, and Signal Processing* **ASSP-24**, 63–71.
- Higgins, R. J. (1976). "Fast Fourier transform: An introduction with some minicomputer experiments," *Am. J. Physics* **44**(8), 766–773.
- Iwamiya, S., Kosugi, K., and Kitamura, O. (1983). "Perceived principal pitch of vibrato tones," *J. Acoust. Soc. Japan (E)* **4**(2), 73–82.
- Kashima, K. L., and Mont-Reynaud, B. (1985). "The bounded-Q approach to time-varying spectral analysis," *Stanford Department of Music Technical Report*, STAN-M-23.
- Kronland-Martinet, R. (1988). "The wavelet transform for analysis, synthesis, and processing of speech and music sounds," *Computer Music J.* **12**(4), 11–20.
- Maher, R. C., and Beauchamp, J. W. (1994). "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *J. Acoust. Soc. Am.* **95**(4), 2254–2263.
- McIntyre, M. E., and Woodhouse, J. (1978). "The acoustics of stringed musical instruments," *Interdisciplinary Science Reviews* **3**(2), 157–173.
- McIntyre, M. E., Schumacher, R. T., and Woodhouse, J. (1981). "Aperiodicity in bowed-string motion," *Acustica* **49**, 13–32.
- Moore, B. C. J. (1989). *An Introduction to the Psychology of Hearing* (Academic Press, San Diego).
- Moore, B. C. J., and Glasberg, B. R. (1990). "Frequency discrimination of complex tones with overlapping and non-overlapping harmonics," *J. Acoust. Soc. Am.* **87**, 2163–2177.
- Nawob, S. H., Quatieri, T. F., and Lim, J. S. (1983). "Signal reconstruction from short-term Fourier transform magnitude," *IEEE Transactions on Acoustics, Speech and Signal Processing* **ASSP-31**, 986–998.
- Oppenheim, A., Johnson, D., and Steiglitz, K. (1971). "Computation of spectra with unequal resolution using the fast Fourier transform," *Proc. IEEE* **59**, 299–301.
- Oppenheim, A. V., and Schafer, R. W. (1975). *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ).
- Petersen, T. L. (1980). "Acoustic Signal Processing in the Context of a Perceptual Model," Doctoral dissertation, University of Utah. *Dissertation Abstracts International* **41/05**, 1831.
- Piszczalski, M., and Galler, B.A. (1979). "Predicting musical pitch from component frequency ratios," *J. Acoust. Soc. Am.* **66**(3), 710–720.

- Scheffers, M. T. M. (1983). "Simulation of auditory analysis of pitch: An elaboration on the DWS pitch meter," *J. Acoust. Soc. Am.* **74**, 1716–1725.
- Schroeder, M. R., and Atal, B. S. (1962). "Generalized short-time power spectra and auto-correlation functions," *J. Acoust. Soc. Am.* **34**, 1679–168.
- Schroeder, M. R. (1968). "Period histogram and product spectrum: New methods for fundamental-frequency measurements," *J. Acoust. Soc. Am.* **43**, 829–834.
- Schumacher, R. T. (1992). "Analysis of aperiodicities in nearly periodic waveforms," *J. Acoust. Soc. Am.* **91**, 438–451.
- Seneff, S. (1985). "Pitch and Spectral Analysis of Speech Based on an Auditory Synchrony Model," Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.
- Shonle, J. I., and Horan, K. E. (1980). "The pitch of vibrato tones," *J. Acoust. Soc. Am.* **67**, 246–252.
- Smith, J. O., and Serra, X. (1987). "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," *Proc. 1987 Int. Computer Music Conf.*, Urbana, IL (Computer Music Association, San Francisco), pp. 290–297. Also available as Stanford Dept. of Music Technical Report STAN-M-43.
- Spiegel, M. F., and Watson, C. S. (1984). "Performance on frequency-discrimination tasks by musicians and nonmusicians," *J. Acoust. Soc. Am.* **76**, 1690–1695.
- Stautner, J. P. (1983). "Analysis and Synthesis of Music Using the Auditory Transform," unpublished masters thesis, Massachusetts Institute of Technology, Cambridge, MA.
- Sundberg, J. (1978). "Effects of the vibrato and the singing formant on pitch," *Musicologica Slovaca* **6**, 51–69.
- Terhardt, E. (1979). "Calculating virtual pitch," *Hearing Research* **1**, 155–182.
- Terhardt, E., Stoll, G., and Seewann, M. (1982). "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *J. Acoust. Soc. Am.* **71**, 679–688.
- Vercoe, B. (1986). *Csound: A Manual for the Audio Processing System and Supporting Programs* (MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA).
- Youngberg, J. E., and Boll, S. F. (1978). "Constant-Q signal analysis and synthesis," *Record of the 1978 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Tulsa, OK (IEEE, New York), pp. 375–378.

Beyond Traditional Sampling Synthesis: Real-Time Timbre Morphing Using Additive Synthesis

LIPPOLD HAKEN, KELLY FITZ, AND PAUL CHRISTENSEN

1 Introduction

Because of its theoretical advantage for making timbral manipulations, sine wave additive synthesis is an attractive alternative to sampling synthesis, which is currently the most popular method for real-time synthesizers. Nevertheless, until recently performers have seldom used additive synthesis because of the practical difficulty of accomplishing these timbral manipulations, which inherently require modification of large numbers of time-varying amplitude and frequency control functions.

While sampling synthesis is easy to use, it suffers from limited nuance. First of all, a sampling synthesizer employs a limited set of source recordings to synthesize each acoustic instrument. When a performer plays a note on a synthesizer, the synthesizer attempts to select and play an appropriate recording, i.e., one that is closest to the intended pitch and dynamic. However, a note that does not correspond to an available source recording must be synthesized by playing a similarly pitched recording at a modified amplitude and sample rate. Even if the synthesizer has a large set of source recordings in its memory, its sound is generally easily distinguishable from that of acoustical instruments. This shortcoming is mainly due to an inability to produce all the spectral variations associated with the dynamic and pitch changes of acoustical instruments. Simply varying the amplitude and the sample rate of a recording, as is done in sampling synthesis, does not capture these changes. The basic problem is that because sampling synthesis operates strictly in the time domain, it is incapable of intelligently interpolating between stored sounds to produce sounds that are not in the source set.

Additive sine wave synthesis, on the other hand, allows independent fine control of the amplitude- and frequency-vs-time characteristics of each partial in a sound. This makes it convenient to implement a wide variety of modifications such as frequency shifting, time stretching, cross synthesis, and timbre morphing. While in the past additive synthesis has been too slow or too expensive for real-time applications, this method is now easily within the capability of current digital (DSP) technology.

In this chapter we present several topics related to the operation of a real-time additive synthesizer we have recently developed. Like traditional sampling synthesizers, our instrument utilizes a set of source recordings for synthesis. These recordings are manipulated in real-time to control synthesis timbre.

In order to implement efficient real-time spectral manipulations, we have developed a stream-based representation of partial envelopes. Envelope parameter streams are our counterpart to the sample streams used in sampling synthesis. They provide amplitude, frequency, and noise information for each partial. Noise envelopes represent noise associated with each partial and constitute an important extension to traditional additive sine wave synthesis. In addition, time envelopes are used to achieve time dilation, i.e., warping of envelope data with respect to time.

Our spectral analysis software extracts amplitude, frequency, and noise envelopes for each partial. This method uses spectral reassignment to improve time and frequency accuracy for the partials.

We use a simple and intuitive interface for the performer. The performer navigates the timbres of the source recordings using a timbre control space, where the dimensions correspond to pitch, loudness, and timbre. Notes that do not correspond to available source recordings are synthesized by combining timbral aspects from recordings that are “nearby” in terms of pitch, loudness, and timbre. We implement morphing synthesis using movements in the timbre control space, which produces continual timbre changes in response to these movements.

Additive synthesis has great promise for performers. As an example of new possibilities, the Continuum Fingerboard, a polyphonic performance instrument that borrows from both the traditional piano keyboard and the fretless fingerboard, has continuous control parameters that are especially suited for additive synthesis. It can also be used to control any MIDI real-time synthesizer.

2 Additive Synthesis Model

Many synthesis systems allow the sound designer to operate on streams of samples. In our real-time implementation we also work with streams of data, but the data are not time-domain samples. Rather, the streams contain parameters for each partial component in additive synthesis.

Much of the strength of systems that operate on sample streams is derived from the uniformity of the data. This data homogeneity gives the sound designer great flexibility with a few general-purpose processing elements. In our encoding of additive parameter streams, data homogeneity is also of prime importance. We have avoided the use of separate models to represent noise and transients. Although hybrid additive models are a proven success (Serra and Smith, 1990), we have developed a single homogeneous model that is well-suited to stream-based processing.

Our streams encode envelope parameters for each partial. The envelope parameters for all the partials in a sound are encoded sequentially. Typically, a stream has

a frame size of 128, which means the parameters for each partial are updated every 128 samples, or every 2.9 ms at a 44.1 kHz sampling rate. Sample streams generally do not have frame sizes associated with them, but this concept is necessary in our additive stream implementation.

Envelope parameter streams are typically generated by traversing files containing data from non-real-time analyses of source recordings. The parameter streams may also be generated by real-time analysis, or by real-time algorithms, but that processing is beyond the scope of this discussion. A parameter stream will typically pass through several processing elements. These processing elements can combine multiple streams in a variety of ways, and they can modify values within a stream. The stream finally reaches a synthesis element that produces a sample stream at its output based on the envelope parameter stream at its input.

The synthesis element implements the sum

$$y(t) = \sum_{k=1}^K (A_k(t) + N_k(t)b(t)) \sin(\theta_k(t)), \quad (3.1a)$$

where

$$\theta_k(t) = \theta_k(t - 1) + 2^{F_k(t)}, \quad t > 0 \quad (3.1b)$$

and where

- y is the time domain waveform of the synthesized sound;
- t is the sample number;
- k is the sinusoid partial number in the sound;
- K is the total number of partials in the sound (usually 20–160);
- A_k is partial k 's amplitude envelope;
- N_k is partial k 's noise envelope;
- b is a zero-mean noise factor variable having a low-pass spectrum (e.g., white noise through a 4-pole low-pass IIR filter with 1 kHz cutoff);
- F_k is partial k 's log frequency envelope;
- $\theta_k(t)$ is the running phase of partial k , which depends on the partial's frequency envelope;
- $\theta_k(0)$ is the initial phase, which is specified.

2.1 Real-Time Synthesis

We have implemented real-time synthesis in Symbolic Sound's Kyma sound design environment (Scaletti, 1987; Hebel and Scaletti, 1994). Together with Symbolic Sound Corporation, we developed Kyma Sound Objects that generate, process, and synthesize envelope parameter streams (Haken, 1995). While it is possible to use processing elements originally designed for sample streams with our envelope parameter streams, the additive synthesis method described in this chapter was implemented on symbolic Sound's Capybara synthesizer module.

2.2 Envelope Parameter Streams

As mentioned above, values for envelopes $A_k(t)$, $N_k(t)$, and $F_k(t)$ are updated from the parameter stream every 2.9 ms. The synthesis element performs linear interpolation between updates, so that A_k , N_k , and F_k are piecewise linear envelopes with 2.9 ms linear segments (Haken, 1992). The sinusoidal portion of the synthesis is implemented using conventional oscillator table lookup with linear interpolation between frames.

2.3 Noise Envelopes

The noise envelope N_k is an important extension to our original additive sine wave model (Fitz and Haken, 1995). Rather than use a separate model to represent noise in our sounds, we define this third envelope to retain a homogeneous data stream. There are several advantages of this representation over the purely sinusoidal representation, which requires many short partials to represent noisy parts of a sound. We simplify the representation of noisy parts of the sound, and, more importantly, we obtain an intuitive parameter for timbre manipulation. Quasiharmonic sounds, even those with noisy attacks, have just one partial per harmonic in our representation.

Noise envelopes allow a sound designer to manipulate noise-like components of sound in an intuitive way, using a familiar set of controls. The control parameters for each partial are amplitude, (center) frequency, and relative noise. These can be used to manipulate and transform both sinusoidal and noise-like components of a sound.

3 Additive Sound Analysis

3.1 Sinusoidal Analysis

With our method, analysis of the sinusoid parts of sounds follows the well-known frequency-tracking algorithm invented by MacAulay and Quatieri (1986) and Smith and Serra (1987) and further developed by Maher (1989) and Fitz and Haken (1995, 2002). In this method spectral peaks are retained and tracked from frame-to-frame to form partial tracks. However, with our method, time and frequency values are enhanced by the method of spectral reassignment, which is discussed below.

3.2 Noise-Enhanced Sinusoidal Analysis

Purely sinusoidal analysis techniques such as McAulay and Quatieri's and our first implementation represent noise as many short partials with widely varying frequencies and amplitudes. These short partials are capable of producing good quality syntheses of the noisy parts of many sounds, but this approach has shortcomings: When noisy sounds are stretched in time, the partials representing the noise are also stretched and can be heard as rapidly modulated sine waves. Noisy

sounds analyzed and stretched in this way may be described as “wormy.” In addition, the noisy character of a sound is carried mostly in the phase contributions from these many short partials. Because time- or frequency-scale modifications inevitably change the phase portraits of the partials, such operations tend to destroy subtle properties of the noise and result in unacceptable quality. Moreover, the representation of noise as a collection of short partials is intuitively unsatisfactory because it provides no means for manipulating useful parameters of the noise, such as noise energy and center frequency, and no means for separating the noise into distinct components.

Serra and Smith (1990) proposed a method for separating a noise component from a sinusoidal representation. Their algorithm performs a sinusoidal analysis and resynthesis of the signal and then computes the spectral difference between the original signal and the resynthesized signal that is inverted to produce a difference signal called the “residual.” The residual may be stored and used in future resyntheses, or its short-time spectra may be stored and synthesis performed using inverse spectral analysis (stochastic modeling). This method yields a very high fidelity synthesis, but Serra and Smith’s noise representation is problematic for our purposes.

The Smith–Serra and other stochastic methods of accommodating noise, including those which represent noise energy in fixed frequency bands, do not provide homogeneous representations of sinusoidal and noise components. With our envelope-parameter streams the noise components of a sound are combined with the same data stream as the deterministic components and are manipulated by introducing a noise envelope for each partial.

In our current analysis program, Loris (Fitz et al., 2000), we divide the short-time frequency spectrum into overlapping regions in order to associate noise energy with nearby sinusoidal components. Each region contains strong-magnitude peaks, selected according to the McAulay–Quatieri (1986) process. However, the total spectral energy of the weak-magnitude peaks in each region is represented as noise energy associated with the strong-magnitude peaks. Thus, a bandwidth-enhanced component results, corresponding to each strong-magnitude peak. The frequency of each component is found using the spectral reassignment method described below. The amplitude of each component is equivalent to that of a sinusoid containing both the strong-magnitude peak’s energy and a proportion of the region’s spectral energy contained in weak-magnitude peaks. The noise factor of each component specifies the ratio of noise energy (from the weak-magnitude peaks) to sinusoidal energy (from the strong-magnitude peaks).

While this method of energy association is not analytically rigorous—i. e., the noise energy associated with each partial is only an approximation of the spectral energy in the partial’s frequency region—the approximation is reasonable for signals having mostly sinusoidal energy, and it preserves both the brightness and the total energy of the overall spectrum.

As mentioned above, we represent quasiharmonic sounds with one partial per harmonic, as shown in Fig. 3.1. The use of noise envelopes in our analysis allows this, even for noisy parts of the sound.

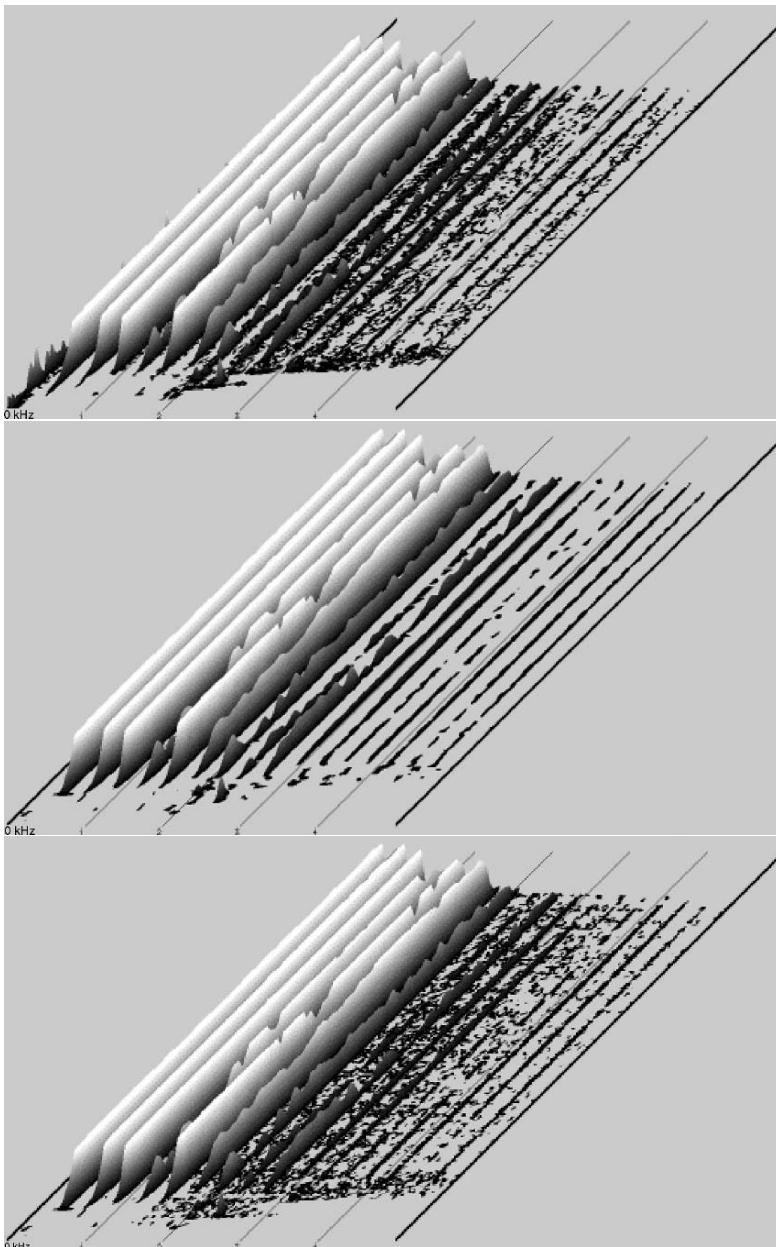


FIGURE 3.1. Spectrograms of an original flute recording (top), sine-only synthesis (N_k set to 0) (middle), and complete synthesis using Eq. (3.1) (bottom). The low-frequency rumble present in the original flute recording was omitted from the synthesis. The horizontal axis is frequency in kHz, the vertical axis is relative amplitude, and the front-to-back axis is approximately 2.3 s of time. Strong low-frequency components are clipped and appear to have unnaturally flat amplitudes due to the high gain used to make low-amplitude high-frequency partials visible. These plots were made using SoundMaker by Alberto Ricci. [From Fitz and Haken (2002), Figs. 13, 14, and 15].

For nonharmonic sounds, any peak that is “left over” (not part of a partial) will be a contributor to the noise of nearby partials. In Loris, one of the analysis parameters controls the maximum density of partials in frequency in order to limit the amount of data obtained using this mechanism. If many peaks occur at a similar time and are too close in frequency, some of them will not become partials but will be contributors to the noise of nearby partials.

3.3 Spectral Reassignment

The analysis and representation of transients is a well-known problem for additive synthesis. The onset of a sound, in particular, is perceptually important (Berger, 1964; Saldanha and Corso, 1964) and is difficult to analyze with sufficient time accuracy. The time-domain shape of an attack deteriorates because the window used in the analysis of the sound cannot be perfectly time-localized. This problem occurs even if, at each window, the analysis guarantees phase-correctness of each partial. Verma et al. (1997) have developed a transient analysis method that may be used together with a deterministic sine model and a stochastic noise model (for related work, also see Chapter 4 by Levine and Smith in this book). In our work, however, we use envelope-parameter streams, which allow us to control amplitude, frequency, and noise envelopes for each partial. We avoid a separate transient model in our implementation by taking a different approach to improving transients within the few parameters of our model. We produce improved envelope-parameter streams by incorporating spectral reassignment into our analysis method.

Spectral reassignment in time and frequency has been used for sharpening blurred speech spectrograms (Auger and Flandrin, 1995; Plante et al., 1998). Each point of the spectrogram is moved to a new point that represents the distribution of the energy in the time-frequency window more accurately. We apply spectral reassignment in our analysis to sharpen attack transients that would otherwise be blurred due to the length of our analysis window.

3.3.1 Time Reassignment

Our analysis first performs a sequence of short-time Fourier transforms. Traditionally, the result of each Fourier transform is assigned to the center of the window in time and frequency. However, this approach has a limitation, as shown in Fig. 3.2. In the case where a window is positioned such that samples in the left-hand portion of the window occur just before the beginning of a sound’s attack, whereas samples in the right-hand portion correspond to the beginning of the attack, traditional analysis methods do not explicitly detect this situation.

As a solution to this problem, the center of gravity of each bin in the transform may be computed. For our example this would show that the sound is only present in the right part of the window, because the center of gravity would occur after the window’s midpoint. We can use this information to resynthesize sharper attacks, thereby avoiding “mushy” or blurry attacks that heretofore have plagued most sinusoid-based analysis systems.

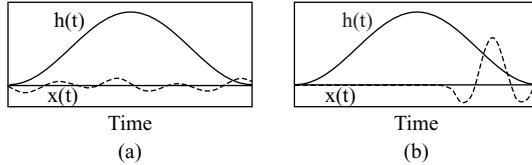


FIGURE 3.2. Two waveforms $x(t)$ (dashed lines) that have the same magnitude spectrum. The sustained quasiperiodic waveform on the left (a) yields time corrections near zero, while the strong transient on the right (b) yields components needing large time corrections (positive in this case because the transient is near the right tail of the window). $h(t)$ is the window function used for analysis.

A reassignment formula for time based on Auger and Flandrin [1995, Eq. (3.26a)] and Plante et al. [1998, Eq. (3.4)] is

$$t_r(x; t, f) = t - \Re e \left\{ \frac{\int_{-\infty}^{\infty} (t - \tau)x(\tau)h(t - \tau)e^{-j2\pi f \tau} d\tau}{\int_{-\infty}^{\infty} x(\tau)h(t - \tau)e^{-j2i\pi f \tau} d\tau} \right\}, \quad (3.2)$$

where $x(t)$ is the signal waveform function and $h(t)$ is the window function used for analysis. (The use of τ in the equation as the dummy integration variable; this has no connection to the use of this symbol for a time-dilation function used later in this chapter.) In actual computer implementations, the integrals would be replaced by finite summations with limits dictated by the window function. t is then the frame time, and the second term on the right-hand side of Eq. (3.2) gives the fraction of the window length needed to correct it. These offset times are different for each frequency bin and can be larger than the frame duration.

The denominator of Eq. (3.2) corresponds to the short-time Fourier transform (STFT) of the signal centered at t , using the following definition:

$$\text{STFT}_h(x; t, f) = \int_{-\infty}^{\infty} x(\tau)h(t - \tau)e^{-j2\pi f \tau} d\tau. \quad (3.3)$$

Note that the numerator of Eq. (3.2) is similar to the denominator, except that the numerator's integrand includes an extra multiplication by $t - \tau$. This time weighting has the effect of emphasizing data in the right (later) part of the STFT window differently than data in the left (earlier) part of the window. The ratio of the numerator to the denominator indicates where within the window the center of gravity of the bin's data is concentrated. The time-corrected value t_r is computed by adding the window's midpoint time to this ratio. Thus, if all the data are contained in the right half of the window, t_r will be greater than t ; and if all the data are in the left part of the window, t_r will be less than t .

If we define a new window function that incorporates time-weighting

$$g(\tau) = \tau h(\tau), \text{ so that } g(t - \tau) = (t - \tau)h(t - \tau), \quad (3.4)$$

then Eq. (3.2) can be rewritten in terms of Fourier transforms to speed up the calculation:

$$t_r(x; t, f) = t - \Re e \left\{ \frac{\text{STFT}_g(x; t, f)}{\text{STFT}_h(x; t, f)} \right\}. \quad (3.5)$$

By using the time correction, we can avoid spreading out attacks over the analysis window length. As we move the window so that it covers more and more of the onset of the sound, the time correction moves from the right side of the window toward the center. Note that because the window size ordinarily exceeds the frame (or hop) size, the time correction can actually exceed the frame size.

3.3.2 Frequency Reassignment

In addition to reassigned time, we can calculate a reassigned frequency. The frequency correction equation is as follows [Plante et al. (1998), Eq. (5)]:

$$f_r(x; t, f) = f + \Re e \left\{ \frac{\int_{-\infty}^{\infty} \xi X(\xi) H(\xi - f) e^{j2\pi\xi t} d\xi}{\int_{-\infty}^{\infty} x(\tau) h(t - \tau) e^{-j2\pi f \tau} d\tau} \right\}, \quad (3.6)$$

where f is the bin center frequency and $X(\xi)$ and $H(\xi)$ are Fourier transforms of $x(t)$ and $h(t)$, respectively. Here, f_r is the frequency center of gravity for each bin. Equation (3.6) is similar to the time reassignment equation [Eq. (3.2)], except that the numerator uses a multiplication by frequency rather than time. It follows that we can define a new window function which takes into account this multiplication by noting that it corresponds to a derivative in the time domain:

$$c(\tau) = \frac{dh(\tau)}{d\tau} \quad (3.7)$$

Auger and Flandrin [1995, Eq. (3.26b)] show that we can rewrite the frequency-correction equation using STFTs:

$$f_r(x; t, f) = f - \Im m \left\{ \frac{\text{STFT}_c(x; t, f)}{\text{STFT}_h(x; t, f)} \right\}. \quad (3.8)$$

An alternative parabolic interpolation method for refining frequency estimates is based on a single FFT (Smith and Serra, 1987). We have yet to do a detailed comparison of the two approaches to refining frequency and are currently using the Auger–Flandrin method.

3.3.3 Spectral-Reassignment Summary

Our time- and frequency-reassigned short-time analysis requires three times as many FFTs as a traditional short-time analysis, assuming the same overlap between

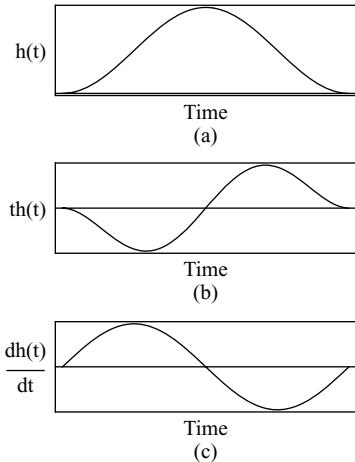


FIGURE 3.3. Window functions used by the three different short-time transforms used to compute reassigned times [Eq. (3.5)] and frequencies [Eq. (3.8)]. Function (a) is the original continuous-time window function $h(t)$, (b) is the time-weighted window function $g(t) = th(t)$, and (c) is the frequency-weighted window function computed in the time domain by $c(t) = dh(t)/dt$.

successive analysis windows. Fig. 3.3 shows the window functions for the three FFTs used to compute a time- and frequency-corrected spectrum for a particular analysis window.

The amplitude, noise, and frequency envelopes for each partial are found by following ridges in the time-frequency surface resulting from successive time- and frequency-reassigned spectra, as shown in Fig. 3.4. Note that the envelope breakpoints are not evenly spaced in time or frequency.

4 Navigating Source Timbres: Timbre Control Space

A real-time additive synthesizer has been implemented that uses a large number of recordings to provide the source material for synthesis. A timbre control space gives the performer a simple and intuitive way to navigate among the available sounds (Haken, 1992).

Pitch, loudness, and timbre are normally defined to be what the listener hears. However, in our case they refer to physical quantities that correspond to what the listener hears in a more-or-less one-to-one fashion.

We define timbre to be the collection of characteristics of a sound, other than its pitch, loudness, and duration, which distinguish it from other sounds. Two sounds with the same fundamental frequency and the same amplitude often have different timbres, even if they are produced by the same instrument. For example, bowing a cello near the fingerboard results in a mellower timbre than bowing near the bridge.

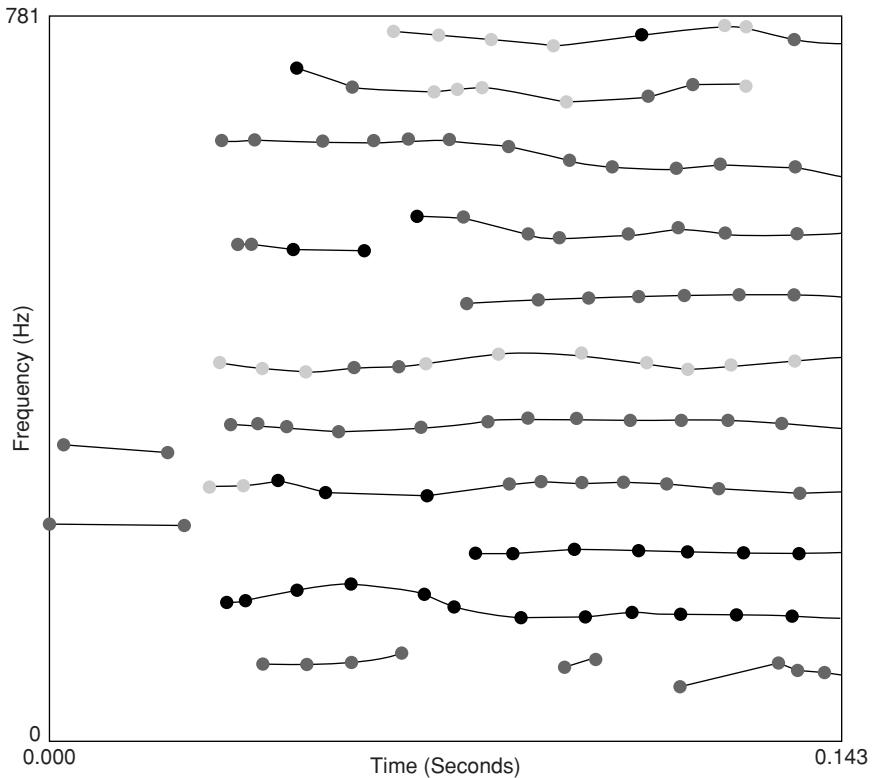


FIGURE 3.4. A portion of a Loris analysis of a low cello tone. Lines indicate ridges present in the time-frequency surface during analysis. Dots indicate time-frequency data points that make up the ridges; all other data points from the time-frequency surface are not shown. Note that the dots are not at regular intervals due to the method of reassignment. Each ridge corresponds to a partial, and is synthesized with a bandwidth-enhanced oscillator as indicated in Eq. (3.1). Darker lines correspond to higher ridges (larger A_k), and lighter lines indicates wider ridges (larger N_k).

For example, a source timbre is a time-varying spectrum derived from a particular instrument tone. A timbral quality refers to a collection of timbres produced by a particular instrument at different pitches and loudnesses. A timbral blend corresponds to a blend between the timbres of two different instruments or possibly the same instrument played two different ways.

On our synthesizer, we generally vary timbre with pitch and loudness, as it does on acoustic instruments. The spectrum of a loud note on a cello, for instance, is not just a scaled version of the spectrum of a quiet cello note. Similarly, the spectrum of a high-pitched cello note is not just a frequency-shifted version of the spectrum of a low-pitched cello note.

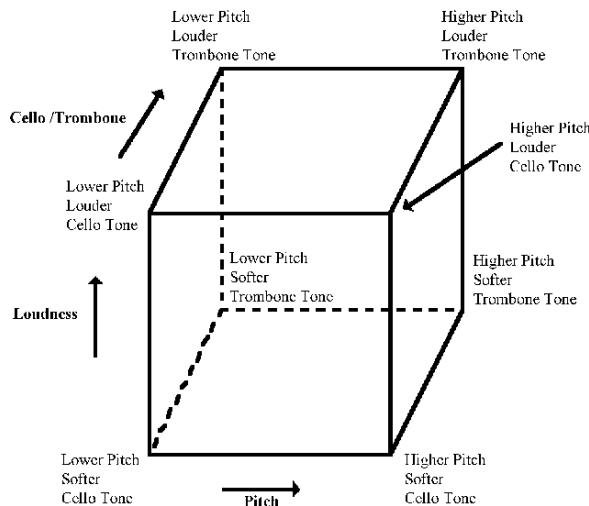


FIGURE 3.5. A timbre control space cube with source timbres derived from four cello and four trombone tones corresponding to the corners of the cube. [From Haken et al. (1998), Fig. 16.]

We define a three-dimensional timbre control space to be a space in which one dimension corresponds to pitch, another to loudness, and the third dimension to timbral quality. During performance, when one moves a point along a pitch axis (where loudness and timbral quality are fixed), the resulting timbres approximate those associated with playing different pitches on the same instrument. The spectrum is not simply shifted in frequency. Similarly, moving a point along the loudness axis (where pitch and timbral quality are fixed and blend is set to one end point) approximates the timbral changes associated with playing at different loudnesses on the same instrument. The spectrum is not simply scaled in amplitude. Moving along the third axis (keeping loudness and pitch fixed) produces timbral changes corresponding to a blend between two source timbres.

It should be noted here that our timbre control space is quite different from a *timbre space* derived from multidimensional (MDS) perception experiments (Grey, 1975; Wessel, 1979; Risset and Wessel, 1982). Our intention is merely to provide an intuitive and practical method for controlling the parameters used to generate each tone rather than to categorize the properties of the resultant timbres. It is quite possible, in fact, for nearby tones in the timbre control space to be located far apart in a perceptual timbre space, although we hope this wouldn't be so.

The performer controls the x , y , and z positions associated with any note played in the timbre control space. We can divide the three-dimensional timbre control space into cubes, with each neighboring cube sharing one face. Figure 3.5 shows one cube of a timbre control space illustrating the use of time-variant spectral analyses of four cello tones and four trombone tones, each performed at two different pitches and two different loudnesses. Each corner of the cube is characterized by a set of amplitude, frequency, and noise envelopes derived from an analyzed tone.

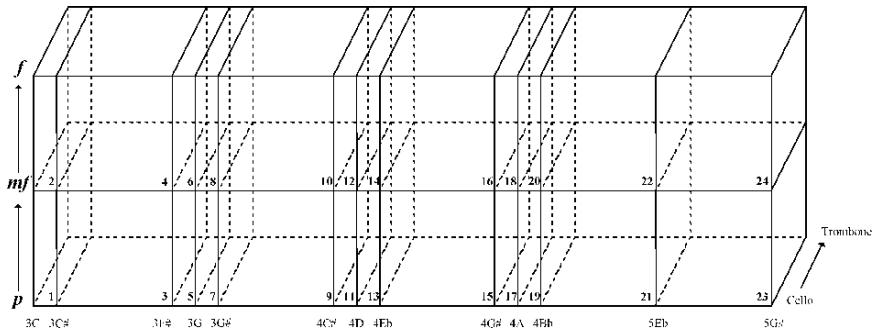


FIGURE 3.6. A timbre control space made up of 24 cubes, based on 78 source timbres (39 trombone tones and 39 cello tones). The cubes have unequal widths because the pitches of the source recordings were not equally spaced. The pitches correspond to the open strings of the cello, a half-step up and down from the open strings, and some higher pitches. [From Haken et al. (1998), Fig. 17.]

These eight sets of envelopes completely define this part of the timbre control space.

The timbre of a note corresponding to a point located within this cube will possess a blend of characteristics of the source timbres at all eight corners of the cube. If the note is exactly at the center of the cube, it should share equally the characteristics of all eight source timbres. The note's location may change over time, corresponding to crescendo, glissando, vibrato, or other performer actions. As the note's location moves toward one face of the cube, the four source timbres of that face should contribute proportionally more to the note's timbre, while the four source timbres of the opposite face contribute proportionally less. If the note's location is exactly at the center of one face of the cube, it should share equally the characteristics of the four source timbres at the corners of that square. In this manner, the timbre control space provides a method for arranging the source timbres into a framework for describing new timbres.

Figure 3.6 shows an example of a complete three-dimensional timbre control space made of 24 cubes. The complete timbre control space is based on the analyses of 78 tones (39 trombone tones and 39 cello tones). When a note is played, the note's x , y , and z location falls within one of the cubes in the timbre control space. For example, if the note's x location corresponds to the pitch $F_4^{\#}$ and its z location corresponds to the dynamic f , it falls into cube 16 in Fig. 3.6. The synthesized sound is created by combining timbral characteristics of the eight preanalyzed source recordings at the eight corners of cube 16.

If the x , y , or z location gradually changes during a note, this corresponds to a gradual change of location, usually within a cube of the timbre control space. If the x , y , and z location changes greatly during a note, the timbre control space location of the sound is likely to travel through the face of one cube into a neighboring cube. In all cases, the timbral changes associated with changing x , y , and z are smooth.

and continuous. This aspect—continuous timbre change—is the motivation for using additive synthesis in place of traditional sampling synthesis.

4.1 Creating a New Timbre Control Space

The simplest timbre control space consists of a single cube based on the analyses of eight recorded sounds. More timbral variation is possible when the timbre control space consists of several cubes. However, in practice, it is difficult to create a cohesive timbre control space because it is difficult to obtain recordings that are perfectly matched in terms of pitch, loudness, duration, and manner of performance, and source timbres of such recordings are needed to define the corners of each cube. For example, it could be desired to have available a set of trombone tone recordings that differ only in pitch. They would be matched in terms of loudness and manner of performance, and the timbre would not change excessively from one note to the next. To some extent, side-by-side listening comparisons together with editing operations (amplitude multiplication, pitch shifting, time compression, expansion, or cutting) can be used to reduce the differences among the recordings (Grey, 1975). Because the perceived similarity between timbres depends on many psychoacoustic effects, this process cannot be completely automatic. Therefore, building a new timbre control space is a time-consuming process. Moreover, each timbre control space defines a very different overall sound and feel for the performer, so, once he or she learns to perform with one timbre control space, learning to play in a new one is not necessarily a trivial undertaking.

4.2 Timbre Control Space with More Control Dimensions

While pitch, loudness, and timbral quality are perhaps the most obvious choices for control dimensions, the choice of the control dimensions in a timbre control space is actually arbitrary. Also, any number of control dimensions can be defined. The meaning of any control dimension depends only on the source recordings that are used and where the source recordings are assigned in the timbre control space.

4.3 Producing Intermediate Timbres: Timbre Morphing

We use a note's position in the timbre control space to determine the timbral blend, or *morph*, between previously analyzed source recordings. Timbre morphing is the process of combining several sounds to create a new sound with intermediate timbre. The process differs from simply mixing sounds, because only a single sound, having some of the characteristics of each original sound, is heard as the morphed sound. Timbre morphing is a topic that has been discussed by many researchers. It is sometimes called “timbre interpolation,” and the resulting sounds are sometimes referred to as “hybrid sounds.” We now list a few previous studies before we present our implementation.

In perhaps the first attempt at timbre morphing, Grey (1975) developed a method for transition between two original sounds to create new intermediate sounds.

Schindler (1984) described sounds as a hierarchical tree of timbre frames, and discussed a morphing algorithm that operates on this representation. Time-varying filters have been used to combine timbres (Peterson, 1975; Depalle and Poirot, 1991). With multiple wavetable synthesis, timbre morphing can be implemented as a straight mix because all harmonics are phase-locked (Horner et al., 1993).

4.4 Weighting Functions for Real-Time Morphing

As described in the previous section, amplitude, frequency, and noise envelopes for each partial of the source recordings are defined at the eight corners of a cube, and weighted averages of these are used to synthesize a sound. This processing is straightforward to implement in real time using envelope parameter stream weighted averages, which are continually computed according to the sound's location within the cube. If we normalize the x , y , and z position within the cube such that $0 \leq X, Y, Z \leq 1$, then interpolation weights for each corner of the cube are defined as

$$W_0 = XYZ, \quad (3.9a)$$

$$W_1 = XY(1 - Z), \quad (3.9b)$$

$$W_2 = X(1 - Y)Z, \quad (3.9c)$$

$$W_3 = X(1 - Y)(1 - Z), \quad (3.9d)$$

$$W_4 = (1 - X)YZ, \quad (3.9e)$$

$$W_5 = (1 - X)Y(1 - Z), \quad (3.9f)$$

$$W_6 = (1 - X)(1 - Y)Z, \quad (3.9g)$$

$$W_7 = 1 - (W_0 + W_1 + W_2 + W_3 + W_4 + W_5 + W_6). \quad (3.9h)$$

Note that W_7 is computed by subtracting the sum of the other weights from 1, in order to avoid roundoff error problems.

These weights implement a rectilinear distance measure to the corners of the cube, not a Cartesian measure. We use the rectilinear measure to avoid discontinuities when the sound's coordinates in the timbre space travel through the face of a cube into an adjacent cube.

4.5 Time Dilation using Time Envelopes

Acoustic instrument sounds vary in their rates of attack. A trumpet sound may have a fast attack while a French horn sound may have a slower attack. If the morphing process simply averaged the envelopes of these two sounds, the new sound would have two averaged attack peaks, rather than a single attack of intermediate speed. To handle this problem, we define a new envelope, called the time envelope, which gives the time rate of change for a partial of a particular source sound. Then, we average these envelopes to produce a single averaged attack rate. We use time-normalized amplitude, frequency, and noise envelopes in this processing.

At synthesis time we compute $\tau(t)$ to index into time-normalized partial envelopes:

$$\tau(t) = \tau(t - 1) + \sum_{q=0}^7 W_q(t) E_q(\tau(t - 1)) \quad (3.10)$$

where

t is the sample number,

$\tau(t)$ is the running index into the time-normalized partial envelopes at sample number t ,

W_q is the weighting for corner (timbre) q of the cube based on current x , y , and z location,

E_q is the time envelope (time dilation function) for corner q of the cube.

These are the steps involved to produce and use time-normalized envelopes:

- (1) All the source recordings in the timbre control space are analyzed with the Loris program, producing a set of analyzed source timbres.
- (2) Corresponding time points are manually specified for each analyzed source timbre. Any number of time points may be used; common ones are start of attack, peak of attack, sustain times, start of release, and end of release.
- (3) At load time, all the source timbres are time-stretched and/or compressed to produce intermediate time-normalized envelopes [$\alpha_{k,q}$, $\phi_{k,q}$, and $\beta_{k,q}$ in Eq. (3.11)] that have time points separated by a normalized amount of time. As part of this process, the time envelope E_q is computed for each analyzed source timbre, to indicate how much time stretch/compression is to be applied in each part of the analyzed source timbre.
- (4) At synthesis time, a weighted average of time envelopes is used to produce the proper final timing according to Eq. (3.10).

Source recordings containing vibrato present further problems. Care must be taken when morphing between differing vibrato rates, to avoid producing an irregular vibrato. Equation (3.10) allows us to produce a regular vibrato because we manipulate time-normalized envelopes in order to perform an important part of vibrato morphing (Tellman et al., 1995). However, for our real-time application, performers often prefer source recordings without vibrato so that they can produce vibrato, if they wish, solely by movements in the timbre control space.

4.6 Morphed Envelopes

The weights from Eq. (3.9) and the running time index from Eq. (3.10) are used to compute the envelope functions A_k , N_k , and F_k of Eq. (3.1) as follows:

$$A_k(t) = \sum_{q=0}^7 W_q(t) \alpha_{k,q}(\tau(t)), \quad (3.11a)$$

$$N_k(t) = \sum_{q=0}^7 W_q(t) \beta_{k,q}(\tau(t)), \quad (3.11b)$$

$$F_k(t) = \sum_{q=0}^7 W_q(t) \phi_{k,q}(\tau(t)), \quad (3.11c)$$

where

- t is the sample number,
- $\tau(t)$ is the running time index into the partial envelopes,
- k is the partial number in the sound,
- $\alpha_{k,q}$ is partial k 's time-normalized amplitude envelope in the source timbre at corner q ,
- $\phi_{k,q}$ is partial k 's time-normalized log-frequency envelope in the source timbre at corner q ,
- $\beta_{k,q}$ is partial k 's time-normalized noise envelope in the source timbre at corner q ,
- W_k is the weighting for corner q of the cube based on the current x, y, z location,
- A_k is the partial's real-time morphed amplitude envelope,
- F_k is the partial's real-time morphed log-frequency envelope,
- N_k is the partial's real-time morphed noise envelope.

4.7 Low-Amplitude Partials

The analysis process cannot accurately determine the frequency of very-low-amplitude partials. Such partials are generally so quiet that they are inaudible in the original sound. A problem occurs, however, when a low-amplitude partial containing inaccurate frequency information is morphed with a high-amplitude partial with accurate frequency information. In this case, the morphed partial will be a medium amplitude partial with audibly inaccurate frequency information.

This problem is avoided by not relying exclusively on the analysis for frequency information. For very quiet partials of quasiharmonic sounds, the frequency of the nearest harmonic (suitably scaled) is used for interpolation. For sufficiently loud partials, the frequency from the analysis is used in interpolation. However, for intermediate-amplitude partials, the frequency used in interpolation is derived from both the analysis frequency and the nearest integer multiple of the time-varying fundamental frequency (Tellman et al., 1995):

$$\phi_k(\tau) = \begin{cases} (1 - \alpha_k(\tau)/\varepsilon) \log_2(f_h(\tau)) + (\alpha_k(\tau)/\varepsilon) \log_2(f_k(\tau)), & \text{if } \alpha_k(\tau) < \varepsilon \\ \log_2(f_k(\tau)), & \text{otherwise} \end{cases} \quad (3.12)$$

where

- τ is the time index into the partial envelopes,
- α_k is the partial number in the sound,
- ϕ_k is the partial's frequency from the analysis,

f_h is the nearest integer multiple of the fundamental frequency,
 ϵ is the amplitude threshold below which analyzed frequencies are unreliable,
 α_k is the partial's amplitude,
 ϕ_k is the partial's log frequency in Eq. (3.11), adjusted for low amplitudes.

As a partial gradually increases in amplitude, more of the frequency used in morphing is taken from the analysis of that component. Consequently, there is no abrupt change between a calculated frequency and an analyzed frequency when a partial reaches the threshold beyond which only the analyzed frequency is used.

5 New Possibilities for the Performer: The Continuum Fingerboard

Real-time additive synthesis can be controlled by a standard MIDI keyboard. If a performer plays our additive sine-wave synthesizer using a MIDI keyboard, the performer's use of aftertouch and pitch bend corresponds to movements in the timbre control space. Aftertouch and pitch bend will result not only in volume and pitch changes, but also in corresponding timbre changes. Pedals or other continuous controllers can correspond to movements in the third dimension of the timbre control space.

As an alternative to a MIDI keyboard, we have developed a new performance device that allows the performer more continuous control than that offered by a MIDI keyboard (Haken et al., 1992, Haken, 1998). The Continuum Fingerboard, shown in Fig. 3.7, resembles a traditional keyboard in that it is approximately the same size and is played with 10 fingers. Like keyboards supporting MIDI's polyphonic aftertouch, it continually measures each finger's pressure. It also resembles a fretless string instrument in that it has no discrete pitches; any pitch may be played, and smooth glissandi are easily produced.



FIGURE 3.7. The Continuum Fingerboard is approximately the same size as a traditional music keyboard, but it has no discrete keys. It has a pitch range of nearly eight octaves. The white markings indicate the white key pattern of a traditional keyboard.

The Continuum Fingerboard tracks an x , y , z position for each finger. The output of the fingerboard can be used to control any synthesis technique. Because of its continuous three-dimensional nature, the output of the fingerboard works well with a three-dimensional timbre control space.

The x (side-to-side) position of each finger provides continuous pitch control for a note. One inch in the x direction corresponds to a pitch range of 160 cents. The performer must place fingers accurately to play in tune and can slide or rock fingers for glissando and vibrato.

The z (pressure) position of each finger provides dynamic control. Tremolo is produced by changing the amount of finger pressure. An experienced performer may simultaneously play a crescendo and decrescendo on different notes.

The y (front-to-back) position of each finger provides timbral control for each note in a chord. By sliding fingers in the y direction while notes are sounding, the performer can create timbral glides.

Depending on the source timbres used in the timbre control space, the y position can produce a variety of effects. One possibility is to select source timbres so that the y position on the Continuum Fingerboard corresponds to the bowing position on a string instrument, where bowing near the fingerboard produces a mellower sound and bowing near the bridge produces a brighter sound. Another possibility is to select source timbres so that the y position morphs between timbres of different acoustic instruments. In either case, the performer can bring out certain notes in a chord not only by playing them more loudly, as on a piano, but also by playing them with a different timbral quality.

5.1 Previous Work

Interest in keyboard instruments with dynamic and timbre control has a long history. Fifteenth century clavichords, although very quiet, provided dynamic control over individual notes as well as a sort of “aftertouch.” By varying the amount of pressure on a key after initially striking it, the performer could produce a vibrato, because, unlike the pianoforte, the clavicord’s plectrum remains in contact with the string after plucking. Eighteenth and nineteenth century pianos were loud enough to fill a recital hall, and they too provided dynamic control over individual notes in a chord. To this day, the piano remains the most popular acoustic keyboard instrument. However, the limitations of the piano action make crescendo and vibrato during sustained notes impossible.

In the twentieth century, analog synthesizers were built with ribbon controllers. These provided one-dimensional continuous control but could not track more than one finger. More recently, electronic keyboards offered key velocity polyphonic aftertouch. These capabilities have been extended by certain experimental keyboards, such as Moog’s Clavier (Moog, 1982), which responds to additional parameters including the exact horizontal and vertical location of the finger on each key. Several other innovative keyboard designs have been developed over the last two decades (Snell, 1983; Johnstone, 1985; Keislär, 1987; Fortuin, 1995). Our work on the Continuum Fingerboard began in the early 1980s, with initial attempts incorpo-

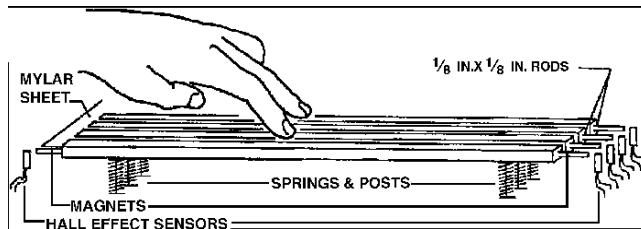


FIGURE 3.8. The Continuum Fingerboard's Hall-effect sensors, which detect the positions of magnets mounted on each rod. [Adapted from Haken et al. (1998), Fig. 7.]

rating photoelastics, conductive rubbers, and capacitive techniques (Haken et al., 1992). These initial attempts were of limited success compared to our more recent approach described below.

5.2 Mechanical Design of the Playing Surface

The Continuum Fingerboard, although roughly the size of a standard piano keyboard, is a continuous playing surface rather than a keyboard. The playing surface is constructed using 256 rods, each 5.75 in. long, mounted on piano-wire springs. The rods are covered by a Mylar and Mylon sheath so that the performer has the impression of a continuous surface rather than discrete keys. A magnet is mounted at both ends of each rod, and the rods are placed between two rows of Hall-effect sensors.

Figure 3.8 shows a portion of the mechanical design of the Continuum Fingerboard's playing surface. The Hall-effect sensors are used to measure the positions of the magnets. When the performer applies finger pressure, the rods under the finger are depressed, and the magnets on those rods move closer to the sensors.

Scanning software running on a controller computer detects finger position by looking for any bar that has normalized pressure values greater than both of its neighboring bars. We call this the center bar, and the neighboring bars the left bar and right bar. The x , y , z coordinates of the finger are calculated from the six sensor values on these three bars as follows: The software tracks the front-to-back position (y position) by summing the normalized sensor values of the back sensors on the left, center, and right bars and then dividing that sum by the sum of the normalized sensor values for all six sensors on these bars. Pitch (x position) and loudness (z position) are estimated using parabolic interpolation. A parabola is assigned to the normalized sensor values of the left, center, and right bars, and the location of the minimum point of this parabola provides the x and z positions for the note.

Figure 3.9 illustrates the use of these parabolas in detecting x variation during vibrato. In this example, the center bar is always the same bar because it is always more depressed than its right and left neighbors. Still, the x position is accurately tracked as the performer rocks the finger back and forth, because the movement of the neighboring bars affects the minimum point of the parabola.

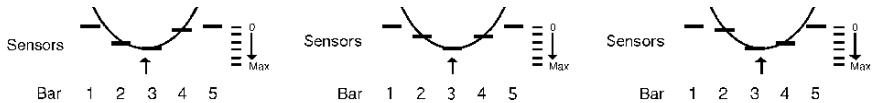


FIGURE 3.9. Continuous pitch tracking using parabolic interpolation: A finger rocking left, centered, and rocking right. [From Haken et al. (1998), Fig. 12.]

6 Final Summary

We have discussed the concepts of timbre control space, envelope parameter streams, noise envelopes, time envelopes, and timbre morphing. We have also described the Continuum Fingerboard, which may be used to control a real-time synthesizer utilizing these concepts possibly in combination with synthesis methods offered by other synthesizers.

We have shown that additive synthesis is a viable alternative to conventional sampling synthesis. Like the sampling synthesizer, our additive system uses a collection of source recordings in synthesis. However, in our case, real-time timbre manipulations of these source timbres are implemented using streams of envelope parameters. The envelope parameters include traditional amplitude and frequency information as well as noise information. This noise information simplifies timbral manipulation of noisy sounds. The performer navigates the source timbres using a timbre control space. We retain much of the generality of sampling synthesis because the timbre control space is completely defined by the source recordings.

The Continuum Fingerboard is a new type of performance device that provides more control over real-time pitch, loudness, and timbre than a traditional MIDI keyboard. However, performing on the Continuum Fingerboard is challenging because performers must rely on audio feedback and manual dexterity to place their fingers for accurate intonation and expression. Like the Theremin, the Continuum Fingerboard requires extensive practice.

A major advantage of our additive approach over traditional sampling is that it allows improved continuous morphs between source timbres. We believe that this approach is not only a practical way to implement real-time additive synthesis, but also one that holds promise for further development.

References

- Auger, F., and Flandrin, P. (1995). "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Trans. Signal Processing* **43**, 1068–1089.
- Berger, K. W. (1964). "Some factors in the recognition of timbre," *J. Acoust. Soc. Am.* **36**, 1888–1891.
- Depalle, P., and Poirot, G. (1991). "SVP: A modular system for analysis, processing and synthesis of sound signals," *Proc. 1991 Int. Computer Music Conf.*, (Int. Computer Music Assoc., San Francisco), pp. 161–164.

- Fitz, K. and Haken, L. (1995). "Bandwidth Enhanced Sinusoidal Modeling in Lemur," *Proc. 1995 Int. Computer Music Conf.* (Int. Computer Music Assoc., San Francisco), pp. 154–156.
- Fitz, K. and Haken, L. (2002). "On the Use of Time-Frequency Reassignment in Additive Modeling," *J. Audio Eng. Soc.* **50** (11), 879–893.
- Fitz, K., Haken, L., and Christensen, P. (2000). "A new algorithm for bandwidth association in bandwidth-enhanced additive sound modeling," *Proc. 2000 Int. Computer Music Conf.* (Int. Computer Music Assoc., San Francisco), pp. 384–387.
- Fortuin, H. (1995). "The clavette: A generalized microtonal MIDI keyboard controller," *Proc. 1995 Int. Computer Music Conf.*, Banff, Canada (Int. Computer Music Assoc., San Francisco), p. 223.
- Grey, J. M. (1975). "An exploration of musical timbre," doctoral dissertation, Stanford University, Stanford, CA. Also available as Dept. of Music Report STAN-M-2, Stanford University, Stanford, CA.
- Haken, L. (1992). "Computational methods for real-time Fourier synthesis," *IEEE Trans. Signal Processing* **40**(9), 2327–2329.
- Haken, L. (1995). "Real-time timbre modifications using sinusoidal parameter streams," *Proc. 1995 Int. Computer Music Conf.* (Int. Computer Music Assoc., San Francisco), pp. 162–163.
- Haken, L., Abdullah, R., and Smart, M. (1992). "The continuum: A continuous music keyboard," *Proc. 1992 Int. Computer Music Conf.*, (Int. Computer Music Assoc., San Francisco), pp. 81–84.
- Haken, L., Tellman, E., and Wolfe, P. (1998). "An indiscrete music keyboard," *Computer Music J.* **22**(1), 30–48.
- Hebel, K. J., and Scaletti, C. (1994). "A framework for the design, development, and delivery of real-time software-based sound synthesis and processing algorithms," *Audio Eng. Soc. Preprint* 3874.
- Horner, A., Beauchamp, J., and Haken, L. (1993). "Methods for multiple wavetable synthesis of musical instrument tones," *J. Audio Eng. Soc.* **41**(5), 336–356.
- Johnstone, E. (1985). "The rolky: A poly-touch controller for electronic music," *Proc 1985 Int. Computer Music Conf.* (Computer Music Assoc., San Francisco), pp. 291–295.
- Keislar, D. (1987). "History and principles of microtonal keyboards," *Computer Music J.* **11**(1), 18–28.
- Maher, R. C. (1989). "An approach for the separation of voices in composite musical signals," doctoral dissertation, Univ. Illinois at Urbana-Champaign, Urbana, IL. *Dissertation Abstracts International-B*, **50/07**, 3074.
- McAulay, R. J., and Quatieri, T. F. (1986). "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech, and Signal Processing* **34** (4), 744–754.
- Moog, R. A. (1982). "A multiply touch-sensitive clavier for computer music systems," *Proc. Int. Computer Music Conf.*, Venice, Italy (Int. Computer Music Assoc., San Francisco), p. 275.
- Peterson, T. L. (1975). "Vocal tract modulation of instrumental sounds by digital filtering," *Proc. Second Annual Music Computation Conf.*, Urbana, IL (Int. Computer Music Assoc., San Francisco), pp. 33–41.
- Plante, F., Meyer, G., and Ainsworth, W. A. (1998). "Improvement of speech spectrogram accuracy by the method of reassignment," *IEEE Trans. Speech and Audio Processing* **6**(3), 282–287.

- Risset, J.-C., and Wessel, D. (1982). "Exploration of timbre by analysis and synthesis," in *The Psychology of Music*, ed. Diana Deutsch (Academic Press, New York), pp. 25–58.
- Saldanha, E. L., and Corso, J. F. (1964). "Timbre cues and the recognition of musical instruments," *J. Acoust. Soc. Am.* **36**, 2021–2026.
- Scaletti, C. (1987). "Kyma: An object-oriented language for music composition," *Proc. Int. Computer Music Conf.*, Urbana, IL (Int. Computer Music Assoc., San Francisco), pp. 49–56.
- Schindler, K. W. (1984). "Dynamic timbre control for real-time digital synthesis," *Computer Music J.* **8**, 28–42.
- Serra, X., and Smith, J. O. (1990). "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music J.* **14**(4), 12–24.
- Smith, J. O., and Serra, X. (1987). "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," *Proc. Int. Computer Music Conf.*, Urbana, IL (Int. Computer Music Assoc., San Francisco), pp. 290–297.
- Snell, J. M. (1983). "Sensors for playing computer music with expression," *Proceedings of the Rochester 1983 Int. Computer Music Conf.*, Eastman School of Music, Rochester, NY (Int. Computer Music Assoc., San Francisco), pp. 113–126.
- Tellman, E., Haken, L., and Holloway, B. (1995). "Timbre morphing of sounds with unequal numbers of features," *J. Audio Engineering Soc.* **43**(9), 678–689.
- Verma, T. S., Levine, S. N., and Meng, T. H. Y. (1997). "Transient modeling synthesis: A flexible analysis/synthesis tool for transient signals," *Proc. Int. Computer Music Conf.*, Thessaloniki (Int. Computer Music Assoc., San Francisco), pp. 164–167.
- Wessel, D. (1979). "Timbre space as a musical control structure," *Computer Music J.* **3**(2), 45–52.

A Compact and Malleable Sines+Transients+Noise Model for Sound

SCOTT N. LEVINE AND JULIUS O. SMITH III

1 Introduction

This chapter describes an audio representation which supports time and frequency scale modifications in a compressed domain. The input audio is segregated into three component representations: sinusoids, transients, and noise. Each component can be individually quantized and/or time-scaled and/or pitch-shifted.

Parametric models of sound are useful in a variety of applications. For the composer using recorded sounds as raw materials in a composition, control parameters are necessary for musically transforming the sound in an intelligible manner. For the telecommunications engineer, parametric sound models can provide a high degree of data compression, with little or no loss of quality, by transmitting sound parameters in place of the sound itself. For the audio engineer, there are many applications for time-scale modification, i.e., speeding up or slowing down a sound playback without changing musical pitch; examples include synchronizing a sound track to a film or providing a high-quality “fast forward” feature in a sound “browser.”

One of the oldest and most successful parametric models for sound is the sinusoidal model. Conceptually, sinusoidal models are rooted in basic Fourier theory, which states that any periodic sound $s(t)$ can be expressed mathematically as a sum of sinusoids:

$$s(t) = \sum_{k=1}^K A_k(t) \cos(\omega_k t + \phi_k(t)) \quad (4.1)$$

where t denotes time; $\omega_k = 2\pi k / P$ the k th harmonic radian frequency, where P is the sinusoidal period in seconds; $A_k(t)$, and $\phi_k(t)$ are the amplitude and phase of the k th harmonic sinusoidal component, and K is the number of the highest audible harmonic. Sinusoidal models are most appropriate for “tonal” sounds such as spoken or sung vowels, or the sounds of musical instruments in the string, wind, brass, and “tonal percussion” families. Ideally, only one sinusoid is needed to represent each harmonic or overtone in the sound. To represent the “attack” and “decay” of natural tones, sinusoidal components are multiplied by an amplitude

envelope that varies over time. That is, the amplitude $A_k(t)$ is a slowly varying function of time. Similarly, to allow pitch variations such as vibrato, the phase ϕ_k may be modulated in various ways. (The frequency deviation from the harmonic frequency may be defined as the time derivative of the phase.)

Sinusoidal models are extremely effective. Perhaps the main reason for this is that the ear focuses most acutely on peaks (amplitude maxima) in the spectrum of a sound. For example, when there is a strong spectral peak at a particular frequency, it tends to mask lower-level sound energy at nearby frequencies. As a result, the ear–brain system is, to a first approximation, a “spectral peak analyzer.” In modern audio coders (Painter and Spanias, 2000), exploiting masking has resulted in an order-of-magnitude data compression, averaged across various styles of popular music, with no loss of quality, according to listening tests (Brandenburg and Bosi, 1997).

However, for noise-like sounds, such as wind, scraping sounds, or breath noise in a flute, sinusoidal models are relatively expensive, requiring many sinusoids across the audio band. It is therefore helpful to combine a sinusoidal model with some kind of noise model, such as pseudo-random numbers passed through a filter (Serra and Smith, 1990).

Another situation where sinusoidal models become inefficient occurs at sudden transients in a sound, such as the click-like onset of a percussive sound. From Fourier theory, we know that transients too can be modeled exactly, but only by using large numbers of sinusoids at exactly the right phases and amplitudes. However, to keep the model compact, it is better to introduce an explicit transient model that works together with sinusoids and filtered noise to represent the sound more parsimoniously. Another advantage of an explicit transient model is that transients can be preserved during time-compression or expansion. That is, when a sound is stretched (without altering its pitch), it is usually desirable to keep the transients sharp (i.e., to preserve their time scales) and simply translate them to new times.

1.1 History of Sinusoidal Modeling

In the 1930s, Russian “futurists” used “syntones” to synthesize film soundtracks by means of a sum of sinusoidal components (Smirnov, 1998). (Fourier transforms for analyzing periodic sounds had to be carried out by hand.) Artificial synthesis of film soundtracks employed professional animators to “draw” sounds directly to produce photographic masks.

Also in the 1930s, the *vocoder* (“voice coder”) was developed by Homer Dudley at Bell Telephone Laboratories as a means of reducing the bandwidth required to transmit speech (Dudley, 1939). The vocoder could be regarded as a sinusoidal *or* a noise model, in that it switched between a tonal and a noise signal depending on whether the speech was voiced or unvoiced. Amplitude-vs-time controls for the vocoder’s band-pass filters used for synthesis were measured by means of amplitude followers at the outputs of the band-pass filters used for analysis. A simplified version of Dudley’s system, called the “voder,” was manually operated by trained technicians and was demonstrated at the 1939 World’s Fair. The name

“vocoder,” however, can be applied to any automatic system that synthesizes speech (or music) based on the results of analysis, i.e., coder-driven synthesis.

In the 1960s, the phase vocoder was introduced by Flanagan and Golden (1966) based on interpreting the classical vocoder filter bank as a sliding short-time Fourier transform. A digital computer made it possible for the phase vocoder to easily support phase modulation of the synthesis oscillators as well as implementing their amplitude envelopes. Thus, in addition to computing the instantaneous amplitude at the output of each (complex) band-pass filter, the instantaneous phase was also computed. Moreover, time-varying phase could be converted to time-varying frequency by taking a time derivative. Complex band-pass filters were implemented by first multiplying the incoming signal by $e^{j\omega_k t}$, where ω_k is the k th channel radian center frequency, and then by low-pass filtering it using convolution with the impulse response of a sixth-order Bessel filter.

The phase vocoder also relaxed the requirement of pitch-following (needed in the vocoder), because the phase modulation computed by the analysis stage automatically fine-tuned each sinusoidal component within its filter bank channel. The main remaining requirement was that only one sinusoidal component be present in any given channel of the filter bank; otherwise, the instantaneous amplitude and frequency computations would be based on “beating” waveforms instead of single sinusoids that produce smooth amplitude and frequency envelopes necessary for good data compression.

In the early 1960s, sine wave summation synthesis (otherwise known as “additive synthesis”) was one of the first general methods of sound synthesis used in computer music. In fact, it is extensively described in the first article of the first issue of the Computer Music Journal (Moorer, 1978). Some of the first high-quality synthetic musical instrument tones using additive synthesis were developed in the 1960s by Jean-Claude Risset (1985).

In the 1970s, the phase vocoder was reimplemented using the FFT for increased computational efficiency (Portnoff, 1976). The FFT window (analysis low-pass filter) was also improved to yield exact reconstruction of the original signal when synthesizing without modifications. Shortly thereafter, the FFT-based phase vocoder was adopted as the analysis method of choice for additive synthesis in computer music (Moorer, 1978). Since then, numerous variations and improvements of the phase vocoder have appeared (Griffin and Lim, 1988; Laroche and Dolson, 1999). For an excellent introductory tutorial, see Dolson (1986). A summary of vocoder research from the 1930s through the mid-1960s is given by Schroeder (1966).

With the phase vocoder, the instantaneous amplitude and frequency are normally computed only for each “channel filter.” A consequence of using a fixed-frequency filter bank is that the frequency of each sinusoid is not normally allowed to vary outside the bandwidth of its channel band-pass filter. Ordinarily, the band-pass center frequencies are harmonically spaced, i.e., they are integer multiples of a base frequency. So, for example, when analyzing a piano tone, the intrinsic progressive sharpening of its overtones leads to some sinusoids falling “in the cracks” between adjacent filter channels. This is not an insurmountable condition because the adjacent bins can be combined in a straightforward manner to provide

accurate amplitude and frequency envelopes (e.g., Horner et al., 1997), but it is inconvenient and outside the original scope of the phase vocoder. Moreover, it is unwieldy to work with the instantaneous amplitude and frequency signals from all of the filter-bank channels.

Many modern sinusoidal models can be thought of as “pruned phase vocoders” in that they follow only the peaks of the short-time spectrum rather than the instantaneous amplitude and frequency from every channel of a uniform filter bank. Peak-tracking with a sliding short-time Fourier transform has a very long history going back almost half a century (Peterson and Cooper, 1957; General Electric Co., 1977). Peak-tracking sinusoidal modeling of speech signals was introduced by McAulay and Quatieri (1984, 1985, 1986), application of this method to musical sounds was initiated by Smith and Serra (1987), and inverse-FFT methods were introduced by Rodet and Depalle (1992).

Historically, both vocoders and sinusoidal models have focused on modeling monophonic sound sources such as a single saxophone tone. By going to multiresolution sinusoidal modeling (described in Section 3), it is possible to encode general polyphonic sound sources with a single unified system (Levine, 1998).

In the late 1980s, Serra and Smith combined sinusoidal modeling with noise modeling to enable more efficient synthesis of the noise-like components of sounds (Serra, 1989; Serra and Smith, 1990, 1991). In this extension, the output of the sinusoidal model is subtracted from the original signal, leaving a residual signal. Assuming that the residual is a random signal, it is modeled as filtered white noise where the magnitude envelope of its short-time spectrum becomes the filter characteristic through which white noise is passed during resynthesis.

A more recent addition to the sines-plus-noise model is transient modeling (Ali, 1996; Verma et al., 1997; Levine, 1998; Levine et al., 1998; Levine and Smith, 1998, 1999). These methods address the principal remaining deficiency in sines-plus-noise modeling which is preserving crisp “attacks,” “clicks,” and the like, without having to use hundreds or thousands of sinusoids to accurately resynthesize the transient. (In general, the noise component cannot be used for transient modeling because no matter how much resolution is provided in the amplitude envelope of the noise, there is usually no guarantee that the noise, being random, will have the desired amplitude at the critical time it is needed.)

A 74-page summary of sinusoidal modeling, including sines-plus-noise modeling is given by Quatieri and McAulay (1998). Additional references related to sinusoidal modeling include George and Smith (1987, 1992); McAulay and Quatieri (1989, 1990, 1991); Roads et al. (1997); Rodet and DePalle (1992); and Wang (1995).

1.2 Audio Signal Models for Data Compression and Transformation

Audio representations for data compression are not always desirable when the main goal is manipulation and transformation of the audio signal. However, when

such a representation is based on a good signal model, it can be quite valuable for transformational purposes as well.

It often happens that the model that is natural from a conceptual (and manipulative) point of view is also very effective from a compression point of view. This is because, in the “right” signal model for a natural sound, the model’s parameters tend to vary slowly. As an example, physical models of the human voice and musical instruments have led to expressive synthesis algorithms that can also represent high-quality sound at much lower bit rates (such as MIDI event rates) than normally obtained by encoding the sound directly (Smith, 1998, 2004).

In the present context, the signal model follows a natural perceptual decomposition of sound into three qualitatively different components: “tones,” “noises,” and “clicks.” A successful signal model should naturally represent the essence of any sound, and a compact representation of sonic essence is simultaneously valuable for purposes of both transformation and compression.

1.3 Chapter Overview

The goal of this chapter is to present a new representation for audio signals that allows for low-bit-rate coding while still allowing for high-quality compressed-domain time-scaling and pitch-shifting modifications. In this system, the target bit-rates are from 16 to 48 kbps, while allowing for high audio bandwidth (44.1 kHz sampling rate) and high-quality time- and pitch-scale modifications. This compares to the $44.1 \text{ kHz} \times 16 \text{ bits/channel} = 705.6 \text{ kbps/channel}$ rate needed for uncompressed 16-bit audio.

To achieve these data compression rates and wide-band modifications, we first segment the audio (in time and frequency) into three separate signals: (1) a signal which models all sinusoidal content with a sum of time-varying sinusoids (Levine et al., 1998), (2) a signal which models all attack transients present using transform coding (Bosi et al., 1997), and (3) a Bark-band noise signal (Zwicker, 1961; Goodwin, 1996) which models all of the input signal not modeled by the sines or transients. Each of these three signals can be individually quantized using psychoacoustic principles pertaining to each representation.

High-quality time-scale and pitch-scale modifications become possible because the signal has been split into sinusoids, transients, and noise. The sinusoids and noise can be stretched or compressed with good results, and the transients can be time-translated while still maintaining their original temporal envelopes. Using phase-matching algorithms, the system can switch between sines and transients seamlessly. In time-scaled (slowed) polyphonic music with percussion or drums, this results in slowed harmonic instruments and voice, with the drums still having sharp attacks.

In the following sections, the system is first described from a high-level point of view, showing how an input audio signal is segmented in time and frequency. Then, each of the three signal models, sines, transients, and noise, is described along with their separate methods of parameter quantization. Finally, the last section is devoted to a particular application: compressed-domain time-scale modifications.

2 System Overview

The analysis/resynthesis system is designed to perform high-quality modifications, such as time-scale modification and pitch-shifting, on full-bandwidth audio while being able to maintain low bit-rates. Before delving into this hybrid system, other successful current data-compression systems are described, including discussions of their advantages and disadvantages.

2.1 Related Current Systems

Current state-of-the-art transform compression algorithms can achieve very high-quality results (perceptually lossless at 64 kbytes/s/channel) but cannot achieve time- or pitch-scale modifications without independent post-processing modification algorithms (Bosi et al., 1997).

The most recent phase vocoders can achieve high-quality time- and pitch-scale modifications, but they currently impose a data expansion rather than a data compression (Laroche and Dolson, 1999). The parameters in this class of modeling method are $\times 2$ -over-sampled FFT coefficients (i.e., the hop size is equal to half the window size). Once expressed in magnitude and phase form, they can be time-scaled and pitch-scaled. Because of the oversampling, there are now twice as many FFT coefficients as original time coefficients [or corresponding modified discrete cosine transform (MDCT) coefficients (Malvar, 1992)]. In addition, it has not been shown how well these time- and pitch-scale modifications will perform if the FFT magnitude and phase coefficients are quantized to very low bit-rates. Phase vocoders do not separately model transients or noise, so they generally suffer from the disadvantages of any purely sinusoidal model.

Sinusoids-plus-noise modeling (Serra and Smith, 1990) has been developed for high-quality time- and pitch-scale modifications for full-band audio, but it is currently limited to single-sound sources and necessitates hand-tweaking of the analysis parameters by the user. This user interaction would be unacceptable for a general purpose audio system. The system also has difficulties modeling sharp, percussive attacks. These attack signals are not efficiently represented as a sum of sinusoids, and the attack time is too sharp for the frame-based noise modeling used in the system. In addition, this method typically gives a data expansion rather than a data compression, because its goal is to achieve a transformable audio representation rather than compression.

Sinusoidal modeling has also been used effectively for very-low bit-rate speech (2 to 16 kbytes/channel) (McAulay and Quatieri, 1986) and audio coding (Edler et al., 1996). In addition, these systems are able to achieve time- and pitch-scale modifications. But these systems were designed for band-limited (0–4 kHz) monophonic (i.e., single source) signals. If the bandwidth is increased or if a polyphonic input signal is used, the results are not of sufficiently high quality.

2.2 Time-Frequency Segmentation

None of the individual algorithms mentioned in the previous section can handle both high-quality compression and modifications. While sinusoidal modeling works well for steady-state signals, it is not the best representation for attack transients or very high frequencies (above 5 kHz). For this reason, we segment the time-frequency plane into three general regions: sines, transients, and noise. In each time-frequency region, we use a different signal representation and a different quantization algorithm.

The first step in the segmentation process is to analyze the signal with a transient detector. (Details of the transient detector are discussed in Section 4.1.) This step time-segments the input signal into attack-transient and non-transient signals. Below 5000 Hz, the non-transients are modeled as multiresolution sinusoids (Levine et al., 1998) (described in Section 3). Also below 5000 Hz, a Bark-band-noise algorithm models the residual signal between the original audio and the sinusoidal data. Above 5000 Hz, the non-transients are completely modeled using only Bark-band-noise envelopes, similar to the techniques developed by Goodwin (1996). (These noise-modeling algorithms are described in Section 5.) Between 0 and 16 kHz, the transient signals are modeled using variants of current transform coding techniques (Bosi et al., 1997) (described in Section 4).

Time-frequency segmentation is illustrated in Fig. 4.1. Overlap regions between the sinusoids and the transients are phase-matched, so no discontinuities can be heard. (This is further discussed in Section 3.) Incremental improvements to the time-frequency segmentation, which allow for lower bit-rates and higher fidelity synthesis, are possible (described later in the chapter).

2.3 Reasons for the Different Models

Sinusoidal modeling is used only for the non-transient segments of the audio signal because attack transients cannot be efficiently modeled by a set of linearly ramped sinusoids. It is possible to model transients with a set of sinusoids, but such a system would typically need hundreds of sinusoidal parameters, consisting of amplitudes, frequencies, and phases. In this system, we attempt to model only steady-state signals with sinusoids, thus allowing for a more efficient representation.

Sinusoidal modeling is only used below 5000 Hz because most music (but not all), seldom contains isolated, definite-pitched sinusoidal components with frequencies above 5000 Hz. This is consistent with results found in the speech world (Laroche et al., 1993). Also, while pitch pipes and glockenspiels certainly have stable high-frequency sinusoids, this is not the case for most popular music. In the future, this cutoff frequency could become signal adaptive, perhaps relying on a measure of *tonality* (Bosi and Goldberg, 2003). Moreover, it is very inefficient to model high-frequency noise with sinusoids, and it is also very difficult to track stable, high-frequency sinusoids reliably, especially when high-amplitude, high-frequency background noise is present. Still, to increase signal modeling flexibility

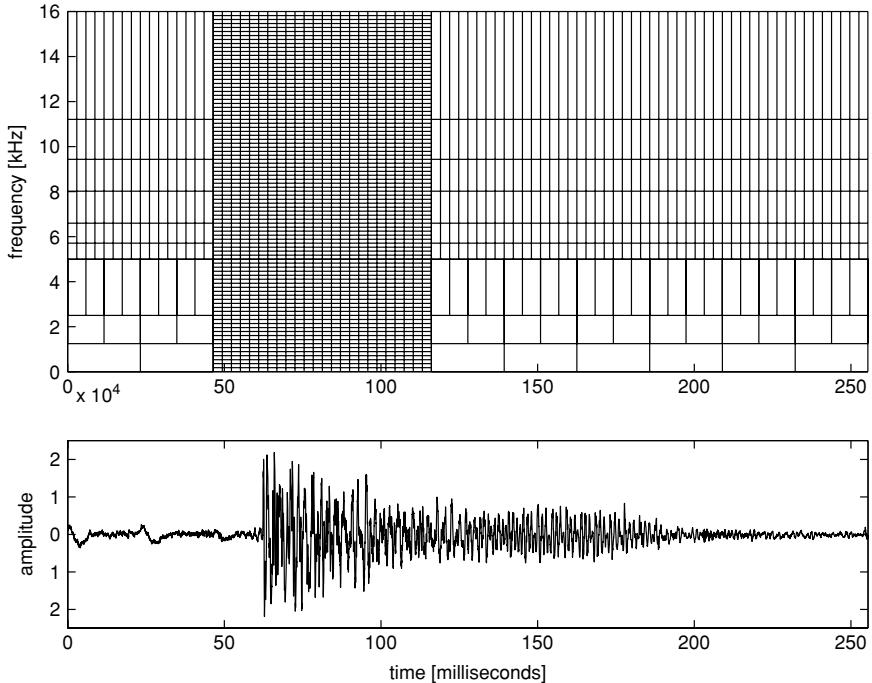


FIGURE 4.1. The lower plot shows 250 ms of a drum attack in a piece of pop music. The upper plot shows the time-frequency segmentation of this signal. During the attack portion of the signal, transform coding is used over all frequencies and for about 66 ms. During the non-transient regions, multiresolution sinusoidal modeling is used below 5 kHz and Bark-band noise modeling is used from 0–16 kHz.

at the expense of the overall data rate, an additional higher octave of sinusoids may be utilized.

For transient modeling, bit allocation is minimized by a method of transform coding (Bosi et al., 1997). Because transform coding is a waveform coder, it can be used to give a high-precision representation over a short time duration (about 66 ms). Whenever an audio signal is to be time-scaled, we simply translate transform-coded, short-time transients to their correct new positions in time. (More details are provided in Section 6.)

3 Multiresolution Sinusoidal Modeling

Sinusoidal modeling has proved to be a good representation for modeling monophonic music (Smith and Serra, 1987) and speech (McAulay and Quatieri, 1985, 1986), but has only recently been used for general purpose wide-band audio compression (Hamdy et al., 1996). Certain problems arise when switching from

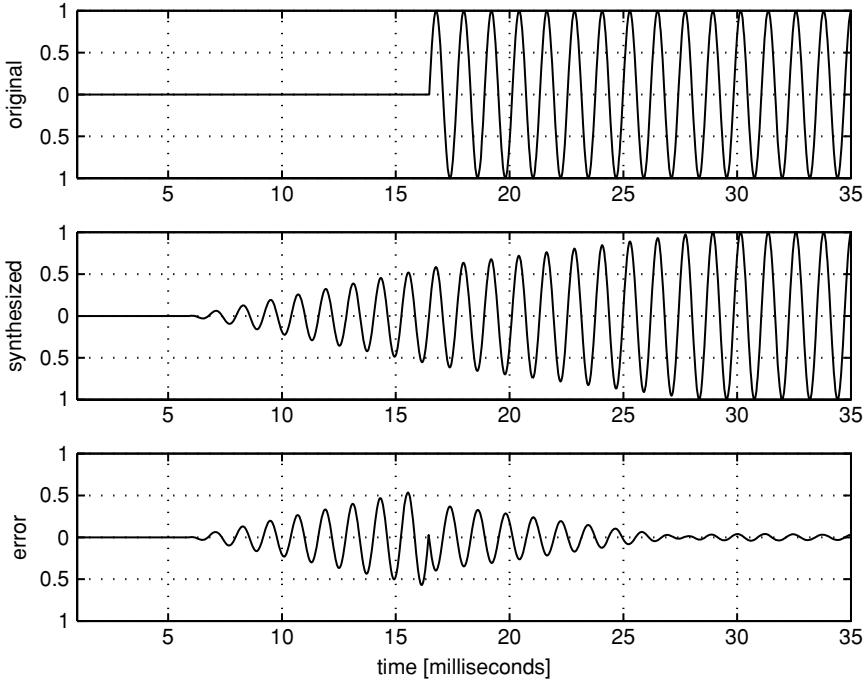


FIGURE 4.2. Pre-echo error resulting from sinusoidal modeling. Because in synthesis the sinusoidal amplitude is linearly ramped from frame to frame, the synthesized onset time is limited by the length of the analysis window.

monophonic speech/audio (i.e., single voice) to polyphonic audio. For example, a single fundamental frequency can no longer be assumed, and thus no pitch-synchronous analysis can be performed in general.

One problem is to choose a proper analysis window length. While long windows guarantee good frequency resolution at low frequencies, short windows tend to reduce pre-echo artifacts (see Fig. 4.2). With pitch-synchronous analysis, one could choose an adaptive window length that is two to three times longer than the current fundamental period.

However, because multiple pitches and instruments may be present, we use a multi-resolution sinusoidal modeling algorithm (Levine, et al., 1998). This algorithm splits the signal into three different frequency ranges and uses different window lengths for each range. Each range uses 50% overlap between window computations. See Table 4.1 for the parameters used in this system.

Figure 4.3 shows how the frequency range segmentation can be visualized in the time-frequency plane. Each rectangle indicates the times at which the sinusoidal {amp, freq, phase} parameters can be updated. For example, in the lowest frequency range sinusoidal parameters are only updated every 23 ms (the hop size in that range). But in the highest range, parameters are updated every

TABLE 4.1. Window Length and Hop Size for Each Frequency Band

Frequency range	Window length	Hop size
0–1250 Hz	46 ms	23 ms
1250–2500 Hz	23 ms	11.5 ms
2500–5000 Hz	11.5 ms	5.75 ms

5.75 ms. Usually, there are about 5 to 20 sinusoids present in each range at any one time.

3.1 Analysis Filter Bank

In order to obtain the multiresolution sinusoidal parameters, a $\times 2$ -over-sampled, octave-spaced, filter-bank front end is used. Each octave output of the filter bank is analyzed separately by the sinusoidal-modeling algorithm with a different window length. The filter outputs are over-sampled by a factor of 2 in order to attenuate the aliasing energy between the octaves below the threshold of audibility. Note that

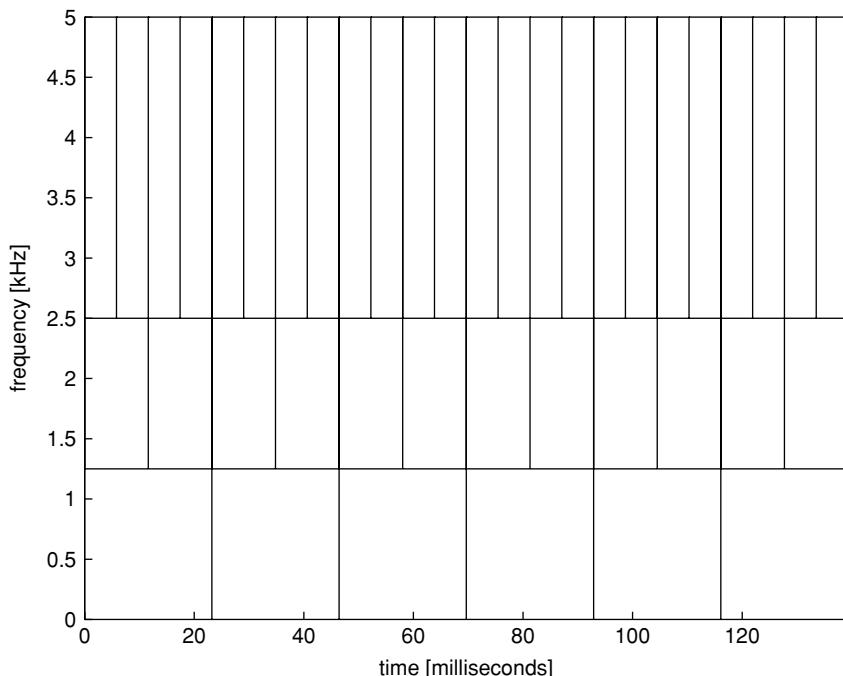


FIGURE 4.3. The time-frequency segmentation of multiresolution sinusoidal modeling. Each rectangle shows the update rate of sinusoidal parameters at different frequencies. In the top frequency range, parameters are updated every 5.75 ms, while in the lowest range the update rate is only 23 ms. Usually, there are 5 to 20 sets of sinusoidal parameters present in any one time-frequency rectangle.

for a critically sampled filter bank, such as a discrete-time wavelet transform, each octave output would contain aliased energy from the neighboring octaves. This aliased energy would introduce errors in the sinusoidal modeling. More details on the filter bank design are given by Fliege and Zolzer (1993) and Levine et al. (1998).

3.2 Sinusoidal Parameters

For each l th frame of analyzed audio, in a given frequency range, the system produces $R[l]$ sets of sinusoidal parameter triads $p_r[l] = \{A_r[l], \omega[l], \phi_r[l]\}$ (amplitude, frequency, phase), where r is the sinusoidal component index, based on maximum likelihood techniques developed by Thomson (1982) and previously used for sinusoidal modeling by Hamdy et al. (1996). Assuming that component amplitudes and frequencies are constant during each frame l and starting phases are given by $\phi_r[l]$, the synthesized sound during the frame is given by

$$s(m + lS) = \sum_{r=1}^{R[l]} A_r[l] \cos(m\omega_r[l] + \phi_r[l]), \quad m = 0, \dots, S - 1 \quad (4.2)$$

where S is a frequency-range-dependent hop size (number of samples per hop), as given by Table 4.1. However, in order to synthesize a signal without discontinuities at frame boundaries, the sinusoidal parameters must be interpolated for each sample m from the observed frame-boundary parameter values occurring at $m = 0$ and $m = S$. While amplitudes are simply linearly interpolated from frame-to-frame, phase and frequency interpolations are more complex and are discussed in Section 3.3.

In Sections 3.2.1–3.2.4, we show first how sinusoids are tracked from frame-to-frame and then give a method for computing a psychoacoustic masking threshold for each sinusoid. Based on this information, decisions are made about which sinusoids to eliminate from the system and how to quantize the remaining sinusoids.

3.2.1 Sinusoidal Tracking

Between frames l and $(l - 1)$, the sets of sinusoidal parameters are processed through a simple peak continuation algorithm. If $|A_i[l] - A_j[l - 1]| < \text{Amp}_{\text{thresh}}$ and $|\omega_i[l] - \omega_j[l - 1]| < \text{Freq}_{\text{thresh}}$, then the parameter triads $p_j[l - 1]$ and $p_i[l]$ are combined into a single sinusoidal trajectory. If a parameter triad $p_i[l]$ cannot be joined with another triad in adjacent frame $\{p_j[l - 1], j = 1, \dots, R[l - 1]\}$ or frame $\{p_k[l + 1], k = 1, \dots, R[l + 1]\}$, then this parameter triad becomes a trajectory of length one. The sinusoidal trajectory lengths coupled with their psychoacoustic masking properties (discussed in the following subsection) will determine which sinusoids are kept and which are discarded from the audio representation.

3.2.2 Masking

The first step in reducing the bit-rate for the sinusoids is to estimate which sine-wave amplitudes are above the psychoacoustic masking threshold for the synthesized

signal. In each frequency range, a separate masking threshold is computed based on the MPEG psychoacoustic model II [see the ISO/IEC 11172-3 standard (ISE/IEC, 1993)]. In each frequency range, the masking threshold is computed on an approximate third-Bark-band scale or Threshold Calculation Partition Domain as defined by the standard. From 0 to 5 kHz, there are about 50 non-uniformly spaced frequency divisions within which the thresholds are computed. Therefore, each i th sinusoidal parameter triad $p_i[l]$ in frame l obtains another parameter, the signal-to-masking threshold $m_i[l]$. This threshold is the difference between the energy of the i th sinusoid (correctly scaled to match the psychoacoustic model) and the masking threshold of its third-Bark band (in dB).

Not all of the sinusoids estimated in the initial analysis are stable (Thomson, 1982). Because we only desire to encode stable sinusoids and *not* model noisy signals represented by many closely spaced short-lived sinusoids, we use a psychoacoustic model that provides a tonality measure (Bosi and Goldberg, 2003) based on the prediction of FFT magnitudes and phases (ISE/IEC, 1993) to double-check the results of the initial sinusoidal estimations.

As can be seen in Fig. 4.4, shorter sinusoidal trajectories have (on average) lower signal-to-masking thresholds. This means that many shorter trajectories will

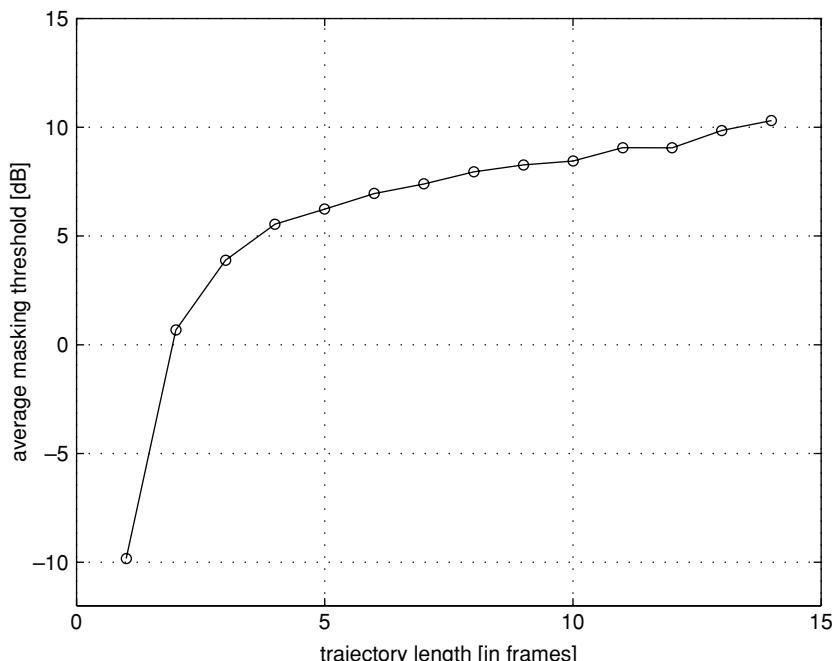


FIGURE 4.4. Average maximum signal-to-masking threshold (in decibels) vs sinusoidal trajectory length. Note that the longer a trajectory lasts, the higher its signal-to-masking threshold. These data were derived from the top frequency range of 8 s of pop music, where each frame length is approximately 6 ms.

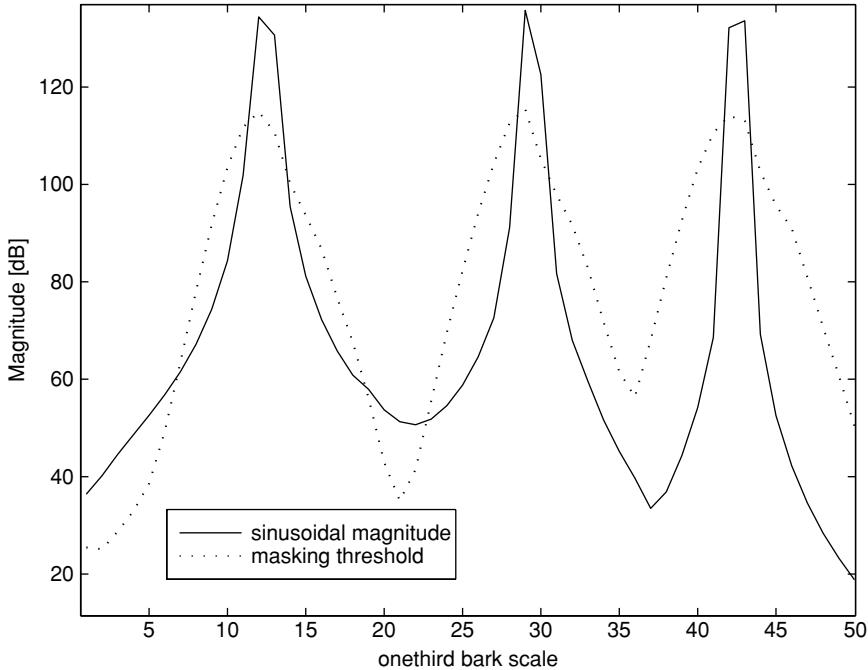


FIGURE 4.5. The original spectral energy vs the masking threshold for three pure sinusoids at frequencies 500, 1500, and 3200 Hz. Note that the masking threshold is approximately 18 dB below each sinusoidal peak.

be masked by those that are longer and more stable. A likely reason for this trend is that the shorter trajectories attempt to model noise, while the longer trajectories model true sinusoids. As illustrated in the IEC/ISO standard (ISE/IEC, 1993), a stable sinusoid typically has a signal-to-masking threshold of -18 dB in its third-Bark band, whereas a noisy signal typically has only a -6 dB masking threshold. Therefore, tonal signals have a lower signal-masking threshold than noisy signals (Zwicker and Fastl 1990). A simple graphical example of the masking thresholds for stable sinusoids can be seen in Fig. 4.5. As mentioned above, these signal-to-masking thresholds and sinusoidal trajectory lengths are important factors for determining which trajectories to eliminate and the number of bits to assign to the remaining parameters.

3.2.3 Sinusoidal Trajectory Elimination

Not all sinusoidal trajectories constructed as described in Section 3.2.1 are retained. For example, a trajectory is eliminated if it is completely masked, meaning its time-averaged energy is below the masking thresholds of the third-Bark bands that contain it. By eliminating the completely masked trajectories, the sinusoidal bit-rate is decreased by approximately 10% in typical audio input signals. Trajectories

that are near the masking threshold and have sufficiently short duration are also eliminated, typically reducing the sinusoidal bit-rate by approximately 40%. Most of these masked (or nearly masked) trajectories have very short trajectory lengths and are most likely attempts to model noise. For more details on the trajectory selection process, see Levine (1998) and Levine and Smith (1999). Section 5 discusses how signal energy corresponding to the eliminated sinusoidal trajectories is modeled by residual noise.

3.2.4 Sinusoidal Trajectory Quantization

Once masked and short-length trajectories have been eliminated, the remaining ones are quantized. In this section we focus only on amplitude and frequency quantization. Phase quantization is discussed in Section 3.3. Initially, amplitudes are quantized to 5 bits, in increments of 1.5 dB, giving a dynamic range of 96 dB. Frequencies are quantized to an approximate just-noticeable-difference frequency (JNDF) scale using 9 bits. Because amplitude and frequency trajectories vary slowly, temporal first-order differences across each trajectory can be efficiently quantized. These are then Huffman-encoded (Huffman, 1952; Ali, 1996).

In the previous section, we discussed how masked or short-length near-masking-threshold trajectories are eliminated while retaining all other trajectories even those whose energies are just barely higher than their Bark-band masking thresholds with longer duration. In principle, these lower-energy trajectories should not be allocated as many bits as the more perceptually important trajectories; i.e., those having energies much higher than their masking thresholds. A solution found to be bit-rate efficient, which did not impair sound quality, was to down-sample the lower-energy sinusoidal trajectories by a factor of 2. Thus, their sinusoidal parameters are updated at half of the original rate. At the decoder, the missing parameters are linearly interpolated. This effectively reduces the bit-rate of these trajectories by 50% and the total sinusoidal bit-rate by an additional 25%.

After testing several different kinds of music, we were able to quantize the three frequency ranges within 0–5 kHz (see Table 4.1) of the multiresolution sinusoids at bit-rates between 12 and 16 kbps. In practice, these numbers depend on how much of the signal from 0 to 5 kHz is encoded using transient modeling, as discussed in Section 4. As a tradeoff, more transients per unit time lowers the sinusoidal bit-rate, while increasing the transient-modeling bit-rate.

3.3 Switched Phase Reconstruction

In sinusoidal modeling, computing and saving correct phase information is usually only necessary for one of two reasons: The first reason is to assist in creating a residual error signal obtained by subtracting the synthesized sinusoids from the original signal (Serra, 1989; Serra and Smith, 1990). If the synthesized phases are not correct, much of original sinusoids will “leak” into the residual. However, this is only required at the encoder, not at the decoder. Thus, we need not transmit phases for this purpose. The second reason phase information is important is for improved

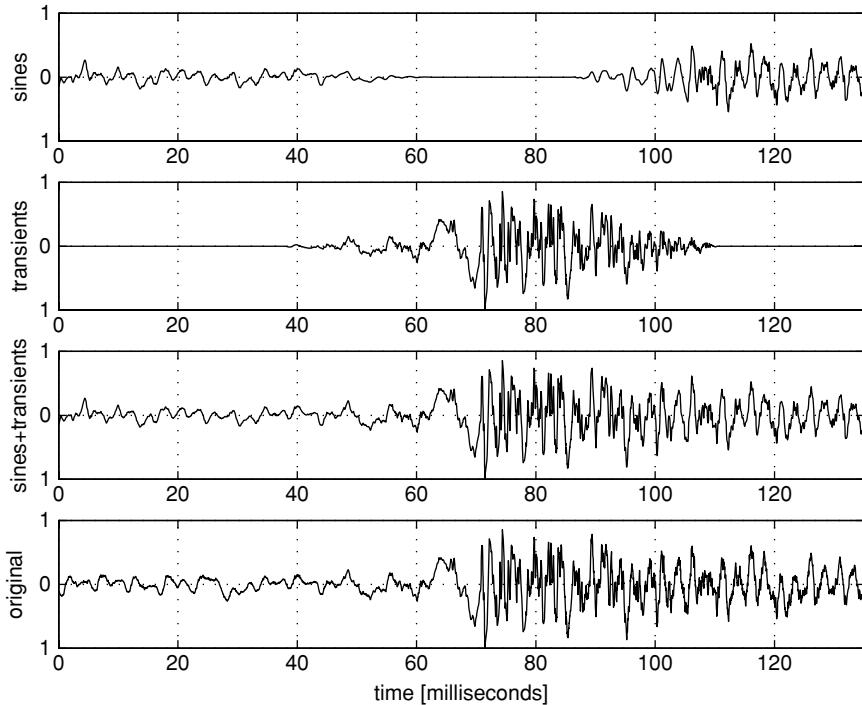


FIGURE 4.6. How sines and transients are combined: The top plot shows the multiresolution sinusoidal modeling component of the original signal. The sinusoids are faded-out during the transient region. The second plot shows a transform-coded transient. The third plot shows the sum of the sines plus the transient. For comparison, the bottom plot is the original signal. The original signal has a sung vowel through the entire section, with a snare drum hit occurring at $t \approx 60$ ms. Note that between 0 and 30 ms, the sines are *not* phase-matched with the original signal, but they do become phase-matched between 30 and 60 ms, when the transient signal is cross-faded in.

modeling of attack transients. During sharp attacks, the phases of sinusoids can be perceptually important. But in our system sharp attacks are not modeled by sinusoids; instead they are modeled by a transform coder. Thus, phase information is not needed for this purpose.

A simple example of switching between sines and transients is depicted in Fig. 4.6. At time $t = 40$ ms, the sinusoids are cross-faded out and the transients are cross-faded in. Near the end of the transients region at time $t = 90$ ms, the sinusoids are cross-faded back in. The trick is to phase-match the sinusoids during the cross-fade in/out times while only transmitting the phase information for the frames at the boundaries of the transient region.

To accomplish this goal, cubic-polynomial phase interpolation (McAulay and Quatieri, 1986) is used at the boundaries between the sinusoidal and transient regions. At all other times, we perform phaseless reconstruction (see Section 3.3.2)

sinusoidal synthesis. Because transient boundaries only occur at most several times a second, the contribution of phase information to the total bit-rate is extremely small.

Next, we describe the cubic-polynomial phase reconstruction and then show the differences between it and phaseless phase reconstruction. Then, we show how we can switch seamlessly between the two methods.

3.3.1 Cubic-Polynomial Phase Reconstruction

As discussed in Section 3.2, at each l th frame, $R[l]$ triad sets of parameters $p_r[l] = \{A_r[l], \omega_r[l]\phi_r[l]\}$ are estimated. These parameters must be interpolated from frame-to-frame to eliminate any discontinuities at the frame boundaries. While the amplitude is simply linearly interpolated from frame-to-frame, the phase interpolation is more complicated. At each sample m the instantaneous phase $\theta_r[l, m]$ is computed as a function of surrounding frequencies $\{\omega_r[l], \omega_r[l - 1]\}$ and surrounding phases $\{\phi_r[l], \phi_r[l - 1]\}$. Because the instantaneous phase is derived from four parameters, a cubic-polynomial interpolation function is used [see McAulay and Quatieri (1986) or Chapter 1 by Beauchamp]. Finally, the reconstruction for frame l becomes

$$s(m + lS) = \sum_{r=1}^{R[l]} A_r[l, m] \cos(\theta_r[l, m]), m = 0, \dots, S - 1 \quad (4.3)$$

where $A_r[l, m] = A_r[l] + m(A_r[l + 1] - A_r[l])$ is the linearly interpolated amplitude and $\theta_r[l, m]$ is the cubic-interpolated phase.

3.3.2 Phaseless Reconstruction

With “phaseless” reconstruction, explicit phase information is not required for signal resynthesis. The resulting signal is not phase-aligned with the original signal, but, on the other hand, it is guaranteed not to have any discontinuities at frame boundaries.

Instead of deriving the instantaneous phase from frame-boundary phases and frequencies, phaseless reconstruction derives instantaneous phase as the cumulative sum of the instantaneous frequency (Serra, 1989). The instantaneous frequency, $\omega_r[l, m]$, is first obtained by linear interpolation from the frame boundary values:

$$\omega_r[l, m] = \omega_r[l] + \frac{(\omega_r[l + 1] - \omega_r[l])}{S}m, m = 0, \dots, S - 1 \quad (4.4)$$

Then, the instantaneous phase for the r th trajectory in the l th frame is

$$\theta_r[l, m] = \theta_r[l, m - 1] + \omega_r[l, m], m = 0, \dots, S - 1, \quad (4.5)$$

where the term $\theta_r[l, m - 1]$ refers to the instantaneous phase at the last sample of the previous sample frame. The signal is then synthesized using Eq. (4.3), but using $\theta_r[l, m]$ from Eq. (4.5) instead of the result of a cubic-polynomial interpolation

function. For the first frame of phaseless reconstruction, the initial instantaneous phase is randomly picked from the range $[-\pi, \pi]$.

3.3.3 Phase Switching

As a simple example of “seamless” switching between the cubic-polynomial or “phaseless” phase reconstruction algorithms, consider the following case: First, all frames $(0, 1, \dots, l - 2)$ are synthesized using the phaseless reconstruction algorithm outlined in Section 3.3.2. Then, if we assume that a transient begins at frame l , our task is to seamlessly interpolate between the estimated parameters $\{\omega_r[l - 1]\}$ and $\{\omega_r[l], \phi_r[l]\}$ during frame $l - 1$ using the cubic interpolation method of Section 3.3.1. Because frame boundary $l - 1$ has no transmitted phases, we let $\phi_r[l - 1] = \theta_r[l - 1, S]$ at the last sample of the instantaneous phase of that frame. Then, in frame l , cubic interpolation is performed between $\{\omega_r[l], \phi_r[l]\}$ and $\{\omega_r[l + 1], \phi_r[l + 1]\}$. Because $\omega_r[l] = \omega_r[l + 1]$, and $\phi_r[l + 1]$ can be derived from $\{\omega_r[l], \phi_r[l], S\}$ as shown by Quatieri and McAulay (1986), we need only the phase parameters, $\phi_r[l]$, for $r = 1, 2, \dots, R[l]$ for each transient onset detected.

To graphically describe this scenario, see Fig. 4.7. Two consecutive 1024 sample frames $l - 1$ and l are shown. A decay transient begins at $m = 1024$ samples relative to the beginning of frame $l - 1$, or the beginning of frame l . The top plot shows a signal with explicit phase parameters transmitted for each frame boundary. The phase within each frame is interpolated using the cubic-polynomial phase reconstruction method as described in Section 3.3.1. The middle plot shows a signal with no explicit phase parameters transmitted except at the transient boundary at time $m = 1024$ samples. At all non-transient times, the phase of this signal is interpolated using phaseless reconstruction as described in Section 3.3.2. During the first 1024 samples of the figure, comprising frame No. 1, the middle signal slowly becomes phase-locked to the top signal. By the beginning of frame No. 2, which is the transient onset frame, the top two signals are phase-locked. The bottom signal shows the difference signal between the two top plots. A similar algorithm is performed at the end of the transient region to ensure that the ramped-on sinusoids are phase-matched to the final ramped-off transient frame.

4 Transform-Coded Transients

Because sinusoidal modeling does not model transients efficiently, transients are instead represented by a short-time transform code. The length of transform-coded sections could be varied, but in the current system it is 66 ms. This assumes that most transients last less than this amount of time and that after an initial attack, most signals become somewhat periodic and can be well-modeled using sinusoids. First, we discuss our transient detector, which decides when to switch between sinusoidal modeling and transform coding. Then, we describe the basic transform

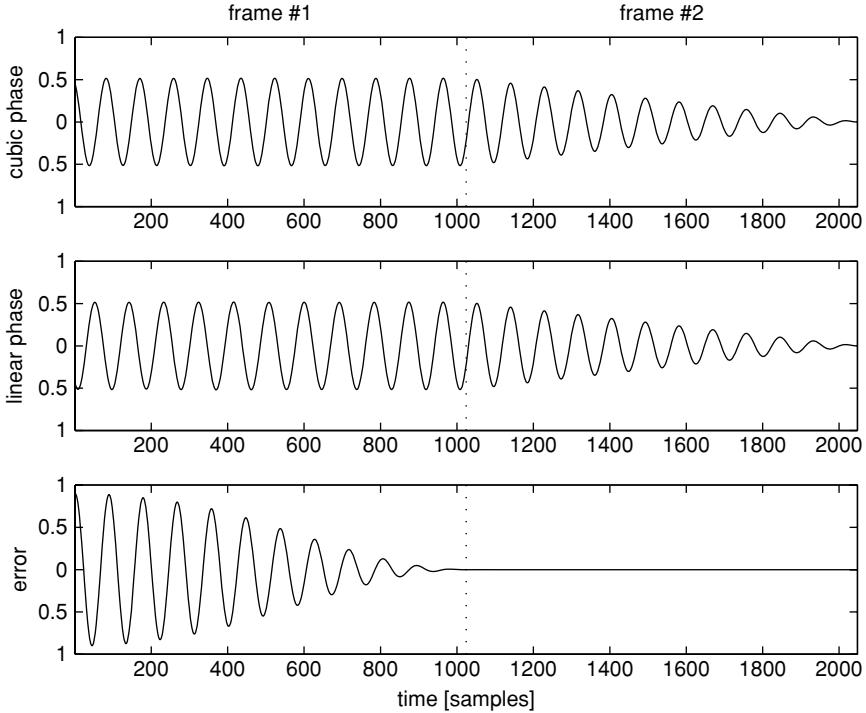


FIGURE 4.7. The top signal shows a signal synthesized with phase parameters, where the phase is interpolated between frame boundaries using a cubic-polynomial interpolation function (McAulay and Quatieri, 1986). The middle signal is synthesized using no explicit phase information except at the transient boundary, which is at $t = 1024$ samples. The initial phase is random, and is otherwise interpolated using the switched method of Section 3.3. Over the time range shown there are two frames, each 1024 samples long. Frame 1 shows the middle signal slowly becoming phase-locked to the signal above. By the beginning of frame 2, the top two signals are phase-locked. The bottom plot is the difference between the top two signals.

coder used in the system. In the following subsection, we then discuss methods to further reduce the number of bits needed to encode the transients.

4.1 Transient Detection

Design of the transient detector is very important to the overall performance of the system. The transient detector should only flag a transient during attacks that are not well modeled with sinusoids. If too many parts of the signal are modeled by transients, the overall bit-rate will become excessive because transform coding inherently requires a higher bit-rate than multiresolution sinusoidal modeling. In addition, this will degrade the quality of an audio signal after time-scale

modification has been applied (as discussed in Section 6). On the other hand, if too few transients are tagged, some attacks will sound dull or exhibit pre-echo problems due to limitations of sinusoidal modeling.

The system's transient detection algorithm combines two methods. The first method is a conventional frame-based energy measure. It looks for a suitably fast-rising edge in the energy envelope of the original signal over short frames. The second method involves a residual signal, which is the difference between the original signal and the multiresolution sinusoidal-modeled signal (with cubic-polynomial-interpolated phase). This method measures the ratio of short-time energies of the residual and the original signal. If the residual energy is very small relative to the original energy, then that portion of the signal is most likely tonal and is modeled well by sinusoidal modeling. On the other hand, if the ratio is high, it concludes that the energy in the original signal was not modeled well by the sinusoids and that an attack transient might be present.

The final transient detector uses both methods, i.e., it takes into account both the rising edge of the original signal's short-time energy and the ratio of the residual to the original short-time energy. The system declares a region to be a transient region when both of these methods agree that a transient is present.

4.2 A Simplified Transform Coder

The transform coder used in this system is a simplified version of the MPEG-AAC (Advanced Audio Coding) system (Bosi et al., 1997). It has been simplified to reduce the system's overall complexity. In this study we did not wish to improve the current state of the art in transform coding, but rather to use transform coding as a tool to encode transient signals. In the future, we plan to further optimize this simplified coder by reducing the bit-rate of the transients and by introducing a bit reservoir to be shared among the sines, transients, and noise modeling algorithms. In this system, a transient signal is defined as the residual that occurs during the duration of a detected transient after the off-ramping and on-ramping sinusoids are subtracted from the original signal. A graphical example of a transient can be seen in the second graph of Fig. 4.6.

In more detail, the transient coding occurs as follows: First, the transient is windowed into a series of short (256-point) segments, using a raised cosine window function. At 44.1 kHz, the current system encodes each transient with 24 short overlapping 256-point windows, for a total duration of 66 ms. There is no window length switching as done with the MPEG-AAC method, because the system has already identified the transient as such. Then, each segment is processed by a modified discrete cosine transform (MDCT) algorithm (Princen and Bradley, 1986) to convert from the time domain to a critically sampled frequency domain. A psychoacoustic model (ISE/IEC, 1993) is performed in parallel on the short segments in order to create the masking thresholds necessary for perceptually lossless subband quantization.

Next, the MDCT coefficients are quantized using scale factors and a global gain as in the AAC system. However, there are no iterated rate-distortion loops. A single

binary search quantizes each scale-factor band of MDCT coefficients, resulting in a mean-squared error just less than the psychoacoustic threshold allows. The resulting quantization noise should now be completely masked. Finally, we use a simplified version of the MPEG-AAC noiseless coding to Huffman-encode the MDCT coefficients, along with the differentially encoded scale factors.

4.3 Time-Frequency Pruning

As mentioned above, transients are encoded in this system for durations of 66 ms. However, in principle, a transient's duration should be frequency-dependent. At lower frequencies, the time duration required to encompass a transient is usually longer than it is at higher frequencies. Note that while we do not have a rigorous definition for determining when a signal should be considered a transient, in general we construe it to be the time during which a signal is quite aperiodic. Therefore, because a single transient does not actually have the same duration at all frequencies, there is no need to encode all 66 ms of the transient in every frequency range. In particular, we can construct a tighter time-frequency range of transform coding around the attack of the transient. While, for example, as shown in Fig. 4.8, our algorithm transform-encodes the signal's 0 to 5 kHz region for a total of 66 ms, it only transform-encodes the 5–16 kHz region for a total of 29 ms. The remaining time-frequency region above 5 kHz is modeled as Bark-band noise (discussed in Section 5).

This pruning of the time-frequency plane greatly reduces the number of bits necessary to encode transients. As will be shown, Bark-band noise modeling offers a much lower bit-rate representation than transform coding. After informal listening tests on many different kinds of music, no differences were detected between using transform coding over all frequency ranges for the full durations of transients vs using the reduced regions in the time-frequency plane.

As shown in Fig. 4.8, we currently only use two frequency regions that have different transform-encoded transient durations. But this could easily be generalized to more bands, octave-spaced bands, or even bands spaced according to a Bark-band scale. By only using transform coding for time-frequency regions where it is required, bit-rates can be lowered further. The remaining time-frequency regions are modeled using multiresolution sinusoidal modeling and Bark-band noise modeling, both of which have lower bit-rate requirements.

5 Noise Modeling

As we previously mentioned, in order to reduce the total system bit-rate, we do not model energy above 5 kHz as tonal (i.e., with sinusoids). Above 5 kHz the signal is either modeled as a transform-coded transient or as Bark-band filtered noise, depending on the state of the transient detector. With Bark-band noise modeling, the original signal's 5–16 kHz range is filtered into six Bark-spaced bands (Goodwin, 1996). Note that for a noisy signal, the ear is sensitive only to the total amount

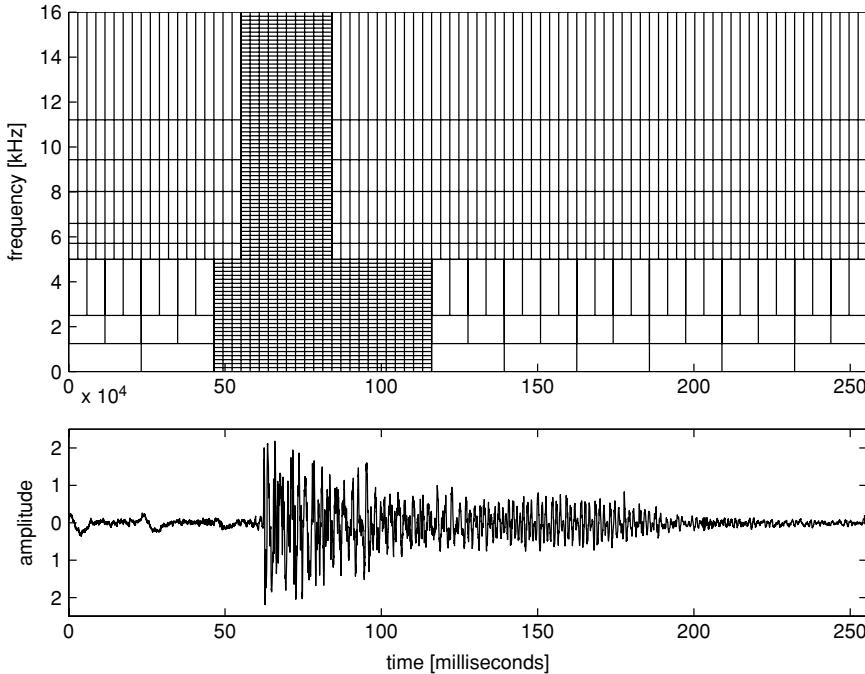


FIGURE 4.8. How to prune the time-frequency plane for transform coding of a transient. Like Fig. 4.1, the lower plot shows 250 ms of a drum attack in a piece of pop music. The upper plot shows the time-frequency segmentation of this signal. During the attack portion of the signal, transform coding is used for about 66 ms between 0 and 5 kHz, but for only 29 ms between 5 and 16 kHz. By reducing the time-frequency region of transform coding, the bit-rate is reduced as well. During the non-transient regions, multiresolution sinusoidal modeling is used below 5 kHz, and Bark-band noise modeling is used from 0 to 16 kHz.

of short-time energy in a Bark band and not the specific distribution of energy within the Bark band. Therefore, every 128 samples (3 ms at 44.1 kHz) an rms amplitude is measured from each of the six Bark-band-pass filters. To synthesize the noise, white noise is passed through the same Bark-spaced filters, and their outputs are amplitude-modulated using the individual rms-amplitude envelopes. Similarly, below 5 kHz the sinusoidal residual signal is modeled as Bark-band noise as well, but it uses a much longer frame duration of 1024 samples.

5.1 Bark-Band Quantization

After some informal listening tests, we determined that the coarsest useable quantization without introducing audible artifacts was achieved by employing 1.5-dB Bark-band amplitude-level steps. An example of a Bark-band amplitude envelope can be seen in the top graph of Fig. 4.9. Then, when we Huffman-encode this

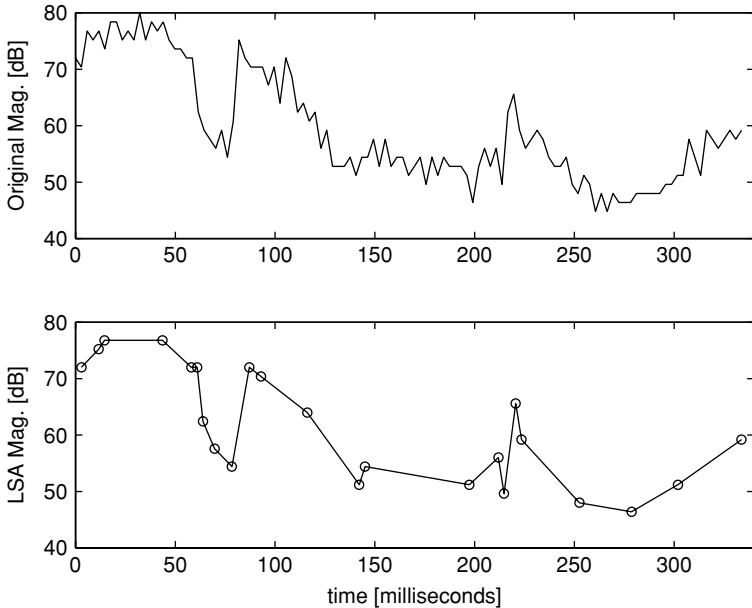


FIGURE 4.9. The top plot shows a Bark-band (8000–9200 Hz) rms-level amplitude envelope (in dB) for about 300 ms. The bottom plot shows the corresponding line-segment-approximated envelope. The circled points are the transmitted envelope points, and the remaining points are linearly interpolated using the transmitted points.

information, the total data rate is reduced to the neighborhood of 10 kbps. However, it does not seem perceptually important to sample the envelope every 128 samples (345 frames/s). It seems more important perceptually to preserve the rising and falling edges of the envelopes. Small deviations in Bark-band amplitude envelopes can be smoothed without audible consequence. The goal is to transmit only a small subset of the original envelope points and linearly interpolate the missing points at the decoder.

5.2 Line-Segment Approximation

Samples of the transmitted Bark-band amplitude-level envelopes are called breakpoints, because they are points at which the straight lines “break” to change slope. A greedy algorithm (Horner and Beauchamp, 1996) is used to iteratively decide where a new breakpoint in the envelope best minimizes the error between the original and approximated envelopes. The number of breakpoints is set to 20% of the length of the envelope itself. We found that while using fewer breakpoints lowered the bit-rate, it introduced audible artifacts in the synthesized noise. An example of an amplitude-level envelope reduced by line-segment approximation can be seen in the lower graph of Fig. 4.9.

There are now two sets of data to quantize: the times and amplitudes of the breakpoints. Time and amplitude differences between consecutive breakpoints are Huffman-encoded (Huffman, 1952), and, in addition, a Huffman table is used to encode the first amplitude of each envelope. The initial time of each envelope is inferred from time information obtained from the preceding transform-coded transient signal. If there is a possibility of losing some data in transmission, the time-differential methods will obviously need to be changed. For most noise signals, quantization of all Bark-bands results in a bit-rate of approximately 6 kbps.

6 Applications

The sines-plus-transients-plus-noise representation allows musicians and engineers to easily modify any input music source, whether it be a simple monophonic harmonic instrument or a complex polyphonic work, using only a relatively small number of meaningful parameters. Time-scale and pitch-scale modifications are relatively simple to perform on the compressed data because the input audio has been segregated into three separate parametric representations, all of which are well-behaved under time-frequency compression/expansion.

In this section we will concentrate on the time-scale modification. For more details on pitch shifting capabilities, see Levine (1998). Because the transients have been separated from the rest of the signal, they can be treated differently than the sines or the noise. To time-scale the audio, the sines and noise components are stretched linearly in time, while the transients are simply translated in time. How each of the three models is time-scale-modified is discussed in detail in the next three subsections. (See Figs. 4.10 and 4.11 for graphical examples and further explanation.) Apart from conventional time- and pitch-scaling modifications, there are many other kinds of transformations available for musical contexts, especially when applied to monophonic sounds, such as discussed by Serra (1989):

- Retuning the individual harmonics of a sound (e.g., to make a harmonic sound inharmonic, or vice versa).
- Making the noise component louder or softer than it was in the original analyzed sound, or independently filtering it.
- Making the transient component louder or softer, or slightly time-shifted.
- Replacing the noise or transients of one sound with the noise/transients of another sound (a new kind of “cross-synthesis”).
- Retuning a pitched sound without altering the transient or noise component.
- Automatically synchronizing an audio soundtrack using the transient location information.

These are only a few examples of what can be accomplished with software utilizing the sines-plus-noise-plus-transients audio representation.

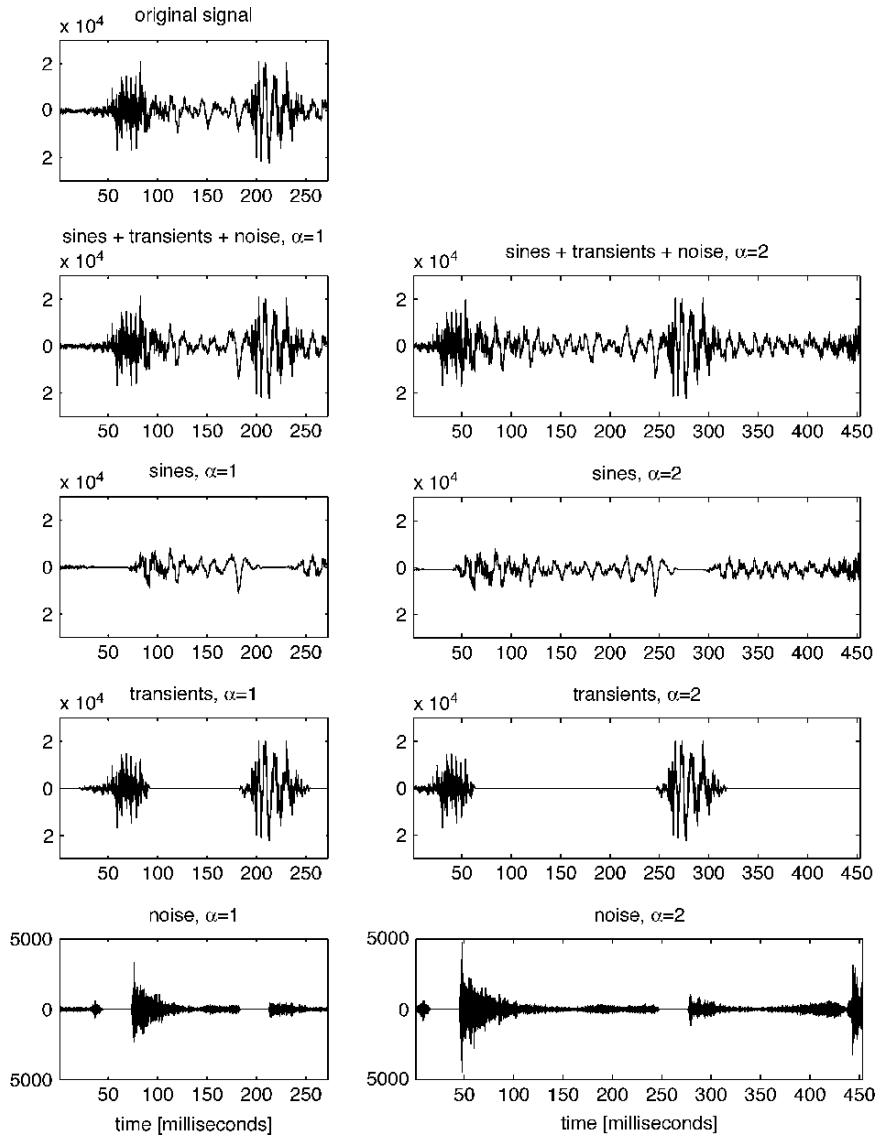


FIGURE 4.10. How time-scale modification is performed: The original signal, shown at top left, shows two transients: first a hi-hat cymbal hit and then a bass drum hit. There are also vocals present throughout the sample. The left-side plots show the full synthesized signal at top and then the sines, transients, and noise independently. They were all synthesized with no time-scale modification, at $\alpha = 1$. The right-side plots show the same synthesized signals, but with the time-scale modified by $\alpha = 2$, or twice as slow with the same pitch. Note how the sines and noise are stretched, but the transients are translated. Also, the vertical amplitude scale on the bottom noise plots are amplified 15 dB for better viewing.

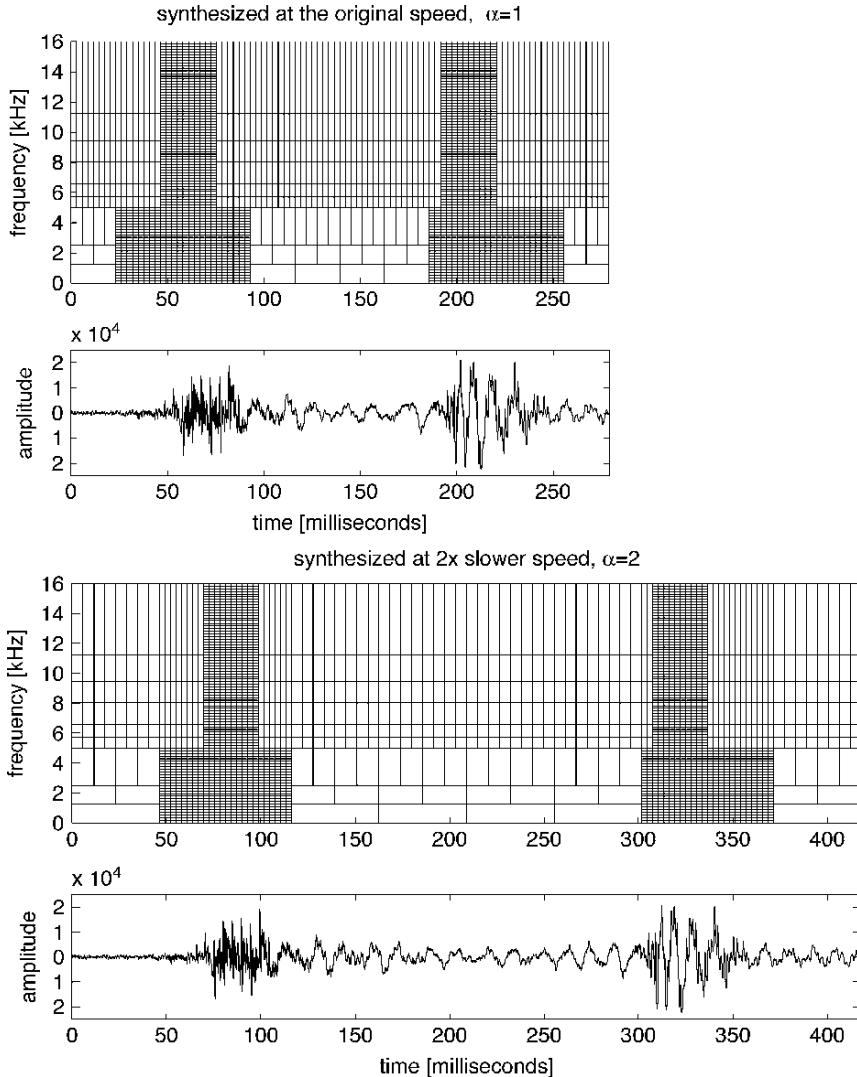


FIGURE 4.11. These figures illustrate the time-frequency plane segmentations used for Fig. 4.10. The figure on the top is synthesized with no time-scaling, $\alpha = 1$. The figure on the right is slowed down by a factor of 2, i.e., $\alpha = 2$. Note how the grid spacing of the transform-coded regions are not stretched, but rather shifted in time. However, the time-frequency regions of the multiresolution sinusoids and the Bark-band noise have been stretched in time in the bottom plot. Each of the rectangles in those regions is now twice as wide in time. The exception to this rule is the Bark-band noise modeled within the time span of the low-frequency transform-coded samples. These Bark-band noise parameters are shifted (not stretched), so that they remain synchronized with the rest of the transient. No sinusoids are used during the transform-coded segments.

6.1 Sinusoidal Time-Scale Modification

Since the earliest sinusoidal modeling systems for speech and audio have become available, methods for time-scaling signals by additive sine-wave synthesis have been quite obvious. For example, the l th frame in Eq. (4.3) can be slightly altered by scaling the hop size S by a time-stretch factor α . Thus, we have

$$s(m + lS\alpha) = \sum_{r=1}^{R[l]} A_r[l, m] \cos(\theta_r[l, m]), \quad m = 0, \dots, \alpha(S - 1) \quad (4.6)$$

When $\alpha = 1$, no time-stretching is applied. When $\alpha > 1$, the playback speed is slowed but the pitch remains the same. Similarly, when $\alpha < 1$, the playback speed is faster with the same pitch. The amplitude parameters are still linearly interpolated, but over a different frame length. In addition, the frequency/phase interpolation described in Section 3.3.3 is computed over a different frame length, matching both frequency and phase and ends of the new frame.

6.2 Transient Time-Scale Modification

To keep transients precise, the transform-coded transients are simply translated in time rather than stretched in time. Therefore, Modified Discrete Cosine Transform frames are moved to their new places in time and played at the original playback speed. Because these signals are so short in time (66 ms), transients sound natural and blend well with time-stretched sinusoids and noise. Thus, attacks are still sharp, no matter how much the music has been slowed down.

While time-scale modification causes the cross-fade regions between sinusoids and transients to appear at different regions in time, phase-locking is preserved by the frequency/phase interpolation algorithm (Section 3.3.3) when the sinusoids overlap with the transient signal.

6.3 Noise Time-Scale Modification

Because the noise has been parametrized by envelopes, it is very simple to time-scale the noise. Breakpoints in the Bark-band temporal envelopes are stretched according to the time-scale factor α . Using linear interpolation between the breakpoints, new stretched envelopes are formed. Six channels of Bark-band filtered noise are then amplitude-modulated by these new stretched envelopes and summed to form the final stretched noise. Similarly, efficient inverse-FFT methods could be used (Rodet and Depalle, 1992; Goodwin, 1996).

7 Conclusions

A system that allows both data compression and high-quality compressed-domain modifications has been described. By providing a separate representation for sines, transients, and noise, large data reductions typical of perceptually based

quantization schemes are obtained, while retaining the ability to perform compressed-domain processing such as time-scaling. In addition, sharp attack transients are preserved, even with large time-scale modification factors. To hear demonstrations of the data compression and modifications described in this chapter, see Levine (1998).

8 Acknowledgment

The first author would like to thank Tony Verma for his sinusoidal modeling software and for many hours of discussions about parametric coders and compression.

References

- Ali, M. (1996). "Adaptive signal representation with application in audio coding," doctoral dissertation, Univ. of Minnesota, Minneapolis, MN, Dissertation Abstracts Int.-B **57-04**, 2739.
- Bosi, M., Brandenburg, K., Quackenbush, S., Fielder, L., Akagiri, K., Fuchs, H., Dietz, M., Herre, J., Davidson, G., and Oikawa, Y. (1997). "ISO-IEC MPEG-2 advanced audio coding," *J. Audio Eng. Soc.* **45**, 789–814.
- Bosi, M. and Goldberg, R.E. (2003). *Introduction to Digital Audio Coding and Standards* (Kluwer Academic, Boston).
- Brandenburg, K. and Bosi, M. (1997), "Overview of MPEG audio: Current and future standards for low-bit-rate audio coding," *J. Audio Eng. Soc.* **45**(1/2), 4–21.
- Dolson, M. (1986). "The phase vocoder: A tutorial," *Computer Music J.* **10**(4), 14–27.
- Dudley, H. (1939). "Remaking speech," *J. Acoustical Soc. Am.* **11**, 169–177.
- Edler, B., Purnhagen, H., and Ferekidis, C. (1996). "ASAC—analysis/synthesis audio codec for very low-bit rates," *100th Convention of the Audio Engineering Society*, Copenhagen, *Audio Eng. Soc. Preprint No. 4179*.
- Flanagan, J. L., and Golden, R. M. (1966). "Phase vocoder," *Bell Syst. Tech. J.* **45**, 1493–1509. [reprinted in *Speech Analysis*, R. W. Schafer and J. D. Markel (eds.), IEEE Press, New York, 1979, pp. 388–404].
- Fliege, N. J., and Zolzer, U. (1993). "Multi-complementary filter bank," *Proc. 1993 Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-93), Minneapolis (IEEE, New York), Vol. 3, pp. 193–196.
- General Electric Co. (1977). "ADEC subroutine description," Technical Report, Heavy Military Electronics Department (General Electric Co., Syracuse, NY).
- George, E. B. and Smith, M. J. T. (1987). "A new speech coding model based on least-squares sinusoidal representation," *Proc. 1987 Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-87), Dallas, TX (IEEE, New York), pp. 1641–1644.
- George, E. B., and Smith, M. J. T. (1992). "Analysis-by-synthesis/Overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones," *J. Audio Eng. Soc.* **40**(6), 497–516.
- Goodwin, M. (1996). "Residual modeling in music analysis/synthesis," in *Proc. 1996 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-96), Atlanta, GA (IEEE, New York), pp. 1005–1008.

- Griffin, D. W., and Lim, J. S. (1988). "Multiband excitation vocoder," *IEEE Trans. on Acoustics, Speech, Signal Processing* **36**(8), 1223–1235.
- Hamdy, K. N., Ali, M., and Tewfik, A. H. (1996). "Low bit rate high quality audio coding with combined harmonic and wavelet representations," *Proc. 1996 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-96), Atlanta, GA (IEEE, New York), pp. 1045–1048.
- Horner, A. and Beauchamp, J. (1996). "Piecewise Linear Approximation of Additive Synthesis Envelopes: A Comparison of Various Methods," *Computer Music J.* **20**(2), 72–95.
- Horner, A., Ayers, L., and Law, D., (1997). "Modeling Small Chinese and Tibetan Bells," *J. Audio Eng. Soc.* **45**(3), 148–159.
- Huffman, D. A. (1952). "A Method for the Construction of Minimum-Redundancy Codes," *Proc. IRE* **40**, 1098–1101.
- ISE/IEC JTC 1/SC 29/WG 11 (1993). "ISO/IEC 11172-3: Information technology—coding of moving pictures and associated audio for digital storage media at up to about 1.5 mbit/s—Part 3: Audio" (Motion Picture Experts Group, Los Angeles, CA).
- Laroche, J., Stylianou, Y., and Moulines, E. (1993). "HNM: A simple, efficient harmonic + noise model for speech," *Proc. 1993 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (WASPAA-93), New Paltz, NY (IEEE, New York), pp. 169–172.
- Laroche, J., and Dolson, M. (1999). "Improved Phase-Vocoder Time-Scale Modification of Audio," *IEEE Trans. Speech and Audio Processing* **7**(3), 323–332.
- Levine, S. N. (1998). "Audio representations for data compression and compressed domain processing," doctoral dissertation, Stanford University, *Dissertation Abstracts Int.-B* **60/04**, 1767. [available for download at <http://www-ccrma.stanford.edu/~scottl/thesis.html>; this site also includes audio examples.]
- Levine, S. N., and Smith, J. O. (1998). "A sines+transients+noise audio representation for data compression and time/pitch-scale modifications," *105th Convention of the Audio Eng. Soc.*, San Francisco, Audio Eng. Soc. Preprint 4781. [available for download at <http://www-ccrma.stanford.edu/~scottl/papers.html>.]
- Levine, S. N., Verma, T. S., and Smith, J. O. (1998). "Multiresolution sinusoidal modeling for wideband audio with modifications," *Proc. 1998 Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-98), Seattle (IEEE, New York), pp. 3585–3588.
- Levine, S. N., and Smith, J. O. (1999). "A switched parametric and transform audio coder," in *Proc. 1999 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-99), Phoenix (IEEE, New York), pp. 985–988. [available for download at <http://www-ccrma.stanford.edu/~scottl/papers.html>.]
- Malvar, H. (1992). *Signal Processing with Lapped Transforms* (Artech House Telecommunications Library, Boston), pp. 175–179.
- McAulay, R. J. and Quatieri, T. F. (1984). "Magnitude-only reconstruction using a sinusoidal speech model," *Proc. 1984 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-84), San Diego (IEEE, New York), pp. 27.6.1–27.6.4.
- McAulay, R. J. and Quatieri, T. F. (1985). "Mid-rate coding based on a sinusoidal representation of speech," *Proc. 1985 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-85), Tampa, FL (IEEE, New York), pp. 945–948.
- McAulay, R. J., and Quatieri, T. F. (1986). "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. on Acoustics, Speech and Signal Processing* **34**, 744–754.
- McAulay, R. J., and Quatieri, T. F. (1990). "Pitch estimation and voicing detection based on a sinusoidal speech model," *Proc. 1990 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-90), Albuquerque, NM (IEEE, New York), pp. 249–252.

- McAulay, R. J., and Quatieri, T. F. (1991). "Sine-wave phase coding at low data rates," *Proc. 1991 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-91), Toronto, Canada (IEEE, New York), pp. 577–580.
- Moorer, J. A. (1978). "The use of the phase vocoder in computer music applications," *J. Audio Eng. Soc.* **26**, 42–45.
- Painter, T. and Spanias, A. (2000). "Perceptual coding of digital audio," *Proc. IEEE* **88**(4), 451–513.
- Peterson, E., and Cooper, F. S. (1957). "Peakpicker: A bandwidth compression device" (abstract), *J. Acoust. Soc. Am.* **29**, 777.
- Portnoff, M. R. (1976). "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Trans. on Acoustics, Speech, Signal Processing* **ASSP-24**, 243–248.
- Princen, J. P., and Bradley, A. B. (1986). "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. on Acoustics, Speech, Signal Processing* **ASSP-34**, 1153–1161.
- Quatieri, T. F. and McAulay, R. J. (1986). "Speech transformations based on a sinusoidal representation," *IEEE Trans. on Acoustics, Speech, Signal Processing* **ASSP-34**, 1449–1464.
- Quatieri, T. F., and McAulay, R. J. (1989). "Phase coherence in speech reconstruction for enhancement and coding applications," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-89), Glasgow, Scotland (IEEE, New York), pp. 207–210.
- Quatieri, T. F., and McAulay, R. J. (1998). "Audio signal processing based on sinusoidal analysis/synthesis," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, eds. (Kluwer, Boston, MA), pp. 343–416.
- Risset, J.-C. (1985). "Computer music experiments, 1964...," *Computer Music J.* **9**(1), 11–18.
- Roads, C. (Ed.). (1989). *The Music Machine: Selected Readings from Computer Music Journal* (MIT Press, Cambridge, MA).
- Roads, C., Pope, S. T., Piccialli, A., and De Poli, G. (eds.). (1997). *Musical Signal Processing* (Swets and Zietlinger, Exton, PA).
- Rodet, X. and Depalle, P. (1992). "Spectral envelopes and inverse FFT synthesis," *93rd Convention of the Audio Eng. Soc.*, San Francisco, CA, Audio Eng. Soc. Preprint 3393.
- Schafer, R. W., and Markel, J. D. (eds.). (1979). *Speech Analysis* (IEEE Press, New York).
- Schroeder, M. R. (1966). "Vocoders: Analysis and synthesis of speech (a review of 30 years of applied speech research)," *Proc. IEEE* **56**, 720–734. [reprinted in *Speech Analysis*, R. W. Schafer and J. D. Markel (eds.), (IEEE Press, New York), 1979, pp. 352–366].
- Serra, X. (1989). "A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition," doctoral dissertation, Stanford University, *Dissertation Abstracts Int.-A*, **51/01**, 18 [also available as Dept. of Music Report No. STAN-M-58, Stanford Univ., 1989].
- Serra, X. and Smith, J. O. (1990). "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music J.* **14**, 12–24.
- Serra, X. and Smith, J. O. (1991). "Soundsheet examples for a sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music J.* **15**, 86–87.
- Smirnov, A. (1998). "Proto musique concrete: Russian futurism in the 10s and 20s and early ideas of sonic art and art of noises," presented at Inventionen 98 Festival, September 28, 1998, Haus des Rundfunks, Berlin, Germany.

- Smith, J. O. and Serra, X. (1987). "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," *Proc. 1987 Int. Computer Music Conf.* (ICMC-87), Urbana, IL (Computer Music Assoc., San Francisco), pp. 290–297. (also available as Dept. of Music Technical Report STAN-M-43, Stanford Univ., 1987.)
- Smith, J. O. (1998). "Principles of digital waveguide models of musical instruments," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, eds. (Kluwer Academic Publishers, Boston), pp. 417–466.
- Thomson, D. J. (1982). "Spectrum estimation and harmonic analysis," *Proc. IEEE* **70**(9), 1055–1096.
- Verma, T. S., Levine, S. N., and Meng, T. H. Y. (1997). "Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals," *Proc. 1997 Int. Computer Music Conf.* (ICMC-97), Thessaloniki, Greece (Int. Computer Music Assoc., San Francisco), pp. 164–167.
- Wang, A. L. (1995). "Instantaneous and frequency-warped techniques for source separation and signal parametrization," *Proc. 1995 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics* (WASPAA-95), New Paltz, NY (IEEE, New York), Paper 2.5.
- Zwicker, E. (1961). "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)," *J. Acoust. Soc. Am.* **33**(2), 248.
- Zwicker, E., and Fastl, H. (1990). *Psychoacoustics, Facts, and Models* (Springer-Verlag, Berlin).

Spectral Envelopes and Additive + Residual Analysis/Synthesis

XAVIER RODET AND DIEMO SCHWARZ

1 Introduction

The subject of this chapter is the estimation, representation, modification, and use of *spectral envelopes* in the context of sinusoidal-additive-plus-residual analysis/synthesis. A spectral envelope is an amplitude-vs-frequency function, which may be obtained from the envelope of a short-time spectrum (Rodet et al., 1987; Schwarz, 1998). [Precise definitions of such an envelope and short-time spectrum (STS) are given in Section 2.] The additive-plus-residual analysis/synthesis method is based on a representation of signals in terms of a sum of time-varying sinusoids and of a non-sinusoidal residual signal [e.g., see Serra (1989), Laroche et al. (1993), McAulay and Quatieri (1995), and Ding and Qian (1997)]. Many musical sound signals may be described as a combination of a nearly periodic waveform and colored noise. The nearly periodic part of the signal can be viewed as a sum of sinusoidal components, called partials, with time-varying frequency and amplitude. Such sinusoidal components are easily observed on a spectral analysis display (Fig. 5.1) as obtained, for instance, from a discrete Fourier transform.

In consequence, some of the first attempts at sound synthesis were based on the additive synthesis method, i.e., the summation of time-varying sinusoidal components [e.g., Risset and Mathews (1969)]. This signal-modeling approach inherits a rich history of signal processing techniques. For example, harmonic or inharmonic partials are easy to characterize and easy to synthesize. Also, there exist many methods to automatically analyze sounds in terms of partials and noise that can then be used directly for additive synthesis [e.g., Serra and Smith (1990)]. Another interesting aspect of additive synthesis is its ease for mapping partial parameters (frequency and amplitude) into the human perceptual space. Also, these parameters are meaningful and easily understood by musicians. Furthermore, because independent control of every component is available in additive synthesis, it is possible to implement models of perceptually significant features of sound such as inharmonicity and roughness. Thus, additive synthesis is accepted as perhaps the most powerful and flexible sound synthesis method available.

A drawback of the classical sinusoidal oscillator (i.e., simple addition of sine waves) implementation of additive synthesis (Moore, 1990) is its computational

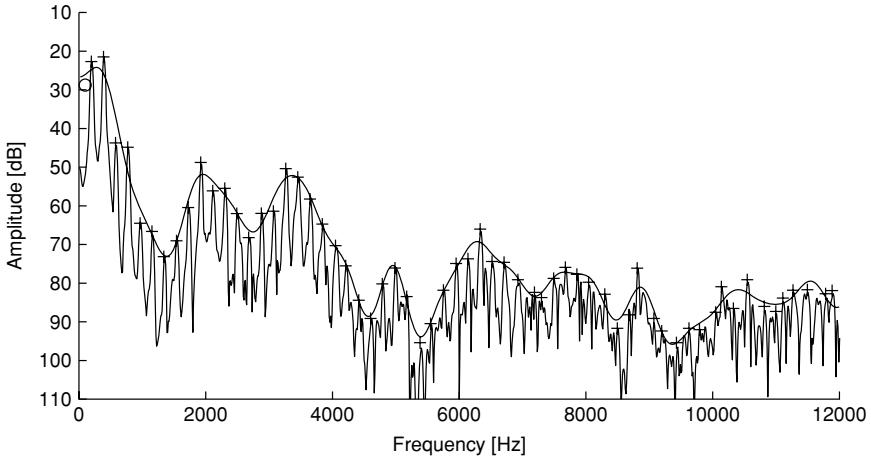


FIGURE 5.1. Spectrum and spectral envelope of the vowel /e/.

cost that can easily be seen by considering a sound such as a low-pitched piano tone, which can sometimes require more than a hundred partials to properly represent it. However, another additive synthesis technique (Rodet and Depalle, 1992) will be examined in Section 6.3. This method, named FFT^{-1} , is based on the inverse fast Fourier transform and allows an efficiency gain of 10–30 compared to the classical method. A second drawback of the oscillator method of additive synthesis is its difficulty for introducing precisely controlled noise components that are very important for realistic sounds and musical timbres such as speech or Japanese shakuhachi flute sounds, which cannot be created without noise. The FFT^{-1} technique and spectral envelopes make noisy components easy to describe and cheap to compute. Last but not least, controlling hundreds of sinusoids is a great challenge for the computer musician. Spectral envelopes render this control more simple, direct, and user-friendly, and are easily implemented with the FFT^{-1} method.

As mentioned above, speech and musical sounds always have random components, often heard as a noise, superposed on the harmonic or inharmonic parts. Since a second assumption underlying the sinusoidal additive model is that the number of sinusoidal partials is limited, a purely sinusoidal model $d(t)$ with slowly varying parameters cannot completely represent a real signal $s(t)$ and therefore must be complemented by a non-sinusoidal residual part $r(t)$ aimed at representing the random components:

$$r(t) = s(t) - d(t). \quad (5.1)$$

Even though the term spectral envelope is commonly used only for the envelope of the magnitude of the short-time spectrum, we will also consider envelopes that include the phase of the STS and even the frequencies of nearly harmonic partials as a function of their harmonic number. Such envelopes are called generalized

spectral envelopes (Rodet et al., 1987). One of the main features of generalized spectral envelopes is that several important properties of sounds are captured in a simple and powerful representation. Note that if a reduction of memory size is required, spectral envelopes need only be defined at a small number of specific times such as at the beginning of an attack, the end of attack, the end of sustain, and so on. Then, at any time, the synthesis algorithm may use envelopes interpolated between the defined envelopes. This procedure also yields an economical and efficient user control.

As explained in Section 2, the usefulness of spectral envelopes is primarily due to theoretical reasons: The concept of spectral envelope is connected with the production models (signal models and physical models) of musical instruments as well as with the perception of musical sounds. Spectral envelopes also offer a simple and concise representation of important sound properties that largely ease the control of synthesis models for musical applications. As an example, in speech or in the singing voice, the spectral envelope is rather independent of the pitch. This concept is violated if we use time-domain resampling to transpose a vowel up by, for instance, an octave, which results in the partials' frequencies being multiplied by 2 while not changing their amplitudes. Then, the spectral envelope will necessarily also be transposed. This effect sounds quite unnatural because all resonances (i.e., *formants*) are shifted up by an octave, corresponding to shrinking the vocal tract to half its length, the size of a young child's vocal tract compared to that of an adult. Obviously, this is not the natural behavior of the vocal tract. To avoid this, the spectral envelope has to be kept constant. This means that the amplitude of a transposed partial is no longer determined by the amplitude of the original partial, but by the value of the spectral envelope at the frequency of the transposed partial.

From the viewpoint of spectral analysis, our interest in spectral envelopes is due to the existence of many envelope estimation techniques. For an arbitrary sound signal, the spectral envelope at a given time is not known *a priori*. Therefore, it has to be estimated by using one of several techniques that will be described in Section 3. This estimation step is crucial because it governs any further use of the estimated spectral envelopes, whether it be for feature extraction, such as timbre characterization, or for resynthesis. As detailed in Section 4, spectral envelopes can be coded in one of several representations that differ by the memory space and the computational power they require. In Section 4.6, transcoding and manipulation of spectral envelopes are explained. It appears that the choice of these methods depends on the chosen representation. The use of spectral envelopes for the synthesis of sinusoidal and of non-sinusoidal parts of the signal are presented in Section 6.

At this point, it is worth noting that sinusoidal partials on one hand and non-sinusoidal components on the other hand are usually created by the human voice, as well as by musical instruments, using different mechanisms. Therefore, the sinusoidal part and the non-sinusoidal part should be treated separately for all steps from estimation to synthesis. That is to say, specific estimation techniques are used for each of these two parts, and each one is attributed a different spectral

envelope. Also, modifications applied to the two types of spectral envelope are usually different. Finally, the synthesis methods used for the two parts are not the same. Some examples of spectral envelope applications for additive-plus-residual analysis and synthesis, in various contexts and software systems, are given in Section 7. Finally, Section 8 gives a few conclusions and perspectives for future development and research concerning spectral envelopes in musical sound signal synthesis.

2 Spectral Envelopes and Source–Filter Models

2.1 *Source–Filter Models*

The concept of spectral envelope is closely related to the concept of the source–filter model (Depalle, 1991). In such a model, one considers that a source signal, or excitation $x(t)$, is the input to a filter or resonator H , the output of which is the signal $s(t)$ under consideration. Source–filter models have been used extensively for speech analysis, processing, and synthesis (Fant, 1970; Flanagan, 1972). Moreover, since the birth of electronic music, sources and filters have been used in many synthesizers and programs, but often in limited and relatively imprecise ways because of strong accuracy limitations that analog filters and controllers impose. Digital signal processing and estimation techniques have allowed many developments and applications of source–filter models, often inspired by speech research (Moorer, 1979; Rodet and Delatre, 1979; Rodet, 1980; Rodet and Depalle, 1986). Source–filter models can be considered as models for the production and the perception of musical sounds. On the one hand, by means of a rather simple signal representation, they take into account some physical properties of musical instruments. On the other hand, they also take into account some of the properties of human perception of musical sounds. Finally, they benefit from the huge developments of theory and applications in the field of digital signal processing.

For the majority of acoustic instruments, there is an exciter and a resonator. Under the assumption (which is not rigorously true but often reasonably valid) that the interaction between the exciter and the resonator modifies little of their individual behaviors, the sound production scheme can be simplified into a source–filter model. From the point of view of human perception, it has been shown that the ear is, above all, sensitive to the short-time spectrum of sounds and to the evolution of this spectrum. More precisely, we can look at the spectral characteristics of a sound on two frequency scales: On a small scale, the fine structure of a spectrum is characterized by sinusoidal partials, which appear as sharp peaks in the spectrum (see Fig. 5.1), and by non-sinusoidal components, which comprise a residual characterized by a random aspect of the spectrum. On a large frequency scale, the spectral envelope, which traces the connection between the peaks (see Fig. 5.1), indicates the broad structure of the spectrum (Rodet, 1984; Marin and McAdams, 1991), giving the distribution of energy in spectral bands to which the ear is very sensitive.

Partials with harmonically related frequencies are found in periodic sound signals with well-defined perceived pitch. Inharmonically distributed partials are characteristic of sounds with several pitches, such as multiphonics of musical instruments, or with no well-defined pitch, such as some metallic percussion sounds. Because peaks corresponding to sinusoidal partials appear at precise frequency values and in limited number, the corresponding spectrum is said to be discrete. Random aspects of the fine structure of the spectrum are characteristic of noise, such as noise due to air turbulence in a wind instrument. The corresponding spectrum is spread continuously on the frequency axis and is said to be continuous.

The importance of spectral envelopes for sound perception has been shown in many circumstances. For example, vowels are essentially perceived according to their spectral envelopes, and, in more detail, according to the frequency positions of some of the peaks of this envelope, called formants. This can be related to the fact, already mentioned in the introduction, that spectral envelopes of speech and singing voice sounds are quite independent of pitch. The vocal source is a train of pulses of air passing through the vibrating vocal folds while the resonator is the vocal tract that acts as a steady filter as long as the articulation remains unchanged. If, ignoring this production model, we transpose the vowel in Fig. 5.1 up by an octave by doubling all partial frequencies and performing additive resynthesis (as discussed in Section 1), the spectral envelope will also be transposed. Figure 5.2 shows this effect in comparison to Fig. 5.1, and the resulting signal sounds quite unnatural. This unnaturalness comes from the fact that the formants are shifted up one octave, corresponding to shrinking the vocal tract to half of its length, obviously not a natural behavior of the vocal tract.

To avoid this unnaturalness, the spectral envelope needs to be kept constant while the partials slide under it, taking on new values. This means that the amplitudes of the transposed partials should not be determined by the amplitudes of the

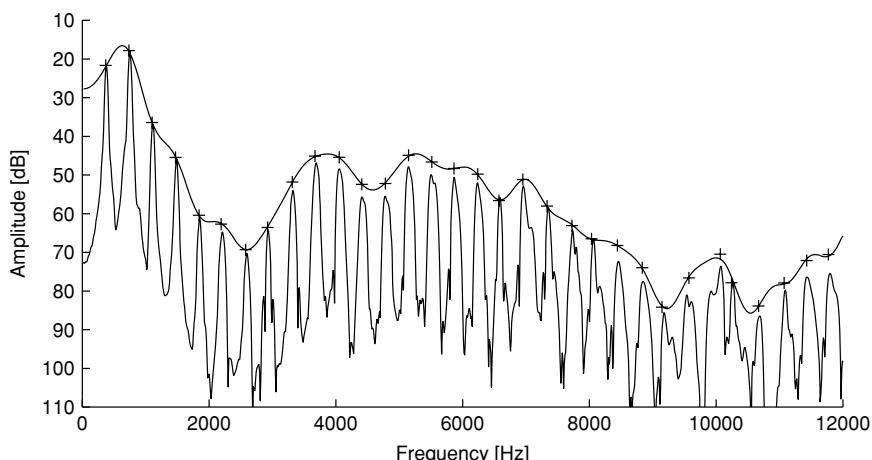


FIGURE 5.2. Transposition of voice without spectral envelope correction.

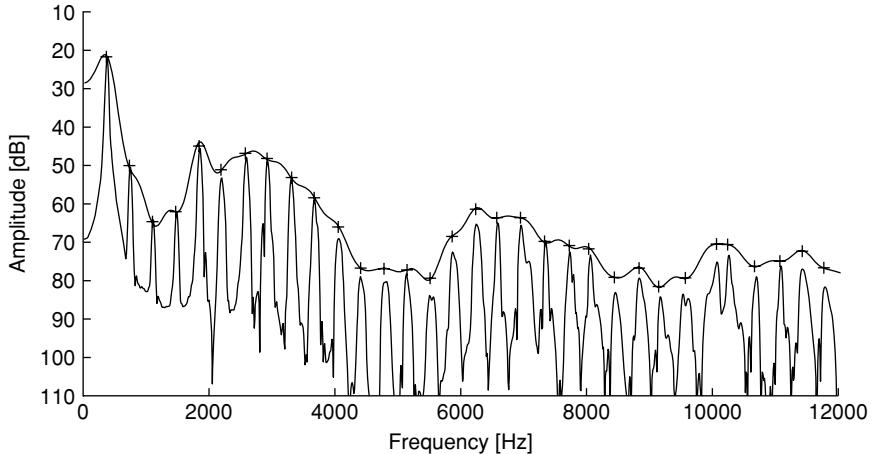


FIGURE 5.3. Transposition of voice with spectral envelope correction.

original partials, but rather by values of the spectral envelope samples at the frequencies of the transposed partials, as shown in Fig. 5.3. This way, only the partial frequencies are shifted, while the spectral envelope and thus the formant locations are preserved, and the resulting vowel sounds natural. For an easier comparison, Fig. 5.4 shows a superposition of the spectral envelopes of the transposed sound, with and without a constant spectral envelope, applied on a frequency grid spaced at 366 Hz, the fundamental frequency of the transposed sound. It can be clearly

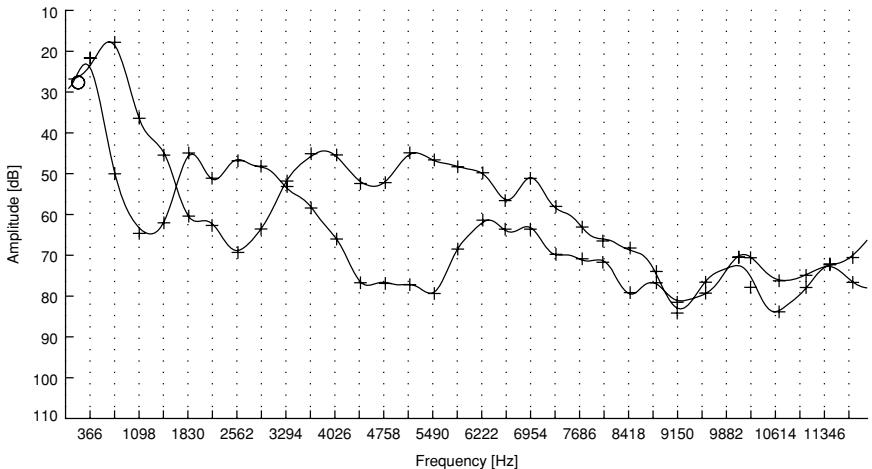


FIGURE 5.4. Transposition of voice: The spectral envelopes Figs. 5.2 and 5.3 are overlaid to show the effect of transposition with and without envelope correction.

seen that, with a fixed spectral envelope applied, each partial receives a different amplitude, and that the spectral envelopes are stretched versions of each other.

2.2 Source–Filter Models Represented by Spectral Envelopes

The properties concerning the production and perception of sounds, explained above, have motivated our study and use of source–filter models for sound synthesis. From production and perceptual viewpoints, the excitation or source signal is important because it implements the fine structure of the spectrum. The source signal consists of sinusoidal partials and noise components, usually considered as random signals. The source signal serves as input to a filter that implements the broad structure of the spectrum, again important from a perceptual point of view. Multiplication of a filter transfer function $H(\omega)$ by the source spectral envelope $X(\omega)$ results in the output sound's spectral envelope $S(\omega)$, i.e.,

$$S(\omega) = H(\omega) \cdot X(\omega). \quad (5.2)$$

Note that $H(\omega)$ is a complex function of radian frequency ω (where $f = \omega/2\pi$ is the frequency in Hz). Its magnitude $|H(\omega)|$ and its phase $\arg(H(\omega))$ are both important as shown below. Even though the term spectral envelope is commonly used for the magnitude only, we will also consider the phase spectral envelope or the complex spectral envelope, which includes both magnitude and phase.

In sinusoidal-additive-plus-residual source–filter synthesis, the source is a sum of sinusoids and random signals, which usually has a flat spectral envelope. Ideally this would be

$$X(\omega) = 1. \quad (5.3)$$

Therefore, the transfer function of the filter directly defines the spectral envelope of the resulting sound:

$$S(\omega) = H(\omega) \quad (5.4)$$

Equation (5.2) shows that, in the frequency domain, the spectral envelope application reduces to a simple multiplication of the amplitude of each source component, at a frequency ω , by the value of the spectral envelope at this frequency, $H(\omega)$. Consequently, the use of spectral envelopes directly in the frequency domain appears particularly cost effective and attractive (see Section 6.2).

As noted earlier, sinusoidal partials and non-sinusoidal components are usually created in voice and musical instruments by different mechanisms. Sinusoidal partials result from a nearly periodic process, such as the oscillation of a reed or the stick-slip cycle of a bow-string interaction. Non-sinusoidal components result from other mechanisms, such as air-flow-turbulence or friction noise. Therefore, the spectral envelopes of the sinusoidal and the non-sinusoidal parts of the signal have to be treated separately at all steps from estimation to synthesis. Indeed, before estimation can be applied, it is necessary to distinguish between the sinusoidal

and non-sinusoidal components. (This step is referred to as “voicing estimation” in the speech field and “sinusoidality” or “tonality” for audio signals in general.) (Griffin and Lim, 1985; Rodet et al., 1987; Peeters and Rodet, 1998). Then, spectral envelopes can be estimated for the sinusoidal and non-sinusoidal parts, with the estimation technique adapted individually to the properties of each part.

The resonator has a strong influence on the produced sound. In the case of the male voice, vocal tract resonances (formants) have mean separations of about 1000 Hz with 3-dB bandwidths on the order of 40–100 Hz or more. Consequently, the corresponding filter transfer function, or spectral envelope, is described by the general outline of the pattern of peaks in the short-time spectrum of the voice sound (Fig. 5.1). It also appears in the variations of the amplitudes of sinusoidal partials as a function of frequency (Maher and Beauchamp, 1990).

As an example of amplitude variations being caused by frequency variations, consider the signal of any sinusoidal partial $p(t)$ of a source with time-varying frequency $\omega(t)$ and constant amplitude y :

$$p(t) = y \sin \left(\int \omega(t) dt \right). \quad (5.5)$$

Then, we can approximate [under the assumption that $\omega(t)$ is slowly varying] the output of the filter $H(\omega)$ as the sinusoidal partial

$$\begin{aligned} q(t) &= y \cdot |H(\omega(t))| \cdot \sin \left(\int \omega(t) dt + \arg(H(\omega(t))) \right) \\ &= b(t) \cdot \sin \left(\int \omega(t) dt + \varphi(t) \right). \end{aligned} \quad (5.6)$$

We observe that the amplitude $b(t)$ of the output partial $q(t)$ is a function of its frequency

$$b(t) = y \cdot |H(\omega(t))|. \quad (5.7)$$

As a consequence, if the filter’s response H and the amplitude of the source are fixed with respect to time, the variation of the amplitude of an output partial as a function of its frequency simply traces a portion of the filter amplitude transfer function, i.e., the amplitude spectral envelope $H(\omega)$ multiplied by the input amplitude y . As an example, let us examine a singing voice signal whose fundamental frequency is varying around f_0 with a vibrato rate β and an excursion α . The instantaneous fundamental frequency is given by

$$F_0(t) = f_0 + \alpha \sin(2\pi\beta t). \quad (5.8)$$

Then, each harmonic partial with harmonic number k has approximately the frequency

$$f_k(t) = \frac{\omega_k}{2\pi} = kF_0(t) = kf_0 + k\alpha \sin(2\pi\beta t), \quad (5.9)$$

which varies between $kf_0 - k\alpha$ and $kf_0 + k\alpha$, and has the time-varying amplitude $y_k |H(\omega_k(t))|$, where y_k is the amplitude of the k th partial of the source. An example of this effect is shown in Fig. 5.5 for a baritone singing voice. The amplitude of each harmonic partial vs its frequency literally traces the spectral

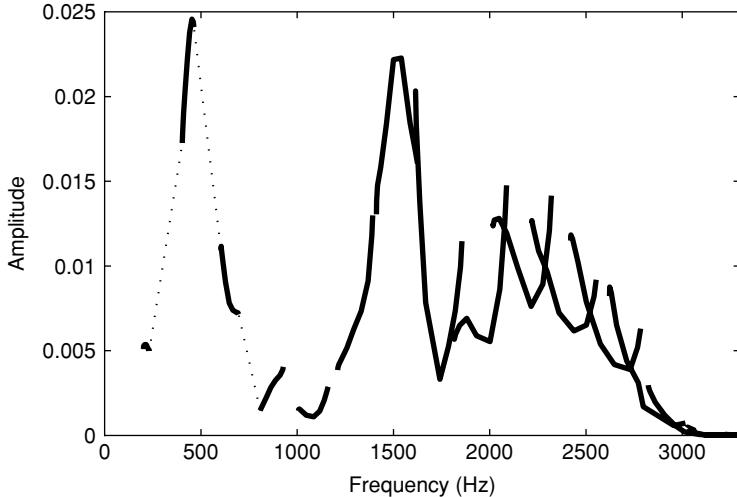


FIGURE 5.5. Amplitude-vs-frequency curves for partials 1 to 16 of a baritone singing voice. The dotted line has been added to better show the complete spectral envelope.

envelope. Note that this spectral envelope would be impossible to obtain from a short-time estimation commonly done on a signal window of some 20 ms. This technique for obtaining the spectral envelope of high pitched sounds has been used for the creation of a synthetic voice singing the The Queen of the Night's aria “*Der Hölle Rache*” in Mozart’s opera *The Magic Flute* (Bennett and Rodet, 1989). Automatic estimation using this technique can be performed by using discrete cepstrum estimation as explained in Section 3.

Finally, partial-indexed spectral envelopes can be defined for nearly harmonic sounds. Guitar and piano strings as well as acoustic tubes have resonances, or modes, whose center frequencies do not fall into an exact harmonic relationship. For a perfectly harmonic distribution with fundamental frequency f_0 , the center frequency of a mode is a linear function of its harmonic partial number k :

$$f_k = f_0 \cdot k \quad (5.10)$$

For nearly harmonic sounds, deviation-vs-partial can be defined as a function g of the partial number, as in

$$f_k = f_0 \cdot k + g(k), \quad (5.11)$$

or in terms of frequency, as in

$$f_k = f_0 \cdot k + g(f_0 \cdot k), \quad (5.12)$$

and can be considered to be a partial-indexed spectral envelope. Such spectral envelopes provide easy control of the distribution of the frequencies of partials. They can also be used in order to algorithmically define the frequencies of inharmonic

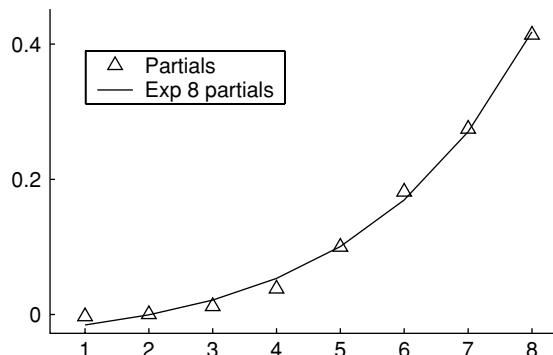


FIGURE 5.6. Relative frequency deviations from harmonicity for eight partials of a piano tone (courtesy J. P. Lambert). The continuous line is an exponential approximation of the relative frequency deviations. This is an example of a generalized envelope for frequency.

partials, such as the modes of metal plates (Benade, 1976; Potard et al., 1986) or stiff strings (Fig. 5.6).

2.3 Spectral Envelopes and Perception

Interestingly enough, the importance for human perception of the spectral envelope outlining the partial amplitudes can be easily demonstrated. It is remarkable that we are able to hear the change of the vocal tract shape of a singer even at very high pitch, i.e., when the spectral envelope would be impossible to obtain from a short-time estimation as explained above. This suggests that perception somehow deduces spectral envelope shape from partial frequency variations (similar to the method explained in Section 2.2). To demonstrate this effect, McAdams and Rodet (1988) tested the perception of synthetic sung vowels with similar spectral envelopes. These envelopes differed only by the magnitude of the envelope segment traced by the second partial when vibrato was applied, crossing exactly at the mean frequency of the second partial. In the absence of vibrato, the value at this frequency was the same for the two envelopes, so that the sounds were identical. However, for one sound the spectral envelope in the neighborhood of the second partial increased with frequency, while, for the other sound, it decreased with frequency. Listening tests showed that a vibrato with an excursion of one percent of the fundamental frequency was sufficient to hear these sounds as two different vowels.

Resonances of a violin or a cello body are more densely distributed and narrower than those of the vocal tract. In consequence, the corresponding filter transfer function or spectral envelope does not appear so directly in the general shape of the short-time spectrum of the violin sound. But its influence on partial amplitudes is similar to that of the voice (Beauchamp, 1974; Mellody and Wakefield, 1997, 2000; Dubnov and Rodet, 1997).

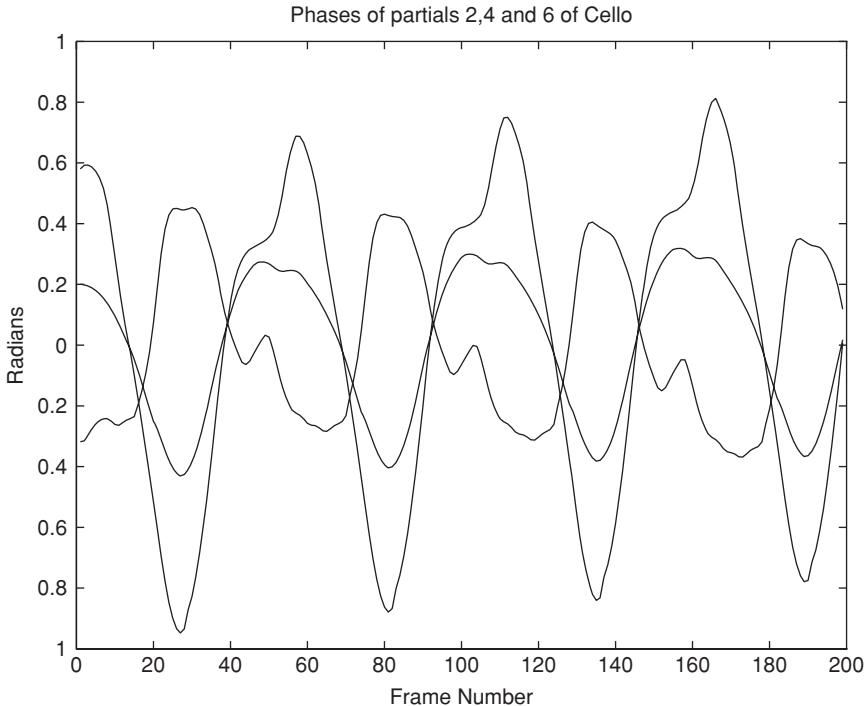


FIGURE 5.7. Relative phase variations of harmonic partials 2, 4, and 6 corresponding to their frequency variations (vibrato) of a cello sound (from Dubnov and Rodet, 1997).

Resonances also affect phase variations (Beauchamp, 1974). From Eq. (5.6), the relative phase ϕ of an output partial similarly varies with its frequency ω according to

$$\varphi(t) = \arg(H(\omega(t))). \quad (5.13)$$

In the vicinity of a resonance center frequency, $\arg(H(\omega))$ changes rapidly with ω . Therefore, the phase of an output partial also changes rapidly with ω , and its variation depends on the position of the resonance center frequency relative to the interval on which ω varies. Different partial frequencies $\omega_k(t)$ may exhibit phase variations that appear uncoupled (Dubnov and Rodet, 1997). An example from the analysis of a cello sound with vibrato is shown in Fig. 5.7, where phase variations of harmonic partials number 2, 4, and 6 are superposed. In a simulation with a source–filter model, similar phase variations have been obtained for the phase of output partials when a similar amount of vibrato is applied (Fig. 5.8). These phase variations are easily perceived and are important features that allow the distinction of sounds from different instruments, such as cello (uncoupled phases) and trumpet (coupled phases) as detailed in Dubnov and Rodet (1997). The importance of the phase spectral envelope is also demonstrated when an inadequate phase spectral

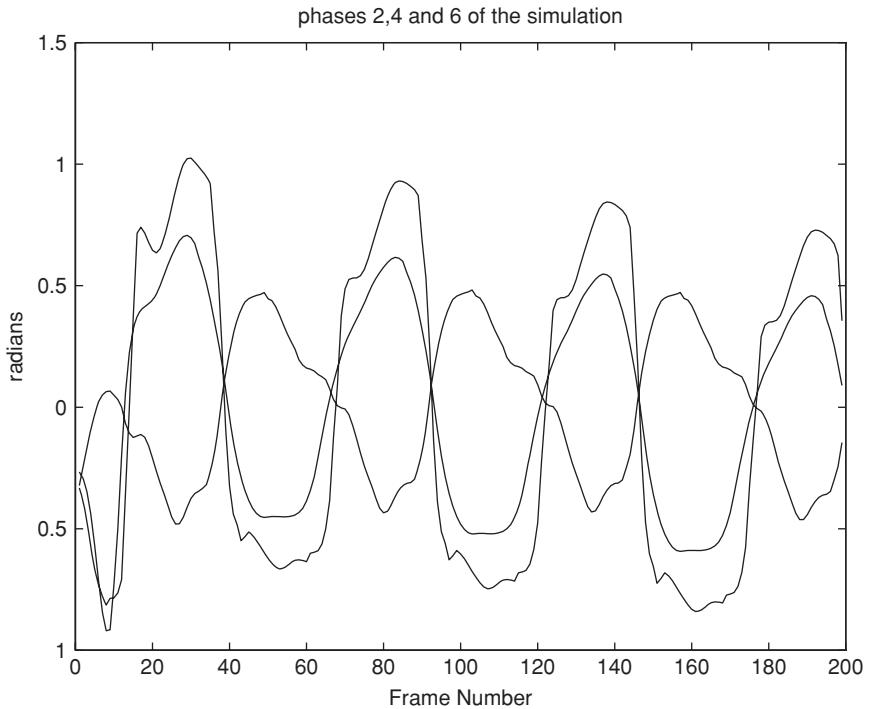


FIGURE 5.8. Relative phase variations of harmonic partials 2, 4, and 6 corresponding to frequency variations (vibrato) of a source–filter model based on the cello sound of Fig. 5.7 [from Dubnov and Rodet (1997)].

envelope is obtained by an estimation technique that does not reflect the underlying physical model (see Giron, 1990, pp. 40–47).

2.4 Source and Spectrum Tilt

In the preceding discussion we have noted that if the source has a flat spectral envelope [i.e., $X(\omega) = 1$], the transfer function of the filter directly defines the spectral envelope of the resulting sound [Eq. (5.4)]. However, acoustic instruments, such as the trumpet and the voice, have source spectra that can vary greatly, especially according to the intensity at which the instrument is played (Beauchamp, 1975, 1980; Benade, 1976; Bennett and Rodet, 1989; Fletcher and Tarnopolsky, 1999). The louder the sound, the stronger the high-frequency components become. This is often referred to as a downward spectrum tilt (or slope) that decreases with loudness. Therefore, instead of being flat, the source should be represented by a spectrum shape which is a function of the intended intensity level l :

$$X(\omega) = \Xi(\omega, l). \quad (5.14)$$

where Ξ is a function whose tilt decreases with increasing l . Then, the spectral envelope is the product:

$$S(\omega) = \Xi(\omega, l) \cdot H(\omega). \quad (5.15)$$

This formula demonstrates how the concept of spectral envelope can permit an easy control of both the resonator and the exciter of simulated instruments, real or imagined.

2.5 Properties of Spectral Envelopes

Three important properties for spectral envelopes are:

Envelope fit: A spectral envelope is a curve that envelopes the spectrum, i.e., it wraps tightly around the magnitude spectrum, linking the peaks (for the sinusoidal or discrete part of the spectrum) or passing close to the maxima (for the residual or continuous part of the spectrum).

Smoothness: A certain smoothness of the curve is required. This means that the spectral envelope does not oscillate too much, but gives a general idea of the distribution of energy of the signal over frequency.

Adaptation to fast spectrum variations: A spectral envelope is defined for a short segment of signal as the envelope of a short-time spectrum (STS). When the signal's STS varies rapidly from one analysis frame to the next, the spectral envelope should precisely follows its fast-time variation. (See Section 3.2.)

Examples of musical instrument and speech spectra with overlaid spectral envelopes are shown in Figs. 5.9–5.13.

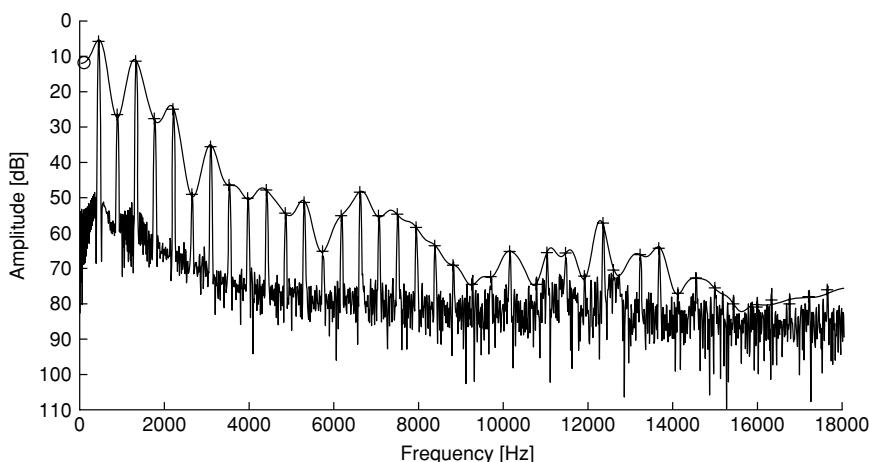


FIGURE 5.9. Spectrum and spectral envelope of a clarinet sound.

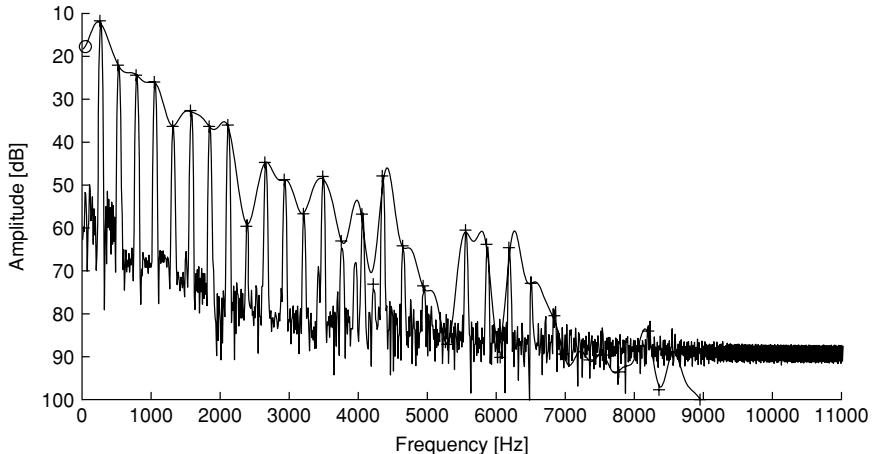


FIGURE 5.10. Spectrum and spectral envelope of a piano sound.

3 Spectral Envelope Estimation Methods

Because, in general, the short-time spectrum of a sound is not stationary, the corresponding spectral envelope varies with time. Estimation of a time-varying spectral envelope is usually obtained by estimating a spectral envelope $S_i(\omega)$ on a short-time window (typically 5–40 ms) centered on a time t_i , then advancing the window by a fraction of its size, δ , to the time $t_{i+1} = t_i + \delta$ where a new estimation $S_{i+1}(\omega)$ is done, and so on. Such a repetitive operation is called sliding window analysis.

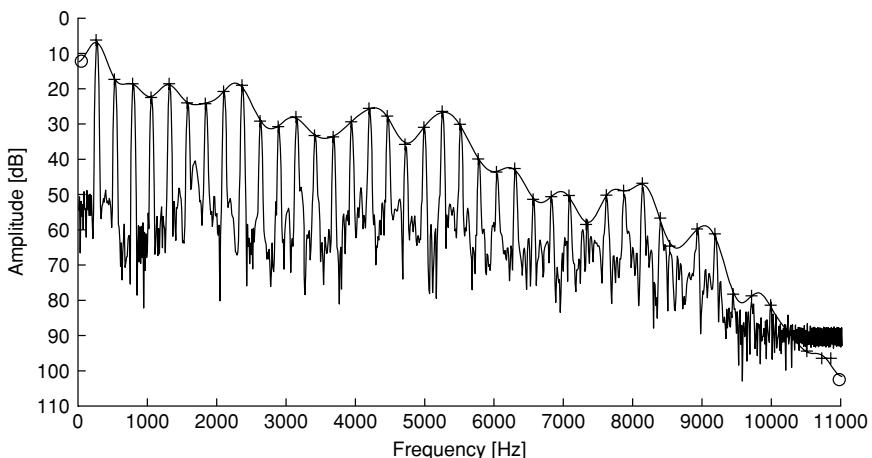


FIGURE 5.11. Spectrum and spectral envelope of a violin sound.

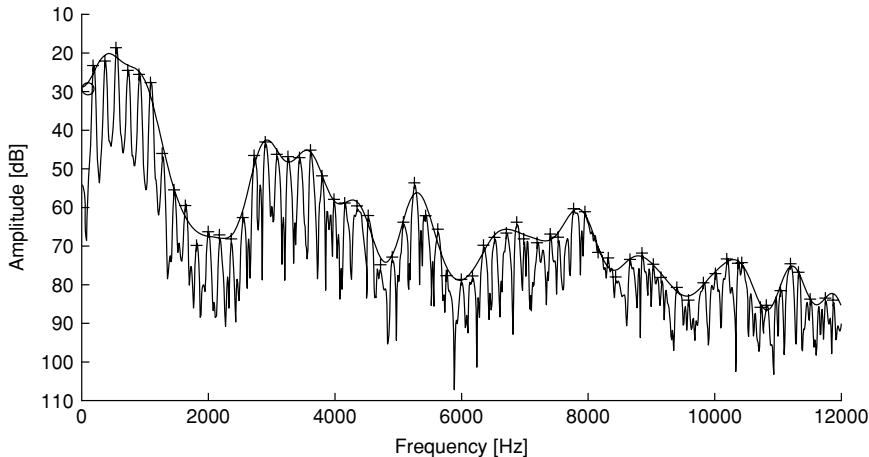


FIGURE 5.12. Spectrum and spectral envelope of the vowel /a/.

The general requirements for spectral envelope estimation are introduced in Section 3.1. Then various estimation methods, autoregression (AR), cepstrum, and discrete cepstrum, are described in the following Sections 3.2–3.5. These descriptions will proceed from a non-formal introduction (what the algorithm does) to a detailed formal development (how it is done).

All of the methods for spectral envelope estimation described in this section have their strong and weak points, depending on the signal and the needs of the user. In particular, some methods are better suited for sinusoidal components (the discrete part of the spectrum) and others are better suited for non-sinusoidal components

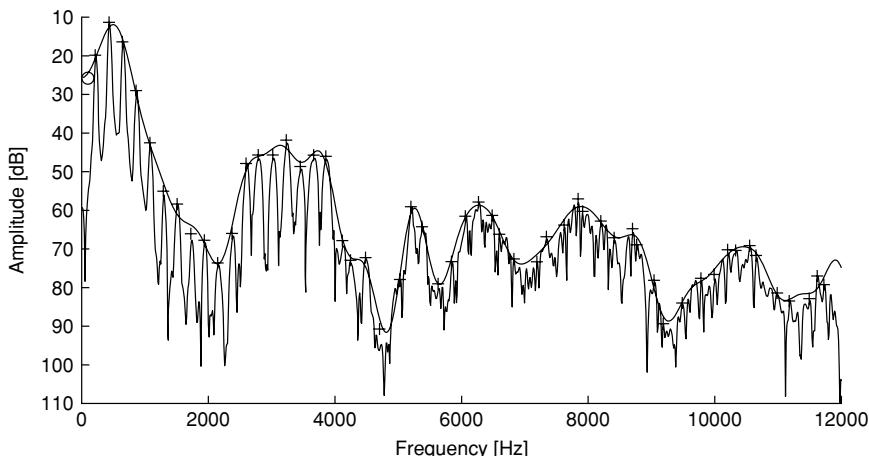


FIGURE 5.13. Spectrum and spectral envelope of the vowel /o/.

(the continuous part of the spectrum). Also, a method is usually more effective if the analysis model takes into account the physical model which has produced the sound.

3.1 Requirements

The requirements for spectral envelope estimation are basically the fulfillment of the three properties of spectral envelopes described in Section 2.5, with some additions and more details.

Exactness: For each sinusoidal partial, the spectral envelope should precisely intersect the point, in the frequency–amplitude plane, defined by the spectral magnitude maximum associated with that partial. In Section 2.5 this was called the envelope fit property, in that the spectral envelope wraps tightly around the magnitude spectrum, linking the peaks. The required degree of exactness is determined by the perceptual abilities of human audition. In the lower frequency range, humans can distinguish differences in amplitude as small as 1 dB (Moore, 1997). For higher frequencies, the sensitivity is a little lower. It may not be necessary to link every peak in a group of peaks close to each other in the upper frequency range. Then, the spectral envelope should find a reasonable intermediate path, e.g., through the center-of-gravity of peaks in each frequency band such as a critical band or a fraction of it. Finally, a spectral envelope should also precisely follow rapid variations of the signal spectrum in time.

Robustness: The estimation method has to be applicable to a wide range of signals with very different characteristics, from high-pitched harmonic sounds with their widely spaced partials to noisy sounds or mixtures of harmonic and noisy sounds. Very often, problems come from additive analysis when very-low-amplitude peaks are identified as sinusoidal partials, although they pertain to the residual noise or even to the noise floor of the recording. This is also a question of choosing the right parameters for spectral analysis, e.g., the threshold for accepting partial amplitudes.

Smoothness: A certain smoothness is required. This means that the spectral envelope must not oscillate too much over its frequency range, but rather it should give a general idea of the distribution of the energy of the signal over frequency. This translates to a restriction on the slope of the envelope (given by its first derivative), which may be dependent on context.

3.2 Autoregression Spectral Envelope

Autoregression (AR) estimation is a well-known digital-signal-processing method (Oppenheim and Schafer, 1975; Oppenheim, 1978; Markel and Gray, 1980). It is widely used for speech transmission and compression under the name linear predictive coding (LPC). Special properties of the method allow it to be used for spectral envelope estimation. The idea behind AR analysis is to represent each sample of a signal $s(n)$ in the time-domain by a linear combination of the preceding

values $s(n - p - 1)$ through $s(n-1)$ (Kay, 1988). The value p is called the order of the AR model. The estimated value $\hat{s}(n)$ is computed from the preceding values using the AR coefficients (also called predictor-coefficients or LPC-coefficients) a_i as follows:

$$\hat{s}(n) = \sum_{i=1}^p a_i s(n-i) \quad (5.16)$$

For each analysis frame, the coefficients a_i are computed in order to minimize, in some sense, the prediction error, or LPC-residual, defined by $e(n) = \hat{s}(n) - s(n)$.

When the residual signal $e(n)$ is minimized, an analysis filter A given by the Z-transform transfer function,

$$A(z) = 1 - \sum_{i=1}^p a_i z^{-i}, \quad (5.17)$$

attempts to attenuate the frequency components in the input signal $s(n)$ that have high magnitudes in order to achieve a maximally flat spectrum (this is sometimes called “whitening” a spectrum). The corresponding synthesis filter is the inverse of the analysis filter, given by

$$\frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^p a_i z^{-i}}. \quad (5.18)$$

This filter restores the amplitudes of the frequency components that have been attenuated by the transfer function of the analysis filter. It is an all-pole filter because its transfer function is defined by a rational function with no zeros in the numerator and p zeros (called *poles*) in the denominator. Most of these poles come in complex-conjugate pairs resulting in the magnitude of the filter transfer function showing several peaks corresponding to these pairs. An LPC analysis/synthesis system block diagram is shown in Fig. 5.14.

As the analysis filter attempts to flatten the spectrum, it adapts to it in such a way that its inverse filter describes the spectral envelope of the signal. As the filter order is increased (i.e., more poles become available), the approximation of the spectral envelope becomes more precise. The envelope obtained with a low order will nevertheless reflect the rough distribution of energy in the spectrum. This can be seen in Fig. 5.15.

Several methods exist for the actual evaluation of the predictor coefficients to minimize the prediction error, such as the autocovariance method and the

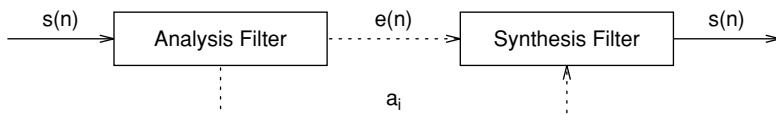


FIGURE 5.14. LPC-analysis and synthesis system used for data transmission.

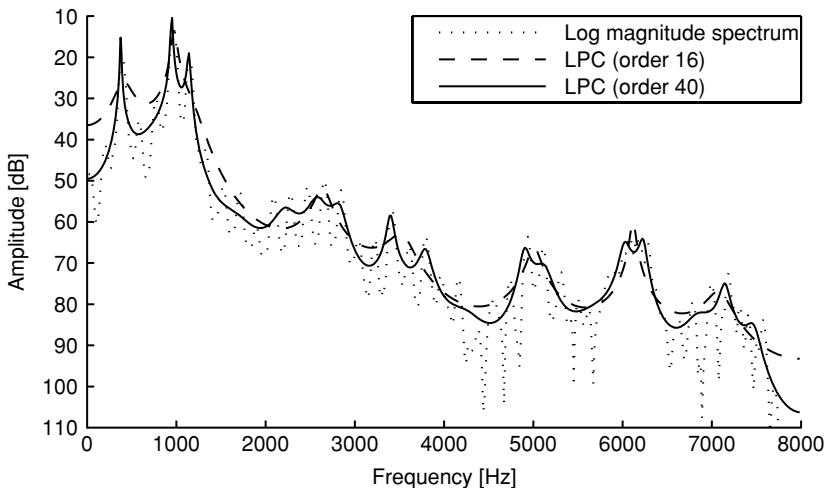


FIGURE 5.15. The LPC spectral envelope for a Mongolian chant spectrum. As the order increases, more poles are available for the model and the spectral envelope follows the spectrum in more detail.

autocorrelation method (Makoul, 1975; Markel and Gray, 1980; Kay, 1988). The autocorrelation method is more widely used and can be efficiently implemented using Durbin–Levinson recursion (Markel and Gray, 1980). We will not elaborate on the methods here, because they are amply described in the literature. However, we will give a typical example used for envelope estimation of musical signals. In order to obtain envelopes as exactly as possible, which adapt rapidly to signal changes, an adaptive method with a particular time-window has been developed by Rodet et al. (1987) and Depalle (1991). They used a recursive-adaptive lattice LPC method proposed by Vishwanathan and Makhoul (1978) combined with the left half of a Blackman–Harris window (Harris, 1978). To accurately model fast transitions, such as in consonants, it is necessary that the LPC’s whitening filter adapt itself as fast as possible. Classically, in adaptive techniques, an exponential sliding window is applied on the error signal. The value of the exponential decay coefficient is usually chosen close to 1.0 (typically 0.995 at 16 kHz).

However, one can sometimes observe, especially when the energy of the signal is abruptly attenuated (for instance, in occlusives), that the filter tends to maintain characteristics of the past, so that the synthesized signal exhibits a kind of reverberant quality. Conversely, if the exponential decay coefficient is too small, the optimization criterion does not remain valid, and the spectral envelope is not representative of the power spectral density of the signal. This motivates the use of a window with better properties for the analysis: i.e., one which is close to the value 1.0 on the right end and smoothly damped to zero on the left end. This window is applied either to the sound signal itself or to the error signal. In the former case, the windowed signal is analyzed by the method cited above. In the latter case, the analysis method is itself modified because the optimization criterion is modified.

This leads to an extremely accurate estimation, even on a segment as short as 20 ms.

In the course of evaluation of the predictor-coefficients, an intermediate set of parameters, the reflection coefficients k_i are obtained, which, in fact, correspond to the reflection of acoustic waves at the boundaries between successive sections of an acoustic tube. These coefficients have advantages for synthesis, and can be interpolated without stability problems for the resulting synthesis filter.

Various other parameter sets exist (Markel and Gray, 1980), e.g., the roots of the analysis filter $A(z)$ and log area ratios (LAR), i.e., the logarithm of the ratios of the areas of the sections of the acoustic tube model given by

$$\frac{A_{i+1}}{A_i} = \frac{1 - k_i}{1 + k_i}. \quad (5.19)$$

Also, there are line spectral pairs (LSP) (Itakura, 1975; Soong and Juang, 1984) and others. Because it is possible to convert between these parameter sets, they do not need to be considered separately for representation. (See also Section 4.2.)

3.2.1 Disadvantage of AR Spectral Envelope Estimation

A disadvantage of the AR method for sound analysis of signals with a limited number of dominant partials is that even though the method will tend to envelope the spectrum as tightly as possible, under certain conditions it will descend down to the level of residual noise in gaps between adjacent partials. As shown in Fig. 5.16,

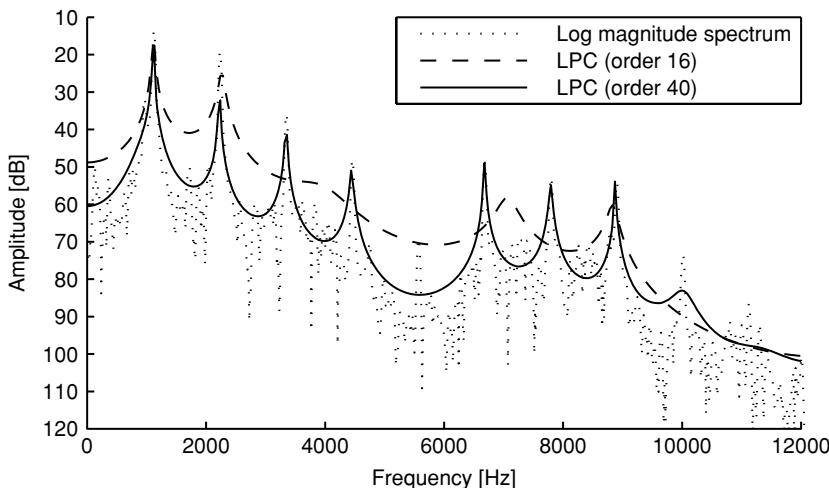


FIGURE 5.16. Problematic behavior of the LPC spectral envelope estimation when the partials are spaced far apart. The order-40 LPC spectral envelope reaches most of the peaks but “hangs down” in between, while the order-16 version reaches only two peaks exactly and describes the average between peaks and residual noise for the rest.

this will happen whenever the space between partials is large, as in high-pitched sounds, or when the order is high enough, i.e., when there are enough poles to correspond to every partial peak. To tackle this difficulty, some methods have been proposed, such as discrete all-pole modeling (El-Jaroudy and Makhoul, 1991; Gallas and Rodet, 1991b). However, these methods have not yet been applied widely in practice.

3.3 Cepstrum Spectral Envelope

To explain the general idea of the cepstrum method used for spectral envelope estimation, two approaches are possible. First, we can envision obtaining the spectral envelope from a Fourier magnitude spectrum by smoothing its curve to eliminate rapid fluctuations. This can be accomplished by applying a low-pass filter to the spectrum, interpreted as a signal, thus letting only the slow fluctuations (low-frequency oscillations of the curve) remain. Second, considering the signal as the convolution of a source signal with a filter impulse response, we can attempt to separate the source spectrum from the filter transfer function, which we assume is a good estimate of the spectral envelope.

According to the source–filter model introduced in Section 2, a signal $s(n)$ can be expressed in terms of the convolution of a source or excitation signal $x(n)$ and the impulse response of a filter $h(n)$ as

$$s(n) = h(n) * x(n). \quad (5.20)$$

In the frequency domain, this convolution becomes the multiplication of the respective Fourier transforms:

$$S(\omega) = H(\omega) \cdot X(\omega). \quad (5.21)$$

Taking the logarithm of the absolute value of the Fourier transforms (the magnitude spectra), the multiplication of Eq. (5.21) is converted to an addition:

$$\log |S(\omega)| = \log |H(\omega)| + \log |X(\omega)|. \quad (5.22)$$

If we now apply an inverse Fourier transform F^{-1} to the log magnitude spectrum, we get the frequency distribution of the fluctuations in the curve of the log magnitude spectrum, which is called the *cepstrum* (Bogert et al., 1963):

$$c = F^{-1}(\log |S|) = F^{-1}(\log |H|) + F^{-1}(\log |X|). \quad (5.23)$$

The independent variable of c is called “quefrency,” and because we have taken two Fourier transforms of the original signal, it is in units of time. Also, because we are using discrete transforms, c is defined at discrete quefrequencies with values $\dots, 0.5c_{-k}, \dots, 0.5c_{-1}, c_0, 0.5c_{-1}, \dots, 0.5c_k, \dots$, where $c_{-k} = c_k$. The $\{c_k, k = 0, 1, \dots\}$ are called the cepstral coefficients.

The independent contributions of $H(\omega)$ and $X(\omega)$ to c are easy to see from Eq. (5.23). Under the reasonable assumption that the source spectrum has only rapid fluctuations, its contribution to c is concentrated in its higher-quefrency regions, while the contribution of H , due to its slow fluctuations, is therefore concentrated

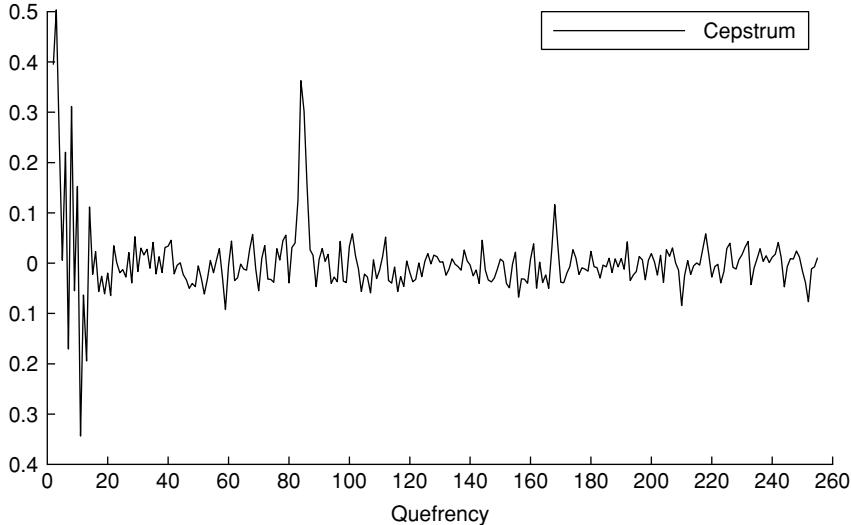


FIGURE 5.17. The cepstrum c of the Mongolian chant magnitude spectrum. The *quefrency* is given as the index k of the cepstral coefficients c_k (the abscissa could be labeled in time as well). It can be seen that most information is concentrated in the left part, up to order 20. The sharp peak at about index 84 corresponds to the distance between the regularly spaced lobes in the log magnitude spectrum corresponding to the harmonic partials. In other words, it is due to the contribution of the source spectrum and indicates the fundamental frequency of the sound.

only in the lower part of c , as can be seen in Fig. 5.17. Thus, the separation of the two components becomes quite trivial.

Normally we retain only the first $p + 1$ cepstral coefficients c_0, c_1, \dots, c_p , and the cepstrum is said to be of order p . These coefficients represent the low-quefrency components, which we assume are due to the slowly changing fluctuations of $|H(\omega)|$. Note that c_0 represents the average energy on the signal frame. By computing the forward Fourier transform of the truncated c , the spectrum $\log|S(\omega)|$ becomes smoothed, resulting in a valid spectral envelope. This smoothing effect can be seen in Fig. 5.18.

Interestingly, the spectral envelope may be obtained directly from the cepstral coefficients. First, let us define the frequencies f_i at which values of the envelope are to be obtained (the bins of the envelope). Usually, one wants M equidistant bin frequencies up to the Nyquist frequency $f_s/2$:

$$f_i = i \frac{f_s/2}{M}, \quad i = 1, 2, \dots, M. \quad (5.24)$$

Then, the equivalent angular frequencies are

$$\omega_i = 2\pi \frac{f_i}{f_s} = \frac{i}{M}\pi. \quad (5.25)$$

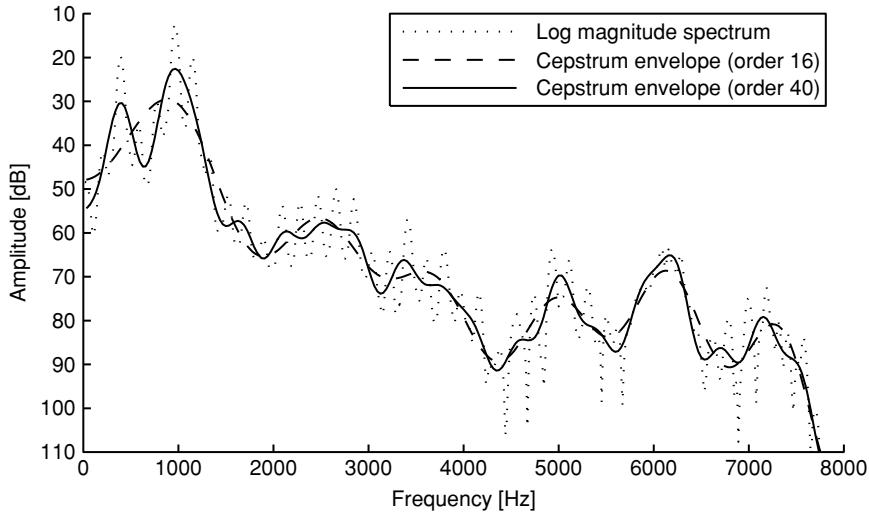


FIGURE 5.18. The cepstrum spectral envelope for the Mongolian chant spectrum. With increasing order, more of the rapid fluctuations of the magnitude spectrum will show up.

Because the truncated cepstrum is an even function, the forward transform consists only of cosine terms. Also, the exponential function can be used to nullify the log. Therefore, the spectral envelope values $H_i = |H(\omega_i)|$ for frequencies f_i are given by

$$H_i = \exp \left(\sum_{k=0}^p c_k \cos(k\omega_i) \right). \quad (5.26)$$

While, in general, the c_k must be recomputed for every analysis frame, the cosine terms can be precomputed as a $(M, p+1)$ matrix Φ with elements

$$\phi_{ik} = \cos(k\omega_i) = \cos(ki\pi/M) \quad (5.27)$$

so that, as a vector-matrix equation, Eq. (5.26) becomes

$$H = \exp(\Phi c). \quad (5.28)$$

3.3.1 Disadvantages of the Cepstrum Method

There are two disadvantages of the cepstrum method of spectral envelope estimation, described in the previous subsection. First, as this method essentially carries out low-pass filtering of the magnitude spectrum interpreted as a signal, it actually averages out the fluctuations of this curve. The effect can be seen in Fig. 5.19, where the envelope curves fall well below the peaks. What we want is for the envelope curves to link the peaks of the spectrum (cf. Section 3.1). Second, similar to the AR method, when analyzing harmonic sounds with partials spaced far apart, as is the case for high-pitched sounds, high-ordered cepstral envelopes follow the

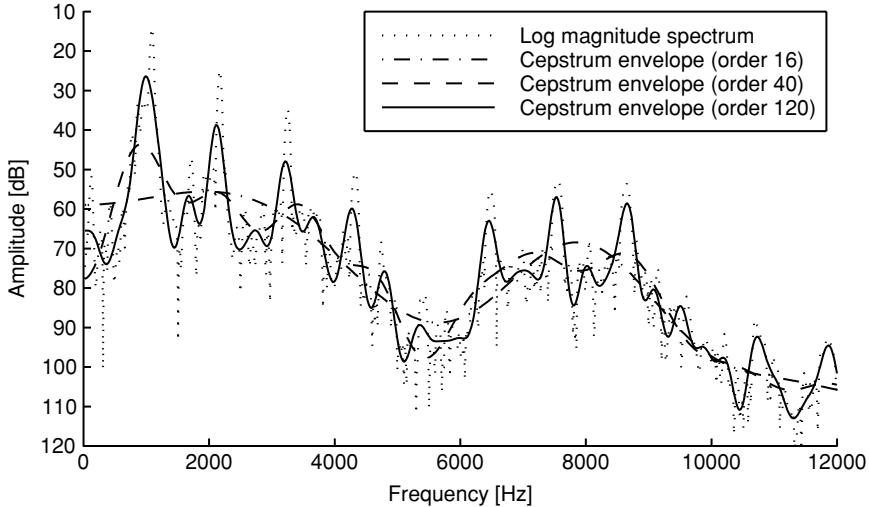


FIGURE 5.19. Problematic behavior of the cepstrum spectral envelope estimation when partials are spaced far apart.

spectra down to the residual noise level in gaps between adjacent partials. Again, see Fig. 5.19 for an example of this behavior.

3.4 Discrete Cepstrum Spectral Envelope

Contrary to the last two methods discussed, AR and cepstrum, which are computed from uniformly sampled representations of the signal, the discrete cepstrum spectral envelope (Galas and Rodet, 1990) is computed from nonuniformly spaced discrete points in the frequency domain. These points correspond to the spectral peaks of a sound, which most often correspond to sinusoidal partials in the sound. As described at the ends of Sections 3.2 and 3.3, the AR and cepstrum spectral envelopes both exhibit the problem of descending down to the level of residual noise between partials that are spaced too far apart, as can be seen in Fig. 5.16 and Fig. 5.19. The discrete cepstrum, on the other hand, adheres only to the underlying sinusoidal partials and generates a smoothly interpolated curve that links the partial peaks, as shown in Fig. 5.20.

Let us briefly explain the discrete cepstrum estimation method (Galas and Rodet, 1991a, 1991b). First, a given set of N spectral peaks (partials) with amplitudes S_i at frequencies ω_i , $i = 1, \dots, N$, defines a magnitude spectrum $S(\omega)$ as

$$S(\omega) = \sum_{i=1}^N S_i \delta(\omega - \omega_i), \quad (5.29)$$

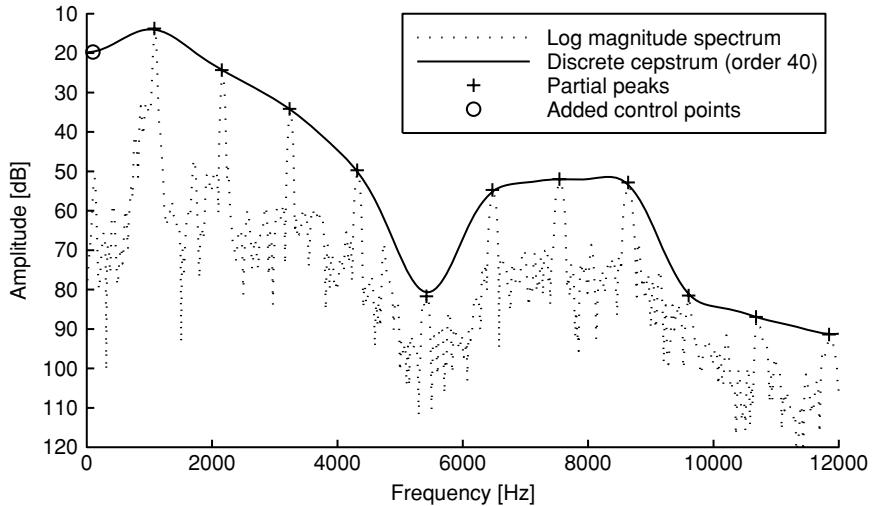


FIGURE 5.20. Example of a discrete cepstrum spectral envelope. (The control points are explained at the end of Section 3.5.2.)

where $\delta(\omega - \omega_i)$ is the Dirac delta function. If we consider $S(\omega)$ to be the result of the product

$$S(\omega) = H(\omega) \cdot X(\omega), \quad (5.30)$$

where $X(\omega)$ is the source spectrum with amplitudes X_i at the same frequencies ω_i as for S , then we can express

$$X(\omega) = \sum_{i=1}^N X_i \delta(\omega - \omega_i). \quad (5.31)$$

Also, by generalization of Eq. (5.26), let us take $H(\omega)$ to be the transfer function of a filter modeled by

$$H(\omega) = \exp \left(\sum_{k=0}^p c_k \cos(k\omega) \right), \quad (5.32)$$

where p is the order of the discrete cepstrum.

We only need to find filter parameters c_k that minimize the quadratic error E between samples of the log spectrum of the model and those of the signal. This error criterion is developed from the idea of a spectral distance:

$$\begin{aligned} E &= \sum_{i=1}^N (\log(X_i H(\omega_i)) - \log(S_i))^2 \\ &= \sum_{i=1}^N \left(\sum_{k=0}^p c_k \cos(k\omega_i) + \log(X_i) - \log(S_i) \right)^2. \end{aligned} \quad (5.33)$$

To achieve this minimization, we use least squares (Press et al., 1992) to formulate a matrix equation

$$\Psi c = b, \quad (5.34)$$

where Ψ is a square matrix of size $p + 1$, given by

$$\psi_{kl} = \sum_{i=1}^N \cos(k\omega_i) \cdot \cos(l\omega_i), \text{ with } k, l = 0, \dots, p, \quad (5.35)$$

c , the column vector of the filter parameters, is

$$c = \begin{bmatrix} c_0 \\ \vdots \\ c_p \end{bmatrix}, \quad (5.36)$$

and b is a column vector given by

$$b_k = \sum_{i=1}^N \log \left(\frac{S_i}{X_i} \cos(k\omega_i) \right), \quad \text{with } k = 0, \dots, p. \quad (5.37)$$

Most often we will only have measurements of the S_i available, in which case we can, without loss of generality, set $X_i = 1, \forall i$, thus simplifying Eqs. (5.33) and (5.37). However, we retain X_i in those equations for those cases where source spectrum measurements are available, as in the case of glottal source measurements.

The matrix Ψ can be computed very efficiently by using an intermediate vector T given by

$$t_k = \frac{1}{2} \sum_{i=1}^N \cos(k\omega_i) \quad \text{with } k = 0, \dots, 2p, \quad (5.38)$$

so that

$$\psi_{kl} = t_{k+l} + t_{|k-l|}. \quad (5.39)$$

The matrix given by Eq. (5.34) can be efficiently solved applying Cholesky decomposition (Press et al., 1992), which factors Ψ such that

$$\Psi = U D U', \quad (5.40)$$

where U is an inferior triangular matrix whose diagonal elements are 1, U' is the transposed matrix of U (i.e., it is a superior triangular matrix), and D is a diagonal matrix. Now the matrix equation can be solved by simple substitution and division.

The asymptotic complexity of the discrete cepstrum method described above is $O(Np + p^3)$, which means that the number of partials N is not a big concern, because the complexity is linear in N , but that the order p has to be kept as small as possible, because of its cubic influence.

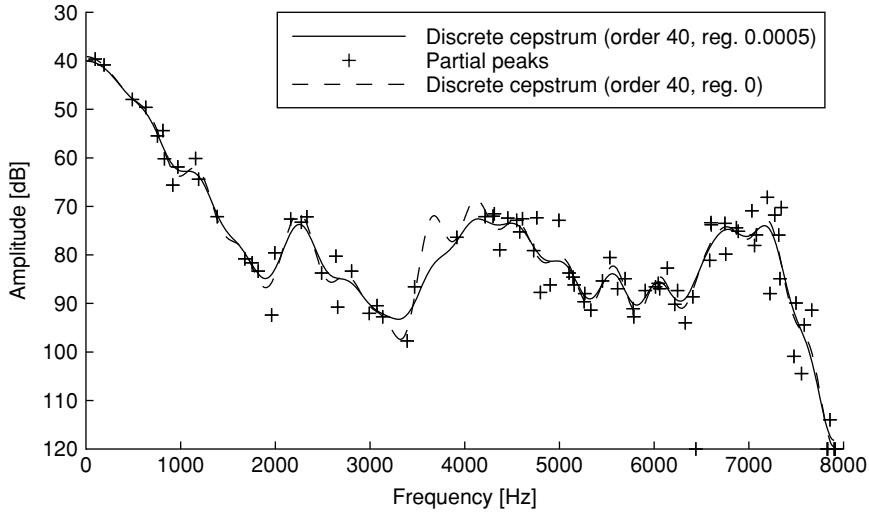


FIGURE 5.21. The effect of regularization: The unregularized discrete cepstrum spectral envelope ($\rho = 0$) shows a large hump between 3500 and 4000 Hz, whereas the curve regularized by a factor $\rho = 0.0005$ behaves nicely.

3.5 Improvements on the Discrete Cepstrum Method

3.5.1 Regularization

The technique of regularization, developed by Campedel-Oudot et al. (2001), improves the smoothness of the spectral envelope. Its idea is to penalize a spectral envelope slope that is too steep by adding a regularization matrix ρR to the matrix Ψ , defined by Eq. (5.35), where ρ is a regularization coefficient and R is a square matrix of size $p + 1$, whose diagonal is defined by

$$r_{kk} = 8\pi^2(k - 1)^2. \quad (5.41)$$

Then, the discrete cepstrum algorithm proceeds as in Section 3.4.

The effect of regularization can be seen in Fig. 5.21. The disadvantage of regularization is that sometimes a steep slope is necessary to reach a single extremely situated peak, as with the low peak at about 3400 Hz in Fig. 5.21. With regularization, the curve falls short of reaching it.

3.5.2 Stochastic Smoothing (the Cloud Method)

The cloud method developed by Galas and Rodet (1990) is a way to avoid abnormal behavior of the spectral envelope that sometimes results from the discrete cepstrum algorithm. The method generates a cloud of points around each partial on the frequency–amplitude plane to give the discrete cepstrum algorithm more freedom trying to fit a curve that links all the partials. Added points are displaced from each original point (ω_i, S_i) at frequency ω_i and amplitude S_i by a frequency shift β

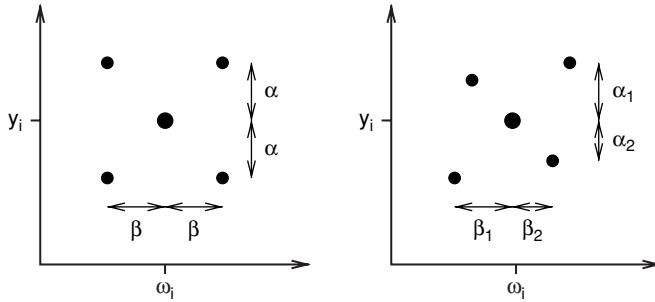


FIGURE 5.22. The cloud of four points around the original partial generated by stochastic smoothing with indifferent slope (left) and with a hint for a rising slope (right).

and an amplitude factor α as shown in Fig. 5.22 (left). Furthermore, if additional information is known, the shape of the cloud can be used to influence the behavior of the spectral envelope, as shown in Fig. 5.22 (right). For example, if a point is known to be situated within the rising slope of a formant, the spectral envelope could be influenced to also prefer a rising slope. However, to avoid too strong a deviation of the spectral envelope from the original point, weighting is introduced in the discrete cepstrum algorithm to attenuate the influence of the added points with respect to the original point. For example, the original point could be weighted with a factor of 5, whereas the added points would be weighted with a factor of 1, as expressed by the thickness of the points in Fig. 5.22. In general, weighting factors w_i can be introduced into the error formula of Eq. (5.33):

$$E = \sum_{i=1}^N w_i (\log(X_i H(\omega_i)) - \log(S_i))^2. \quad (5.42)$$

Thus, Eq. (5.35) becomes

$$\psi_{kl} = \sum_{i=1}^N w_i \cos(k\omega_i) \cos(l\omega_i), \quad (5.43)$$

and Eq. (5.37) is changed to

$$b_k = \sum_{i=1}^N w_i \log \left(\frac{S_i}{X_i} \cos(k\omega_i) \right). \quad (5.44)$$

From a more formal point of view, the cloud method is in fact a replacement of each original spectral peak, (ω_i, S_i) , by a probability distribution $\pi_i(\omega, S)$. This is necessary because of the uncertainty of the precise position of spectral peaks. The uncertainty is reflected by a probability distribution instead of a perfect knowledge of the spectral peak.

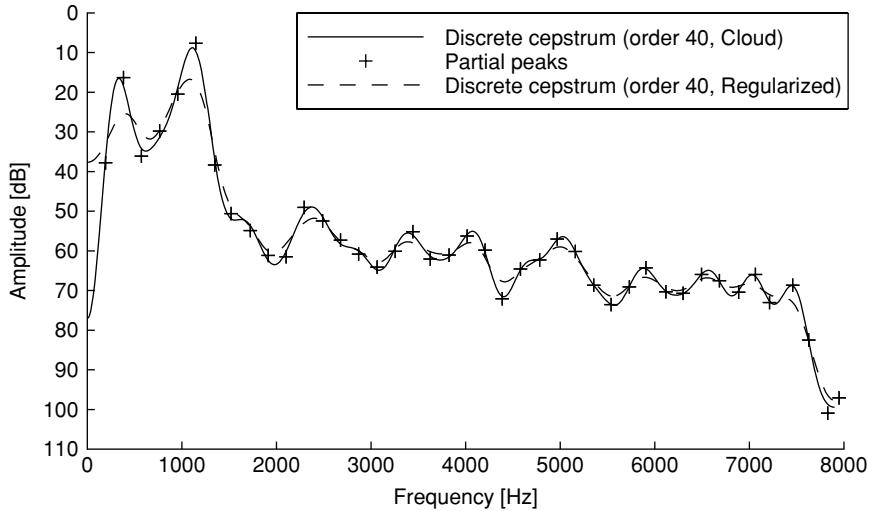


FIGURE 5.23. Improvement of discrete cepstrum spectral envelope estimation with stochastic smoothing. The envelope smoothed by the cloud algorithm reaches the two peaks below 1500 Hz, while the regularized envelope is too restrained.

Thus, the new error criterion becomes

$$E = \sum_{i=1}^N \int \int \pi_i(\omega, S) (\log(X_i H(\omega_i) - \log(S_i))^2 d\omega dS. \quad (5.45)$$

The distribution π_i can be sampled; that is, each spectral peak (ω_i, S_i) can be replaced by a set of peaks (ω_{ik}, S_{ik}) , to yield the cloud of points described at the beginning of this section.

Figure 5.23 shows the improvement of the discrete cepstrum spectral envelope estimation when stochastic smoothing is employed. The cloud method can also be combined with regularization, described in Section 3.5.1, to further improve the results.

Complementary points can be added to the partials before discrete cepstrum estimation in order to control the resulting spectral envelope. For example, it is advised to add points at the zero and Nyquist frequencies and between the highest partial and the Nyquist frequency (Schwarz, 1998) in order to prevent an unjustified oscillation in the spectral envelope, which would disturb the smoothness.

3.5.3 Nonlinear Frequency Scaling

When estimating the spectral envelope, a nonlinear frequency scale, similar to the *Mel* or the *Bark* scale, is appealing because it reflects some properties of human perception. As we have seen in Section 3.4, the discrete cepstrum algorithm is of cubic complexity in p , the order of the discrete cepstrum. This means that in order to keep computation times short and the amount of data small, we must try to

reduce the order necessary for a good estimation of the spectral envelope. One way to achieve this is to judiciously concentrate the precision or resolution where it is most needed and reduce it where it is not so important, in accordance with the properties of the human auditory system.

Owing to the approximately logarithmic frequency resolution of human hearing, we do not need to be very exact with the spectral envelope in higher-frequency ranges. Whereas in the low frequencies, very slight deviations in frequency and amplitude are perceptible, in the higher frequencies it suffices to represent the rough location of energy. Therefore, as suggested by Galas and Rodet (1991b), we can introduce a frequency-warping function which is linear below a given break frequency f_b , and logarithmic above. Our frequency warping function γ is defined by

$$\gamma(f) = \begin{cases} \alpha f/f_b, & f \leq f_b \\ \alpha (1 + \log_{10}(f/f_b)), & f > f_b \end{cases} \quad (5.46)$$

where α is a normalization factor given by

$$\alpha = \frac{f_s/2}{1 + \log_{10}(f_s/2f_b)}. \quad (5.47)$$

The effect of nonlinear frequency scaling can be seen in Fig. 5.24 with f_b taken to be 2500 Hz.

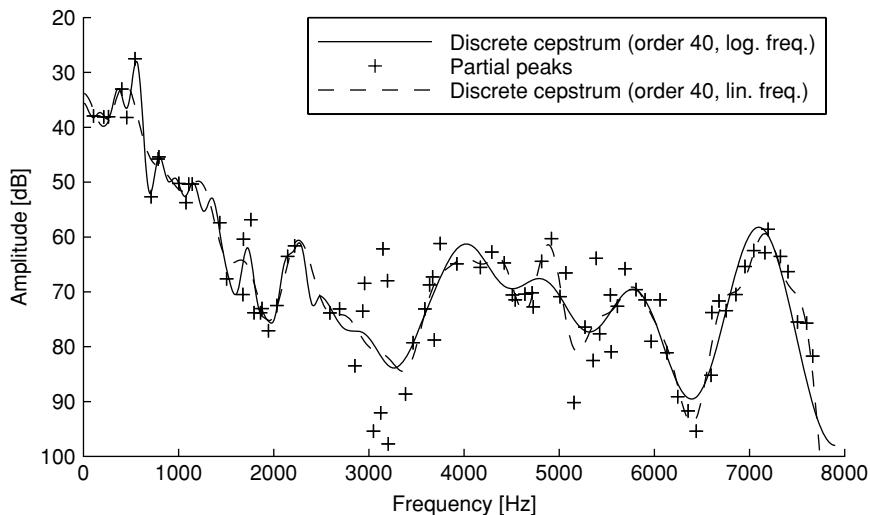


FIGURE 5.24. Effect of nonlinear frequency scaling with a break frequency of 2500 Hz on discrete cepstrum estimation. The higher-frequency part of the spectral envelope is rather inaccurate, while accuracy is much better in the low-frequency range below the break frequency.

As an additional advantage, for a given amount of precision, spectral envelopes represented in terms of logarithmically spaced frequencies require fewer matching parameters, reducing the space necessary for storage or transmission. Also, the complexity of synthesis can be reduced when fewer points are needed to represent a spectral envelope.

To obtain the spectral envelope from cepstral coefficients based on nonlinear frequency scaling, the frequencies f_i of the spectral-envelope samples [see Eq. (5.24)] are converted using

$$f'_i = \gamma(f_i) \quad (5.48)$$

and then converted to the corresponding angular frequencies,

$$\omega_i = 2\pi \frac{f'_i}{f_s}. \quad (5.49)$$

Computation then proceeds with Eqs. (5.26)–(5.28).

There is one pitfall to avoid in the application of nonlinear frequency scaling: Performing the nonlinear transformation before applying the cloud deteriorates the results slightly. To see why this is so, remember that the cloud algorithm (Section 3.5.2) adds points with a constant linear shift around each peak frequency, which will subsequently be stretched for the linear part or asymmetrically converted to the logarithmic scale for the rest, thus distorting the underlying probability distribution.

3.6 Estimation of the Spectral Envelope of the Residual Signal

The non-sinusoidal residual signal $r(t)$ [see Eq. (5.1)] is often considered as a random signal. Therefore, its spectrum is continuous and AR estimation techniques are well suited for this kind of spectrum. Estimation of the spectral envelope $G(\omega, t)$ of the residual signal around time t can be done with any of the well-known AR estimation techniques [see Section 3.2 (Kay, 1988)]. Such a technique provides the p coefficients $a_i(t)$ of an all-pole filter with magnitude transfer function $G(\omega, t)$. In practice, the $a_i(t)$ are only estimated around successive times t_k , $k = 1, 2, 3, \dots$, with steps $t_{k+1} - t_k$ in the order of 5–20 ms. At the resynthesis stage, the coefficients $a_i(t)$ are well suited for computing the residual by time-domain filtering of a white-noise signal. Cepstral estimation can also be used and provides cepstral coefficients that are well suited for frequency domain filtering of a noise at the synthesis stage.

However, a correct spectral-envelope estimation of the residual usually requires that the sinusoidal partials have been completely separated from the random component, and such separation is not a simple task (Peeters and Rodet, 1998). Another difficulty is that, as we have seen, the estimation methods are different for the sinusoidal and residual components. Therefore, Oudot et al. (1997) have proposed a method for estimating a unique envelope, simultaneously taking into account sinusoidal partials and nonsinusoidal components with a convenient estimation criterion for each type.

For easy envelope estimation of the residual, some authors (Freed, 1995; Goodwin, 1996) have proposed to simply represent the short-time spectrum magnitude $|R(\omega, t)|$ of the signal $r(t)$ by its mean value $G(\omega, t)$ taken over frequency for each of several channels distributed on a nonlinear scale. This representation also is well suited for frequency-domain filtering at the synthesis stage, because it only requires multiplication of a white-noise signal's STFT by $G(\omega, t)$.

4 Representation of Spectral Envelopes

Representation of spectral envelopes is essential for their use in musical synthesis. As we have seen in the previous section, various estimation methods result in very different parametrizations of spectral envelopes. However, the choice of a single canonical representation is essential for the flexibility of further processing (see Section 7).

Also, the choice of a good canonical representation for spectral envelopes is crucial for their applicability to a specific task. Important concerns include the ability to manipulate the envelopes in a useful and easy way as well as the speed of synthesis, both of which depend heavily on the representation. Toward achieving these goals, the requirements of locality, flexibility, speed of synthesis, and space are laid out in Section 4.1. They are then tested against the different possible representations, filter coefficients (4.2), sampled representation (4.3), geometric representation (4.4), and formants (4.5). Finally, a comparative table gives an overview of the fulfillment of the requirements by the different representations in Section 4.6.

4.1 Requirements

Precision: Naturally, the representation must describe an arbitrary spectral envelope as precisely as possible, whether it is obtained by estimation or given manually. Methods that do not fulfill this basic requirement have not been considered here.

Stability: The requirement of stability mandates that the representation be resilient to small changes in the data to be represented. Small changes, e.g., due to introduced noise, must not lead to big changes in the representation, but must result in equally small changes in the representation. Stability is of great importance, especially if we consider that the data to be represented may result from different estimation methods, such as cepstral or AR analysis, or even from manual input, and that some noise is always present.

Locality in frequency: This requirement states that it must be possible to easily achieve a local change of the spectral envelope by a change in a small subset of the representation parameters. Here, “local” means “without affecting the intensity of frequency components far away from the point of manipulation.”

Flexibility and ease of manipulation: A representation must allow easy manipulation for achieving an exactly defined outcome, such as the production of a certain formant in voice synthesis. For manipulation to be really useful

in musical applications, the relationship between manipulation parameters and spectral effect must be easily understood.

Synthesis speed: As much as possible, the representation should be directly usable for sound synthesis, without first having to be converted to a different form at high computational cost. This requirement is heavily dependent on the type of synthesis, e.g., additive synthesis or filtering. While no ideal solution can be presented, a compromise can be found which is not the fastest choice for each synthesis type, but which does not penalize speed or quality too much, even in the worst case.

Memory space: It is important for file storage and even more so for transmission that the representation not take up too much memory space.

Manual input: Finally, the representation should be easy to specify manually, e.g., either by drawing a curve, selecting primitive shapes, or by parameter text input.

4.2 Filter Parameters

When a spectral envelope is directly estimated from a signal, the most straightforward representation is the set of filter parameters that characterizes the output of the estimation, whether it be cepstral coefficients c_i (Section 3.3), or one of the AR coefficient sets, a_i, k_i, l_i (Section 3.2). These are, in general, very precise representations. Filter parameters are advantageous for fulfilling the space requirement (only order p values) and for being very efficient for filtering synthesis, because they can be directly used for fast time-domain filtering.

However, there are spectral-envelope shapes that are not easily represented by filter parameters. An extreme example is the ideal rectangular low-pass filter. In these cases, representation by filter parameters is stable, but not local. The non-locality is due to the fact that these methods essentially represent a spectral envelope as a time-domain impulse response or reflection function (by one of several possible filter models). It is easy to demonstrate that changing one filter parameter will change the spectral envelope's values at all frequencies. Filter parameters are also not easy to manipulate for obtaining a desired effect, especially when the effect is specified in the frequency domain. In addition, they are costly for evaluation of additive synthesis parameters, because at least p cosines have to be computed for each frequency selected from the spectral envelope.

Various types of filter structures can be used (see Section 6.1). Each structure leads to different types of filter parameters.

4.3 Frequency Domain Sampled Representation

A sampled representation gives the amplitude value $v(f_i)$ of a spectral envelope at M equidistant or logarithmically spaced frequency points f_1, \dots, f_M . (Each of the M grid points is also called a frequency bin.) It is obtained by either sampling a continuous spectral envelope obtained by estimation or by directly using given values. Care must be taken to ensure that M is high enough to avoid aliasing of the

sampled spectral envelope due to rapidly varying components of the continuous spectral envelope.

This representation is as stable as the filter parameter representation and can be derived directly from the parameters. It obviously satisfies the locality criterion because the amplitude at each frequency can be changed independently from the others. It is the most flexible representation (due to the high locality), but not very easy to manipulate, because locality demands that values at all frequencies within a certain frequency range be specified. Especially when we think of an application for the singing voice, the preferred manipulations are changes of the position and bandwidth of formants, which means that new amplitude values must be specified for the range of frequencies occupied by the formant.

This approach is fastest for additive synthesis and fast for filtering in the frequency domain. It is reasonably compact, because the data required can be as low as 100 points, even less when a logarithmic frequency scale is used, and manual input is easy.

4.4 Geometric Representation

Starting from a sampled spectral-envelope representation, a geometrical representation can be derived that attempts to describe the amplitude curve of the spectral envelope in the frequency domain with fewer points not spaced at equidistant frequencies. The geometrical representation can be of the form of a break-point function or of splines, as described as follows:

Break-point functions: A break-point (or piecewise linear) function (BPF) is a general method of representing a function, be it in the time or frequency domain, by a set of connected linear segments. It consists of n break points P_i at (x_i, y_i) . In our frequency-domain case, x_i is the frequency and y_i is the amplitude. The $n - 1$ segments between the break points are interpolated linearly. [For a discussion of BPFs in the time domain see Horner and Beauchamp (1996).]

Splines: These are similar to break-point functions, but they provide for quadratic or cubic interpolation of each section between the points P_i given, using a polynomial of degree 2 or 3 (Unser et al., 1993). The value and slope should be continuous at each point. The slope can also be accessible as a parameter for manipulation.

It is useful to apply splines to spectral-envelope representations in such a way that the points P_i are placed on the maxima and minima of the sampled representation, where the slope is zero, and on the inflection points, where the curvature changes direction.

In general, a weak point of geometric representations is that they do not model spectral envelopes in a way relevant to the signals from which they are derived, but rather simply as curves in Euclidean space. Especially, they do not take into account interdependencies between given points that arise from the time-varying character of the spectral envelope.

The stability of geometric representations is seriously disturbed by the fact that small parameter changes can cause sudden changes of the maxima found. These changes are quite local and can be made more stable by manually adding points. In general, the representations are flexible, easy to manipulate, and always result in smooth curves. However, there is a tradeoff between ease of manipulation and precision: A parameter which is useful to manipulate but affects a broad region often causes precision to suffer because of the large stretch of the spectral envelope that must be interpolated.

Regarding speed, geometric representations are slightly more costly for synthesis than the sampled representation is. For splines, evaluation of interpolating polynomials must be taken into account. The amount of space needed is less than for the sampled representation and even less if redundant points are pruned (again at the cost of precision).

Finally, geometric representations are very well suited for specifying spectral envelopes manually by drawing.

4.5 Formants

Maxima of voice spectral envelopes are known to convey most of the perceptual information concerning the vocal tract. Therefore, spectral envelopes can be conveniently coded in terms of their maxima or peaks. It is assumed that these peaks result from vocal tract resonances, and we will, for simplicity, refer to them as “basic formants.” In general, a resonance is characterized by a complex transfer function $H(\omega)$, which has a magnitude and a phase. In the case of a basic formant, we only consider the magnitude. Because, normally, several formant functions are needed to represent a spectral envelope, how they combine to form the total envelope must be decided. For example, formant functions can be added or multiplied. Note that these two algorithms correspond to two different synthesis filter structures: Addition corresponds to the parallel structure and multiplication corresponds to the serial or cascade structure. These two structures have different properties, which have been much discussed in the literature [see for instance Klatt (1980), Holmes (1983), and Section 6.2]¹.

Three convenient ways to represent formants are detailed in the following sections (see Fig. 5.25 for an overview).

4.5.1 Formant Wave Functions

An FOF, from the French *Forme d’Onde Formantique* (Rodet, 1984), was originally a waveform used in voice synthesis and in general sound synthesis. It constitutes the basic synthesis model of the CHANT synthesis system (Rodet et al., 1984) and

¹ Use of a *parallel* formant synthesizer leads to well-known problems such as *zeros* created between two formants. The skirts of formants are also not easy to control in the parallel structure. In a *serial* formant synthesizer, the amplitude of formants are difficult to handle. Furthermore, when using automatic formant extraction, formants may appear or disappear during transitions, and this can hardly be handled by a serial synthesizer.

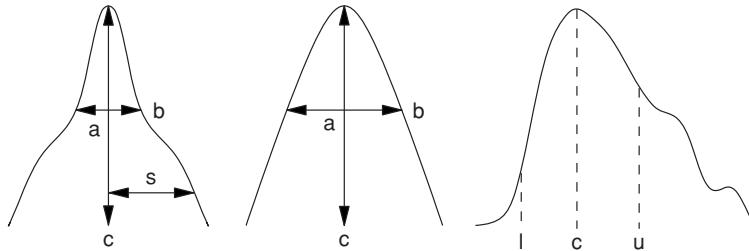


FIGURE 5.25. An FOF (left), a basic formant (middle), and a fuzzy formant (right), with their frequency-domain parameters.

corresponds to the time-domain representation of a single formant as an elementary waveform, i.e., to the impulse response of a resonance. Several FOFs can be added up to build a desired spectrum (typically five to seven are used for voice synthesis).

The FOF has parameters in both the frequency domain and the time domain. The frequency-domain parameters, shown in Fig. 5.25, are center frequency c , amplitude a , bandwidth b , and skirt width s (which can be controlled independently from the bandwidth). The time-domain parameters are phase, excitation time, and attenuation time. Although FOFs give a very precise way to define a spectrum for singing voice and general music synthesis, they contain more information than required to represent a spectral envelope.

4.5.2 Basic Formants

A basic response v_k may be described in terms of parameters c_k (center frequency), b_k (bandwidth), and a_k (maximum amplitude in dB). Bandwidth specifies the -3 dB frequency width of the formant (the frequency range over which the formant's response is not less than 3 dB below the maximum amplitude). With these parameters, the spectral envelope for the k th formant can be defined as a function of frequency f using the formula

$$v_k(f) = \frac{10^{\frac{a_k}{20}}}{1 + \left(10^{\frac{3}{20}} - 1\right) \left(\frac{c_k - f}{b_k/2}\right)^2}. \quad (5.50)$$

This function approximates very well the magnitude transfer function of a two-pole filter,² the usual model for a resonance. The final spectral envelope is then obtained by summation of the basic formant functions $v_k(f)$.

4.5.3 Fuzzy Formants

Representing real-life spectral envelopes precisely in terms of formants is often difficult. However, the approximate locations and bandwidths of formants, in vowels

² The smaller skirt width of $v_k(f)$ compared to that of the magnitude transfer function of a two-pole filter is in fact an advantage, because it increases the locality of the representation.

for instance, are fairly well known. This motivates an augmentation of the sampled representation (Section 4.3) by defining a fuzzy formant as a formant region within a sampled spectral envelope where it is believed or known that a formant lies. This knowledge can come from observation of spectral data or from source material labels (e. g., a recording of the voice with annotations on what phonemes were uttered), or from an automatic formant estimation that includes an estimate of the uncertainty of the estimate.

A fuzzy formant is specified by three frequency parameters, the lower bound l , the upper bound u , and the center c , if known. The center corresponds to the frequency location of the formant peak in the spectrum. Additionally, a bookkeeping parameter is used to identify the formants, so that they can be associated into (fuzzy) formant tracks.

4.5.4 Discussion of Formant Representation

Because a small spectral dip can suddenly create a new formant candidate in the estimation procedure, the formant representation is not stable. However, with the fuzzy formant representation, such instabilities are not damaging. They are local, flexible, and very easy to manipulate. Synthesis is reasonably fast, both for the frequency domain and, except for fuzzy formants, for the time domain. If a pure formant representation is sufficient, FOFs and basic formants are very compact for storage. Otherwise, they need a residual spectral envelope, which is the difference between the complete spectral envelope and the spectral envelope expressed with formants, in sampled representation to be stored along with them.

In summary, formant representations are very well suited for specifying spectral envelopes manually, especially for convincing synthesis of the voice.

4.6 Comparison of Representations

Table 5.1 (Schwarz and Rodet, 1999) shows a condensed comparison of the representations discussed in the preceding sections. The scores (++, +, o, -, -) indicate the authors' judgment of the degree to which the requirements from Section 4.1 have been fulfilled. The precision requirement is not listed, as it is fulfilled by all methods.

TABLE 5.1. Comparison of spectral envelope representations.

Representation	Stability	Locality	Flexibility / ease of manipulation	Speed of synthesis TD / FD	Memory space	Manual input
Filter coefficients	++	-	-- / -	++ / o	+	--
Sampled	++	++	++ / +	- / ++	o	+
Geometric	-	+	+ / ++	- / +	+	++
Formants	-	+	++ / ++	+ / o	++	++

5 Transcoding and Manipulation of Spectral Envelopes

“Transcoding” (Section 5.1) is the conversion of a spectral envelope from one representation to another representation, with the least possible change of its form. “Manipulation” (Section 5.2) is the deliberate change of it for musical purposes. “Morphing” (Section 5.3) is a special kind of manipulation that gradually changes one spectral envelope into another.

5.1 *Transcodings*

Transcoding of a spectral envelope from any type of representation into the sampled representation (Section 4.3) is performed simply by sampling the curve generated by the defining equation of its original representation. Transcoding from the sampled representation to other types of representations can be achieved by re-estimating the spectral envelope with an appropriate estimation method. Direct conversion of a sampled representation into other representations, or conversions among other representations, are less trivial, and only one example is given here.

5.1.1 Converting Formants to AR-Filter Coefficients

For voice synthesis, the ability to directly calculate AR-filter coefficients from spectral envelope data represented in terms of basic formants is important. From the formant parameters, we first compute the magnitude transfer function $|H(\omega)|$ that should be applied to an excitation function. It is the sum of n individual magnitude transfer functions corresponding to each formant (Rodet, 1984). A serial filter P with magnitude transfer function $|P(\omega)|$, which is the product of individual formant transfer functions (each corresponding to a conjugate pair of poles according to the formant frequency and bandwidth), yields peak amplitudes somewhat different from the desired ones. But, owing to the similarity of peaks in $|P(\omega)|$ and $|H(\omega)|$, the ratio $Q(\omega) = |H(\omega)|/|P(\omega)|$ is a smooth function of ω . We can then easily compute a few autocorrelation coefficients a_i , associated with $Q(\omega)$. Then, by using the Durbin–Levinson method (Markel and Gray, 1980), we can derive coefficients d_i for a filter with a magnitude transfer function $|D(\omega)|$ very close to $Q(\omega)$. Hence, the product filter $D \cdot P$ has a magnitude transfer function very close to the desired magnitude transfer function $|H(\omega)|$ (Depalle, 1991).

5.1.2 Formant Estimation

Estimating the parameters of formants from the sampled representation has been studied by Depalle (1991) using spectral peaks and inflection points, followed by hidden Markov model (HMM) tracking of formant paths (Depalle et al., 1993).

Formants can also be determined from all-pole filters obtained by autoregressive analysis. A large number of methods have been developed for formant estimation

from speech signals [see, for instance Schafer and Rabiner, 1970; Olive, 1971; Atal and Hanauer, 1971; Atal, 1974; Chandra and Lin, 1974; McCandleem, 1974; Markel and Gray, 1980; Rodet and Depalle, 1985; Kopec, 1986; Sandler, 1989; and Niranjan and Cox, 1994)]. As an example, starting from an all-pole filter $1/A(z)$, we can find the roots of the denominator. Then, the roots are separated into two sets: The first has only real poles, the global contribution of which represents the tilt of the source spectrum. The p complex pole-pairs of the second set are sorted into m classes ($m \leq p$) corresponding to m maxima, because several pole-pairs may contribute to the same maximum, corresponding to a unique “formant.”

5.2 Manipulations

Manipulation of spectral envelopes is at the heart of the creative process. It allows composers and musicians to surpass the limitations of recorded sounds, either by creating sounds extremely different than the originals, to subtly modify given sounds, or to merge characteristics of different sounds.

While ordinary amplification or attenuation of a spectral envelope is easily implemented by multiplication by a constant, frequency-selective amplification or attenuation can be implemented by multiplying by another spectral envelope, which is equivalent to applying a filter:

$$v'(f) = v(f) \cdot v_a(f). \quad (5.51)$$

One example is to modify the *spectral tilt*, the overall slope of a speech or instrument spectrum. For speech, it is one of the acoustic correlates of intensity. For the singing voice, it is related to vocal effort. For instruments, it can be dependent on performance dynamic or the relative force with which an instrument is played. To tilt a spectral envelope by T decibels between $f_1 > 0$ and $f_2 > f_1$, we can multiply it by a frequency ramp $v_t(f)$ (Bennett and Rodet, 1989):

$$v_t(f) = \begin{cases} 1, & f < f_1 \\ \exp\left(T \frac{\ln(10)}{20} \frac{\ln(f/f_1)}{\ln(f_2/f_1)}\right), & f \geq f_1 \end{cases} \quad (5.52)$$

For example, if $T = -20$ dB, $v_t(f_2) = 10^{-1}$.

5.3 Morphing

In general, morphing means performing a gradual transition from one parameter set to another, in our case moving from one spectral envelope to another. The simplest method for morphing between envelopes is linear interpolation, i.e., computing a weighted sum of the spectral envelopes. If the envelopes are given as $v_1(f)$ and $v_2(f)$ at frequency f , and the interpolation factor is m , then

$$v'(f) = (1 - m)v_1(f) + mv_2(f) \quad (5.53)$$

is the linearly interpolated envelope.

5.3.1 Shifting Formants

When dealing with the spectral envelope of speech or the singing voice, we want to preserve the formant structure of the envelope. Therefore, to morph between two spectral envelopes, we do not want to linearly interpolate the amplitudes at each frequency as in Eq. (5.53), but rather shift the formants from their place in the original spectral envelope to that in the target spectral envelope. In fact, we want to simulate the effect of morphing the articulatory parameters of the vocal tract. Figure 5.26 compares straightforward linear interpolation with true formant shifting.

The prerequisites for properly shifting formants are that we know, first, the original formant locations and, second, which formant in the original spectral envelope is associated with which formant in the target spectral envelope. The former is not at all obvious and is a question of formant detection. The latter is equally difficult for a formant-tracking algorithm without providing manual input, i.e., the ability to label the formants of successive time frames to define the tracks. However, an automatic procedure for matching formants between two spectral envelopes has been proposed in Laura and Rodet (1989).

Fortunately, for some applications, we know *a priori* where the formants should be. For example, when treating the voice in a piece with given lyrics, it is known which vowels are sung at which moment, and thus we can look up the formant center positions and bandwidths in the phonetics literature. These would be used to partition the spectral envelope into formant regions so that we can obtain a fuzzy formant representation, as described in Section 4.5.3.

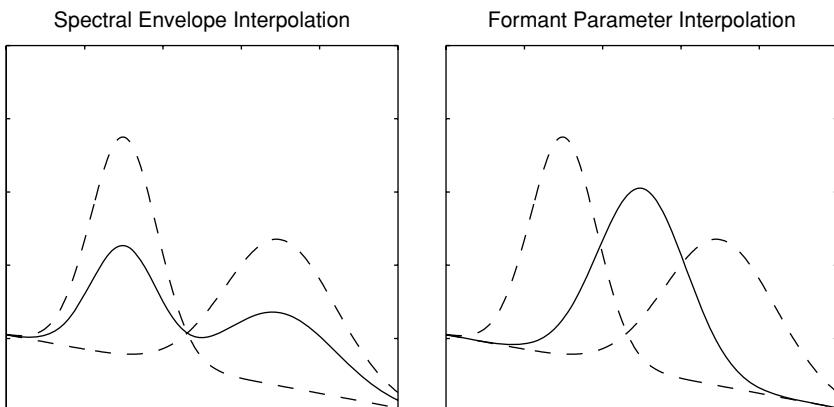


FIGURE 5.26. Formant interpolation versus formant shift: The dashed curves in both figures show two spectral envelopes consisting of one formant each. In the left figure we see the result of direct interpolation of the spectral envelopes, which is a weighted sum of the two curves, here with an interpolation factor of 0.5. The figure to the right shows what is really desired: interpolation of the parameters of the formants, resulting in a formant frequency shift.

5.3.2 Shifting Fuzzy Formants

The fuzzy formant representation of spectral envelopes consists of an envelope in sampled representation partitioned into several formant regions, which are indexed for identification. Given two spectral envelopes each having two fuzzy formants with the same indices, it is nevertheless not trivial to determine exactly how the intermediate spectral envelopes, with their formants moving from their positions in the original envelope to those in the target envelope, are to be generated.

Fortunately, for the special case of two sample-represented envelopes, each having a single formant, there is an alternative automatic morphing method which does not require formant indexing. The idea is to first integrate over the envelopes and then to linearly interpolate between the integrals. Finally, we retrieve the interpolated formant by subsequent differentiation of the interpolated integral. The result is that the envelopes are morphed to appropriately shift the single formant. How this idea works is illustrated in Fig. 5.27.

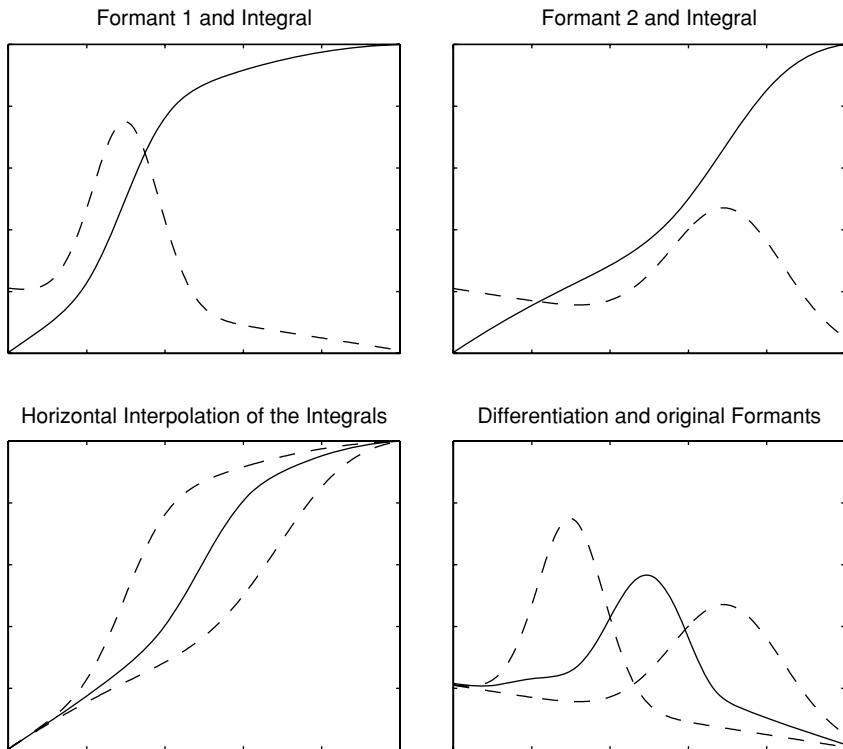


FIGURE 5.27. Interpolation of formants by linear interpolation of the spectral integral. All amplitudes are normalized. The upper row shows the integrals (the cumulative sum) of the two formants shown as dashed spectral envelopes. The lower left figure shows the linear interpolation by a factor of 0.5 of the integrals, drawn again as dashed lines. Taking the derivative of the result in the lower right figure reveals an almost perfectly shifted formant. (The original formants are shown again in dashed lines for clarity.)

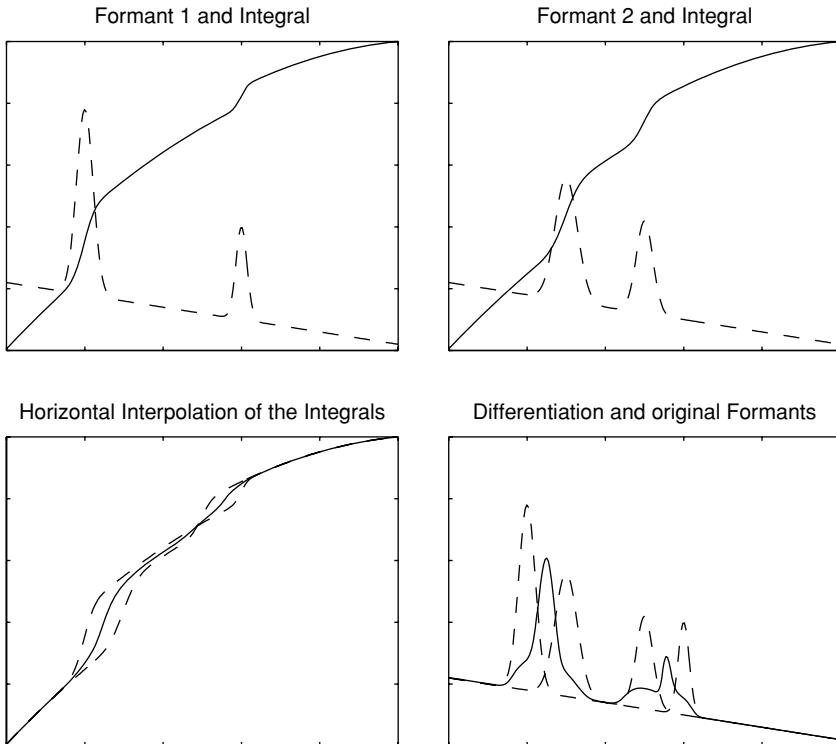


FIGURE 5.28. Interpolation of two formants by linear interpolation of the integral. Obviously, the result in the lower right graph does not correspond to the interpolation of the formant parameters.

Unfortunately, spectral integral morphing fails when there is more than one formant, as can be seen in Fig. 5.28, where two formants are attempted. Nevertheless, we could do better if we used formant region information. In this case, we could restrict the technique of linear interpolation of the integral independently to each of the given formant regions, with an appropriate fade-in and fade-out applied at the region borders.

5.3.3 Morphing Between Well-Defined Formants

If both the original and target spectral envelopes to be morphed are well defined in terms of formants with indices, center frequencies, amplitudes, and bandwidths given as parameters, vocal-tract-like morphing becomes trivial. Simply the formant parameters of formants with the same index need to be linearly interpolated, as shown in Fig. 5.26 (right half).

5.3.4 Summary of Formant Morphing

We can recognize a hierarchy in the spectral-envelope representations in regard to formant morphing. The hierarchy is, from highest to lowest degree of

structure:

1. Well-defined formants: can be morphed perfectly.
2. Fuzzy formants: can be morphed reasonably well.
3. Sampled representations of envelopes: can be morphed well only if both original and target spectral envelopes are characterized by a single formant.

With each step down, we lose some information necessary for formant interpolation. This means that when two spectral envelopes to be morphed have representations of different hierarchy, we must convert the higher one down to the lower one's representation, discarding the higher one's formant information.

6 Synthesis with Spectral Envelopes

In synthesis by rule (i. e., using a synthesis model rather than analysis data), a spectral envelope may be given directly as part of the synthesis parameters. With resynthesis, an input signal may be modified, while taking into account the desired spectral envelope. Depending on the structure of the synthesis system, several methods may be used to apply the spectral envelope.

6.1 Filter Synthesis

Various filter implementations have been used for sound synthesis, such as all-pole filters (Moorer, 1979), simple second-order sections (Beauchamp, 1979, 1982; Horner and Beauchamp, 1995), second-order sections in cascade (Pierucci and Paladin, 1997) or in parallel (Klatt, 1980; Holmes, 1983; Rodet et al., 1984; Allen et al., 1987; Sandler, 1989), and a combination of poles and zeros (Massie and Stonick, 1992). We will not detail these different types which are abundantly described in the literature [see for instance Hamming (1977) or Smith (1985)].

If the sound to be changed is in the form of a signal (i.e., not as a spectrum), the spectral envelope to be applied should be converted to filter parameters. The filter can be given by its impulse response $h(n)$ for time-domain filtering by convolution or as a transfer function $H(\omega)$ for filtering in the frequency domain.

If the spectral envelope is represented as AR coefficients, the coefficients a_i can be directly used for time-domain filtering. The transfer function $H(\omega)$ is defined proportional to $1/A$ with gain factor g :

$$H(\omega) = \frac{g}{A(\omega)} = \frac{g}{1 - \sum_{i=1}^p a_i e^{-j\omega_i}} \quad (5.54)$$

For time-domain filtering of an input signal $x(n)$, we can directly apply the predictor coefficients to recursively obtain the output samples $s(n)$:

$$s(n) = gx(n) + \sum_{i=1}^p a_i s(n-i) \quad (5.55)$$

Alternatively, by using reflection coefficients k_i , we can apply a preferred lattice filter structure [see Markel and Gray, 1980].

If the spectral envelope is given in terms of cepstral coefficients c_k , then H becomes

$$H = \exp(F^{-1}(c)) \quad (5.56)$$

which by the properties of cepstrum estimation (Section 3.3) is evaluated at the desired bin frequencies ω_i as

$$H(\omega_i) = \exp\left(\sum_{k=1}^p c_k \cos k\omega_i\right). \quad (5.57)$$

If the spectral envelope is defined by any representation other than filter coefficients (e. g., by conversion, especially from formants, to filter coefficients), the frequency-domain filter $H(\omega_i)$ is given directly by the evaluation of the appropriate spectral-envelope formula at the desired frequency bins ω_i . The time-domain filter h can then be obtained by inverse Fourier transform:

$$h(t) = F^{-1}(H(\omega)) \quad (5.58)$$

6.2 Additive Synthesis

In additive synthesis, the synthetic signal is a sum of the sinusoidal partials,

$$d_i(t) = y_i(t) \sin\left(\int \omega_i(t) dt\right), \quad (5.59)$$

whose amplitudes are specified by the sinusoidal spectral envelope, plus a residual noise whose spectral density (in squared-amplitude per Hertz) is given by a noise spectral envelope. The residual can be easily synthesized by filtering white noise. For the sinusoidal part, the amplitude for each partial is equal to the value of the spectral envelope taken at the frequency of the partial.³

6.3 Additive Synthesis with the FFT^{-1} Method

Additive synthesis is usually done with one sinusoidal oscillator for each partial (Moore, 1990). The cost of this oscillator method is high for sounds that have a large number of partials, such as a low-pitched piano tone. To alleviate the computational

³ For resynthesis, instead of imposing the spectral envelope, a mixture (weighted sum) between the original partial amplitudes, and one or more spectral envelopes is possible, governed by a mix factor m . Considering the mix factor as a function $m(f)$, dependent on frequency, allows frequency-selective application of spectral envelopes. If the input partials and the modifying spectral envelope are from different sounds, this is usually called cross-synthesis, because it crosses the characteristics of two distinct sounds: the partial structure (presence and frequency location of partials and their development in time) of the input sound with the spectral envelope estimated from the partial amplitudes of the other sound.

cost of the oscillator method, one can use the so-called FFT^{-1} method (Rodet and Depalle, 1992), based on the short-term Fourier (STF) model of sound signals, which allows an efficiency gain of 10–30 compared to the oscillator method. It is implemented in various musical sound synthesis systems (Freed et al., 1993; Serra et al., 1997; Wanderley et al., 1998). In the FFT^{-1} method, computation of partials is done by an inverse fast Fourier transform, a transformation of each short-term spectrum (STS), $S_l(k)$ for frame l and FFT-bin k , into the corresponding time-domain signal frame $s_l(n)$.

To implement this, let N be the number of partials of the signal to be computed at a certain sample n , and let $f_i(n)$, $y_i(n)$, and $\phi_i(n)$ be the frequency, the amplitude, and the phase of the i th partial, where $i = 1, 2, \dots, N$. Assuming that these functions vary slowly in time, for each short time frame l they can be replaced by their mean values, say $f_{i,l}$, $y_{i,l}$ and $\phi_{i,l}$. From these values a good approximation of the short-time spectrum $S_l(k)$ can be constructed at low cost. The frame l of the signal, $s_l(n)$, is then computed by an inverse fast Fourier transform, and the signal $s(n)$ is obtained by overlap-add of successive frames $l - 1, l, l + 1, \dots$ (Rodet and Depalle, 1992).

With the FFT^{-1} synthesis method, introducing noise components in any narrow or wide frequency band and with any amplitude is easy and inexpensive. It suffices to add random complex values to the corresponding bins of the short-time spectrum under construction before performing the inverse Fourier transform. Let k be the bin of the STS where noise should be added with an amplitude $r_{k,l}$. Then $r_{k,l} e^{j\phi_{k,l}}$ is simply added to $S_l(k)$, where $\phi_{k,l}$ is a random phase. There exist several ways to obtain $e^{j\phi_{k,l}}$, which differ by a more or less precise noise distribution and by their computational and memory cost (Rodet and Depalle, 1992; Freed, 1999).

Similarly, applying a spectral envelope is easy and inexpensive. For each partial (or for each noise band) at frequency $f_{i,l}$, it suffices to compute the value of the spectral envelope $v(f_{i,l})$ at this frequency and to use this value as the amplitude $y_{i,l}$ of the partial (or of the bin of the short time spectrum) in the FFT^{-1} algorithm.

7 Applications

This section will present some applications of spectral envelopes for sound transformation and synthesis. Note that using the standardized, open, and extensible Sound Description Interchange Format (SDIF) (Wright et al., 1998; Virolle et al., 2001) there is now a way to exchange spectral-envelope data with well-defined semantics (Schwarz, 1998) among programs, hardware architectures, and institutions.

7.1 Controlling Additive Synthesis

Additive analysis/synthesis is a powerful way to parametrize a sound event into sinusoidal partials with their frequencies, amplitudes, and phases. This benefit is also its curse: It puts every minute detail of a sound event at our disposal, but leaves us with the task to control and manipulate this mass of parameters in a sensible way.

So far, control is done by specifying the change of every single parameter over time by break-point functions [e.g. Fitz et al. (1995) and Horner and Beauchamp (1996)]. Because the number of partials can easily rise into the hundreds, modifications are tedious. Moreover, doing valid manipulations with regard to signal processing and from a musical perspective is not obvious, and, what is more, the parameters are interdependent (e.g., changing the frequency of the partials changes the spectral envelope, often with undesirable results, as shown in Section 2.1).

Freed et al. (1993) suggest using spectral envelopes to control the amplitudes of the partials for resynthesis. This drastically reduces the number of parameters, provides parameter sets that are easily understandable (e.g., formants), and renders frequency and amplitude control independent from each other.

Also, modeling the residual noise part by filtering white noise with spectral envelopes makes this component of sound accessible to manipulation. This has not been possible before in the sampled signal representation of the residual.

The most significant advantage, however, lies in the unified handling of the noise and sinusoidal parts, because the spectral envelopes of the two parts are represented in the same way. Therefore, the very same manipulation can affect both parts synchronously, if this is desired (Rodet et al., 1995).

7.2 *Synthesis and Transformation of the Singing Voice*

One of the primary applications of spectral-envelope control is high-quality synthesis of the singing voice. Within the additive-synthesis paradigm, synthesis is often a resynthesis of the previously analyzed and modified voice signal. For modifications to be effective, constraints posed by the human vocal apparatus should be taken into account.

For example, as demonstrated in Section 2.1, pitch transpositions of the voice sound very unnatural when spectral envelopes are not corrected, because they reflect the configuration, especially the length, of the vocal tract. To avoid this, it is necessary to estimate the spectral envelope of the original sound and reconstitute it by applying it to the transposed sound.

Also, many aspects of the expressivity of the singing voice depend on the spectral envelope, such as timbral variations due to changes of spectral tilt, rather than on pitch and amplitude alone.

With the methods of morphing between spectral envelopes and formants described in Section 5.3, a new type of high-quality additive synthesis of the voice is possible. It uses two representations, each one suited for a specific part of the voice. The first one follows a very general harmonic-sinusoids-plus-noise model (Laroche et al., 1993; Oudot, 1998), controlled by envelopes in a sampled representation, to preserve rapid changes in transients (e.g., plosives) and noise spectral envelopes in fricatives. The second one represents spectral envelopes in terms of formants in order to preserve precise formant characteristics in the steady part of vowels. It is then possible to combine the two representations and to interpolate between precise formants and spectral envelopes with marked formant regions (i.e., fuzzy formants, see Section 4.5.3). This combines the excellent generation

of vowels by precise formant synthesis [which is available, e.g., in the CHANT synthesizer (Rodet et al., 1984)] with the flexibility of general additive synthesis [e.g., in the graphical generalized diphone control and synthesis program DIPHONE (Rodet et al., 1988; Rodet and Lefèvre, 1997)].

8 Conclusions

In Section 4.1 we stated that a good representation should offer spectral-envelope parameters that are flexible and easy to manipulate. However, what does it mean for parameters to be easy to manipulate? For singing voice and speech applications this is quite clear, but for the large multitude of possible musical applications a good guess is to offer the greatest possible flexibility. As more applications for spectral envelope manipulation appear, the concept of “best representation for manipulation” will become more clear.

In the context of computer music, the control of spectral envelopes offers the possibility of influencing a sound’s timbre to a great degree, allowing composers to obtain a desired effect or characteristic of a sound by the use of a flexible, unconstrained representation. To the performer, the real-time application of spectral-envelope manipulation greatly enhances expressivity through easily understandable and “musically significant” parameters, i.e., parameters that pertain to a model (e.g., the source–filter model) that is valid for many musical instruments.

Between the creation of completely new sounds and the modification of existing sounds lies the possibility for combining features of different sounds. Cross-synthesis using spectral envelopes can be used to combine characteristics of two distinct sounds: For example, the partial-frequency structure may be taken from one sound and the spectral envelope from another.

Finally, we would like to consider the application of spectral-envelope manipulation for the creation of music. To this end, we asked computer music composers and performers what types of operations for manipulating spectral envelopes they would like to have available, and they invariably came up with ideas about changing spectral envelopes in time. Therefore, for worthwhile artistic applications, we must raise our point of view above the one-dimensional perspective adopted in most of this chapter, where the richness and complexity of sound has only been viewed through the keyhole of a single time-frame. Alas, this is beyond the scope of this chapter. Although we have developed here several mechanisms for controlling time variations, the question of how to apply and manipulate them will have to be answered elsewhere.

9 Summary

In this chapter, we gave a definition of spectral envelopes and their relation to source–filter models and perception (Section 2). We examined various methods

for estimating, representing, transcoding, and manipulating spectral envelopes, and their application to sines-plus-residual analysis and synthesis of musical sounds.

For estimation of spectral envelopes (Section 3), we first stated requirements for exactness, robustness, and smoothness, and then described the AR, cepstrum, and discrete cepstrum methods in detail. Also, we examined various distinct possibilities for improving the discrete cepstrum method: regularization, stochastic (or probabilistic) smoothing, nonlinear frequency scaling, and adding control points to the envelope.

After defining requirements for the representation of spectral envelopes (Section 4), we examined several representations including those which use filter parameters, frequency-domain sampling, geometric representations (break-point functions and splines), and formant representations. A sampled representation, combined with indications of the regions of formants (called “fuzzy formants”) was defined to allow combining spectral envelopes with precise formant descriptions (FOFs and basic formants).

Methods of transcoding between the different representations of spectral envelopes, and some types of manipulations were examined in Section 5. Special attention has been given to morphing between spectral envelopes including those with formants. Other manipulations, based on primitive operations on amplitudes of spectral envelopes were covered.

For applying spectral envelopes to sound synthesis, two cases of filter synthesis (AR and cepstral) and additive synthesis (direct and FFT^{-1}) were examined in Section 6. For the former, methods for converting different representations to time-domain or frequency-domain filters were given.

Appendix: List of Symbols

t	continuous time
n	sample number
$s(t), s(n)$	continuous, discrete signal
$x(t), x(n)$	continuous, discrete excitation or source signal
$h(t), h(n)$	filter impulse response
$d(t), d(n)$	continuous, discrete sinusoidal signal
$r(t), r(n)$	continuous, discrete residual noise signal
$p(t), q(t)$	sinusoid partial continuous signal
$S(\omega)$	signal spectral envelope
$X(\omega)$	excitation or source spectral envelope
$H(\omega)$	filter transfer function
$A(\omega), A(z)$	all-pole filter transfer function
a_i	AR predictor coefficients
k_i	AR reflection coefficients
g	LPC gain factor
c_k	k th cepstral coefficient

p	order of spectral envelope estimation
N	number of partials
M	number of frequency bins
ω, f	angular, Hertz frequencies
x_i, y_i	i th sinusoid partial amplitudes
$v_i, v(f)$	spectral envelope bin, in frequency–amplitude plane
$\delta(\cdot)$	Dirac delta function
F, F^{-1}	forward, inverse Fourier transform
φ	phase of sinusoid partial
Φ	cosine matrix
Ψ	cosine product matrix
α	normalization factor
$\gamma(f)$	frequency-scaling function
$\pi_i(\omega, y)$	probablity distribution function
T	spectrum tilt (in decibels)

References

- Allen, J., Hunnicutt, M. S., and Klatt, D. (1987). *From Text to Speech, The MITalk System* (Cambridge University Press, New York).
- Atal, B. S. and Hanauer, S. L. (1971). “Speech analysis and synthesis by linear prediction of the speech wave,” *J. Acoust. Soc. Am.* **50**, 637–655.
- Atal, B. S. (1974). “Recent advances in predictive coding-applications to speech synthesis,” in *Speech Communication: Proceedings of the Speech Communication Seminar, Stockholm*, G. Fant, ed. (John Wiley, New York), pp. 27–31.
- Beauchamp, J. W. (1974). “Time-variant spectra of violin tones,” *J. Acoust. Soc. Am.* **56**(30), 995–1004.
- Beauchamp, J. W. (1975). “Analysis and synthesis of cornet tones using nonlinear interharmonic relationships,” *J. Audio Eng. Soc.* **23**(10), 778–795.
- Beauchamp, J. W. (1979). “Practical sound synthesis using a nonlinear processor (wave-shaper) and a high pass filter,” *Computer Music J.* **3**(3), 42–49.
- Beauchamp, J. W. (1980). “Analysis of simultaneous mouthpiece and output waveforms of wind instruments,” *66th Conv. Audio Engineering Soc.*, Los Angeles, Audio Eng. Soc. Preprint 1626.
- Beauchamp, J. W. (1982). “Synthesis by spectral amplitude and ‘brightness’ matching of analyzed musical instrument tones,” *J. Audio Eng. Soc.* **30**(6), 396–406.
- Benade, A. H. (1976). *Fundamentals of Musical Acoustics* (Oxford University Press, New York).
- Bennett, G., and Rodet, X. (1989). “Synthesis of the Singing Voice,” in *Current Directions in Computer Music Research*, M. V. Mathews and J. R. Pierce, eds. (MIT Press, Cambidge, MA), pp 19–44.
- Bogert, B., Healy, M., and Tukey, J. (1963). “The Quefrency Alanysis of Time Series for Echoes,” *Proc. Symp. on Time Series Analysis*, M. Rosenblatt, ed. (J. Wiley, New York), Ch. 15, pp. 209–243.
- Campedel-Oudot, M., Cappé, O., and Moulines, E. (2001). “Estimation of the spectral envelope of voiced sounds using a penalized likelihood criterion,” *IEEE Trans. on Speech and Audio Processing* **9**(5), 469–481.

- Chandra, S., and Lin, W. C. (1974). "Experimental comparison between stationary and non-stationary formulations of linear prediction applied to voiced speech," *IEEE Trans. Acoustics, Speech Signal Processing ASSP-22*, 403–415.
- Depalle, P. (1991). "Analyse, modélisation et synthèse des sons basées sur le modèle source/filtre," Doctoral dissertation, Université du Maine, Le Mans, France.
- Depalle, P., Garcia, G., and Rodet, X. (1993). "Tracking of partials for additive sound synthesis using hidden Markov models," *Proc. 1993 Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-1993)*, New Paltz, NY (IEEE, New York), pp. 225–228.
- Ding, Y. and Qian, X. (1997). "Sinusoidal and residual decomposition and residual modeling of musical tones using the QUASAR signal model," *Proc. 1997 Int. Computer Music Conf.*, Thessaloniki, Greece (Int. Computer Music Assoc., San Francisco), pp. 35–42.
- Dubnov, S., and Rodet, X. (1997). "Statistical modeling of sound aperiodicities," *Proc. 1997 Int. Computer Music Conf.*, Thessaloniki, Greece (Int. Computer Music Assoc., San Francisco), pp. 43–50.
- El-Jaroudy, A., and Makhoul, J. (1991). "Discrete all-pole modeling," *IEEE Trans Signal Processing* **39**, 411–423.
- Fant, G. (1970). *Acoustic Theory of Speech Production with Calculations Based on X-Ray Studies of Russian Articulations* (Mouton, The Hague).
- Flanagan, J. L. (1972). *Speech Analysis, Synthesis and Perception* (Springer-Verlag, Berlin).
- Fletcher, N. H. and Tarnopolsky, A. (1999). "Blowing pressure, power, and spectrum in trumpet playing," *J. Acoust. Soc. Am.* **105**(2), Pt. 1, 874–881.
- Freed, A., Rodet, X., and Depalle, P. (1993). "Synthesis and control of hundreds of sinusoidal partials on a desktop computer without custom hardware," *Proc. 1993 Int. Computer Music Conf.*, Tokyo, Japan (Int. Computer Music Assoc., San Francisco), pp. 98–101.
- Freed, A. (1995). "Bring your own control to additive synthesis," *Proc. 1995 Int. Computer Music Conf.*, Banff, Canada (Int. Computer Music Assoc., San Francisco), pp. 303–306.
- Freed, A. (1999). "Spectral line broadening with transform domain additive synthesis," *Proc. 1999 Int. Computer Music Conf.*, Beijing, China (Int. Computer Music Assoc., San Francisco), pp. 78–81.
- Fitz, K., Haken, L., and Holloway, B. (1995). "Lemur—A tool for timbre manipulation," *Proc. 1995 Int. Computer Music Conf.*, Banff, Canada (Int. Computer Music Assoc., San Francisco), pp. 158–161.
- Galas, T. and Rodet, X. (1990). "An improved cepstral method for deconvolution of source–filter systems with discrete spectra: Application to musical sound signals," *Proc. 1990 Int. Computer Music Conf.*, Glasgow, Scotland (Int. Computer Music Assoc., San Francisco), pp. 82–84.
- Galas, T., and Rodet, X. (1991a). "Generalized discrete cepstral analysis for deconvolution of source–filter systems with discrete spectra," *Final Program and Paper Summaries: 1991 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-1991)*, New Paltz, NY (IEEE, New York), Paper No. 3.2.
- Galas, T. and Rodet, X. (1991b). "Generalized functional approximation for source–filter system modeling," *Proc. 1991 European Conf. on Speech Communication and Technology*, Genoa, Italy, pp. 1085–1088.
- Giron, F. (1990). "Analyse et synthèse de sons de Shakuachi," rapport de stage de DEA d'Acoustique de l'université du Maine, IRCAM, October, 1990.
- Goodwin, M. (1996). "Residual modeling in music analysis-synthesis," *Proc. 1996 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP '96)*, Atlanta, GA, (IEEE, New York), pp. 1005–1008.

- Griffin, D. W., and Lim, J. S. (1985). "A new model-based speech analysis/synthesis system," *1985 Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP '85), Tampa, FL, (IEEE, New York), pp. 513–516.
- Hamming, R. W. (1977). *Digital Filters* (Prentice-Hall, Englewood Cliffs, NJ).
- Harris, F. J. (1978). "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE* **66**(1), 51–82.
- Holmes, J. N. (1983). "Formant synthesizers: Cascade or parallel," *Speech Communication* **2**(4), 251–273.
- Horner, A. and Beauchamp, J. W. (1995). "Synthesis of trumpet tones using a wavetable and a dynamic filter," *J. Audio Eng. Soc.* **43**(10), 799–812.
- Horner, A., and Beauchamp, J. W. (1996). "Piecewise linear approximation of additive synthesis envelopes: A comparison of various methods," *Computer Music J.* **20**(2), 72–95.
- Itakura, F. (1975). "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Am.* **57**, 535 (abstract).
- Kay, S. M. (1988). *Modern Spectral Estimation: Theory and Application* (Prentice Hall, Englewood Cliffs, NJ).
- Klatt, D.H. (1980) "Software for cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.* **67**, 971–995.
- Kopec, G. E. (1986). "Formant tracking using hidden Markov models and vector quantization," *IEEE Trans. on Acoustics, Speech and Signal Processing* **34**, 709–729.
- Laroche, J., Stylianou, Y., and Moulines, E. (1993). "HNS: Speech modification based on a harmonic + noise model," *Proc. 1993 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP '93), Minneapolis, MN, Vol. 2 (IEEE, New York), pp. 550–553.
- Laura, C. and Rodet, X. (1989). "Appariement de Pics Spectraux et règles pour la synthèse de la parole par concaténation de dipphones," *Actes du 1er Congrès français d'acoustique*, Lyon, France, pp. 531–536.
- Maher, R. C. and Beauchamp, J. W. (1990). "An investigation of vocal vibrato for synthesis," *Applied Acoustics* **30**(4), 219–245.
- Makoul, J. (1975). "Linear prediction: A tutorial review," *Proc. IEEE* **63**(4), 561–580.
- Marin, C. and McAdams, S. (1991). "Segregation of concurrent sounds. II: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width," *J. Acoust. Soc. Am.* **89**, 341–351.
- Markel, J. D. and Gray, A. H., Jr. (1980). *Linear Prediction of Speech* (Springer-Verlag, Berlin).
- Massie, D. C. and Stonick, V. L. (1992). "The musical intrigue of pole-zero pairs," *Proc. 1992 Int. Computer Music Conf.*, San Jose, CA (Int. Computer Music Assoc.: San Francisco), pp. 22–25.
- McAdams, S. and Rodet, X. (1988). "The role of FM-induced AM in dynamic spectral profile analysis," *Basic Issues in Hearing: Proc. 8th Int. Symposium on Hearing*, H. Duifhuis, J. Horst, and H. Wit, eds. (Academic Press, London) pp. 359–369.
- McAulay, R. J. and Quatieri, T. F. (1995). "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K.K. Paliwal, eds. (Elsevier Science, Amsterdam), pp. 121–173.
- McCandleem, S. S. (1974). "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. on Acoustics, Speech and Signal Processing ASSP-22*, 135–141.

- Mellody, M. and Wakefield, G. H. (1997). "A modal distribution study of violin vibrato," *Proc. 1997 Int. Computer Music Conf.*, Thessaloniki, Greece (Int. Computer Music Assoc., San Francisco), pp. 465–468.
- Mellody, M. and Wakefield, G. H. (2000). "The time-frequency characteristics of violin vibrato: Modal distribution analysis and synthesis," *J. Acoust. Soc. Am.* 107(1), 598–611.
- Moore, B. C. J. (1997). *An Introduction to the Psychology of Hearing* (Academic Press, San Diego).
- Moore, F. R. (1990). *Elements of Computer Music* (Prentice Hall, Englewood Cliffs, NJ).
- Moorer, J. A. (1979). "The Use of Linear Prediction of Speech in Computer Music Applications," *J. Audio Eng. Soc.* 27(3), 134–140.
- Niranjan, M. and Cox, I. J. (1994). "Recursive tracking of formants in speech signals," *Proc. 1994 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP '94)*, Adelaide, South Australia, Vol. 2 (IEEE, New York), pp. 205–208.
- Olive, J. P. (1971). "Automatic formant tracking by a Newton-Raphson technique," *J. Acoust. Soc. Am.* 50(2), 661–670.
- Oppenheim, A. V. (1978). "Digital processing of speech," in *Applications of Digital Signal Processing*, A. V. Oppenheim, ed., (Prentice-Hall, Englewood Cliffs, NJ), pp. 117–168.
- Oppenheim, A. V., and Schafer, R. W. (1975). *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ).
- Oudot, M., Cappé, O., and Moulines, E. (1997). "Robust estimation of the spectral envelope for 'harmonics+noise' models," *Proc. 1997 IEEE Workshop on Speech Coding for Telecommunications*, Pocono Manor, PA (IEEE, New York), 11–12.
- Oudot, M. (1998). "Analyse/synthèse des signaux de parole à partir d'un modèle de sinusoides et de bruit. Application au codage bas débit et aux transformations prosodiques [Speech analysis/synthesis using harmonic sinewaves and noise. Application to low-bit-rate-coding and prosodic transformations]," Ecole Nationale Supérieure de Télécommunications.
- Peeters, G. and Rodet, X. (1998). "Signal characterization in terms of sinusoidal and non-sinusoidal components," *Proc. First COST-G6 Workshop on Digital Audio Effects (DAFX98)*, Barcelona, Spain.
- Pierucci, P., and Paladin, A. (1997). "Singing voice analysis and synthesis system through glottal excited formant resonators," *Proc. 1997 Int. Computer Music Conf.*, Thessaloniki, Greece (Int. Computer Music Assoc., San Francisco), pp. 168–171.
- Potard, Y., Baisnée, P. F., and Barrière, J. B. (1986). "Experimenting with models of resonance produced by a new technique for the analysis of impulsive sounds," *Proc. 1986 Int. Computer Music Conf.*, The Hague, Netherlands (Int. Computer Music Assoc., San Francisco), pp. 269–274.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. (Cambridge University Press, New York).
- Risset, J.-C., and Mathews, M. V. (1969). "Analysis of musical-instrument tones," *Physics Today* 22(2), 23–30.
- Rodet, X. and Delatre, J. (1979). "Time-domain speech synthesis by rules using a flexible and fast signal management system," *Proc. 1979 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP '79)*, Washington, D.C. (IEEE, New York), pp. 895–898.
- Rodet, X. (1980). "Time-domain formant-wave-function synthesis," *Spoken Language Generation and Understanding: Proc. NATO Advanced Study Institute*, Bonas, France, J.C. Simon, ed. (D. Reidel Pub. Co., Dordrecht, Holland), pp. 429–441.

- Rodet, X. (1984). "Time-domain formant-wave-function synthesis," *Computer Music J.* **8**(3), 9–14.
- Rodet, X., Potard, Y., and Barrière, J. B. (1984). "The CHANT Project: From synthesis of the singing voice to synthesis in general," *Computer Music J.* **8**(3), 15–31.
- Rodet, X., and Depalle, P. (1985). "Synthesis by rule: LPC diphones and calculation of formant trajectories," *Proc. 1985 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP '85), Tampa, FL, Vol. 2 (IEEE, New York), pp. 736–739.
- Rodet, X., and Depalle, P. (1986). "Use of LPC spectral estimation for music analysis, processing and synthesis," *Final Program and Paper Summaries for the 1986 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (WASPAA-1986), New Paltz, NY (IEEE, New York), Paper No. 5.5.
- Rodet, X., Depalle, P., and Poirot, G. (1987). "Speech analysis and synthesis methods based on spectral envelopes and voiced/unvoiced functions," *European Conf. on Speech Technology*, 1987, Edinburgh, U.K., pp. 155–158.
- Rodet, X., Depalle, P., and Poirot, G. (1988). "Diphone sound synthesis based on spectral envelopes and harmonic/noise excitation functions," *Proc. 1988 Int. Computer Music Conf.*, Cologne, Germany (Int. Computer Music Assoc., San Francisco), pp. 313–321.
- Rodet, X. and Depalle, P. (1992). "Spectral envelopes and inverse FFT synthesis", *93rd Convention of the Audio Eng. Soc.*, San Francisco, CA, Audio Eng. Soc. Preprint No. 3393.
- Rodet, X., Depalle, P., and Garcia, G. (1995). "New possibilities in sound analysis and synthesis." *Proc. 1995 Int. Symposium on Musical Acoustics*, Dourdan, France.
- Rodet, X., and Lefèvre, A. (1997). "The Diphone program: New features, new synthesis methods and experience of musical use," *Proc. 1997 Int. Computer Music Conf.*, Thessaloniki, Greece (Int. Computer Music Assoc., San Francisco), pp. 418–421.
- Sandler, M. B. (1989). "Auto Regressive modelling and synthesis of acoustic instruments," *86th Convention of the Audio Eng. Soc.*, Hamburg, Germany, Audio Eng. Soc. Preprint 2758.
- Schafer, R. W. and Rabiner, L. R. (1970). "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Am.* **47**, 634–648.
- Schwarz, D. (1998). *Spectral Envelopes in Sound Analysis and Synthesis*, Universität Stuttgart, Fakultät Informatik, Diplomarbeit Nr. 1622, Stuttgart, Germany, June 1998.
- Schwarz, D. and Rodet, X. (1999). "Spectral envelope estimation and representation for sound analysis-synthesis," *Proc. 1999 Int. Computer Music Conf.*, Beijing, China (Int. Computer Music Assoc., San Francisco), pp. 351–354.
- Serra, X. (1989). "A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic plus Stochastic Decomposition," Doctoral dissertation, Stanford University, Stanford, CA. *Dissertation Abstracts International-A*. **51/01**, 18.
- Serra, X. and Smith, J. O. (1990). "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music J.* **14**(4), 12–24.
- Serra, X., Bonada, J., Herrera, P., and Loureiro, R. (1997). "Integrating complementary spectral models in the design of a musical synthesizer," *Proc. 1997 Int. Computer Music Conf.*, Thessaloniki, Greece (Int. Computer Music Assoc., San Francisco), pp. 152–159.
- Smith, J. O. (1985). "Introduction to digital filter theory," in *Digital Audio Signal Processing: An Anthology*, J. Strawn, ed. (William Kaufmann, Los Altos, CA), pp. 69–135. Also available as Stanford University Dept. of Music Technical Report STAN-M-20.

- Soong, F. K., and Juang, B.-H. (1984). "Line spectrum pair (LSP) and speech data compression," *Proc. 1984 IEEE Int. Conf. on Acoustics, Speech and Digital Processing* (ICASSP '84), San Diego, CA (IEEE, New York), pp. 1.10.1–1.10.4.
- Unser, M., Aldroubi, A., and Eden, M. (1993). "B-spline signal processing: I—Theory," *IEEE Trans. Signal Processing* **41**, 821–833.
- Virolle, D., Schwarz, D., and Rodet, X. (2001). *SDIF: Sound Description Interchange Format*. [see <http://recherche.ircam.fr/sdif/>].
- Vishwanathan, R., and Makhoul, J. (1978). "Adaptive lattice methods for linear prediction," *Proc. 1978 IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (ICASSP '78), Tulsa, OK (IEEE, New York), pp. 83–86.
- Wanderley, M. M., Schnell, N., and Rovan, J. (1998). "ESCHER—Modeling and performing composed instruments in real-time," *Proc. 1998 IEEE Symposium on Systems, Man, and Cybernetics*, San Diego, CA (IEEE, New York), pp. 1080–1084.
- Wright, M., Chaudhary, A., Freed, A., Wessel, D., Rodet, X., Virolle, D., Woehrmann, R., and Serra, X. (1998). "New applications of the sound description interchange format," *Proc. 1998 Int. Computer Music Conf.*, Ann Arbor, MI (Int. Computer Music Assoc., San Francisco), pp. 276–279.

A Comparison of Wavetable and FM Data Reduction Methods for Resynthesis of Musical Sounds

ANDREW HORNER

1 Introduction

An ideal music-synthesis technique provides both high-level spectral control and efficient computation. Simple playback of recorded samples lacks spectral control, while additive sine-wave synthesis is inefficient. Wavetable and frequency-modulation synthesis, however, are two popular synthesis techniques that are very efficient and use only a few control parameters.

The term “wavetable synthesis” is currently often used synonymously with “sampling synthesis” in the music industry. However, throughout this chapter the classical computer music meaning of “wavetable synthesis” is used, based on the use of oscillator tables loaded with sums of harmonic sinusoids and indexed by phase functions that depend on a fundamental frequency. With multiple wavetable synthesis, several wavetables can be independently amplitude-controlled and summed (wavetable indexing) or simply crossfaded one after the other (wavetable interpolation) to produce time-varying spectral changes.

There are several types of frequency-modulation (FM) synthesis, including formant FM with multiple carriers, double FM with multiple parallel modulators, and nested FM with serial modulators (see Fig. 6.1). During the height of FM’s popularity in the 1980s, synthesizers such as Yamaha’s DX7 allowed users great flexibility in mixing and matching with these models.

A fundamental problem of computer music is to automatically generate good parameters for resynthesizing recorded instrument tones. Recent work has also shown how to optimize wavetable and FM parameters (Serra et al., 1990; Horner et al., 1993a, b; Horner and Beauchamp, 1996; Horner, 1996a, b; Horner, 1998) for best matching resynthesized tones to specific original tones.

So, which type of wavetable or FM synthesis is best? Which uses the least memory, and which uses the least computation? This chapter compares methods for matching harmonic instrument tones with various wavetable and FM models. Section 2 describes the optimization methods used and the metric employed to measure how well a synthesis method matches instrument time-varying spectra. Section 3 reviews various wavetable and FM synthesis and parameter-matching

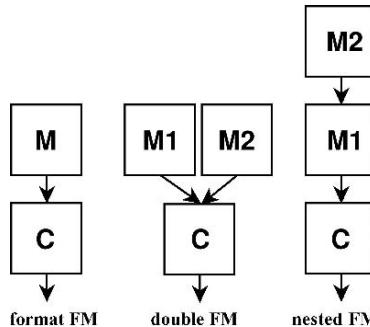


FIGURE 6.1. Block diagrams of three basic FM synthesis methods.

methods. Section 4 gives results for trumpet, tenor voice, and Chinese pipa sounds. In each case, wavetable and FM results are compared and conclusions about their effectiveness for matching acoustic instrument tones are given in terms of the number of modules required for a given synthesis error and the amount of table lookups required.

2 Evaluation of Wavetable and FM Methods

The performance of a synthesis method can be measured in terms of how well it can resynthesize particular musical instrument tones. Synthesis parameters are evaluated according to methods developed in previous parameter-matching studies using genetic algorithms (GA) (e.g., Horner et al., 1993a, b). Figure 6.2 shows the evaluation procedure overview.

First, original sounds are transformed into corresponding time-varying frequency-domain spectra using phase vocoder (Allen, 1977; Dolson, 1986) or frequency-tracking (McAulay and Quatieri, 1986) short-time spectral analysis techniques. Beauchamp (1993) gives more details on the application of these methods for the analysis of quasiharmonic sounds.

Next, parameter sets for a particular synthesis method are postulated and corresponding spectra are computed. For each parameter set synthetic and original spectra are compared. The collection of parameter sets to be examined is called a parameter space. Wavetable and FM parameter spaces are generally too big to allow brute-force enumeration of all the possible parameter combinations. For example, if each of 20 harmonics in a wavetable were allowed to take on relative amplitudes between 0.00 and 1.00 with a resolution of 0.01, 100^{20} combinations would have to be tested, a task not possible in a lifetime, even with current technology. Special optimization algorithms are necessary to arrive at parameters that allow low-cost synthesizers to closely approximate an original complex sound.

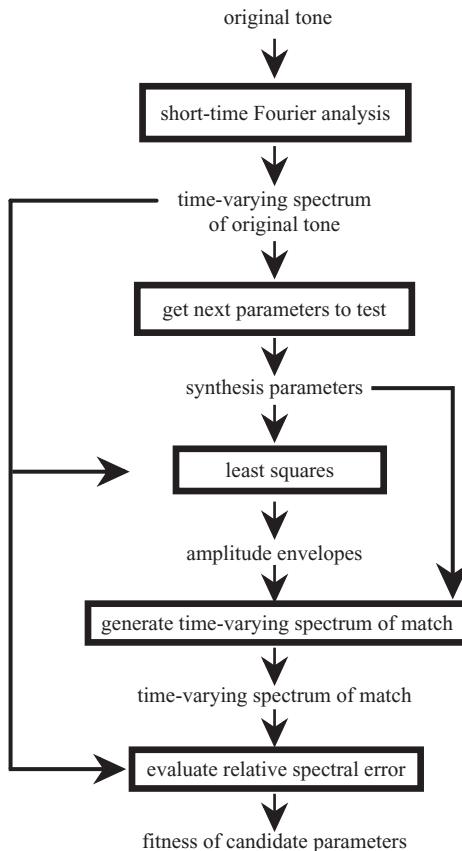


FIGURE 6.2. Overview of the analysis, matching synthesis, and evaluation procedure.

An efficient approach is to choose certain fixed (time-invariant) synthesis parameters using GA optimization (Holland 1975; Goldberg 1989) or by using an enumerative method in the few instances when the search space is small enough to explore by brute force. However, to match a time-varying spectrum with fixed parameters (which determine wavetable or FM carrier outputs), the time-varying amplitude envelopes associated with each wavetable or FM carrier must be computed. Fortunately, it turns out that calculating amplitude envelopes is a linear problem that can be solved by the method of least-squares (Press et al., 1985), thus easily providing a set of amplitude envelopes that minimize the spectral error between the original and synthetic signals (Horner et al., 1993b).

Next, the time-varying spectrum of the synthetic signal is determined from the synthesis parameters and amplitude envelopes. A *relative-amplitude spectral error* function is then used to evaluate the difference between the original and synthetic

signals, averaged over time:

$$\varepsilon_{rel} = \frac{1}{N_{frames}} \sum_{i=1}^{N_{frames}} \left[\frac{\sum_{k=1}^{N_{hars}} (b_k(t_i) - b'_k(t_i))^2}{\sum_{k=1}^{N_{hars}} b_k^2(t_i)} \right]^{1/2}, \quad (6.1)$$

where $b_k(t)$ and $b'_k(t)$ are the time-varying k th harmonic amplitudes of the original and synthetic spectrum, respectively, the t_i are times at which these functions are sampled, N_{frames} is the number of time values used in the measurement, and N_{hars} is the number of harmonics.

The times t_i used in the error calculation yield a relatively small number of representative test spectra and are not necessarily equally spaced. Judicious choice of these times allows the error function to be weighted appropriately. For example, times taken from the attack portion of a tone are very good choices, because the attack is a perceptually critical and a fast-changing portion of the tone (Clark et al., 1963; Berger, 1964; Grey and Moorer, 1977). After some experimentation, a simple method was determined where half of the times are taken (equally spaced) from the attack portion of the tone (where peak RMS amplitude defines the end of the attack) and the others (again, equally spaced) are taken from the rest of the tone. In practice, 20 error measurement times are taken altogether, with 10 equally spaced times selected from each of the two time regions.

Note that the relative-amplitude spectral error returns zero if an exact match occurs, and an error of 0.1 represents a 10% average spectral error. An error of 1.0 results from comparing silence to the original tone. The relative-amplitude spectral error is not necessarily a perfect measure of a match's subjective quality. It is possible that a match with slightly more error may sound more like an original tone than one with less error. However, Eq. (6.1) is reasonably accurate and very efficient.

3 Comparison of Wavetable and FM Methods

This section gives a brief overview of the generalized wavetable, wavetable indexing, wavetable interpolation, group additive, formant FM, double FM, and nested FM matching synthesis methods. Section 4 shows how well each of these methods can simulate three different original musical instrument tones.

Figure 6.3 shows the relation of the different parameter spaces to one another. These spaces are subsets of all spectra possible in generalized wavetable space. For example, FM is a subset of this space because a wavetable can simulate the spectrum produced by an FM module with fixed modulation indices and integer carrier-modulator ratios. Also, double and nested FM contain formant FM as a special case at their intersection, where the second modulator has a modulation index of zero. There may be some overlap between spaces such as wavetable indexing and FM, but that depends on the original tone under consideration.

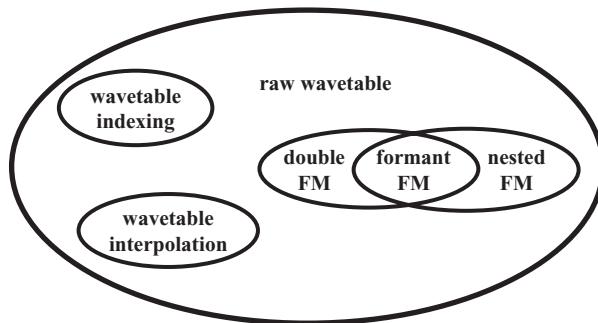


FIGURE 6.3. Spectral subspace relationships of the various wavetable and FM synthesis methods.

3.1 Generalized Wavetable Matching

The most straightforward approach to wavetable matching is to attempt direct optimization of the harmonic amplitudes used as basis spectra (Horner et al., 1993b), a method called generalized wavetable matching (also called “raw wavetable” matching), since it considers solutions throughout the entire wavetable space. Figure 6.4 shows the generalized wavetable parameters in a multiple wavetable block diagram. Because it is impractical to consider all possible matches, even for the one wavetable case, some sort of optimization procedure is always necessary.

3.2 Wavetable-Index Matching

Wavetable-index matching refers to a method of selecting spectra (or corresponding wavetables) from an ensemble of spectra that represent a sound. For this method, an index corresponds to a frame number of a time-varying spectrum used in the matching process. Instead of postulating spectra using some arbitrary method, this method selects a limited number of spectra from the original tone’s time-varying spectrum as basis spectra. This approach is intuitive, guarantees an exact match at the times corresponding to these selected “spectral snapshots,” and usually makes excellent matches at neighboring points as well. Wavetable indexing in a sense “cheats” by taking parameters directly from the original sound, rather than postulating some general synthesis parameters. Figure 6.5 illustrates a multiple wavetable synthesizer containing three hypothetical basis spectra. These are first converted into three waveforms, then they are amplitude-controlled by three corresponding envelope functions, and finally they are summed to form the final output. Note that with wavetable indexing there is no restriction on the envelope functions other than that the resulting time-varying spectrum should be as close as possible to that of the original sound.

With typically 500–2000 spectral snapshots to choose from for each tone, brute-force enumeration of indices can actually be used to test all the possible index choices for 1- or 2-table matches in a reasonable amount of time. For example, for

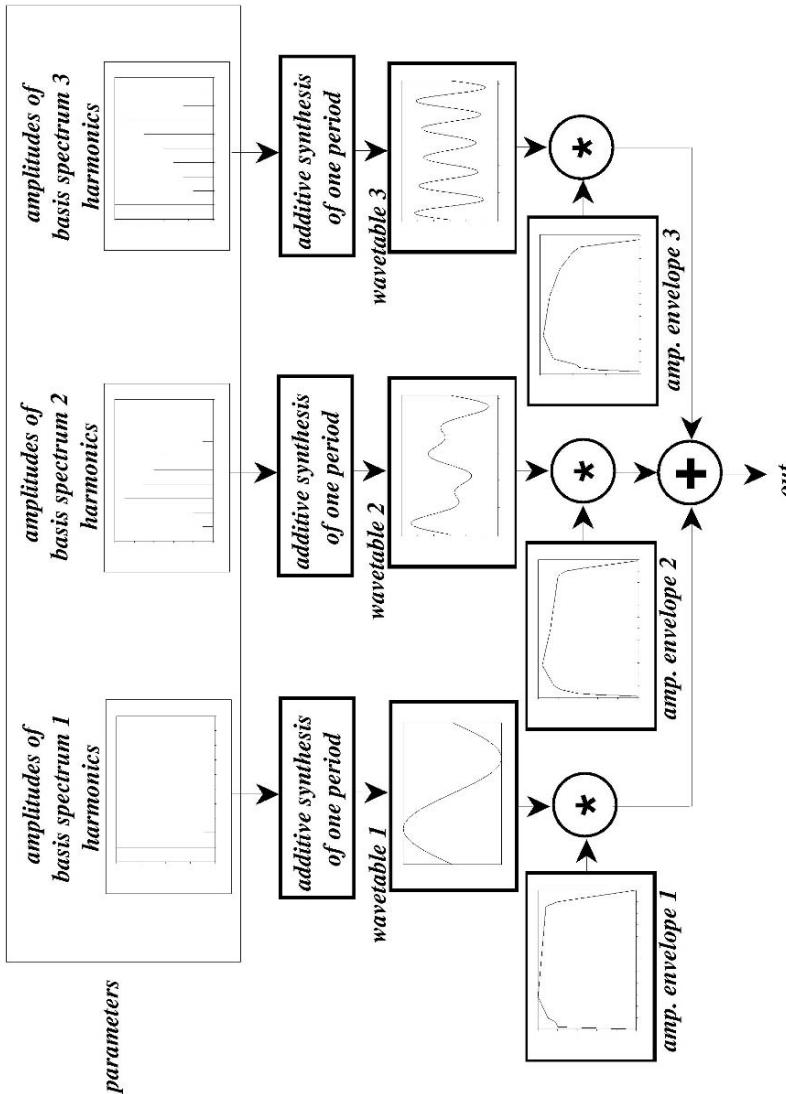


FIGURE 6.4. Generalized wavetable-synthesis block diagram (with three basis spectra) and typical parameters. Basis spectra are arbitrary and do not necessarily correspond to spectra contained in the original sound.

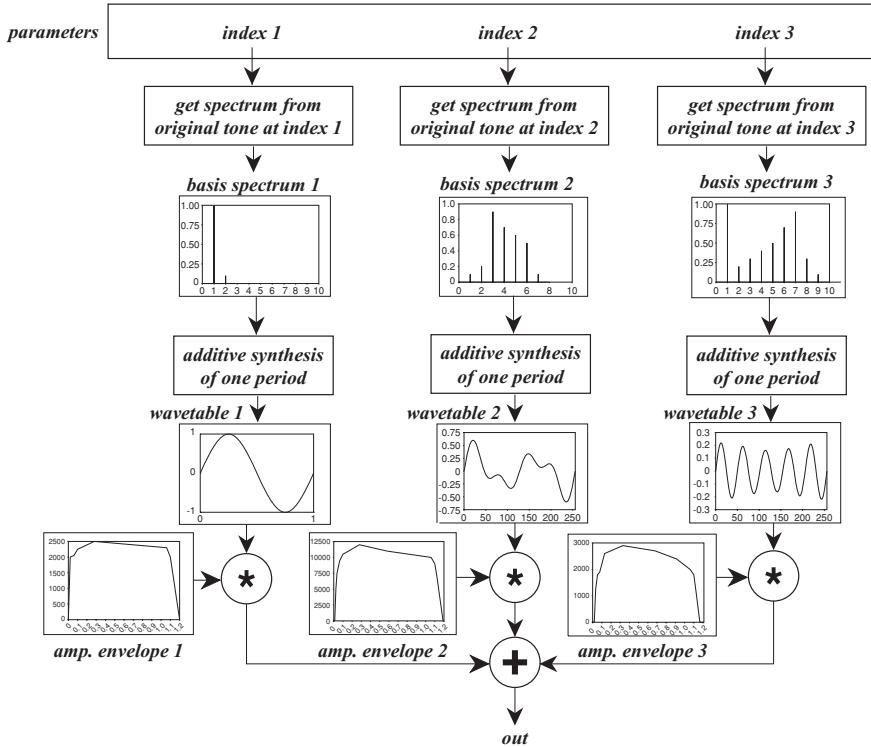


FIGURE 6.5. Wavetable-index-synthesis block diagram (with three basis spectra) and typical parameters. Basis spectra are taken from original sound.

the 1-table case, it is simply a matter of determining which frame gives the least error according to Eq. (6.1). In this case, the synthetic harmonic amplitudes $b'_k(t)$ would vary in time, but their ratios would always correspond to the single selected spectrum. For the 2-table case, some spectral variation is possible. [Recall that each harmonic amplitude corresponds to a sum of weights derived from a least-squares calculation multiplied by the fixed harmonic amplitudes of the basis wavetable spectra. See Horner et al. (1993b) for details.] However, for three or more tables, optimization is required.

3.3 Wavetable-Interpolation Matching

Wavetable interpolation is a special case of wavetable indexing where wavetables cross-fade two-at-a-time (Serra et al., 1990; Horner and Beauchamp, 1996). Again, the wavetable spectra are taken directly from the original sound's time-varying spectrum, so like wavetable indexing, wavetable interpolation “cheats” by limiting the wavetable space to waveforms in the actual signal. In addition, only linear cross-fade interpolation, the simplest interpolation

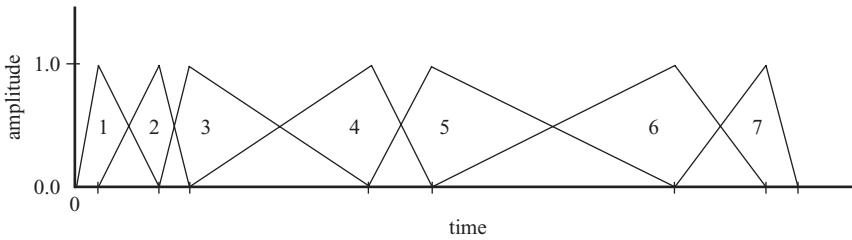


FIGURE 6.6. Example set of cross-fading amplitude envelopes used for wavetable-interpolation synthesis. The numbers refer to the wavetable numbers.

method, is considered in this chapter. Figure 6.6 shows a typical set of linear-wavetable-interpolation amplitude envelopes, showing that each new wavetable begins to fade in at the same time the previous one begins to fade out. Figure 6.7 depicts a wavetable interpolation synthesizer with a complete set of parameters for the three-wavetable case. Note that it is identical to Fig. 6.5 except for the

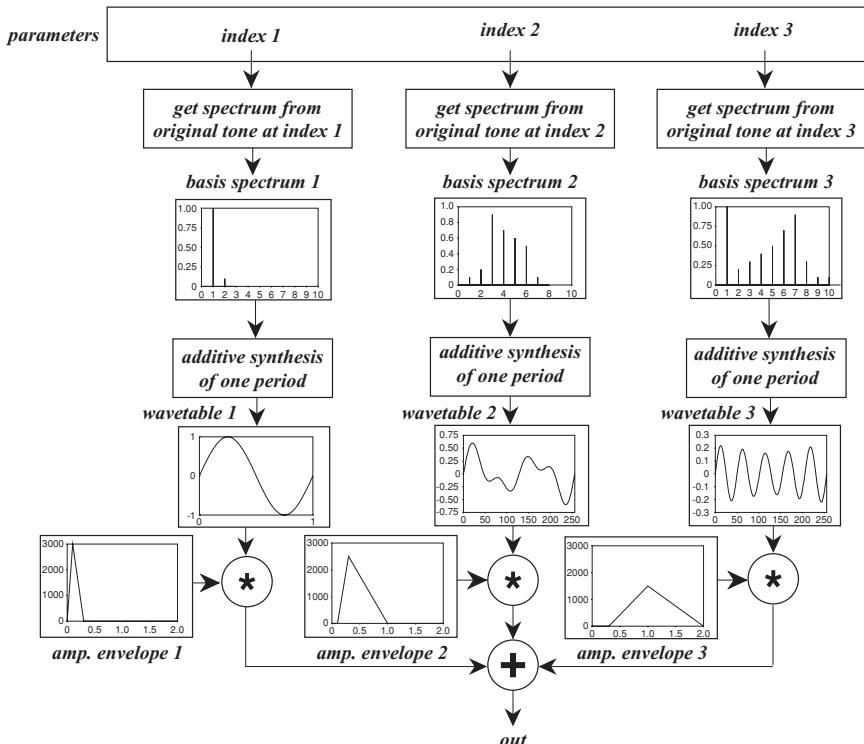


FIGURE 6.7. Wavetable-interpolation-synthesis block diagram (with three basis spectra) and typical parameters. Basis spectra are taken from original sound.

amplitude envelopes. However, for the same synthesis accuracy, the total number of basis spectra usually must be much larger for wavetable interpolation than for wavetable indexing.

As a more general approach, wavetable indexing generally performs better than simple wavetable interpolation when the same number of total wavetables is used. This is because wavetable indexing can use more than two active wavetables at any given time, and its amplitude envelopes are not restricted to linear cross-fades (they can even go negative). However, wavetable interpolation is more intuitive for these same reasons. Also, because wavetable interpolation only uses two active wavetables at any given time, it is better for situations with limited computation capability. Still, can wavetable interpolation achieve a match almost as good as wavetable indexing, or is it always much worse? Or, on the other hand, can wavetable indexing achieve almost the same level of efficiency as wavetable interpolation?

Like wavetable indexing, wavetable-interpolation matching initially selects from typically 500–2000 spectral snapshots of an original tone and then, in synthesis, cross-fades between a few spectra chosen from this group. As mentioned previously, brute-force enumeration can be used to pick one or two basis spectra (wavetables), but special optimization methods are required for the case of three or more wavetables.

3.4 Formant-FM Matching

Formant FM is achieved by a single modulator oscillator driving one or more carrier oscillators whose carrier frequencies are integer multiples of the modulator frequency (Chowning, 1973, 1980). The name stems from the fact that when modulation indices are low and the modulation frequency is equal to or lower than a carrier frequency, each carrier produces a relatively narrow band of components corresponding to a spectral resonance or formant. While many FM-synthesis patches (including the DX7 patches) use time-varying modulation indices, spectral oscillations resulting from time-varying indices make it difficult to match acoustic instruments (Horner et al., 1993a). Therefore, only fixed modulation indices are considered here. The GA optimization procedure restricted modulator frequency ratios to integer values between 0 and 15 and modulation indices to values between 0 and 12.7, in increments of 0.1.

FM-generated spectra often produce negative-amplitude partials, corresponding to 180° phase shifts from their positive amplitude counterparts (see Fig. 6.8). Unfortunately, negative-amplitude components can cause spectral differences due to cancellations between carriers even if the absolute values of the components match the originals exactly. Whether components are positive or negative must be determined before the amplitude envelopes can be constructed and a match evaluated. More details about FM signs are given in previous papers (Horner et al., 1993a; Horner, 1996a). Figure 6.9 shows a formant FM block diagram with typical parameters. For this FM case a relative-amplitude spectral error function somewhat

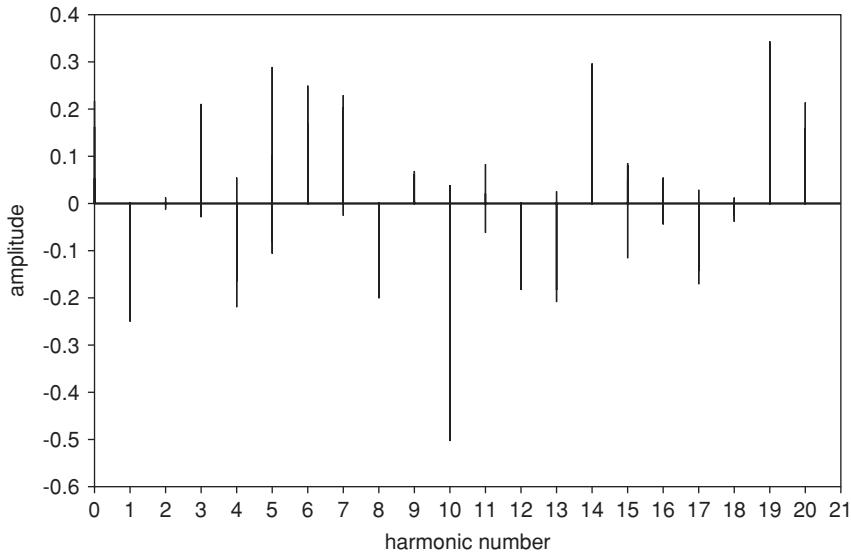


FIGURE 6.8. FM-generated spectrum with both positive and negative amplitude components.

different than that of Eq. (6.1) was constructed. It takes longer to compute than the wavetable fitness functions because of the importance of spectral sign selection.

3.5 Double-FM Matching

Double-modulator FM (Schottstaedt, 1977; LeBrun, 1977) uses two modulators for each carrier instead of one as used in formant FM. Recent work has shown how to compute double-FM parameters (Tan and Lim, 1996; Horner, 1996a) for matching acoustic instrument sounds. Again, only fixed-modulation indices are considered, as in the case of formant FM. Figure 6.10 shows a double-FM block diagram and parameters for the three-carrier case. Note that in this configuration each carrier, whose frequency is tuned to an integer multiple of the fundamental frequency f_1 , receives phase information from one common modulator tuned to f_1 and one independent modulator tuned to an integer multiple of f_1 . Thus, there are two indices and two integer frequency ratios to determine for each carrier.

The GA-optimization procedure restricts each carrier's second modulator to take on integer frequency ratios between 1 and 4, because modulation-frequency ratios greater than about 4 produce spectra with strong formants in upper harmonics, not a common feature of acoustic instruments. It allows both modulation index values to range between 0 and 12.7, in increments of 0.1.

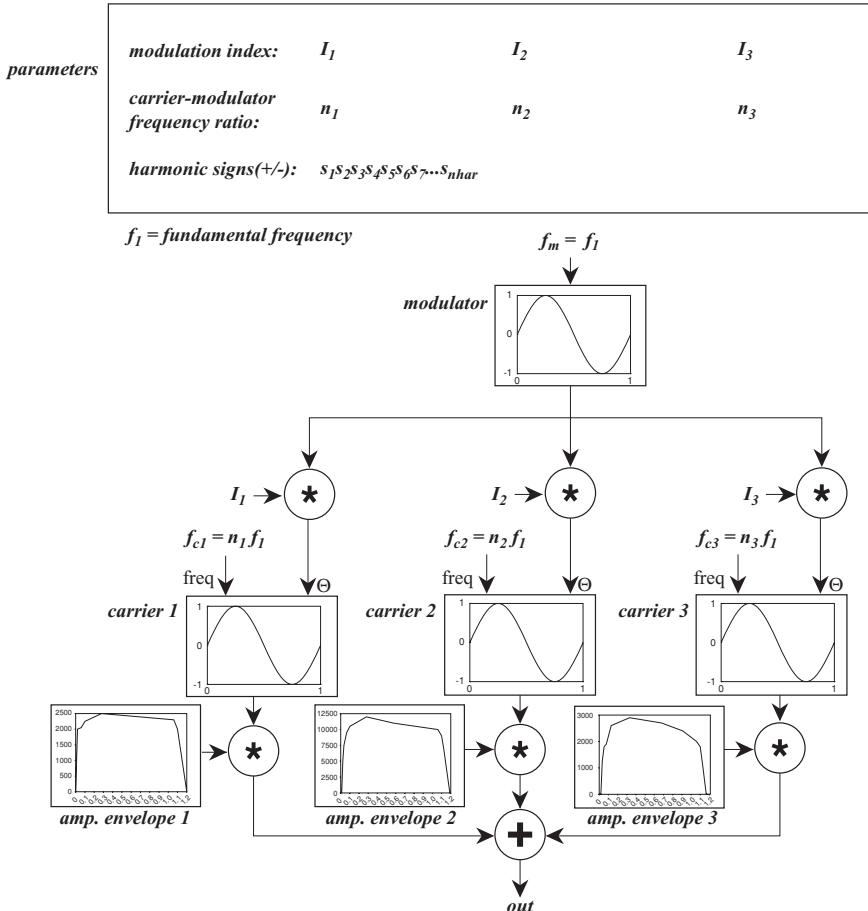


FIGURE 6.9. Formant-FM synthesis block diagram (with three carriers) and typical parameters.

3.6 Nested-FM Matching

Justice (1979) introduced a nested-modulator FM model, which utilized two modulators connected in serial rather than parallel. Payne (1987) extended Justice's model to a pair of carriers with nested modulators. The second carrier was added so that the two carriers could contribute to independent frequency regions. While this allows a more accurate match than a single carrier, it doubles the amount of computation required for resynthesis. Horner (1998) described how to match nested-FM parameters for an arbitrary number of carriers. Figure 6.11 shows a nested FM block diagram and parameters for the case of three carriers. Note that each carrier's independent modulator receives phase information from a common modulator tuned to the fundamental f_1 . Again, two indices and two integer frequency ratios must be determined for each carrier.

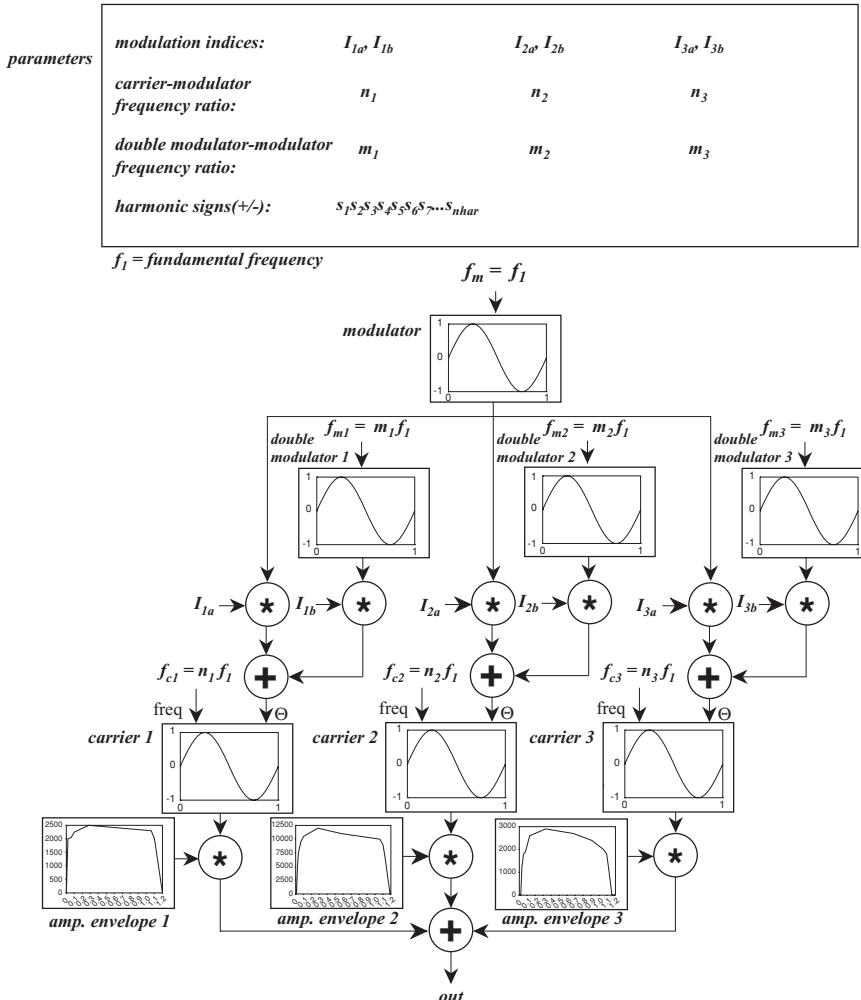


FIGURE 6.10. Double-FM synthesis block diagram (with three carriers) and typical parameters.

The number of modulators is not necessarily restricted to two in nested modulator FM. In fact, Justice's method interactively added nested modulators. Of course, each modulator adds more complexity to the model, making optimization more difficult. In this chapter, investigation is limited to a single nested modulator.

Again, only fixed modulation indices are used. Both modulators are restricted to take on integer frequency ratios between 1 and 4, allowing the carrier's formant to range up to the fourth harmonic. Smaller frequency ratios tend to be the most useful in nested FM, because they help concentrate the energy in the lower harmonics as in most instrument tones. Both modulation index values range between 0 and 6.3, in increments of 0.1.

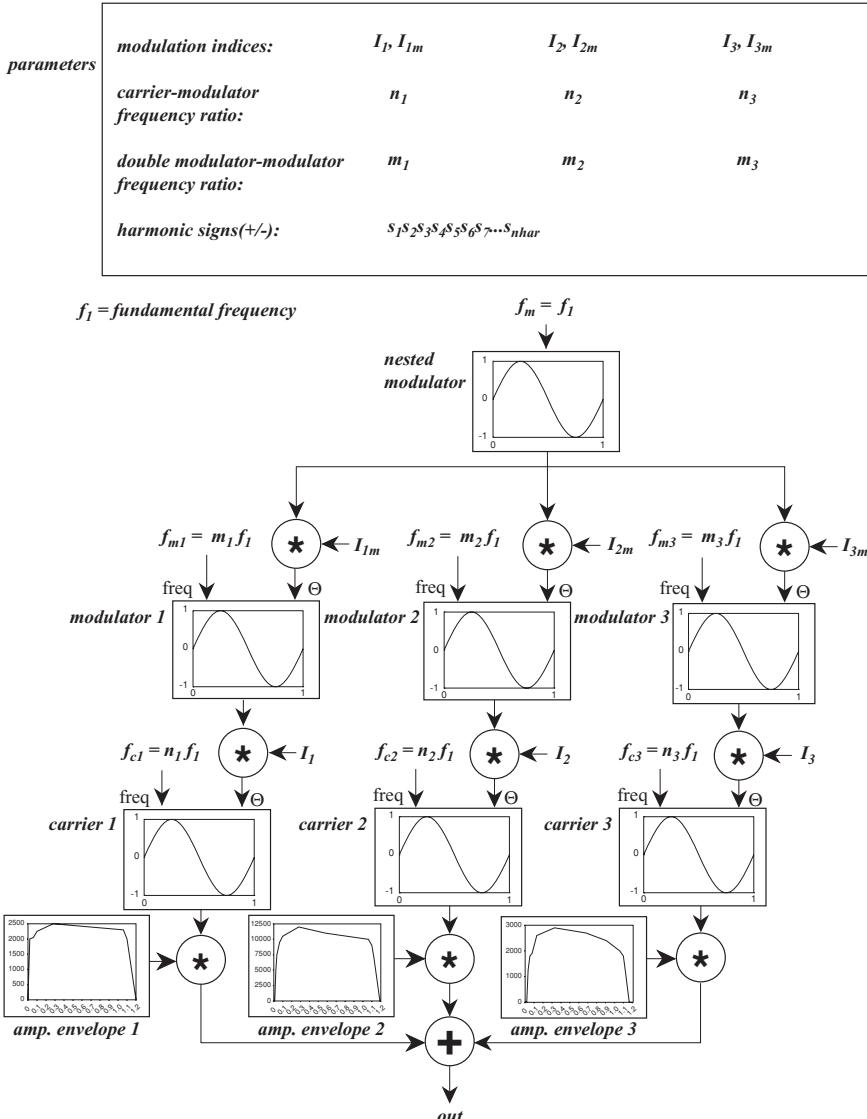


FIGURE 6.11. Nested-FM synthesis block diagram (with three carriers) and typical parameters.

4 Results

The wavetable and FM matching procedures were tested on several musical instruments sounds including those of the trumpet, tenor voice, and Chinese pipa. For each sound, the relative-amplitude spectral error is plotted against the number of wavetables/carriers to show how the error decreases as more units are added to the

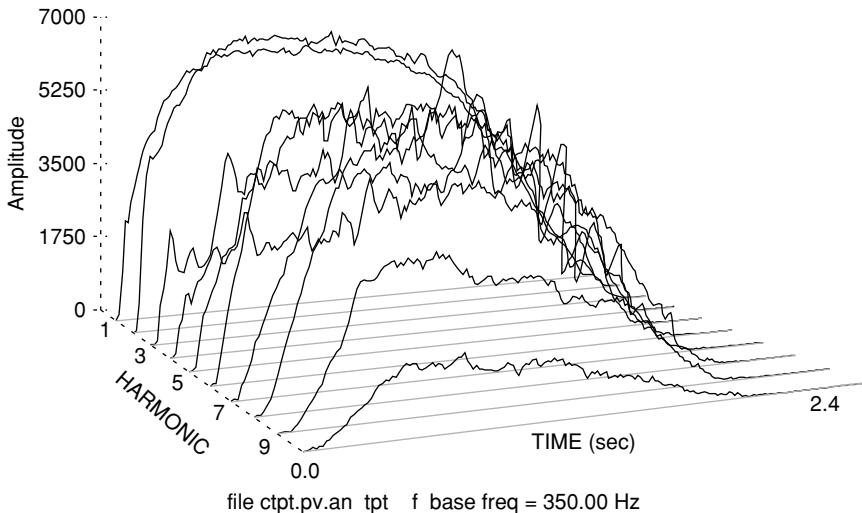


FIGURE 6.12. Time-varying amplitude spectrum of a 350 Hz (F₄) trumpet tone.

synthesis model. To compare the computational efficiency of the various wavetable and FM models, error is also plotted against the number of table lookups for each method. For each instrument sound, 20 harmonics were used in the matching process.

4.1 The Trumpet

Figure 6.12 shows the amplitude-vs-time envelopes of an F₄ (350 Hz) trumpet tone's first 10 harmonics. Note that harmonic envelopes 3 and 4 have different shapes from the others. Also, the higher harmonics reach their peak more slowly and decay faster than the lower harmonics, a common characteristic of brass instruments.

Graphs of relative-amplitude error vs number of wavetables or FM carriers used by the various synthesis methods for the trumpet tone are overlayed in Fig. 6.13. The knees of the curves generally occur between four and five wavetables or carriers. The wavetable index method consistently returns the lowest error, with nested FM the next best for three or more carriers. Figure 6.13 compares the methods for cases where hardware is already available for wavetable or FM synthesis, and it provides useful data for making a decision on the number of wavetables or carriers necessary to achieve a desired spectral error. However, it does not make a fair comparison in terms of total synthesis computation.

Figure 6.14 compares the computational efficiency of the different synthesis methods by plotting relative-amplitude error against the number of table lookups required to compute each output sample. Note that wavetable interpolation uses only two lookups no matter how many wavetables are installed and is therefore the best choice if computation is the main concern, as is often

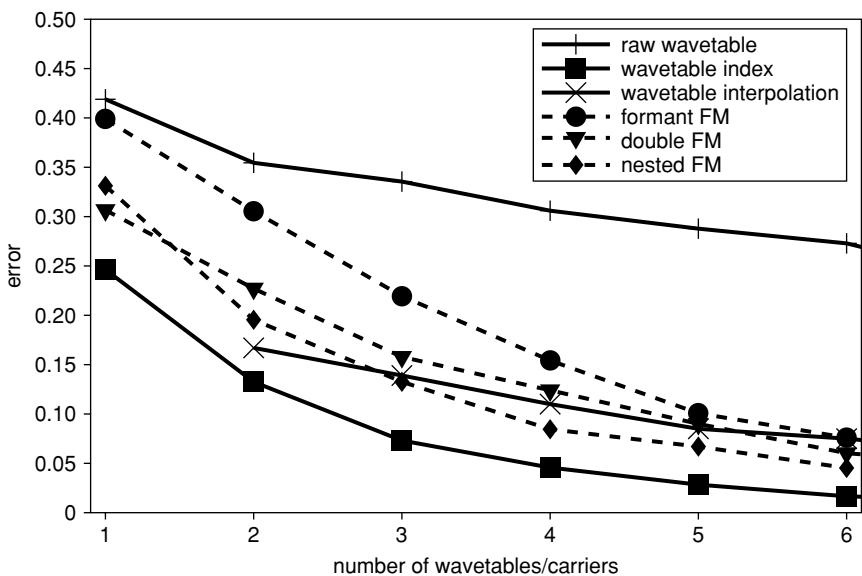


FIGURE 6.13. Convergence of relative error for different numbers of wavetables/carriers for the trumpet tone.

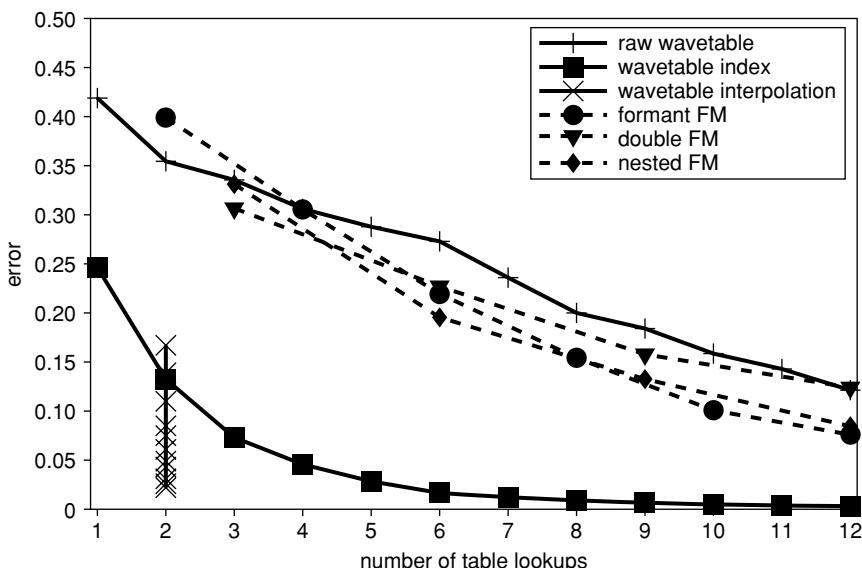


FIGURE 6.14. Convergence of error for different numbers of table lookups for the trumpet tone.

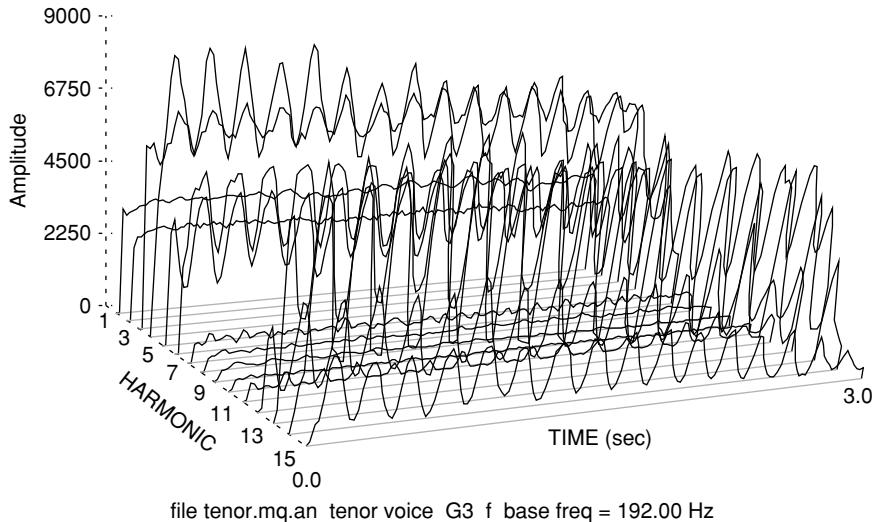


FIGURE 6.15. Time-varying amplitude spectrum of a G3 tenor voice tone.

the case in software synthesis applications. For the trumpet tone, wavetable interpolation was able to reduce the error to 2% with 12 wavetables. Wavetable indexing required six wavetables to reach the same accuracy level, corresponding to about three times the amount of computation as wavetable interpolation.

4.2 The Tenor Voice

Figure 6.15 shows the time-varying amplitude spectrum of a G_3 (192 Hz) tenor voice tone. This tone has a wide frequency vibrato with accompanying amplitude modulation which is very strong for harmonics 3–6 and 12–15, while having little effect on harmonics 1–2 and 7–11. There is a prominent formant resonance around harmonics 13 and 14 (approximately 2600 Hz), corresponding to the “singing formant” (Sundberg, 1974).

Figure 6.16 shows relative error plotted against numbers of wavetables or FM carriers. Wavetable indexing again gives the best results, but wavetable interpolation is a close second. This makes sense because as the tenor’s harmonic frequencies swing back and forth they cause the harmonic amplitudes to swing between two different sets of points on a fixed spectral envelope (Maher and Beauchamp, 1990). Thus, wavetable interpolation can easily cross-fade back and forth in synchronization with the tenor’s amplitude modulation.

Figure 6.17 shows relative error plotted against number of table lookups. Wavetable indexing with two wavetables requires about the same computation as wavetable interpolation. In general, wavetable indexing and interpolation perform

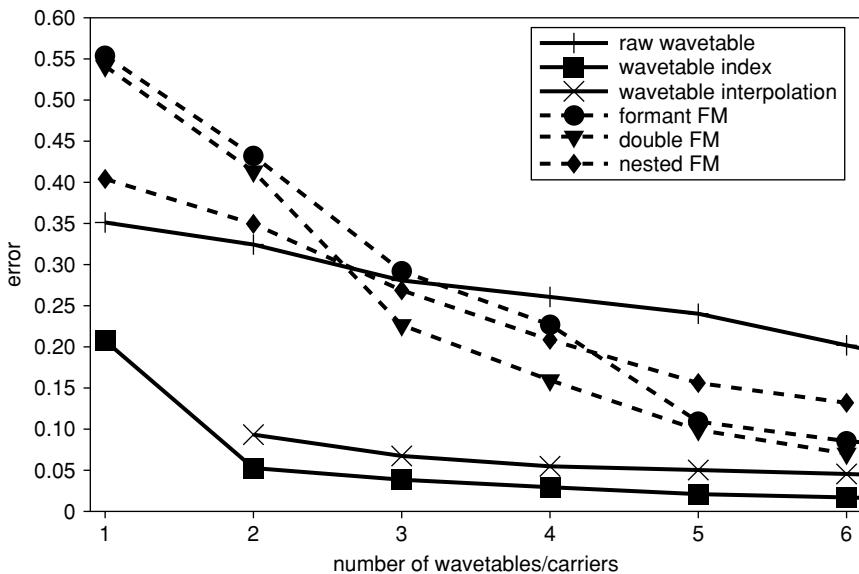


FIGURE 6.16. Convergence of error for different numbers of wavetables/carriers for the tenor voice tone.

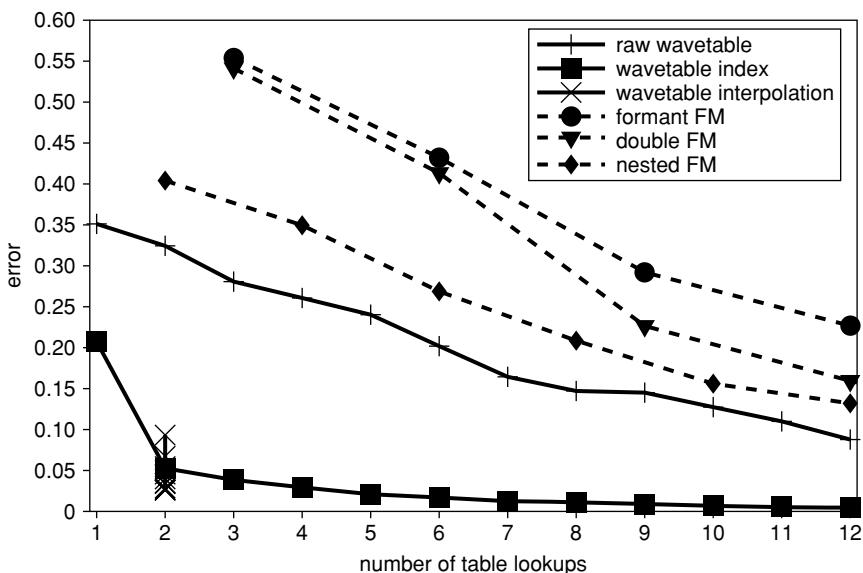


FIGURE 6.17. Convergence of error for different numbers of table lookups for the tenor voice tone.

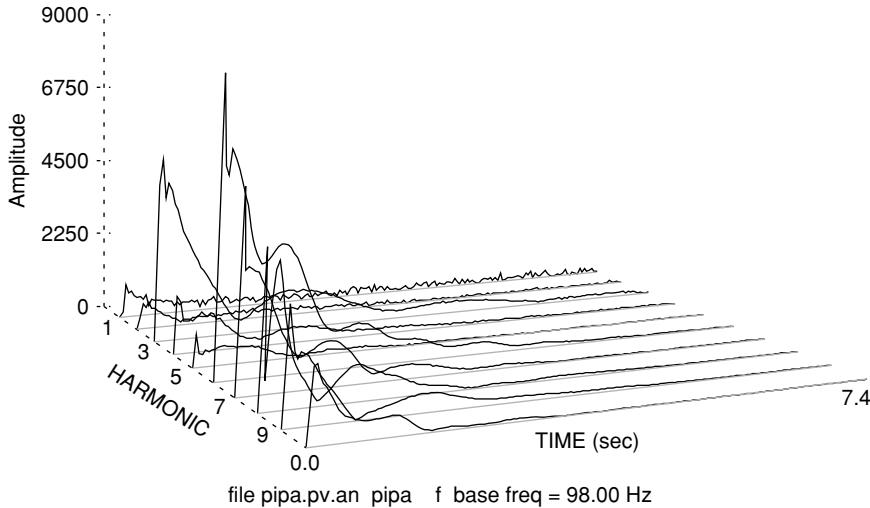


FIGURE 6.18. Time-varying amplitude spectrum of a G₂ Chinese pipa tone.

very comparably on this tenor sound. Even generalized wavetable synthesis performs better than the FM methods, indicating that the tenor tone is difficult to simulate with FM.

4.3 The Pipa

The pipa is a classical Chinese instrument that looks a little like a guitar but has a “twanging” sound more like an Indian sitar. Figure 6.18 shows the time-varying amplitude spectrum of a G₂ (98 Hz) pipa sound, which like a guitar decays exponentially albeit with some slowly varying oscillations (which may account for the twanging). These oscillations are out of phase with one another, making this a more difficult matching problem than the previous two examples.

Figure 6.19 shows the decrease of relative error with more GA-optimized wavetables or FM carriers for the pipa sound. The wavetable errors are much higher for the pipa than for the trumpet and tenor because of the difficulty of matching the spectral oscillations. Again, wavetable indexing performs best.

Figure 6.20 shows relative error plotted against number of table lookups. Results for the pipa are similar to those of the trumpet, with wavetable interpolation able to reduce the error down to about 10% by using 12 wavetables, while wavetable indexing requires about three times as much computation to achieve the same result.

5 Conclusions

The ability of various wavetable and FM synthesis methods to match the dynamic spectra of musical instrument tones such as those of a trumpet, a tenor

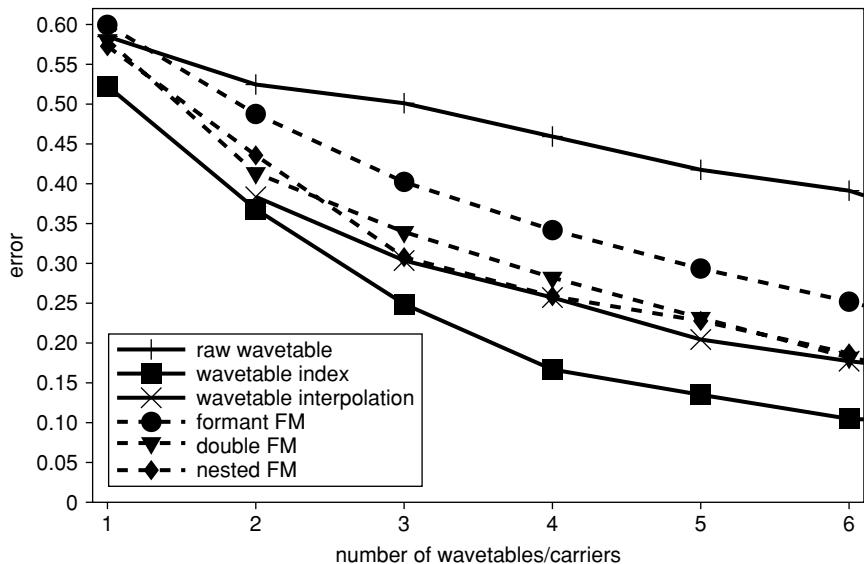


FIGURE 6.19. Convergence of error for different numbers of wavetables/carriers for the Chinese pipa tone.

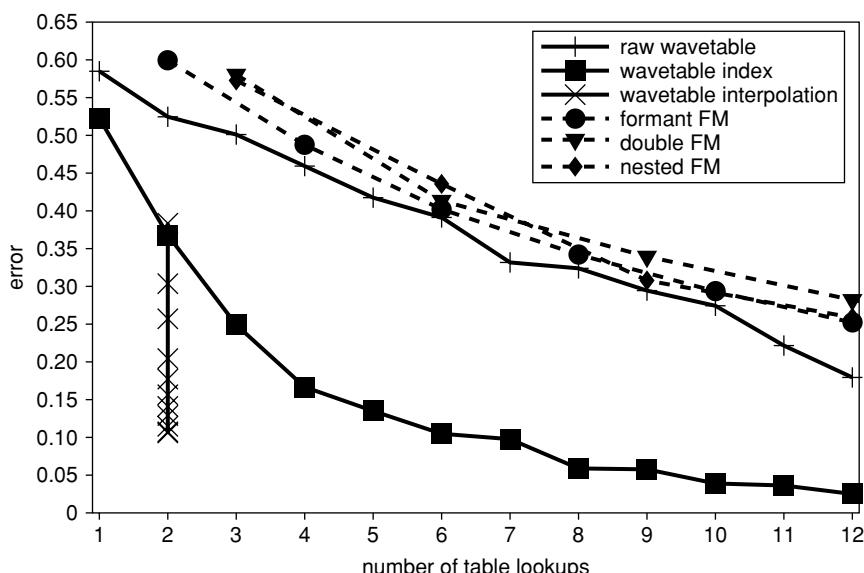


FIGURE 6.20. Convergence of error for different numbers of table lookups for the Chinese pipa tone.

voice, and a Chinese pipa have been compared. Wavetable indexing and interpolation consistently perform the best in terms of memory and computation.

In terms of the number of wavetables or carrier oscillators required for a given error level, wavetable indexing yields the best matches. This indicates that when computational resources are modest, wavetable indexing is probably the best all-around method.

However, for the same number of table lookups per sample computation, wavetable interpolation consistently yields the best matches. This indicates that wavetable interpolation is a good choice for situations where computation is overwhelmingly the main factor, such as with PC sound cards affording limited computation that must be shared between several voices.

The FM methods give much worse results for matching the three tones investigated. Because these tones are quite representative of the variety one would expect to encounter, it appears that FM methods are not intrinsically as well suited for simulating acoustic instruments as wavetable synthesis. However, efficiency and hardware issues can increase the desirability of such apparently inferior synthesis techniques. For example, a big advantage of FM is that it requires little wavetable memory (only one sine-wave table), making it especially useful in sound cards with limited on-board memory and in other real-time systems. The memory savings might well be worth the cost of using extra carriers to achieve more accuracy. Also, wavetable and FM synthesis each have certain types of sounds they can do especially well (and not so well). Adding more wavetables or carriers always improves the match.

High-quality matching to original sounds may not be important if one desires to mutate a sound into something exotically different. Both wavetable and FM matching provide interesting points of departure for instrument designers in applications such as timbre hybridization (Beauchamp and Horner, 1998).

Acknowledgments

This work was supported in part by the Hong Kong Research Grant Council's Projects HKUST729/96E, HKUST6073/97E, and HKUST6136/98E. Thanks to James Beauchamp, whose SNDAN program was used to generate the spectral plots in this chapter.

References

- Allen, J. B. (1977). "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Trans. on Acoustics, Speech, and Signal Processing ASSP-25*(3), 235–238.
- Beauchamp, J. (1993). "Unix workstation software for analysis, graphics, modification, and synthesis of musical sounds," *94th Convention of the AES*, Berlin, Audio Eng. Soc. Preprint 3479.

- Beauchamp, J. and Horner, A. (1998). "Spectral modelling and timbre hybridisation programs for computer music," *Organised Sound* **2**(3), 253–258.
- Berger, K. W. (1964). "Some factors in the recognition of timbre," *J. Acoust. Soc. Am.* **36**(10), 1888–1891.
- Chowning, J. M. (1973). "The synthesis of complex audio spectra by means of frequency modulation," *J. Audio Eng. Soc.* **21**(7), 526–534.
- Chowning, J. (1980). "Computer synthesis of the singing voice," *Sound Generation in Wind, Strings, Computers. Papers by Benade, Chowning, Hutchins, Jansson, Alonso, Moral given at Seminars of The Committee for the Acoustics of Music*. Royal Swedish Academy of Music No. 29, pp. 4–13.
- Clark, M., Luce, D., Abrams, R., Schlossberg, H., and Rome, J. (1963). "Preliminary experiments on the aural significance of parts of tones of orchestral instruments and on choral tones," *J. Audio Eng. Soc.* **11**(1), 45–54.
- Dolson, M. (1986). "The phase vocoder: A tutorial," *Computer Music J.* **10**(4), 14–27.
- Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, Reading, MI).
- Grey, J. and Moorer, J. (1977). "Perceptual evaluations of synthesized musical instrument tones," *J. Acoust. Soc. Am.* **62**(2), 454–462.
- Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems* (University of Michigan Press, Ann Arbor, MI).
- Horner, A., Beauchamp, J., and Haken, L. (1993a). "Machine tongues XVI: Genetic algorithms and their application to FM matching synthesis," *Computer Music J.* **17**(4), 17–29.
- Horner, A., Beauchamp, J., and Haken, L. (1993b). "Methods for multiple wavetable synthesis of musical instrument tones," *J. Audio Eng. Soc.* **41**(5), 336–356.
- Horner, A. (1996a). "Double modulator FM matching of instrument tones," *Computer Music J.* **20**(2), 57–71.
- Horner, A. (1996b). "Computation and memory tradeoffs with multiple wavetable interpolation," *J. Audio Eng. S.* **44**(6), 481–496.
- Horner, A. and Beauchamp, J. W. (1996). "Piecewise linear approximation of additive synthesis envelopes: A comparison of various methods," *Computer Music J.* **20**(2), 72–95.
- Horner, A. (1998). "Nested modulator and feedback FM matching of instrument tones," *IEEE Transactions on Speech and Audio Processing* **6**(4), 398–409.
- Justice, J. H. (1979). "Analytic signal processing in music computation," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **ASSP-27**(6), 670–684.
- Le Brun, M. (1977). "A derivation of the spectrum of FM with a complex modulating wave," *Computer Music J.* **1**(4), 51–52.
- Maher, R. C. and Beauchamp, J. W. (1990). "An investigation of vocal vibrato for synthesis," *Applied Acoustics* **30**(2&3), 219–245.
- McAulay, R. and Quatieri, T. (1986). "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. on Acoustics, Speech, and Signal Processing* **34**(4), 744–754.
- Payne, R. G. (1987). "A microcomputer based analysis/resynthesis scheme for processing sampled sounds using FM," in *Proc. 1987 Int. Computer Music Conf.*, Urbana, IL (Int. Computer Music Assoc., San Francisco), pp. 282–289.
- Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (1985). *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, Cambridge, UK).

- Schottstaedt, B. (1977). "The simulation of natural instrument tones using frequency modulation with a complex modulating wave," *Computer Music J.* **1**(4), 46–50.
- Serra, M.-H., Rubine, D., and Dannenberg, R. (1990). "Analysis and synthesis of tones by spectral interpolation," *J. Audio Eng. Soc.* **38**(3), 111–128.
- Sundberg, S. (1974). "Articulatory interpretation of the singing formant," *J. Acoust. Soc. Am.* **55**(4), 838–844.
- Tan, B. T. G. and Lim, S. M. (1996). "Automated parameter optimization for double frequency modulation synthesis using the genetic annealing algorithm," *J. Audio Eng. Soc.* **44**(1/2), 3–15.

The Effect of Dynamic Acoustical Features on Musical Timbre

JOHN M. HAJDA

1 Introduction

Timbre has been an important concept for scientific exploration of music at least since the time of Helmholtz ([1877] 1954). Since Helmholtz's time, a number of studies have defined and investigated acoustical features of musical instrument tones to determine their perceptual importance, or salience (e.g., Grey, 1975, 1977; Kendall, 1986; Kendall et al., 1999; Luce and Clark, 1965; McAdams et al., 1995, 1999; Saldanha and Corso, 1964; Wedin and Goude, 1972). Most of these studies have considered only nonpercussive, or *continuant*, tones of Western orchestral instruments (or emulations thereof). In the past few years, advances in computing power and programming have made possible and affordable the definition and control of new acoustical variables. This chapter gives an overview of past and current research, with a special emphasis on the time-variant aspects of musical timbre. According to common observation, "music is made of tones in time" (Spaeth, 1933). We will also consider the fact that music is made of "time in tones."

The famous music psychologist Carl Seashore recognized that, of the four major perceptual attributes of tone—pitch, loudness, duration, and timbre—timbre is "by far the most important aspect of tone and introduces the largest number of problems and variables" (Seashore, 1938/1967, p. 21). There are many facets to the complexity of timbre, one of these being the dual categorical and continuous nature of timbre as it is used in real-life musical situations. We categorize familiar musical instruments when we hear them: "that's a piano" or "that's a trumpet." Also, even if we do not know the exact name of an instrument, we can often categorize its sound into its correct instrument family, such as bowed string, woodwind, or brass (Clark et al., 1964). However, musical timbres can also be placed along continua, or dimensions, such that one timbre is said to have more or less of a particular perceptual attribute (or simply attributes) than another does. This concept is more elusive than simple categorization but can be easily demonstrated by auditory morphing [e.g., Slaney et al. (1995)]. Finally, we can assign a considerable range of sounds to the same instrument label; consider, for example, the *chalumeau* versus *clarino* registers of the clarinet, or *sul tasto* versus *sul ponticello* playing on

the violin. Sandell (1998) posits that an instrument's characteristic aural signature, or macrotimbre, is learned by exposure to that instrument playing a variety of spectra over different pitches. A single note may be insufficient to satisfactorily code a macrotimbre after a listener has been exposed to different performances, pitches, loudnesses, and durations.

Researchers have used two basic sets of methods for studying the categorical and continuous nature of timbre. The first set of methods falls under the global term classification, which, as its basic operation, is the partitioning of a collection of objects into groups (Estes, 1994). Therefore, categorization, recognition, and identification are all subsets of classification. The second set of methods utilizes what may be called relational measures. Here, an interval or ratio measure allows for comparisons between classes of objects. A measure of similarity, in which a subject hears a pair of sounds and rates them along a scale between "similar" and "not similar," is one such example. Another example is Verbal Attribute Magnitude Estimation (Kendall and Carterette, 1993a), in which a subject rates a sound along a scale that is anchored by a verbal attribute and its negation, such as "nasal" and "not nasal." Although the boundaries between classification and certain relational measures such as similarity become blurred in theories of cognition [e.g., Estes (1994)], from the point of methodological operations the distinction is still useful.

As mentioned above, most previous research has considered only single, isolated, continuant tones. Researchers have investigated the relative salience of both global time-envelope and spectral characteristics of these tones. In general, the global time-envelope constituents are the attack, the steady state, and the decay. The spectral characteristics are more varied, but generally include the relative energy of upper- and lower-frequency components, frequently measured by the spectral centroid; a feature of the spectral envelope shape called spectral irregularity; and various measures of how the individual frequency components change through time, including mean coefficient of variation and spectral flux. The following section will consider each of these parameters.

2 Global Time-Envelope and Spectral Parameters

What we know is largely determined by what we ask and how we ask it (Kendall and Carterette, 1992). In empirical studies of musical timbre, the types of tones researchers choose to investigate and the way in which their parameters are operationally defined can lead to ambiguous—or even conflicting—results. This is illustrated in the ongoing debate regarding the relative perceptual importance of the global envelope constituents of continuant tones.

2.1 *Salience of Partitioned Time Segments*

In most research, the global time envelope of an isolated continuant tone consists of its attack, steady state, and decay segments. With regard to the attack

and steady-state segments, past and current findings have supported one of the following three hypotheses:

1. The attack is more salient than the steady state.
2. The attack and steady state are equally salient.
3. The steady state is more salient than the attack.

In many studies, tone segments are artificially created by the imposition of a constant time interval from the beginning (for the attack) or from the end (for the decay) of the musical signal. These time intervals are determined *a priori* and sometimes arbitrarily; most researchers choose either a time from onset that is well into the steady-state portion of each stimulus or a time from onset that covers the longest global amplitude rise-time (e.g., time from onset to the first “significant” local maximum) among the stimuli.

Generally, stimuli are presented one at a time to subjects over loudspeakers or headphones, and subjects employ a classification procedure. For a number of studies that date from the 1960s and 1970s, subjects were asked to name—with or without the aid of a word list—the instrument that most likely produced the tone that they heard. In the literature, this procedure is commonly referred to as identification, although, unless the number of choices is equal to the number of stimuli, a more proper term in experimental psychology is name categorization.

In the early identification studies (Berger, 1964; Clark et al., 1963; Elliott, 1975; Saldanha and Corso, 1964; Wedin and Goude, 1972), the durations of attack- and decay-time segments varied from study to study and were usually on the order of a few hundred milliseconds or less. These segments were imposed on every instrument tone, regardless of the type of instrument. These researchers assumed, for the most part, that attack or decay transient segments occurred within these specified segments; the remainder of the signal was considered to be the steady state. Overall, they found that the removal of the attack segments hindered identification, whereas the removal of the decay segments did not affect identification.

Iverson and Krumhansl (1993) examined the role of onsets in similarity-type judgments. Subjects heard consecutive pairs of tones and rated along a scale of “a little” to “a lot” the degree to which they would have to “change the first sound to make it sound like the second sound” (Iverson and Krumhansl, 1993, p. 2597). Three different stimulus contexts were used: the complete tones, onsets only (the segment measured as 80 ms from the beginning of the signal); onsets removed (the complete tone minus the 80-ms onset segment). The authors found that mean subject ratings for all three contexts corresponded highly with one another. They concluded that “the attributes that are salient for timbral similarity judgments are present throughout tones” (Iverson and Krumhansl, 1993, p. 2602). They surmised that the reason their findings did not jibe with those of the earlier identification studies might have been the difference in subject task.

Campbell and Heller (1978, 1979) introduced the influence of melodic context into the onset role issue. Their stimuli were generated from performances of two-note legato phrases (F_4 at 349.2 Hz to A_4 at 440 Hz) played on six different instruments, including piano. The transitional segment between the two notes was

called the legato transient. This transient was operationally defined as a constant time segment before the start of the second steady state, applied uniformly to each instrument recording. The length of the time segment varied from 20 to 110 ms. They also created constant attack-alone and steady-state-alone contexts—generated from the first tone of the sequence. The authors found that the 110-ms legato transients yielded higher identification than either the attacks, steady states, or any of the other shorter legato transients.

Kendall (1986) pursued this issue in two unique ways: (1) He compared the role of transients and steady state across single-note and legato musical phrase contexts, and (2) he included signal characteristics for each stimulus as bases for his operational definitions of transients. Because he also tested for the effect of musical training (musicians vs nonmusicians), Kendall used a non-verbal matching procedure instead of identification. In musical (melodic) contexts, the steady-state-alone contexts—with the attack and legato transients removed—were matched at a mean level (81%) that was statistically equivalent to the unaltered signals (84%). However, in the single-note contexts, both the steady-state-alone (50%) and the attack-alone (51%) contexts were matched at the same level as the unaltered single tones (54%). In comparing his results to those of the earlier identification studies, Kendall (1986, p. 210) concluded that “the perceptual importance of transients in defining the characteristic sounds of instruments has been overstated.”

The contradictory results given by the myriad of studies that have explored the salience of time-envelope characteristics—with the exception of Kendall (1986)—are most likely directly due to the lack of robust operational definitions based on signal characteristics. The attack is not a duration; it is a transient part of the signal that lasts from onset until a more-or-less stable periodicity and modes of vibration are established. This “steady state” is generally achieved well before the end of the initial rise time, as determined by amplitude. Contemporary with many of the identification studies in the 1960s, Luce (1963) descriptively examined the characteristic attacks and steady states for 14 nonpercussive instruments of the Western orchestra. Notes were recorded across the entire range of each instrument. His associate, William Strong used two methods to calculate the attack durations (Luce, 1963, p. 90):

1. Amplitude transient: the time from onset to the time when the amplitude reached 90% of the amplitude of the steady state.
2. Structure transient: the time from onset to the time when the waveform had essentially the same shape or structural characteristics as the steady state.

For every instrument except the tuba, the structure transient was measured as shorter than the amplitude transient was. In the case of the flute, the structure transient could not be ascertained because “rather large intensity modulations were present” (p. 92). Strong’s measurements for 13 instruments (piccolo was excluded) are presented in abbreviated form in Table 7.1.

On average, Strong’s structure transients in Table 1 are 53% as long as the amplitude transients. Luce and Clark (1965) modified the amplitude transient definition to the time necessary for the amplitude to reach 50% (-6 dB IL) of the amplitude

TABLE 7.1. Mean Durations of Amplitude and Structure Attack Transients^a

Instrument	Mean-amplitude transient (ms)	Mean-structure transient (ms)
Violin	218	88
Viola	106	41
Cello	350	124
Double bass	96	84
Oboe	21	16
English horn	52	29
Bassoon	41	30
Clarinet	60	42
Flute	179	not measured
Trumpet	96	24
French horn	34	24
Trombone	51	36
Tuba	73	95

^aData adapted from William Strong (Luce, 1963, Table 8.1.1., p. 91). [From Hajda (1999); used by permission.]

at a point 133 ms further into the signal. So if the measured amplitude transient was 30 ms, the amplitude at 30 ms was equal to 50% of the amplitude at 163 ms. In general, this modification brought the new transients into closer concordance with Strong's structure transients. It is likely, therefore, that contemporary researchers who identify the attack as the time from onset to the global or first "significant" local maximum (e.g., McAdams, et al., 1995; Sandell, 1998) have included a sizable segment of the tone in which periodicity (i.e., pitch) and characteristic harmonic relationships (i.e., timbre) are discernable, even though they have based their operational definition on signal characteristics. It is important to note that the effect of using an amplitude transient over a structure transient depends on the subjective tasks and the manner in which the stimuli were constructed.

Ideally, every constituent segment of a musical tone has a structural element in its operational definition; in other words, the evolution of both global amplitude and spectral components should be considered. In addition, the operational definitions of these segments must be perceptually relevant. Hajda et al. (1997) proposed such a model for the signal partitioning of continuant tones. Part of the impetus for this model, called the "amplitude/centroid trajectory" (ACT), was the observation by Beauchamp (1982) that, for certain continuant signals, RMS amplitude and spectral centroid have a monotonic relationship throughout the steady-state portion of a tone.

The ACT model considers the relationship of amplitude and spectral centroid throughout the duration of a tone. Hajda et al. (1997) identified four consecutive contiguous partitions that are evident in the analyses of most continuant musical instrument signals:

1. Attack: that portion of the signal in which the global RMS amplitude is rising and the spectral centroid is falling after an initial maximum.

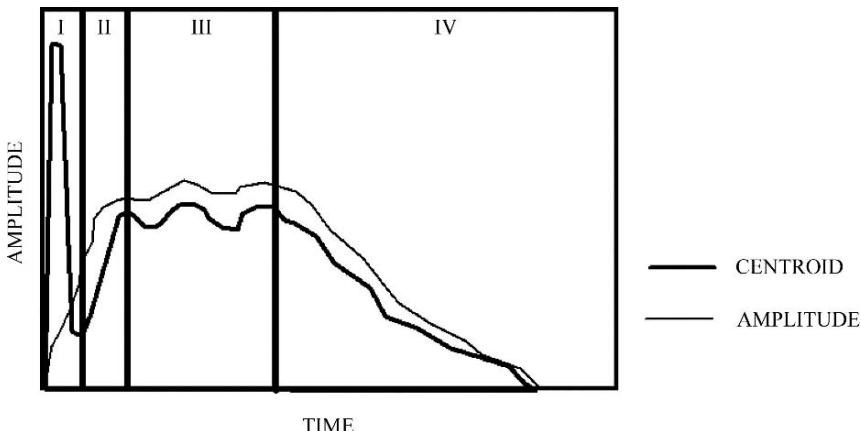


FIGURE 7.1. The RMS-amplitude and spectral-centroid trajectories for a contrived continuant tone. I: Attack; II: Attack/Steady-State Transition; III. Steady State; IV. Decay. [From Hajda (1998), used by permission.]

2. Attack/steady-state transition: the segment from the end of the attack to the first local RMS amplitude maximum.
3. Steady state: the segment during which the amplitude and the centroid both vary around mean values.
4. Decay: the final segment during which the amplitude and centroid both rapidly decrease.

Figure 7.1 illustrates the four ACT segments for a contrived instrument.

Hajda (1996, 1997, 1999) tested the efficacy of this model in a controlled experiment that used single isolated tone stimuli consisting of six “impulse tones” (performed on classical guitar, marimba, piano, pizzicato violin, tubular bell, and xylophone) and six continuant tones (performed on clarinet, flute, oboe, tenor saxophone, trumpet, and bowed violin). The tones were played at concert B_4^b (approximately 466 Hz) in an auditorium and digitally recorded. Two tones from each continuant instrument were used: sustained (about 3.5 s) and staccato (about 600 ms). One tone was recorded from each impulse instrument; because of their different acoustical dampings, the durations of these tones varied. There were 18 unedited tones in all; 12 continuant and 6 impulse. All of the continuant tones used in this study except one manifested characteristics that were consistent with the ACT model. The one exception was a sustained violin tone that was played without an articulated attack.

Continuant tones were partitioned based on three different definitions of attack: (1) fixed attack time from onset to 80 ms into the signal; (2) attack time based on 50% of the average steady-state RMS amplitude, adapted from the operational definition given by Luce and Clark (1965); and (3) the ACT model (Hajda et al., 1997). The partitions for the first two conditions can be described as attack alone

and remainders alone. The partitions for the ACT condition included all possible combinations of the four segments—attack, attack/steady-state transition, steady state, and decay—plus each segment alone. The continuant tones were also subjected to two reverse playback conditions: the entire tone and a 500 ms segment extracted from sustained tones beginning one second after onset.

Nine subjects identified each of the 246 randomly presented stimuli by selecting from a list of the 12 instruments used in the experiment (forced-choice). The probability for a “chance” identification of each stimulus was 8.3%.

The results for continuant tones can be summarized as follows:

1. The unedited signals were correctly identified 93% of the time. The overall results were the same for the unedited sustained and unedited staccato signals, although individual instruments yielded slightly different identifications for different tone durations.
2. For the sustained continuant tones, all three attack-removed conditions yielded a higher percentage of correct identifications than the attack-alone conditions. In addition, the attack-removed conditions yielded results that approached those for the unedited signals. Based on these data, we can conclude that, for these sustained tones, the remainders are more salient than the attacks.
3. For the staccato continuant tones, divergent results were found. For the fixed-80-ms-attack condition, the attacks-alone were identified at a much higher rate than the remainders. In previous studies, the researcher might assume that the removal of the attack adversely affected identification. However, an examination of the raw data showed that the remainders of many of the short signals were confused with impulse instruments (classical guitar, marimba, piano, and pizzicato violin). In fact, removal of the attack was tantamount to imposing an impulse envelope on the staccato tones. In this case, the poor identification results were due to a confounding variable, not experimental control.
4. Therefore, the discussion of the effect of ACT-editing is restricted to the sustained tones. For the sustained ACT conditions, the steady-state-alone edits were identified best. Only the steady-state-alone edits approached the identification rate of the unedited sustained signals (85%–93%). Given all of the above discussion, Hajda (1996, 1997, 1999) concluded that the time-variant steady-state alone is necessary and sufficient for the identification of these isolated sustained continuant tones.
5. For the sustained continuant tones, reverse playback never affected identification.

It seems clear that the process of human identification of an instrument from one of its tones is complex. Listeners can apply a number of strategies, based on the information available. Many of these strategies are determined by the listener’s previous knowledge of the instruments’ capabilities. Other strategies may stem from basic, seemingly pre-musical distinctions, such as distinguishing an impulse from a continuant envelope. Even in these contrived contexts, it is clear that a single rule will not apply between classes of instruments.

Given the above caveat, it seems that, for sustained continuant tones, the time-variant steady state usually provides sufficient and necessary information for the identification of an instrument. The co-evolution of the amplitude and spectral centroid seems important here, but the direction (i.e., regular vs reverse playback) does not.

The acoustical analyses conducted for this study indicate that when one considers the universe of timbres produced by musical instruments, the issue of attack vs steady state bears little relevance, because impulse instruments cannot be usefully partitioned in such a manner. However, the global RMS amplitude and spectral centroid trajectories and their functional relationship are characteristics of all musical (i.e., time-variant) sounds. Research by Hajda (1998, 1999) focused on the salience of these—and other—trajectories; the results of this preliminary work are reported in the following section.

A caveat should be issued regarding the nature of the attacks of nonpercussive instrument tones. A plethora of measurements made by Luce and his colleagues showed that “the duration of the attack transients depends upon the instrument played, upon the note played on the instrument, and upon the performer, but very little on the dynamic marking at which the instrument is sounded or the duration of the notes played, or whether or not the instrument is played with vibrato” (Luce and Clark, 1965, p. 199). We can add other variables that will probably affect the duration of attack transients, including characteristics of the musical phrase (legato, staccato, etc.), musical style, texture (counterpoint, homophony, heterophony), and other musical contexts.

Finally, although various classification paradigms are simple to operationalize and implement in laboratory experiments, we should question the relevance of classification to the “real world” of musical timbre. To what extent do performers or listeners recognize, categorize, or even identify timbres in the course of their musical experience, Benjamin Britten’s *Young Person’s Guide to the Orchestra* (1946) notwithstanding? Certainly, orchestration requires a high level of knowledge regarding the timbral characteristics of each instrument of the ensemble. However, more often than not in Western music, timbres are heard in combination. It is not enough to “know” the timbre of a B^b trumpet that is playing “open middle C”; the orchestrator must know how that trumpet tone will sound in the context of a brass quintet, or as part of a jazz ensemble, or part of a marching band. Musical timbre does not operate as a series of unrelated, isolated entities. Every ensemble operates within its own timbral framework, or palette (see Martens, 1985), in a manner analogous to a painter’s palette of color. Even a solo instrumentalist manipulates timbre in order to produce “coloristic” effects.¹ More often than

¹ This is particularly true for instruments with multiple degrees of freedom. Consider the classical guitar, on which a given note can be alternatively fingered (stopped) on several strings. Each fingering produces a slightly different timbre due to the physical characteristics (thickness, winding) of the different strings and resonant properties of the instrument. In addition, the right hand can produce a myriad of tonal qualities by plucking the string with different combinations of flesh and nail as well as varying locations relative to the bridge.

not, however, the physical correlates for a palette of timbre are more difficult to determine than those for visual color.

2.2 *Relational Timbre Studies*

Relational measures have been used since the middle part of the 20th century in a variety of experimental contexts. Although this review is by no means exhaustive, it is intended to give the reader an idea about the types of timbre studies that have been conducted as well as a convergence of the findings.

In order to find a palette (or representative geometric structure) for timbre, we must be able to determine its dimensions. Such a determination has been made for pitch. Shepard (1982) has summarized and demonstrated models of Western musical pitch structures that can be expressed in two dimensions (circle of fifths), three dimensions (simple helix), four dimensions (double helix wrapped around a torus), and even five dimensions (double helix wrapped around a helical cylinder)! For a number of reasons, the dimensions for timbre are not nearly so well delineated.

Researchers have used two basic approaches to uncovering the structure of timbre. The first is to directly measure specified attributes of timbre by means of a subject's assignment of a value along a scale of adjectival polar opposites, such as "dullness" and "brightness." This technique, commonly known as the semantic differential (Osgood et al., 1957), is considered a measurement of the meaning of a stimulus and has been used to study other facets of music besides timbre. Lichte (1941) and von Bismarck (1974) used versions of this approach. They constructed steady-state synthetic stimuli with varying spectral characteristics and constant temporal envelopes in order to isolate verbal factors that would identify salient perceptual features. Lichte (1941) found a primary relationship between "brightness" and the midpoint of the energy distribution among frequency partials; von Bismarck (1974) found a similar primary relationship for his stimuli and "sharpness." In their study with dyads produced by recording natural wind instrument performances, Kendall and Carterette (1993a) used English translations of von Bismarck's (1974) semantic differential. They found that these adjectives did not significantly differentiate their stimuli. They replicated the study but replaced the semantic differential adjectives, for example, "dull" and "sharp," with an adjective and its negation, such as "sharp" and "not sharp." This procedure, known as Verbal Attribute Magnitude Estimation (VAME), was used on the same stimuli in a subsequent experiment (Kendall and Carterette, 1993b). This time, the verbal attributes came from a descriptive text on orchestration (Piston, 1955). These final ratings produced the most interpretable results, among them a primary

Other instruments, such as the trumpet, maintain timbral control by the prolonged coupling between the energy source (player) and vibrating body. Some of these instruments can also take advantage of additional physical couplings, such as a mute, in order to significantly alter their aural characteristics. From this perspective, instruments such as the piano are impoverished in terms of their degrees of freedom with respect to timbre.

relationship between “nasality” and the relative amount of steady-state energy in the upper frequency partials as compared to the fundamental.

The second approach to determining timbral structures is based on obtaining perceptual qualitative relationships between stimuli, as opposed to directly measuring timbral attributes. Subjects’ rating scores are obtained from a direct method of similarity analysis (Ekman, 1965). After hearing a pair of consecutively presented tones, the subject rates how similar those tones sound in relation to the other pairs in the stimuli set. The ratings for every possible paired comparison are then mathematically transformed into distances in a geometrical (usually Euclidean) space. This statistical analysis is commonly referred to as multidimensional scaling, or MDS. There are a number of MDS algorithms, each of which differs slightly in its intricacies.² The basic purpose of these procedures is the same: Produce a geometric configuration in which stimuli that are similar appear close together and those that are dissimilar appear far apart. Then, it is up to the researcher to interpret this configuration in terms of the characteristics of the stimuli.

In general, the nature and number of the stimuli limit the number of interpretable dimensions. In a paired-comparisons paradigm, the number of judgments that a subject must make is

$$n = \frac{s(s \pm 1)}{2}, \quad (7.1)$$

where s is the number of stimuli. The quantity $(s + 1)$ is used for experiments that include identities—stimulus x is paired with itself—and $(s - 1)$ without identities.³ Therefore, a paired-comparison similarity experiment with 25 stimuli requires 325 judgments by a subject with identities, 300 without. If each stimulus is 3 s and a subject requires 5 s for each response, the entire experiment will take about 1 h, not including the time needed for instructions and any practice experiments. In this author’s experience, many subjects cannot remain focused for such a duration. In fact, most of the similarity studies conducted for musical timbre have used between 10 and 20 stimuli. The MDS spaces for these experiments have produced interpretable solutions for two or three dimensions. Such is the case with Fig. 7.2, a space generated by the similarity ratings for 11 continuant instruments of the Western orchestra (Kendall et al., 1999).

The interpretation of the dimensions of an MDS space requires a good deal of intuition on the part of the researcher. In general, researchers attempt to find musical and extramusical correlates with each dimension of the solution. The musical correlates might include proximity groupings by instrument family (Wessel, 1973; Grey, 1975), pitch (Miller and Carterette, 1975), or the degree of blend for two simultaneously produced timbres (Kendall and Carterette, 1993c, Sandell, 1995). The extramusical variables are typically verbal attributes (Faure et al., 1996;

² For an overview of MDS and related procedures, see Kruskal and Wish (1978) and Arabie et al. (1987).

³ Although the case of stimulus I paired with itself is obviously trivial, it may be advantageous to include such a pairing in order to identify subjects who produce outlying data.

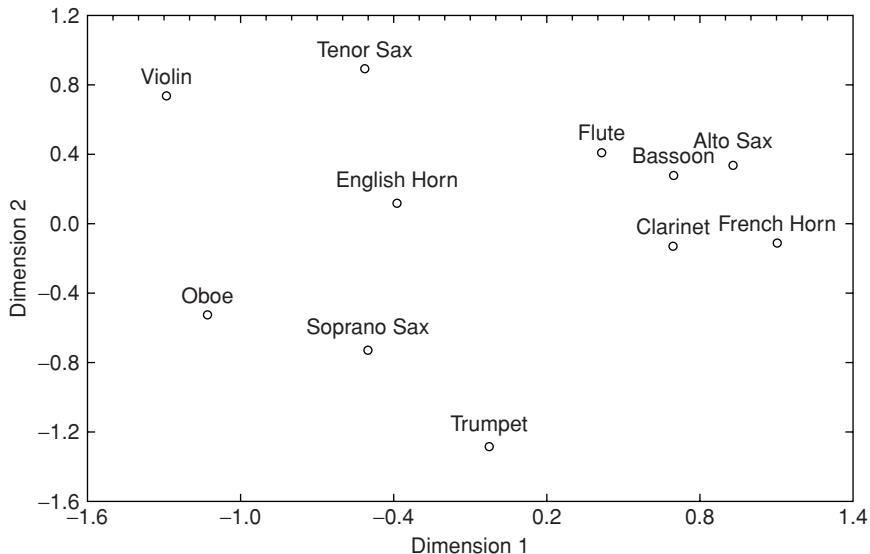


FIGURE 7.2. Two-dimensional MDS solution for similarity ratings of eleven natural instrument tones played at concert B_4^b (ca. 466 Hz). [Reprinted from Kendall et al. (1999). ©1999 by The Regents of the University of California. All rights reserved. Used with permission.]

Kendall et al., 1999) or acoustical parameters. The focus here will be on the correlation of acoustical parameters to the dimensions of MDS solutions.

In general, three acoustical parameters repeatedly appear as correlates to dimensional solutions in timbre studies:

1. Amplitude-vs-time (temporal) envelope, usually expressed in terms of attack or rise times.
2. Spectral energy distribution across frequency components.
3. Spectral variance in terms of the amplitudes of frequency components.

2.2.1 Temporal Envelope

In studies that include both continuant and impulse stimuli, the amplitude-vs-time envelope (aka temporal envelope or amplitude envelope)—in one manifestation or another—is the acoustical correlate to the primary perceptual dimension (Krumhansl, 1989; Iverson and Krumhansl, 1993; McAdams et al., 1995; Kendall et al., 1999). For the most part, researchers have characterized the envelope phenomenon as an issue of attack time; after all, impulse instruments have very brief attacks (less than 10 ms) in comparison to continuant instruments. Therefore, most measures of attack should yield high correlations with a dimension that separates percussive from nonpercussive stimuli. Krimphoff (1993) and McAdams et al. (1995) found precisely such a relationship when they correlated the primary dimension of an MDS space generated by the similarity scaling of impulse and

continuant timbres [from Krumhansl (1989)] with the log-rise-time (“logarithme du temps de montée”) of each stimulus. They defined log-rise-time as

$$\text{Log-rise-time} = \log_{10}(t_{\max} - t_{\text{thresh}}), \quad (7.2)$$

where t_{\max} is the time from onset to maximum RMS amplitude and t_{thresh} is the time from onset to a threshold taken as 2% of the amplitude at t_{\max} .

2.2.2 Spectral Energy Distribution

Many acousticians have described the steady-state portion of continuant tones in terms of a long-time-average spectrum. The amplitude and frequency components of the two-dimensional spectrum are analogous to a series of weights and distances along a beam. The point at which the sum of moments (weight \times distance) equals zero is the fulcrum, or, in the case of the spectrum, the spectral centroid. Such an index for measuring the “quality of a musical instrument” was first described by Knopoff (1963, p. 229).⁴ To this author’s knowledge, the first correlations of spectral centroid and a perceptual dimension were published by Ehresman and Wessel (1978) and Grey and Gordon (1978). Although the formulas vary in detail, these and later studies use a representative long-time average spectrum such that

$$f_{\text{centroid}} = \frac{\sum_{n=1}^N f_n \cdot A_n}{\sum_{n=1}^N A_n}, \quad (7.3)$$

where f_n is the frequency and A_n is the amplitude (usually linear) of the n th partial of a spectrum with N frequency components. This equation yields a measure in frequency units, which will suffice in instances where the fundamental frequencies of the stimuli are the same. It is also possible to produce a unitless measure by (1) replacing f_n with the harmonic number or (2) multiplying the denominator by the fundamental frequency.

The Pearson correlation of spectral centroid with Dimension 1 of the two-dimensional MDS space shown in Fig. 2 is 0.9 (Kendall et al., 1999). This result is consistent with other research that has yielded strong correlations between spectral centroid and the primary perceptual dimension of MDS spaces for continuant stimuli [e.g., Ehresman & Wessel (1978); Grey & Gordon (1978)] and secondary perceptual dimension of spaces for mixed impulse and continuant stimuli [e.g., McAdams et al. (1995); Lakatos (2000)].

⁴ Knopoff (1963) used the term *center of gravity* (from engineering statics) instead of *spectral centroid*. In fact, his measure involved taking the ratio of (1) a theoretical center of gravity calculated by replacing the amplitude of each frequency partial with the moment of that frequency in the original signal, and (2) the center of gravity from the original signal.

2.2.3 Spectral Time Variance

The individual amplitudes of frequency components for many continuant signals vary significantly throughout the duration of a tone. This dynamic feature has been given a number of labels, among them: Spectral Fluctuation (Grey, 1977); Spectral Variation (Ehresman and Wessel, 1978); Spectral Flux (Krumhansl, 1989); and Time Variance (Kendall and Carterette, 1993b). In spite of the number of phenomenological observations made since Grey (1975), spectral time variance was not quantified until the 1990s.

Kendall and Carterette (1993b) calculated a mean coefficient of variation (MCV):

$$MCV = \frac{\sum_{n=1}^{N=9} \frac{\sigma_n}{\mu_n}}{N}, \quad (7.4)$$

in which σ_n is the standard deviation of the amplitude of frequency component n across time, μ_n is the mean amplitude of component n , and N is the number of frequency components analyzed, in this case $N = 9$. The mean coefficient of variation yielded a moderately strong correlation ($r = 0.7$) with the second dimension of the perceptual space generated by Kendall et al. (1999) shown in Fig. 2.

Krimphoff (1993) examined three different measures of spectral flux (“flux spectral”) in order to find the strongest relationship with the third dimension of an MDS space generated by Krumhansl (1989). The first, Spectral Variation (“variation spectrale”), was determined by taking the correlation of respective harmonics of adjacent instantaneous spectra (each corresponding to a single window of analysis of duration $\Delta t = 16$ ms). The absolute values of these correlations were summed and averaged across the entire duration of the tone. The second parameter, Flux (“flux”), was measured as the mean deviation of the spectral centroid of each analysis window with respect to the long-time average measure of spectral centroid. The final parameter, Coherence (“cohérence”), is a measure of the difference in onset times for each harmonic. The term, however, is a bit misleading because a signal in which every harmonic has the same time-to-onset has a coherence value equal to zero; a signal in which harmonics do not have the same time-to-onset has a coherence value greater than zero.

Krimphoff (1993) also examined the relationship of two measures of Fine Spectral Structure (“structure fine du spectrale”) to the third dimension of the Krumhansl (1989) MDS space. The first measure was taken from Guyot (1992). It is essentially a ratio with the sum of the energy in the odd-numbered harmonics above the fundamental taken to be the numerator and the sum of the energy in the fundamental plus the energy in the even-numbered harmonics taken to be the denominator. The final parameter, which Krimphoff (1993) called the Spectral Deviation (“déviation”), is the sum of deviations of each harmonic log-amplitude from the mean of three consecutive harmonic log-amplitudes (centered on that harmonic), normalized by a global mean log-amplitude. This parameter, which yielded the highest correlation with Krumhansl’s perceptual dimension, has been renamed by Krimphoff et al.

(1994) and McAdams et al. (1995) as Spectral Irregularity and most recently as Spectral Envelope Smoothness by McAdams et al. (1999). Kendall and Carterette (1996) used the following linear version of Krimphoff et al.'s (1994) log-based formula to calculate the linear spectral irregularity (LSI) of static synthetic stimuli:

$$LSI = \frac{\sum_{n=2}^{N-1} \left| A_n - \frac{A_{n+1} + A_n + A_{n-1}}{3} \right|}{\sum_{n=1}^N A_n}, \quad (7.5)$$

where A_n is the linear amplitude of the n th harmonic and N is the number of harmonics. A spectral smoothing paradigm used by McAdams et al. (1999) also used linear amplitudes.

In summary, depending on the nature of the stimuli, both long-time average (spectral centroid, spectral irregularity) and time-variant (rise time, mean coefficient of variation) acoustical measures are principal correlates with perceptual spaces generated by relational measures. The experimental control of these acoustical variables has only begun in recent years. Kendall and Carterette (1996) determined difference thresholds for synthetic timbres that varied only in spectral centroid. Jeong and Fricke (1998) found an effect of listening position and reverberation on these difference thresholds. In a separate study, Kendall and Carterette (1996) synthesized timbres with the same centroid but different spectral shapes. These timbres were compared in a separate relational study; as might be expected, spectral irregularity [defined by Eq. (7.5)] correlated very highly with the principal MDS dimension.

3 The Experimental Control of Acoustical Variables

Two recent studies have examined—at least in part—the experimental control of time-variant acoustical variables for tones that were originally produced by acoustical instruments.

McAdams et al. (1999) applied six basic data simplifications and five combinations of these simplifications to seven instrument tones. Five of the instruments were continuant—clarinet, flute, oboe, trumpet, and violin—and two were impulse—harpsichord and marimba. The simplifications are briefly described as follows:

1. Amplitude-Envelope Smoothing: removal of micro time-variations of harmonic amplitudes over the steady-state and decay portions of the tone.
2. Amplitude-Envelope Coherence (spectral envelope fixing): removal of spectral flux while preserving the average spectrum and global RMS envelope over the entire duration of the tone.
3. Spectral-Envelope Smoothness: linear smoothing of the jaggedness or irregularity of a spectral envelope over the entire duration of the tone.

4. Frequency-Envelope Smoothness: removal of micro time-variations of the frequencies of harmonics over the entire duration of the tone.
5. Frequency-Envelope Coherence (harmonic frequency tracking): removal of inharmonicity over the entire duration of the tone.
6. Frequency-Envelope Flatness: removal of frequency variations and inharmonicity over the entire duration of the tone.

Of these six data reduction techniques, numbers 1, 2, 4, and 6 remove a certain amount of time-variance. In all, McAdams et al. (1999) tested the salience of 11 methods of signal simplification: the six methods mentioned above and five combinations of these methods. Listeners were asked to discriminate between (1) sounds that were resynthesized with simplified data and (2) reference sounds that were synthesized versions of the original signal. All analyses and syntheses were conducted with phase-vocoder analysis and oscillator-bank additive synthesis algorithms contained in the SNDAN music sound analysis/synthesis package (Beauchamp, 1993). Overall, the authors found that only amplitude envelope coherence, or the removal of spectral flux, yielded a “very good” proportional mean discrimination (0.91) among the variables that controlled for time-variance. The means of discrimination for other time-variant variables ranged between 0.66 and 0.71; the probability of discrimination due to chance was 0.50. The highest mean discrimination was for spectral envelope smoothing (0.96). In general, edits that combined methods of simplification yielded means of discrimination that were equal to or slightly higher than those for the most salient individual method.

Hajda’s pilot study (1998, 1999) investigated the effects of controlling certain time-variant acoustical parameters of continuant tones. The 10 instrument tones used for this research come from the McGill University Master Samples, or MUMS, set of digital recordings (Opolko and Wapnick, 1989): alto flute, cello, clarinet, C trumpet,⁵ English horn, French horn, flute, oboe, trombone, and violin. The sustained tones were played at concert B_4^b , or approximately 466 Hz. This pitch is within the normal playing range of all of these instruments although it is toward the high end of the range for some of the instruments.

Three time-variant parameters were controlled in this experiment: global RMS amplitude, spectral amplitude envelope, and frequency deviation for each spectral component. The MUMS signals were trimmed by imposing a 40 dB threshold below the maximum amplitude so that noise floor effects would be minimized when the experimental controls were implemented. Segments of 1.1 s duration were extracted for each of the nine edits beginning 500 ms into each signal. The rationale for this was the finding that relevant timbral information is present in the steady-state portions of sustained continuant tones (Hajda, 1996, 1997, 1999). Linear 50 ms fade-ins and fade-outs were imposed on each edit. The original digital signal was edited in the same fashion for experimental control purposes.

⁵ The more common B^b trumpet is not available from the MUMS recordings.

TABLE 7.2. Summary of edits used in Hajda (1998)^a

Simplification	Frequency deviation	Spectral flux	Global RMS amplitude
SYNTH	Varies	Varies	Varies
FRQ	Controlled	Varies	Varies
SPC	Varies	Controlled	Varies
AMP	Varies	Varies	Controlled
FR/SP	Controlled	Controlled	Varies
FR/AM	Controlled	Varies	Controlled
AM/SP	Varies	Controlled	Controlled
S. S.	Controlled	Controlled	Controlled

^aSYNTH = full phase-vocoder synthesis; FRQ = remove all frequency deviations; SPC = remove spectral flux; AMP = remove global amplitude variation; FR/SP = combined removal of frequency deviations and spectral flux; FR/AM = combined removal of frequency deviations and global amplitude variation; AM/SP = combined removal of global amplitude and spectral flux; S.S. = true steady state. [From Hajda (1999); used by permission.]

The following spectrotemporal simplifications were made using SNDAN (Beauchamp, 1993, 1998):

1. SYNTH: full (unmodified) phase-vocoder resynthesis.
2. FRQ: replace all frequency deviations by a fixed average frequency for each harmonic.
3. SPC: remove spectral flux by imposing an average spectrum for the duration of the signal during which relative amplitudes of the harmonics are fixed, but the overall RMS amplitude time-variation is preserved.
4. AMP: remove global amplitude variation by imposing a fixed average RMS amplitude on the overall signal while allowing the harmonic relationships to vary relatively as in the original sound.
5. FR/SP: combination of 2 and 3.
6. FR/AM: combination of 2 and 4.
7. AM/SP: combination of 3 and 4.
8. S.S.: combination of 2, 3, and 4 (a steady-state condition).

These simplifications are summarized in Table 7.2.

A relational procedure was employed in which seven subjects rated the dissimilarity of the original digital tone with each of the eight synthesized edits. A zero rating indicated no discriminable difference between the original tone and the synthesized edit. A 100 rating indicated maximum dissimilarity (among all 90 comparisons).

Figures 7.3 and 7.4 show the mean dissimilarity ratings for the alto flute and clarinet edits. For the alto flute (Fig. 7.3), zeroing frequency deviation (FRQ) has no real effect on subject ratings, fixing global RMS amplitude (AMP) has a moderate effect, and removing spectral flux (SPC) has the strongest effect. Multiple controls increase the dissimilarities between the original and edited tones. By comparison, none of the edits for the clarinet tone (Fig. 7.4) has a significant effect on dissimilarity ratings. Informal listening indicated that the alto flute was played with a deep vibrato while the clarinet tone was played without vibrato.

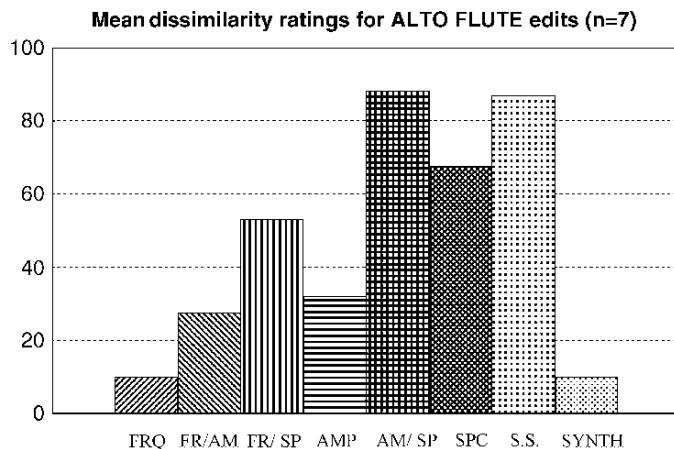


FIGURE 7.3. Mean dissimilarity ratings of seven subjects for the comparison of the original alto flute tone with nine synthetic edits. FRQ = removal of all frequency deviations; FR/AM = combined removal of frequency deviations and global amplitude variation; FR/SP = combined removal of frequency deviations and spectral flux; AMP = removal of global amplitude variation; AM/SP = combined removal of global amplitude and spectral flux; SPC = removal of spectral flux; S.S. = true steady state; SYNTH = full phase-vocoder resynthesis. [From Hajda (1999); used by permission.]

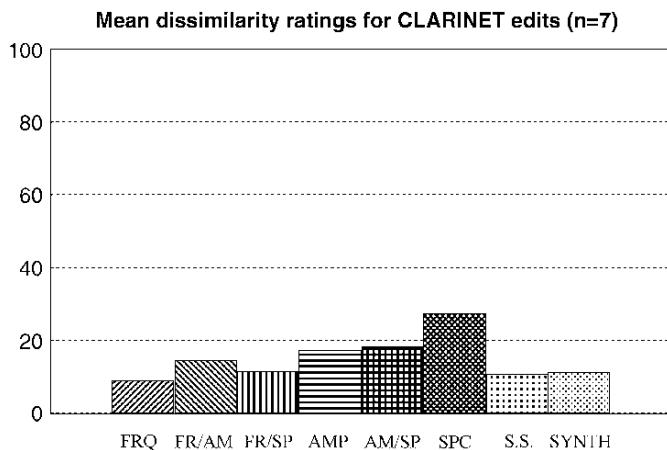


FIGURE 7.4. Mean dissimilarity ratings of seven subjects for the comparison of the original clarinet tone with nine synthetic edits. FRQ = removal of all frequency deviations; FR/AM = combined removal of frequency deviations and global amplitude variation; FR/SP = combined removal of frequency deviations and spectral flux; AMP = removal of global amplitude variation; AM/SP = combined removal of global amplitude and spectral flux; SPC = removal of spectral flux; S.S. = true steady state; SYNTH = full phase-vocoder resynthesis. [From Hajda (1999); used by permission.]

Data analysis indicates the following trends:

1. As one might expect, instruments played with vibrato were affected the most by the acoustical simplifications. However, several instruments played without vibrato—the English horn, oboe, and C trumpet—were affected a moderate amount by the controls. Other instruments played without vibrato—namely, the clarinet, French horn, and trombone—were not affected by the controls.
2. Averaged across all 10 instruments, the mean dissimilarity ratings for zeroing frequency deviations ($\mu = 20.0$) and global amplitude variations ($\mu = 22.5$) are not much different from those of the full resyntheses ($\mu = 16.0$). Removal of spectral flux has a much greater effect on the dissimilarity ratings ($\mu = 42.7$), and, as one might expect, the greatest effect occurs with the steady-state condition ($\mu = 47.5$).

These results are consistent with the findings of McAdams et al. (1999); this is especially interesting given the difference in method (dissimilarity rating versus discrimination).

4 Conclusions and Directions for Future Research

At this point, we can conclude that spectral flux (time variation of the normalized spectrum) is the most salient time-variant parameter of natural continuant tones (Hajda, 1998; Kendall et al., 1999; McAdams et al., 1999). McAdams et al. (1999) found that discrimination of a controlled acoustical variable was strongly correlated to the extent to which it actually varied in the original signal. By a common sense extension, if a parameter varies significantly in a signal, we can hypothesize that a signal resynthesized with the parameter made static will be perceived as significantly different from the original.

In spite of current advances, the salience of time-variant parameters in musical tones is far from fully understood. Part of this is due to the complexity of the musical instrument as a vibrational system, especially in instances in which the performer (driver) maintains a coupling with the generator and resonator. Such is the case with continuant instruments, where the performer controls the time-variant aspect of timbre in an expressive fashion that itself varies from one performance to another.

The next logical extension of this line of research involves musical context. Campbell and Heller (1979) and Kendall (1986) have already conducted work regarding the effect of legato melodic phrases on the classification of timbre. While it is clear that the connection of notes in a melody is important, the manner by which these notes connect has not been investigated in a systematic and controlled fashion. The roles of the time-variant aspects of timbre in a host of other musical contexts, such as expressiveness, dynamics, style, etc., have not been addressed. In addition, orchestral instruments rarely play in an isolated context. The effect of time variance in the presence of vertical combinations of timbres must also be considered.

To this point, the time-variant parameters of impulse signals have not been discussed. This is due to the lack of systematic research on this class of tones. Hajda (1995, 1996, 1997, 1999) found that impulse tones differ from continuant tones in several important ways:

1. Operational definitions of tone segments for continuant signals do not apply to impulse tones, since impulse signals contain no steady state (Hajda, 1996, 1997, 1999).
2. The identification of impulse signals is significantly affected by reverse playback; the identification of continuant signals is not (Hajda, 1996, 1997, 1999).
3. The identification of impulse tones is not affected by any type of partitioning, whether the segment that is presented to listeners is taken from the beginning or middle of a signal; the identification of continuant tones is affected by such signal editing (Hajda, 1997, 1999).
4. The long-time average spectral centroid is the strongest correlate to the primary perceptual dimension of an MDS space generated from the ratings of continuant tones; the *change* in centroid over time is one of several correlates for the primary perceptual dimension of an MDS space generated from the ratings of impulse tones (Hajda, 1995).

The above findings do not jibe entirely with other research (Freed, 1990; Serafini, 1995; Lakatos, 2000). Even if they did, the paucity of research would not warrant generalizations to the entire class of impulse instruments.

As stated by McAdams et al. (1999), two overall goals of research on the time-variant parameters of musical instrument tones are:

1. To facilitate realistic sounding resyntheses with a minimum of control variables.
2. To increase our understanding of the perception of timbre.

As such, musicians of diverse genres—from electronic music composers to orchestrators to music theorists—may benefit from these studies. However, because of the interdisciplinary nature of the research questions, musicians must team with physicists, engineers, and psychologists in order to unravel the mysteries of the “time in tones.”

References

- Arabie, P., Carroll, J. D., and DeSarbo, W. S. (1987). *Three-Way Scaling and Clustering*. Sage university papers. Quantitative applications in the social sciences; no. 07–065. (Sage Publications, Beverly Hills and London).
- Beauchamp, J. W. (1982). “Synthesis by spectral amplitude and ‘brightness’ matching of analyzed musical instrument tones,” *J. Audio Eng. Soc.* **30**(6), 396–406.
- Beauchamp, J. W. (1993). “Unix workstation software for analysis, graphics, modification, and synthesis of musical sounds,” *94th Convention of the Audio Engineering Society*, Berlin, *Audio Eng. Soc. Preprint 3479*.
- Beauchamp, J. W. (1998). “Methods for measurement and manipulation of timbral physical correlates,” *Proc. 16th International Congress on Acoustics and 135th Meeting of the*

- Acoustical Society of America*, 1998, Seattle, Vol. 3, P. K. Kuhl and L. A. Crum, eds. (Acoustical Society of America, Woodbury, NY), pp. 1883–1884.
- Berger, K. W. (1964). “Some factors in the recognition of timbre,” *J. Acoust. Soc. Am.* **36**(10), 1888–1891.
- Britten, B. (1946). *Variations and Fugue on a Theme of Henry Purcell (The Young Person’s Guide to the Orchestra)* (Boosey & Hawkes, London and New York).
- Campbell, W. C. and Heller, J. J. (1978). “The contribution of the legato transient to instrument identification,” *Proc. Research Symposium on the Psychology and Acoustics of Music, 1978*, University of Kansas, Lawrence, KS, pp. 30–44.
- Campbell, W. and Heller, J. (1979). “Convergence procedures for investigating music listening tasks,” *Bull. Council for Res. Music Educ.* **59**, 18–23.
- Clark, M., Jr., Luce, D., Abrams, R., Schlossberg, H., and Rome, J. (1963). “Preliminary experiments on the aural significance of parts of tones of orchestral instruments and on choral tones,” *J. Audio Eng. Soc.* **11**(1), 45–54.
- Clark, M., Jr., Robertson, P. T., and Luce, D. (1964). “A preliminary experiment on the perceptual basis for musical instrument families,” *J. Audio Eng. Soc.* **12**(3), 199–203.
- Ehresman, D., and Wessel, D. (1978). *Perception of Timbral Analogies*, IRCAM Technical Report 13/78 (IRCAM, Centre Georges Pompidou, Paris).
- Ekman, G. (1965). “Two methods for the analysis of perceptual dimensionality,” *Perceptual and Motor Skills* **20**, 557–572.
- Elliott, C. A. (1975). “Attacks and releases as factors in instrument identification,” *J. Res. Music Educ.* **23**(1), 35–40.
- Estes, W. K. (1994). *Classification and Cognition* (Oxford University Press, New York).
- Faure, A., McAdams, S., and Nosulenka, V. (1996). “Verbal correlates of perceptual dimensions of timbre,” *Proc. 1996 Int. Conf. on Music Perception and Cognition*, Montreal (Faculty of Music, McGill University, Montreal), pp. 79–84.
- Freed, D. J. (1990). “Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events,” *J. Acoust. Soc. Am.* **87**(1), 311–322.
- Grey, J. M. (1975). *An Exploration of Musical Timbre* (Report STAN-M-2, CCRMA, Dept. of Music, Stanford University, Stanford, CA).
- Grey, J. M. (1977). “Multidimensional perceptual scaling of musical timbres,” *J. Acoust. Soc. Am.* **61**(5), 1270–177.
- Grey, J. M., and Gordon, J. W. (1978). “Perceptual effects of spectral modifications on musical timbres,” *J. Acoust. Soc. Am.* **63**(5), 1493–1500.
- Guyot, F. (1992). “Etude de la pertinence de deux critères acoustiques pour caractériser la sonorité des sons à spectre réduits,” Unpublished D.E.A. thesis, Université du Maine, Le Mans, France.
- Hajda, J. M. (1995). “The relationship between perceptual and acoustical analyses of natural and synthetic impulse signals,” masters thesis, University of California, Los Angeles, 1995, *Masters Abstracts International*, **33**(6). (University Microfilms International Publications No. 13–61, 681)
- Hajda, J. (1996). “A new model for segmenting the envelope of musical signals: The relative salience of steady state versus attack, revisited,” *101st Convention of the Audio Engineering Society*, Los Angeles, Audio Eng. Soc. Preprint 4391.
- Hajda, J. M. (1997). “Relevant acoustical cues in the identification of Western orchestral instrument tones” (abstract), *J. Acoust. Soc. Am.* **102**(5), pt. 2, 3085.
- Hajda, J. M., Kendall, R. A., Carterette, E. C., and Harshberger, M. L. (1997). “Methodological issues in timbre research,” in *Perception and Cognition of Music*, I. Deliège and J. Sloboda, eds. (Psychology Press, Hove, UK), pp. 253–306.

- Hajda, J. M. (1998). "The effect of amplitude and centroid trajectories on the timbre of percussive and nonpercussive orchestral instruments," *Proc. 16th International Congress on Acoustics and 135th Meeting of the Acoustical Society of America*, Vol. 3, Seattle (Acoustical Society of America, Woodbury, NY), pp. 1887–1888.
- Hajda, J. M. (1999). "The Effect of Time-Variant Acoustical Properties on Orchestral Instrument Timbres," doctoral dissertation, University of California, Los Angeles. UMI number 9947018.
- Helmholtz, H. L. F. ([1877], 1954). *On the Sensations of Tone as a Psychological Basis for the Theory of Music* (Dover, New York).
- Iverson, P., and Krumhansl, C. L. (1993). "Isolating the dynamic attributes of musical timbre," *J. Acoust. Soc. Am.* **94**(5), 2595–2603.
- Jeong, D., and Fricke, F. R. (1998). "The dependence of timbre perception on the acoustics of the listening environment," *Proc. 16th Int. Congress on Acoustics and 135th Meeting of the Acoustical Society of America*, Vol. 3, Seattle (Acoustical Society of America, Woodbury, NY), pp. 2225–2226.
- Kendall, R. A. (1986). "The role of acoustic signal partitions in listener categorization of musical phrases," *Music Perception* **4**, 185–214.
- Kendall, R. A. and Carterette, E. C. (1992). "Convergent methods in psychomusical research based on integrated, interactive computer control," *Behavior Research Methods, Instruments, and Computers* **24**(2), 116–131.
- Kendall, R. A. and Carterette, E. C. (1993a). "Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives," *Music Perception* **10**(4), 445–468.
- Kendall, R. A. and Carterette, E. C. (1993b). "Verbal attributes of simultaneous wind instrument timbres: II. Adjectives induced from Piston's 'Orchestration,'" *Music Perception* **10**, 469–502.
- Kendall, R. A. and Carterette, E. C. (1993c). "Identification and blend of timbres as a basis for orchestration," *Contemp. Music Rev.* **9**(1/2), 51–67.
- Kendall, R. A. and Carterette, E. C. (1996). "Difference thresholds for timbre related to spectral centroid," *Proc. 4th Int. Conference on Music Perception and Cognition*, Montreal, Canada, (Faculty of Music, McGill University, Montreal), pp. 91–95.
- Kendall, R. A., Carterette, E. C., and Hajda, J. M. (1999). "Perceptual and acoustical features of natural and synthetic orchestral instrument tones," *Music Perception* **16**(3), 327–363.
- Knopoff, L. (1963). "An index for the relative quality among musical instruments," *Ethnomusicology* **7**(3), 229–233.
- Krimphoff, J. (1993). "Analyse acoustique et perception du timbre," unpublished D.E.A. thesis, Université du Maine, Le Mans, France.
- Krimphoff, J., McAdams, S., and Winsberg, S. (1994). "Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique," [Characterization of the timbre of complex sounds. 2. Acoustic analysis and psychophysical quantification.] *J. de Physique* **4**(C5), 625–628.
- Krumhansl, C. L. (1989). "Why is musical timbre so hard to understand?," in *Structure and Perception of Electroacoustic Sound and Music: Proceedings of the Marcus Wallenberg Symposium held in Lund, Sweden, on 21–28 August 1988*, S. Nielzen and O. Olsson, eds. (Excerpta Medica, Amsterdam), pp. 43–53.
- Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*, Sage university papers, Quantitative applications in the social sciences, no. 07–011 (Sage Publications, Beverly Hills and London).
- Lakatos, L. (2000). "A common perceptual space for harmonic and percussive timbres," *Perception & Psychophysics* **62**(7), 1426–1439.

- Lichte, W. H. (1941). "Attributes of complex tones," *J. Exp. Psych.* **28**, 455–480.
- Luce, D. A. (1963). *Physical Correlates of Nonpercussive Musical Instrument Tones*, unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.
- Luce, D. and Clark, M. (1965). "Durations of attack transients of nonpercussive orchestral instruments," *J. Audio Eng. Soc.* **13**(3), 194–199.
- Martens, W. L. (1985). "Palette: An environment for developing an individualized set of psychophysically scaled timbres," *Proc. 1985 International Computer Music Conference*, Simon Fraser University, Burnaby, British Columbia, (Computer Music Association, San Francisco), pp. 355–365.
- McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psych. Res.* **58**(3), 177–192.
- McAdams, S., Beauchamp, J. W., and Meneguzzi, S. (1999). "Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters," *J. Acoust. Soc. Am.* **105**(2), 882–897.
- Miller, J. R. and Carterette, E. C. (1975). "Perceptual space for musical structures," *J. Acoust. Soc. Am.* **58**, 711–720.
- Opolko, F. and Wapnick, J. (1989). *McGill University Master Samples User's Manual* (Faculty of Music, McGill University, Montreal).
- Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The Measurement of Meaning* (University of Illinois Press, Urbana, IL).
- Piston, W. (1955). *Orchestration* (W. W. Norton, New York).
- Saldanha, E. L. and Corso, J. F. (1964). "Timbre cues and the identification of musical instruments," *J. Acoust. Soc. Am.* **36**, 2021–2026.
- Sandell, G. J. (1995). "Roles for spectral centroid and other factors in determining 'blended' instrument pairings in orchestration," *Music Perception* **13**, 209–246.
- Sandell, G. J. (1998). "Macrotimbre: Contribution of attack and steady state," *Proc. 16th Int. Congress on Acoustics and 135th Meeting of the Acoustical Society of America*, Vol. 3, Seattle (Acoustical Society of America, Woodbury, NY), pp. 1881–1882.
- Seashore, C. E. ([1938], 1967). *The Psychology of Music* (Dover, New York).
- Serafini, S. (1995). "Timbre judgments of Javanese gamelan instruments by trained and untrained adults," *Psychomusicology* **14**, 137–153.
- Shepard, R. N. (1982). "Structural representations of musical pitch," in *The Psychology of Music*, D. Deutsch, ed. (Academic Press, New York), pp. 334–390.
- Slaney, M., Covell, M., and Lassiter, B. (1995). "Automatic Audio Morphing," *Proc. 1996 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP-96)*, Vol. 2 (IEEE, New York), pp. 1001–1004.
- Spaeth, S. G. (1933). *The Art of Enjoying Music* (McGraw-Hill, New York).
- von Bismarck, G. (1974). "Timbre of steady tones: A factorial investigation of its verbal attributes," *Acustica* **30**, 146–159.
- von Helmholtz, H. L. F. (1877). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. F. Vieweg and Sohn, Braunschweig. English translation by A. J. Ellis, "On the Sensations of Tone as a Physiological Basis for the Theory of Music (2nd ed., 1885)," reprinted by Dover Publications, New York, 1954.
- Wedin, L. and Goude, G. (1972). "Dimension analysis of the perception of instrumental timbre," *Scandinavian J. Psych.* **13**(3), 228–240.
- Wessel, D. L. (1973). "Psychoacoustics and music: A report from Michigan State University," *PAGE: Bulletin of the Computers Arts Soc.* **30**, 1–2.

Mental Representation of the Timbre of Complex Sounds

SOPHIE DONNADIEU

“Un des paradoxes les plus frappants à propos du timbre est que, lorsqu’on en savait moins sur lui, il ne posait pas beaucoup de problèmes . . .”

[One of the most striking paradoxes concerning timbre is that when we knew less about it, it didn’t pose much of a problem . . .]
Philippe Manoury (1991)

1 Timbre: A Problematic Definition

Timbre, in contrast to pitch and loudness, remains a poorly understood auditory attribute. Persons attempting to understand it may be confused as much by its nature as its definition. Indeed, timbre is a “strange and multiple” attribute of sound (Cadoz, 1991, p. 17), defined by what it is not: it is neither pitch, nor loudness, nor duration. Consider the definition proposed by the American National Standards Institute (1973, p. 56): “Timbre is that attribute of auditory sensation in terms of which a subject can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar.” Therefore, timbre is that perceptual attribute by which we can distinguish the instruments of the orchestra even if they play the same note with the same dynamics.

The absence of a satisfactory definition of timbre is primarily due to two major problems. The first one concerns the multidimensional nature of timbre. Indeed, it is timbre’s “strangeness” and, even more, its “multiplicity” that make it impossible to measure timbre along a single continuum, in contrast to pitch (low to high), duration (short to long), or loudness (soft to loud). The vocabulary used to describe the timbres of musical instrument sounds indicates the multidimensional aspect of timbre. For example, “attack quality,” “brightness,” and “clarity” are terms frequently used to describe musical sounds. The second problem concerns timbre as a concept that refers to different levels of analysis. Schaeffer (1966, p. 232) observed that one can talk about “the timbre of a sound without attributing it to a given instrument, but rather in considering it as a proper characteristic of this sound, perceived per se.” He noted that “we shouldn’t confuse two notions of timbre: one related to

the instrument, an indication of the source that is given to us by ordinary listening, and the other related to each of the objects provided by the instrument, appreciation of the musical effects in the objects themselves, effects desired by musical listening as well as by musical activity. We have even gone further, attaching this word timbre to an element of the object: timbre of the attack, distinguished from its stiffness." So, the concept of timbre is much more general than the ability to distinguish instruments. The problem is that only one term refers to many different notions: Timbre can be described in terms of (1) a set of sounds of an instrument and also of the specific timbre of each sound of a particular instrument, (2) an isolated sound, (3) a combination of different instruments, (4) the composition of a complex sound structure, or (5) in the case of timbres produced by analysis/resynthesis, hybrid timbres or chimeras, sounds never heard before, which can be associated with no known natural source. For the purposes of this chapter, we refer to timbre in terms of sound sources or multidimensional perceptual attributes.

Timbre conveys the identity of a sound source. In other words, the timbre of a complex sound comprises the relevant information for identifying sound sources or events, even in a musical context. As Schaeffer (1966) said: "It is denying the evidence to believe that pure music can exempt the ear from its principal function: to inform humans about the events that are occurring" (cited by Cadoz, 1991, p. 17). In the same way, we do not have any difficulty knowing that someone is playing a violin in the neighboring room or that a car has suddenly arrived behind us. This capacity to identify sound objects is necessary to our survival. Indeed, when we hear a motor noise while crossing a street, our reaction is to immediately step back onto the sidewalk to avoid an accident. Most certainly, in everyday life, we use all the sensory systems at the same time. However, the events mentioned above can be identified even if they occur outside our visual field and outside any context likely to facilitate our interpretation of the sound objects (McAdams, 1993).

Most studies of musical timbre have used single, isolated instrument tones, which are easy to manipulate for experimentation. Our discussion of these studies is organized by the theoretical models adopted by the researchers. The first model is information processing (Lindsay and Norman, 1977), which describes the perceptual dimensions of timbre in terms of abstract attributes of sounds. In other words, the acoustical parameters (spectral, temporal, and spectrotemporal) of the signal are processed by the sensory system, and the perceptual result is the timbre of complex sounds. Multidimensional scaling has been fruitful in determining these different perceptual dimensions of timbre. The second approach, based on *ecological theory* proposed by Gibson (1966, 1979), has only recently resulted in systematic experimentation in auditory perception. According to this viewpoint, timbre perception is a direct function of the physical properties of the sound object. The aim of these studies is to describe the physical parameters that are perceptually relevant to the vibrating object.

2 The Notion of Timbre Space

2.1 Continuous Perceptual Dimensions

Multidimensional scaling (MDS) has been a effective tool for studying the timbral relationships among stimuli possessing multiple attributes. The principal advantage of this exploratory technique is that *a priori* hypotheses concerning the number of dimensions and their psychophysical nature are not required. Generally, MDS is used in an auditory study in the following manner: A set of sound stimuli—in this case, the sounds of musical instruments—are presented in all possible pairs. The listener's task is to judge the dissimilarity between the timbres for each pair of sounds. The dissimilarity is measured generally on a numerical scale (for example, 1 to 9, with 1 being very similar and 9 being very dissimilar) or on a bounded, continuous scale (for example, indicated with a cursor varied continuously on a scale between “very similar” and “very dissimilar,” which is subsequently coded numerically). The pitch, subjective duration, and loudness of all the sounds are usually equalized so that the subject's ratings concern only timbral differences. At the end of the experiment, a dissimilarity matrix is tabulated. The aim of MDS is to produce a geometric configuration that best represents, in terms of metric distances, the perceptual dissimilarities between the timbres of the sounds. So, two timbres judged on average to be very similar should appear close together in the space, and two timbres judged to be very dissimilar should appear far apart in the space. The number of dimensions required for the spatial solution is determined by using a goodness-of-fit measure or statistical criterion.

The last step in the MDS analysis is the psychophysical interpretation. The goal is to find a relationship between some acoustical parameters and the perceptual dimensions of the MDS solution. Typically, we measure a number of physical parameters, such as spectral envelope, temporal envelope, and so on, for all of the stimuli. Then we compute correlations between the positions of the timbres relative to the perceptual axes and the physical parameters.

2.1.1 Spectral Attributes of Timbre

Scientists have devoted themselves to the psychophysical analysis of musical sounds for several decades. These studies showed that spectral characteristics have an important influence on timbre. The influence of such spectral factors is revealed by multidimensional analyses. Plomp (1970, 1976) used multidimensional techniques to study synthesized steady-state spectra derived from recordings of musical instrument tones. He found a two-dimensional solution for a set of synthetic organ-pipe stimuli and a three-dimensional solution for a set of wind and bowed-string stimuli. He did not give a psychoacoustical interpretation of the individual MDS axes, but he showed that the spectral distances (calculated as differences in energy levels across a bank of 1/3-octave filters) were similar to those for the dissimilarity ratings for each stimulus set. This result suggests that global activity level present in the human auditory system's array of frequency-specific nerve fibers

may constitute a sufficient sensory representation from which a small number of perceptual factors related to the spectral envelope may be extracted. De Brujin (1978) found a correlation between the spectral envelope of synthesized tones and dissimilarity judgments.

Preis (1984) asked listeners to judge the degree of dissimilarity between synthetic and original musical instrument tones. In this case, a correlation was observed between the metric distances separating the tones and a measure of the degree of dissimilarity between the tones' spectral envelopes. In the same way, Wedin and Goude (1972) observed that spectral-envelope properties explained the three-dimensional perceptual structure of similarity relations among musical instrument tones (winds and bowed strings). In one of their experiments on synthesized tones, Miller and Carterette (1975) varied the number of harmonics, a spectral property. This spectral property corresponded with two of three perceptual dimensions. The remaining acoustical variables employed corresponded with the third perceptual dimension. These were the amplitude-vs-time envelope (temporal) and the pattern of onset asynchrony of the harmonics (spectrotemporal). The results of this study suggested a perceptual predominance of spectral characteristics in timbre judgments. In the same way, Samson et al. (1996) observed a two-dimensional space in which the organization of timbres reflected spectral and temporal differences. Nine hybrid synthetic sounds were created, derived from crossing three levels of spectral change corresponding to a change in the number of harmonics. (The tones were comprised of one, four, or eight harmonics.) The authors observed that the positions of tones along one of the dimensions corresponded closely to the number of harmonics. These results suggest that the manipulation of certain parameters influences subjects' perception of complex sounds.

Grey (1975, 1977) and Wessel (1979) observed similar multidimensional spaces with relatively complex synthesized tones meant to imitate conventional musical instruments (winds, bowed strings, plucked strings, or mallet percussion). Figure 8.1 shows the timbre space constructed by Grey (1975). The first axis is interpretable in terms of the spectral energy distribution. At one extreme, instruments like the French horn or the cello had low spectral bandwidths and concentrations of low-frequency energy. At the other extreme, the oboe has a very wide spectral bandwidth and less concentration of energy in the lowest harmonics.

Grey and Gordon (1978) were the first to propose a quantitative interpretation of spectral energy distribution. They found that the centroid of a loudness function based on time-averaged amplitudes of stimulus harmonics correlated strongly with the first dimension of MDS models for tones interpolated acoustically between Grey's (1975, 1977) original acoustic instrument tones and their spectral modifications of some of these tones. Iverson and Krumhansl (1993), using complete synthetic tones, those with attack portion only, and those with attacks removed, gave a similar interpretation of the second dimension of their three spaces.

Krimphoff (1993) and Krimphoff et al. (1994) conducted acoustical analyses on the set of 21 sounds created by Wessel et al. (1987) and used by Krumhansl (1989) in an MDS timbre study. Most of these synthetic sounds imitated traditional

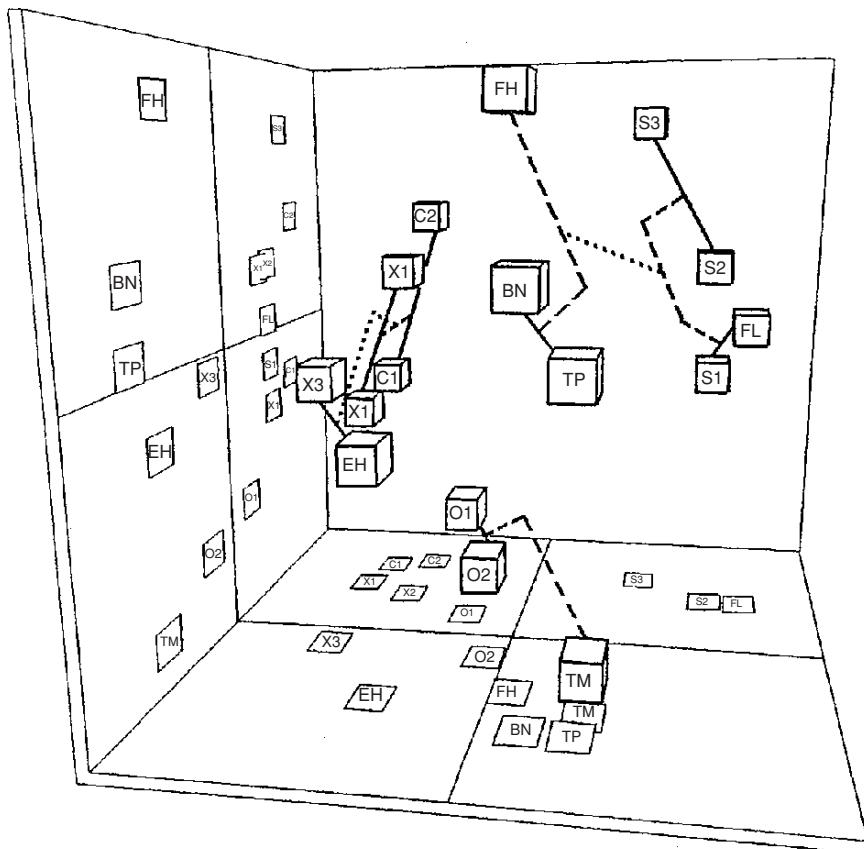


FIGURE 8.1. Three-dimensional INDSCAL solution derived from similarity ratings for 16 musical instrument tones. Two-dimensional projections of the configuration appear on the wall and the floor. Abbreviations for the instruments: O1 and O2, two different oboes; C1 and C2, E^b and bass clarinets; X1 and X2, alto saxophone playing softly and moderately loud, and X3, soprano saxophone, respectively; EH, English horn; FH, French horn; S1, S2, and S3, cello playing with three different bowing styles: *sul tasto*, *normale*, *sul ponticello*, respectively; TP, trumpet; TM, muted trombone; FL, flute; BN, bassoon. Dimension I (top-bottom) represents spectral envelope or brightness (brighter sounds at the bottom). Dimension II (left-right) represents spectral flux (greater flux to the right). Dimension III (front-back) represents degree of presence of attack transients (more transients at the front). Hierarchical clustering is represented by connecting lines, decreasing in strength in the order: solid, dashed, and dotted. [From Grey (1977), Fig. 1, used by permission of Acoustical Society of America.]

instruments, but some were chimerical hybrids (e.g., a “trumpar” created by combining spectrotemporal characteristics of the trumpet and the guitar). Krumhansl (1989) did not attempt to give a quantitative interpretation of her MDS solution, but she intuitively interpreted each of its axes according to the positions

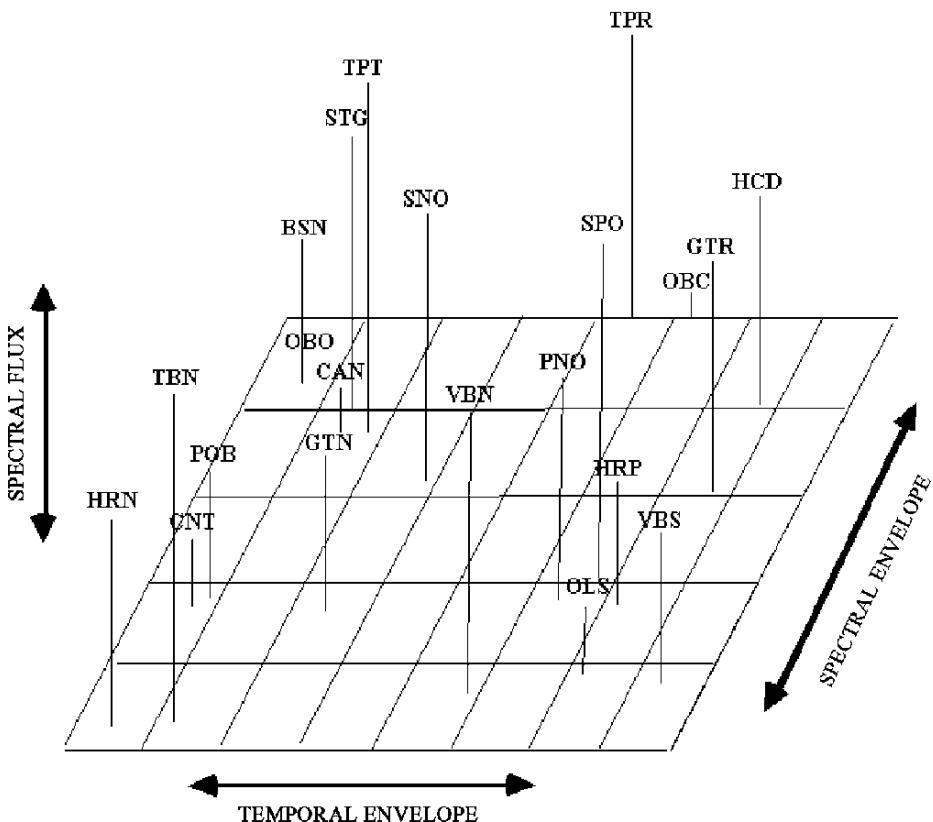


FIGURE 8.2. Three-dimensional EXSCAL solution derived from dissimilarity ratings for 21 synthesized musical instrument tones. Abbreviations for the instruments: BSN, bassoon; CAN, English horn; CNT, clarinet; GTN, guitarnet (hybrid between GTR and CNT); GTR, guitar; HCD, harpsichord; HRN, French horn; HRP, harp; OBC, obochord (hybrid between OBO and HCD); OBO, oboe; OLS, oboleste (hybrid between OBO and celeste); PNO, piano; POB, bowed piano; SNO, striano (hybrid between STG and PNO); SPO, sampled piano; STG, string; TBN, trombone; TPR, trumpar (hybrid between TPT and GTR); TPT, trumpet; VBN, vibrone (hybrid between VBS and TBN); VBS, vibraphone. Dimension I (left-right) represents the Temporal Envelope or attack quality of the sounds (blown-bowed sounds at the right and plucked-struck sounds on the left). Dimension II (front-back) represents the Spectral Envelope of the sounds (brighter sounds at the back). Dimension III (top-bottom) represents Spectral Flux (more spectral flux on the top). [From Krumhansl (1989), Fig. 1, used by permission of Excerpta Medica]

of the different timbres (see Fig. 8.2). Krimphoff aimed to find the acoustic parameters that correlated most strongly to the three dimensions that Krumhansl qualitatively referred to as Temporal Envelope, Spectral Envelope, and Spectral Flux. Thus, two of the three dimensions were expected to correlate with spectral characteristics. Krimphoff found that Dimension 2 (Spectral Envelope) correlated

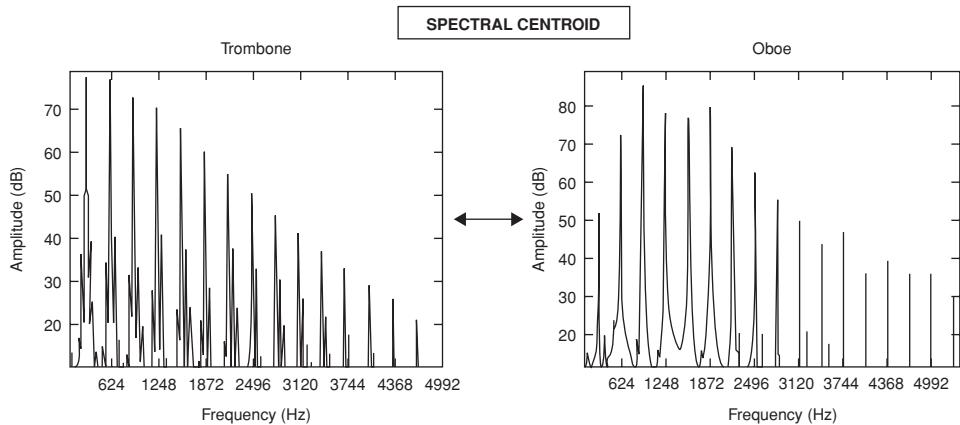


FIGURE 8.3. Spectra of two extreme sounds positioned along the second perceptual dimension of timbre spaces in Figs. 8.1 and 8.2 illustrating the “spectral centroid” parameter. On the left a trombone spectrum has a lower spectral centroid value, and on the right an oboe spectrum has a higher spectral centroid value.

very strongly ($r = 0.94$) with the *spectral centroid* (measured as the time-average of the instantaneous spectral centroid over the duration of the tone. A comparison of spectra with low and high spectral centroids is shown in Fig. 8.3.) However, as discussed further in Section 2.1.3, none of Krimphoff’s several measures of spectral variation over time corroborated Krumhansl’s suggestion that the third dimension could be interpreted in terms of “spectral flux,” a variation of the spectrum over time. Krimphoff’s best measure of spectral flux explained only 34% of the variance ($r = 0.59$).

In an attempt to quantify the acoustic nature of Krumhansl’s third dimension, Krimphoff proposed two new acoustic parameters related to the spectral envelope. First, he tested an acoustic parameter proposed by Guyot (1992) that measures the ratio between the amplitudes of even and odd harmonics. The clarinet, for example, has a high value for this parameter, because its odd-numbered spectral components have higher energy than its even-numbered ones. On the other hand, the trumpet’s value for this parameter is low, because its spectrum is more homogeneous with regard to the amplitudes of the various harmonics. Krimphoff found that the odd/even parameter explained 51% ($r = -0.71$) of the MDS variance for the third dimension. However, a second parameter corresponding to a measure of the *spectral irregularity* of the spectrum (taken as the log of the standard deviation of component amplitudes from a global spectral envelope derived from a running mean of the amplitudes of three adjacent harmonics) yielded a stronger correlation, explaining 73% ($r = -0.85$) of the variance along Krumhansl’s third dimension. (A comparison of spectra with low and high spectral irregularity is shown in Fig. 8.4.) Krimphoff’s spectral envelope result suggested a new interpretation of the third dimension.

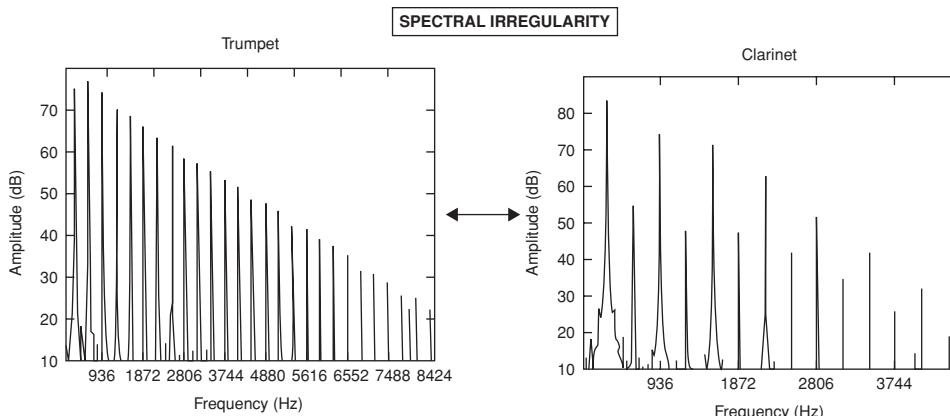


FIGURE 8.4. Spectra of two extreme sounds positioned along the second perceptual dimension of the timbre space in Fig. 8.2 illustrating the “spectral irregularity” parameter. On the left a trumpet spectrum has a lower spectral irregularity value, and on the right a clarinet spectrum has a higher spectral irregularity value (i.e., a more jagged spectral envelope).

One of the aims of a study by McAdams et al. (1995) was to replicate the Krumhansl (1989) study with a large set of listeners having varying degrees of musical training and to check whether any of the acoustic correlates described by Krimphoff (1993) and Krimphoff et al. (1994) could explain the resulting dimensions of the timbre space. Figure 8.5 shows the three-dimensional timbre space produced by McAdams et al. (1995). They correlated several acoustical parameters with derived MDS dimensions for 18 sounds (drawn from the 21 sounds used by Krumhansl and Krimphoff). They found that spectral centroid accounted for 88% of the variance ($r = -0.94$) along Dimension 2 of the figure. However, spectral irregularity did not correlate well with Dimension 3 (only $r = 0.13$), whereas spectral flux gave the highest Dimension 3 correlation ($r = 0.54$).

Grey and Moorer (1977) and Charbonneau (1981) used a different approach, where controlled modifications of acoustical analyses of instrument tones were used as the basis for resynthesis. Grey and Moorer used a computer resynthesis technique based on a heterodyne-filter analysis method to first produce a set of intermediate data for additive synthesis consisting of time-varying amplitude and frequency functions for the set of partials of each tone. Then, from those data they produced synthetic musical instrument stimuli that were used to evaluate the perceptual discriminability of original and resynthesized tones taken from a wide class of orchestral instruments. Sixteen versions of each tone were presented to listeners: (1) original tones; (2) tones resynthesized with line-segment approximations of the amplitude and frequency variations; (3) line-segment approximations with deletion of initial transients; and (4) line-segment approximations with flattening of the frequency variations. Instrument tones from the string, woodwind,

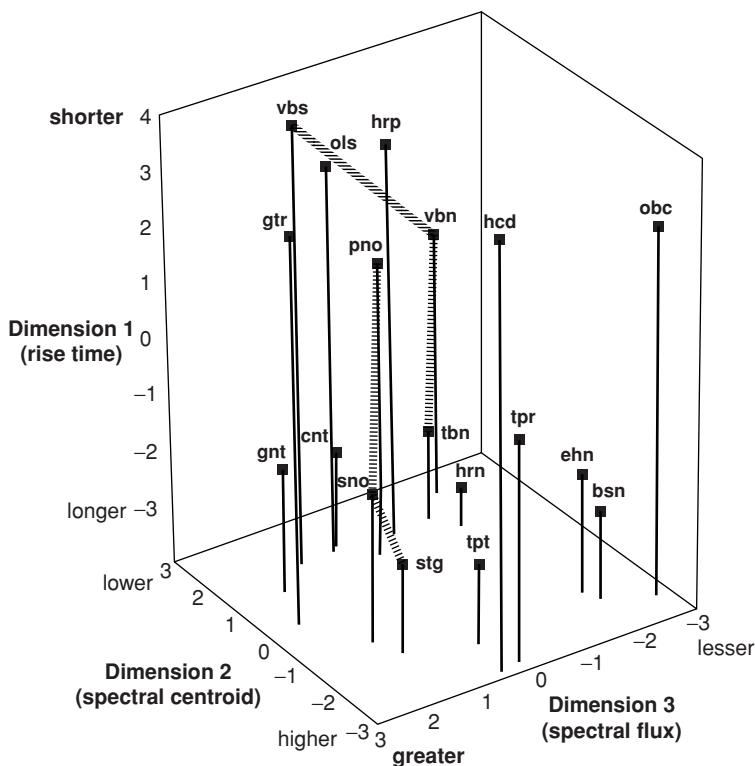


FIGURE 8.5. Three-dimensional CLASCAL solution with specificities and five latent classes derived from dissimilarity ratings on 18 timbres by 88 subjects. The acoustic parameters correlated to the dimensions are indicated in parentheses. Hashed lines connect two of the hybrid timbres (vbn and sno) to their progenitors. [From McAdams et al. (1995), Fig. 1, used by permission of Springer-Verlag.]

and brass families were modified. The pitch, subjective duration, and loudness of these tones were equalized.

Three identical tones and one different tone were presented in an AA-AB vs AB-AA discrimination procedure. Musically trained listeners were asked to discriminate which tone pair was “different” and to rate how different it was on a numerical scale. The data showed that: (1) simplifying the pattern of variation of the amplitudes and frequencies of individual components in a complex sound had an effect on discrimination for some instruments but not for others; (2) tones in which the attack transients were removed were easily discriminated from the originals; and (3) tones in which frequency variations were suppressed were easily discriminated as well. These results suggest that microvariations in frequency and intensity functions are not always essential to timbre and that a reduction of the data can be applied without affecting the perception of some sounds.

Charbonneau (1981) extended Grey and Moorer's study by constructing instrumental sounds that maintained their original global structure, while simplifying the microstructure of the amplitude and frequency envelopes of each harmonic partial. Listeners were asked to evaluate the timbral differences between original sounds and three types of simplifications: (1) replacing the harmonics' amplitude-vs-time envelopes so that each had the same amplitude shape (calculated as the average harmonic-amplitude envelope) but scaled to preserve its original peak value and start- and end-times; (2) replacing the frequency-vs-time envelopes so that each had the same relative frequency variation as the fundamental, meaning that the sound remained perfectly harmonic throughout its duration; and (3) fitting the start- and end-time data to fourth-order polynomials. Results indicated that the amplitude-envelope simplification had the greatest effect. However, as with the Grey and Moorer study, the strength of the effect depended on the instrument. These studies showed that simplifications performed on temporal parameters, and specifically on time-varying functions of amplitude and frequency, influence to a greater or lesser degree the discrimination of musical sounds.

McAdams et al. (1999) attempted to determine the extent to which simplified spectral parameters, without the use of straight-line approximations, affected the perception of synthesized instrumental sounds produced by instruments of various families of resonators (air column, string, or bar) and types of excitation (bowed, blown, or struck). Listeners were asked to discriminate sounds resynthesized with full data from sounds resynthesized with six basic data simplifications: (1) harmonic-amplitude variation smoothing; (2) coherent variation of harmonic-amplitudes over time; (3) spectral-envelope smoothing; (4) coherent harmonic-frequency variation; (5) harmonic-frequency variation smoothing; and (6) harmonic-frequency flattening. (Methods 2 and 4 were similar to Charbonneau's methods 1 and 2.) The results showed very good discrimination for spectral-envelope smoothing and coherent harmonic-amplitude variation, demonstrating, in a negative way, the importance of spectral-envelope detail and spectral flux. However, for coherent harmonic-frequency variation, harmonic-frequency variation smoothing, harmonic-frequency flattening, and harmonic-amplitude variation smoothing, discrimination was moderate to poor in decreasing order.

These techniques appear to be important for the study of timbre perception because they allow modification of the different spectrotemporal parameters of sound in order to reveal which are most important for timbre perception.

2.1.2 Temporal Attributes of Timbre

The classical point of view associates timbre with the spectrum of a sound signal. However, this point of view remains limited because it ignores the importance of temporal factors in timbre. Indeed, instrumental tones physically and perceptually evolve over time. Moreover, the classical conception runs into serious obstacles because musical instruments can be recognized or identified even when their spectra are seriously distorted. This happens in the case of mediocre recordings and when instruments are performed in normal reverberant rooms, where the spectra

of sounds vary a great deal throughout the space. Indeed, when we move about in a room, timbres are not transformed as much as we would expect if they depended exclusively on the precise structure of the source spectra. Nevertheless, spectral factors are undeniably important in timbre, while temporal factors seem to play a role only in certain contexts or for certain instruments.

Let us first examine the extent to which temporal factors are important in the timbre of musical sounds. We often consider musical sounds as composed of three parts: an initial attack portion, a middle sustain portion, and a final decay (the sustain portion being absent, of course, in resonant percussion sounds). The temporal shape of the sound of a piano is an important factor in the definition of its timbre. This is proven by listening to a sound presented in reverse-time. While its long-term average spectrum is identical to that of the original sound, the time-reversed version is often totally unrecognizable (George, 1954; Schaeffer, 1966). In the same way, Berger (1964) showed that suppressing the initial portion of sounds perturbs their recognition. Listeners were asked to discriminate between original musical instrument tones and modified versions with either their initial portions (attacks) or their final ones (decays) removed. Identification was poorest for sounds without attack. Also, Saldanha and Corso (1964) evaluated the relative importance of onset and offset transients, spectral envelope of the sustain portion, and vibrato for identifying musical instruments playing isolated notes. Identification was particularly affected when the attack portions were removed. However, identification was affected less if the instruments were performed with vibrato than if they were performed without vibrato. These results suggest that the attack plays a major role in the identification of instruments, but in the absence of the attack, additional information still exists in the sustain portion (McAdams, 1993). The studies of Grey and Moorer (1977) and Charbonneau (1981) described above also demonstrated the importance of such temporal factors. For example, tones with the attack transients removed were easily discriminated from originals in their studies.

As part of a multidimensional analysis, Samson et al. (1996) produced a two-dimensional space in which each dimension corresponded to temporal factors. The authors observed that the duration of the attack (1, 100, or 190 ms) correlated strongly with one of the perceptual dimensions. Grey (1977) and Wessel (1979) also observed a dimension of this nature. Wessel (1979) determined that the second dimension of his perceptual space corresponded to “attack rapidity.” Grey (1977) interpreted two of his three dimensions to be related to attack features, the second of which corresponded to the “presence of inharmonic transients in the high frequencies just before the onset of the main harmonic portion of the tone.” Strings, flutes, and clarinet, for example, have low-amplitude, high-frequency energy near their tone onsets, contrary to those of the bassoon or the English horn. Krimphoff (1993) and Krimphoff et al. (1994) confirmed this finding in their interpretation of the “Temporal Envelope” dimension of Krumhansl’s (1989) space (see Fig. 8.2). The positions of timbres along this axis were strongly correlated ($r = 0.94$) with the logarithm of the rise time of the temporal envelope (where rise time was measured as the difference between the time at which the amplitude reaches a threshold of 2% of the maximum amplitude to the time it attains maximum

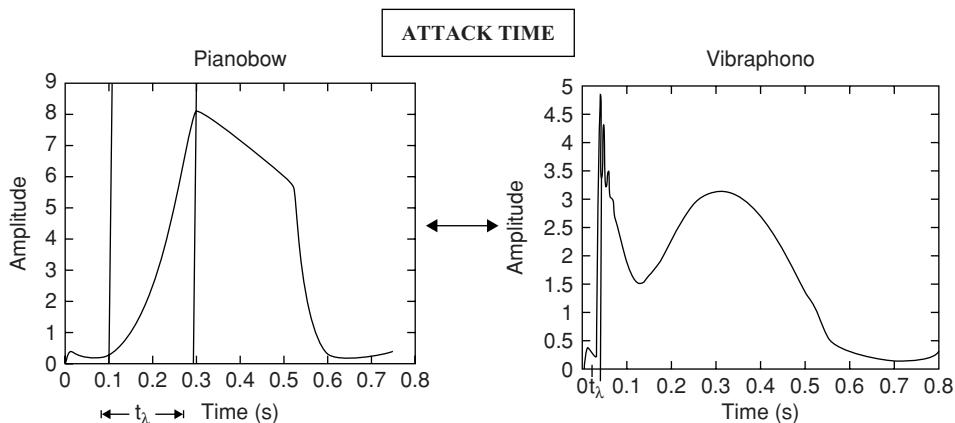


FIGURE 8.6. Temporal envelope of two extreme sounds positioned along the first perceptual dimension of the timbre spaces shown in Figs. 8.2 and 8.5 illustrating the “log-attack time” parameter. On the left, the pianobow has a long attack time (about 190 ms) similar to those for wind and bowed string instruments. On the right, the vibraphone has a short attack time (about 4 ms) similar to those for the set of struck and plucked instruments.

amplitude). The first dimension of McAdams et al.’s (1995) timbre space (Fig. 8.5) is also strongly correlated to this acoustical parameter with 88% of the variance ($r = -0.94$) explained by it. Examples of the measurement of rise time are shown in Fig. 8.6.

2.1.3 Spectrotemporal Attributes of Timbre

Multidimensional scaling in timbre studies has often revealed three perceptual dimensions. While two of these are often easily characterized by acoustical parameters, the third one remains poorly defined. This lack of satisfactory interpretation is probably due to the variability in stimulus sets or listener characteristics across studies. Not all studies have found a valid third dimension (e.g., Wessel, 1979), and those that have interpreted this perceptual axis differently from one study to the next. Some authors have proposed that this dimension corresponds to a spectral factor other than spectral centroid (Krimphoff et al., 1994; McAdams et al., 1995), and others have proposed that it corresponds to a temporal variation in the spectral envelope (Grey, 1977; Krumhansl, 1989) (see Figs. 8.1, 8.2, and 8.5).

Up to this point, this chapter has presented the influence of temporal and spectral factors, considered independently, on timbre. Nevertheless, these factors are not generally independent, and their association may also play a role in musical timbre. Risset and Mathews (1969) notably observed that synthesized trumpet sounds with static spectra and a common amplitude-vs-time envelope, applied synchronously to all frequency components, did not give a satisfactory perceptual result. They demonstrated the necessity of taking into account the variations of the different spectral components over time for certain timbres. Grey (1977) also suggested

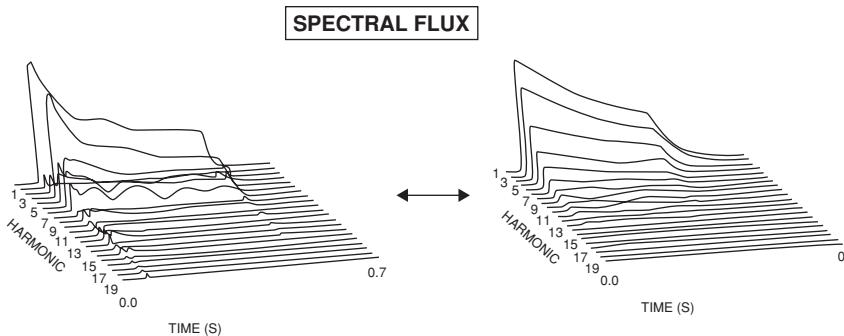


FIGURE 8.7. Time-frequency perspective plots illustrating the “spectral flux” parameter for two extreme sounds positioned along the third perceptual dimension of the timbre space of Fig. 8.2. On the left, the obochord (hybrid between the oboe and the harpsichord) has a high spectral flux value and on the right, the striano (hybrid between a string and a piano) has a low value.

that the physical nature of one of the perceptual dimensions of timbre could be a spectrotemporal factor. The interpretation of the second dimension of his solution was a combination of the degree of fluctuation in the spectral envelope over the duration of a tone and the synchrony of onset of its different harmonics. The woodwinds were at one extreme and tended to have upper harmonics that reached their maximum during the attack but were often in close alignment during the decay. Also, their spectra tended to have little fluctuation over time contrary to the strings or brass situated at the other extreme of this axis.

As mentioned in Section 2.1.1, Krumhansl (1989) named the first two dimensions obtained in her MDS study Temporal Envelope and Spectral Envelope, and the third dimension was called Spectral Flux, because the distribution of timbres along this dimension was presumed to correspond to the degree of spectral variation over time. (Time-variant spectra with high and low spectral variation are compared in Fig. 8.7). This interpretation agreed for the most part with the one proposed by Grey (1977) for simplified, resynthesized instrument sounds. The psychophysical interpretation proposed by Krimphoff (1993) and Krimphoff et al. (1994) for the first two dimensions agreed with the qualitative interpretation of Krumhansl, as previously discussed.

For analysis of Krumhansl’s third dimension, Krimphoff (1993) tested three acoustical parameters that quantified spectral fluctuation over the duration of a sound. These parameters were: (1) “spectral flux,” defined, in this case, as the rms variation of instantaneous spectral centroid around the mean spectral centroid; (2) “spectral variation,” defined as the average of correlations between amplitude spectra in adjacent time windows (note that the smaller the degree of variation of the spectrum over time, the higher the correlation); and (3) “coherence,” defined as the standard error of the onset times across all harmonics. Correlations observed between these three parameters and the third dimension of Krumhansl’s (1989)

solution were not significant, except for spectral flux, which only explained 34% ($r = 0.59$) of the variance along this dimension. Krimphoff (1993) and Krimphoff et al. (1994) found that spectral irregularity, a spectral rather than spectrotemporal parameter, best explained Krumhansl's third dimension. On the contrary, spectral irregularity was not best correlated to the third dimension of the McAdams et al.'s (1995) timbre space, which used 18 of the same 21 sounds in Krumhansl's study. Indeed, spectral variation was the only acoustical parameter in McAdams et al.'s (1995) study that significantly correlated with the third dimension, even though it accounted for only 29% ($r = 0.54$) of the variance along this dimension. When four of the timbres (clarinet, trombone, guitarnet, and vibrone) were removed, the variance increased to 39%, and their removal did not affect the correlations of attack time and spectral centroid with Dimensions 1 and 2.

2.2 *The Notion of Specificities*

The degree of variability in similarity data from the early scaling studies on timbre leads us to think that two or three common dimensions are not enough to describe the perception of timbre. Moreover, one may question the validity of the assumption that two or three dimensions can explain all the differences among extremely complex sounds like musical instrument tones. To take into account this complexity, some authors suggest that each timbre may also be defined by unique characteristics (Krumhansl, 1989; McAdams, 1993). On the other hand, it will be important to take these specificities into account in the modeling of the mental structure of timbre because they might play a major role in the identification of musical instruments. For example, when the spectral envelope is unique (e.g., clarinet vs trumpet), it seems to contribute more to identification than when the temporal envelope is distinguished (e.g., flute vs trombone). This suggests that listeners use characteristics that specify the instrument with the least ambiguity and that they are not constrained to listening for a single cue across all possible sources (Strong and Clark, 1967a,b). For example, in a study on string instruments, Mathews et al. (1965) found an initial inharmonic frequency component corresponding to the irregular vibration that appears when the bow first sets the string into vibration. Such details can be characteristic of particular sound sources, and the auditory system seems to be sensitive to these identifying details.

Timbre may thus be defined by not only two or three common, continuous dimensions but also by distinguishing features or dimensions that are specific to a given sound. To test this notion, Krumhansl (1989) applied an extended Euclidean model developed by Winsberg and Carroll (1989). By postulating the existence of unique features for certain timbres, this model was designed to provide an explanation of the variability in similarity judgments that could not be attributed to the three principal MDS dimensions derived from dissimilarity judgments based on 21 synthesized imitations and hybrids of conventional Western musical instruments. Globally, 60% of the timbres yielded non-zero specificity values. Specific examples are the harpsichord, the clarinet, and some of the hybrid timbres such as the "pianobox" (bowed piano), the "guitarnet" (guitar/clarinet hybrid), and the

“vibrone” (vibraphone/trombone hybrid) that yielded high values of specificity. While no attempt was made to interpret these specificity values by systematically relating them to acoustic properties, Krumhansl conjectured that the specificities of certain instruments reflected specific mechanical characteristics that could be important for their identification. For example, the return of the jack in the harpsichord mechanism or the cylindrical geometry of the air column of clarinet could have important perceptual ramifications.

McAdams et al. (1995) attempted to find a qualitative interpretation of the specificities captured by their model on the same set of sounds. First, the authors noted a monotonic relationship between the specificity values and the perceptual strength of the specificities. However, this relationship was not tested systematically. Second, the authors distinguished: (1) continuous features that varied by degree (such as “raspiness” of attack, inharmonicity, “graininess” deviation of pitch glide, and “hollowness” of tone color); and (2) discrete features that varied by perceptual strength (such as a high-frequency chiff on the onset, a suddenly damped or pinched offset, or the presence of a clunk or thud during the sound). The authors concluded that such specificities may account for both additional continuous dimensions and discrete features of variable perceptual salience.

Another hypothesis was that specificities may reflect unfamiliarity of sounds to listeners, and, therefore, hybrid timbres should yield a high value of specificity. However, on average, in the two models (Krumhansl, 1989; McAdams et al., 1995), hybrid timbres did not yield higher specificities than those of conventional instruments. Actually, half of the hybrid timbres tested yielded lower specificities than the average value. Moreover, this hypothetical relationship between specificity and familiarity was not supported by the (very familiar) piano timbre, which yielded a high value in both studies. In fact, the piano is probably one of the most familiar instruments to the primarily European listeners who participated in these studies.

To conclude, these results suggest that structural sound characteristics influence dissimilarity judgments made by subjects. These characteristics may be common to all the timbres within a stimulus set or specific to some timbres. A classical Euclidean model could not take these specific features into account and an extended model is, therefore, more appropriate. Acoustical analyses must still be conducted in order to give a psychoacoustical interpretation of the specificities that were found.

2.3 Individual and Group Listener Differences

Most of the timbre spaces described above were derived exclusively from musician listeners (Grey, 1977; Wessel, 1979; Krumhansl, 1989). A few studies have tried to determine whether perceptual differences between auditory classes correspond to biographical factors, such as the level of musical training or cultural origin, but they have found no systematic differences related to musical training (Miller and Carterette, 1975; Wedin and Goude, 1972). However, whereas most of us, musician or not, can distinguish a guitar from a clarinet, we might suppose that

the mental structure of the perceptual relations among different timbres would not be the same depending on the musical competence of the listener.

Musical competence potential differences might be found by analysis of weight patterns attributed to the different dimensions and specificities of a common space. The weights' interpretation could be based on biographical factors such as musical experience. The INDSCAL (INdividual Differences SCALing) model, proposed by Carroll and Chang (1970), can account for such individual perceptual differences. Serafini (1993) used individual-differences scaling to test two groups of Western musician listeners on a set of Javanese percussion sounds (xylophones, gongs, and metalophones) and a plucked-string sound. One group was familiar with Indonesian gamelan music (they had played Javanese Gamelan music for at least two years), and the other was unfamiliar with this type of music. The task was to judge the dissimilarity between pairs of isolated notes and pairs of melodies played by these instruments. Stimulus and subject INDSCAL two-dimensional solutions yielded one dimension (Dimension 1) corresponding to the spectral centroid of the attack portion of tones and a second dimension (Dimension 2) to the mean amplitude level of the resonant portion of the tone (a dimension related to loudness). For isolated tones (see Figs. 8.8a and 8.8b), no differences were found between the two groups of listeners. However, for melodies, the group unfamiliar with gamelan music gave equal weight to the two dimensions, whereas gamelan players weighted the attack dimension more heavily (see Figs. 8.8c and 8.8d).

McAdams et al. (1995) conducted a study on a large number of listeners of varying levels of musical training with an analysis of *latent-class structure*. The aim was to examine whether listeners could be sorted into different classes according to their perceptual data and whether a relation between the class structure and musical training of the listeners could be found. For musical pitch, Shepard (1982) had observed a dimensional structure that was different for musicians and non-musicians. The structure was richer, i.e., had higher dimensionality, for musicians than for non-musicians. This result led McAdams et al. (1995) to hypothesize that the same type of result could be observed for timbre perception: either the number of dimensions would be greater for the musicians' dimensional structure, or the weights on the dimensions would be more evenly distributed. However, in fact, musicians, amateurs, and non-musicians did not fall into separate latent classes even if some differences were observed in the proportional distribution of biographical factors. The analysis of the different weights across dimensions and specificities showed that two of the five classes observed among 98 listeners gave roughly equal weights across dimensions and specificities, while the other classes gave high weights on two dimensions, or on one dimension and the specificities, and low weights on the others, respectively (see Fig. 8.9).

Two different interpretations of this weight pattern were proposed by the authors. First, the weight pattern observed could reflect a strategy difference between subjects over the course of the experimental session. Equal weights across the three dimensions and specificities observed for two of the five classes could be due to the subjects in these classes shifting their attention among the different dimensions and specificities, while the subjects in the other classes may have adopted more

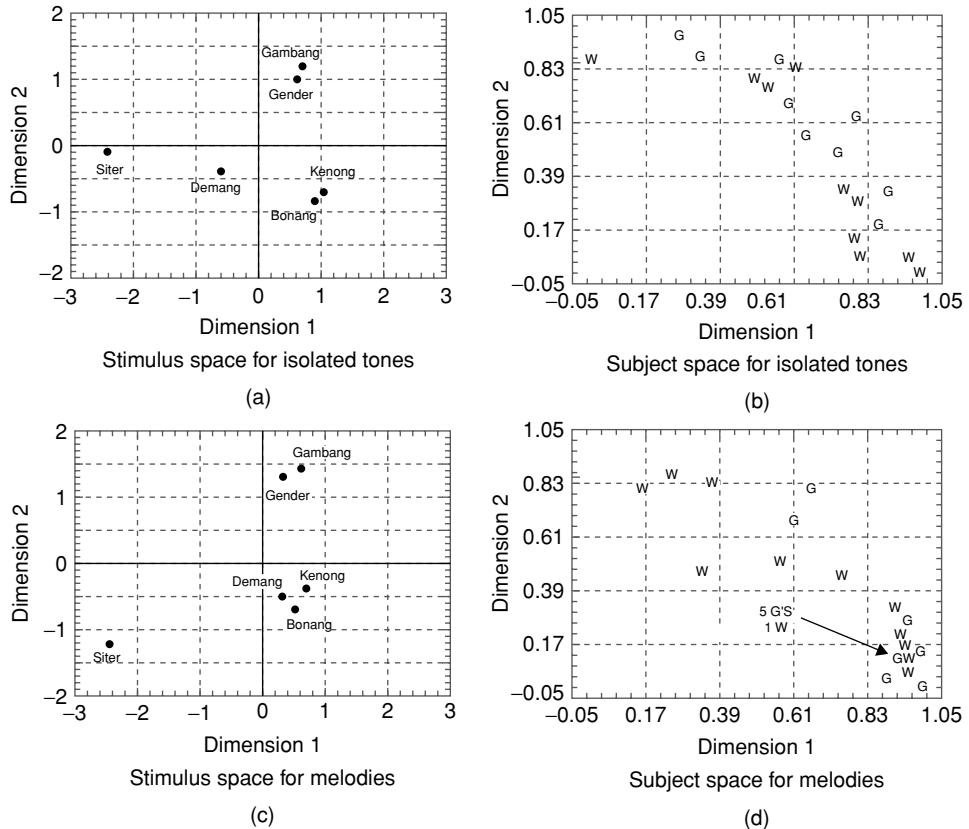


FIGURE 8.8. (a) Two-dimensional stimulus space derived from an INDSCAL analysis on similarity ratings on six isolated gamelan sounds. (b) Subject space observed for the six isolated gamelan sounds. (c) Two-dimensional stimulus space derived from an INDSCAL analysis on similarity ratings on six melodies played by six gamelan sounds. (d) Subject space observed for the six melodies. (“G” refers to listeners familiar with Indonesian gamelan music.) “W” refers to “Western” listeners unfamiliar with gamelan music. [From Serafini (1993), adapted with permission of Waterloo University.]

consistent strategies of judgment that focused on a smaller number of dimensions and stuck to them throughout the experimental session. The second interpretation suggested a difference between subjects in different classes in their cognitive capacity to process different aspects of sounds in parallel. According to this interpretation, subjects in the two classes who equally weighted the dimensions and specificities were able to focus on more dimensions at a time than could members of the other classes, and one might predict *a priori* that these would be principally musicians. However, the authors observed that both musicians and non-musicians were able either to equally weight all dimensions or to give special attention to some dimensions like the attack time or the spectral centroid. Thus, the distribution of

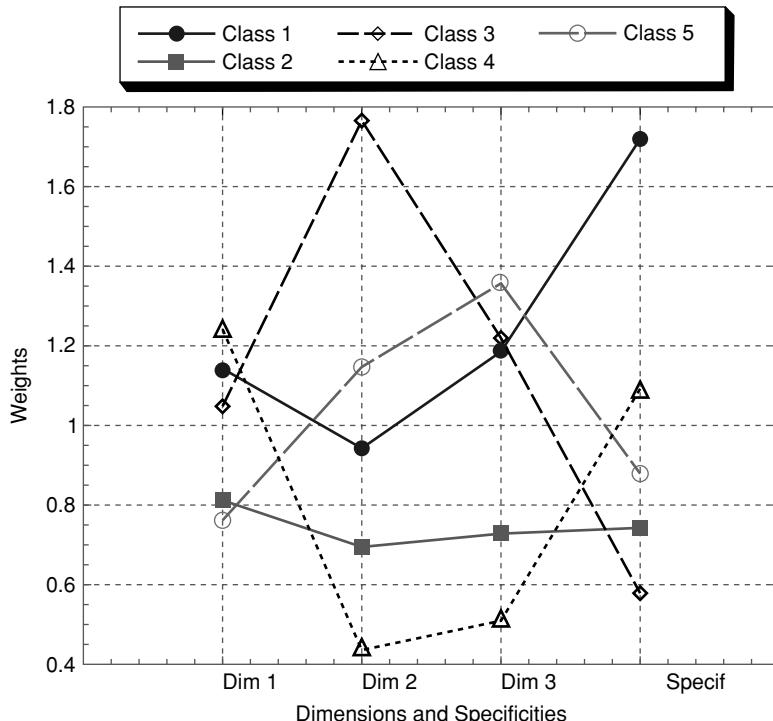


FIGURE 8.9. Class weights (mean weights across dimensions and specificities for each class) for spatial model plotted for each of five classes. Weights were estimated in a three-dimensional space for five latent classes. [Derived from McAdams et al. (1995), Table 4 used by permission of the author.]

the three original classes of listeners (musician, amateur, and non-musician) within each latent class was roughly equivalent to their distribution in the whole subject population employed. The pattern of weighting of a given subject cannot be predicted simply from the biographical data related to that subject concerning their degree of musicianship or their years of music training, performing, or listening.

The only differences observed between musicians, amateurs, and non-musicians were the variances about the model distances, observed for the solutions computed separately for musician and non-musician groups: The variance for non-musicians and amateurs combined was greater than that for musicians. However, the variances observed for individual latent classes, composed of musicians, amateurs, and non-musicians, were less than the variance of the musician group, suggesting that the inclusion of class weights in the dimensional models is justified in terms of model fit because it reduces the overall variance. This pattern of results suggests that the effect of musicianship is, among other things, one of variance. Latent classes do not differ with respect to variance, but musicians and non-musicians do. So musicianship appears to affect judgment precision and coherence.

2.4 Evaluating the Predictive Power of Timbre Spaces

In some studies that attempted to evaluate timbre space as a predictive model, the explicit aim was to determine the validity of the model. In others, the idea was to see if timbre space could be used to test other hypotheses. Four types of research will be discussed that support the validity and utility of such models.

2.4.1 Perceptual Effects of Sound Modifications

An assumption of the timbre space model is that specific acoustic properties underlie the continuous perceptual dimensions. If we modify the acoustic properties for a single perceptual dimension in a systematic way, we should observe perceptually interpretable changes of the positions of stimuli along that dimension. A study conducted by Grey and Gordon (1978) confirmed this assumption. They exchanged the spectral envelopes of pairs of instruments drawn from the Grey (1975, 1977) study, while trying to preserve other properties, and conducted a new multidimensional study with half of the original sounds modified and the other half intact. The hypothesis was that the positions of the original and hybrid sounds should change along the dimension that best correlated with a measure of the spectral envelope. The results demonstrated that in all cases the tones exchanged places along the “brightness” (or spectral-centroid) dimension, although in some cases displacements along other dimensions also occurred. These displacements still respected the nature of the perceptual dimensions: Temporal-envelope changes resulting from the way the spectral envelope varied with time resulted in appropriate changes along the dimension that best correlated with spectral flux.

On the other hand, the most natural way to move in a timbre space would be to attach the handles of control directly to the different dimensions of the space. Wessel and colleagues (1979, 1983, 1987) examined such a control scheme in a real-time context. A two-dimensional timbre space was represented on a computer graphics terminal allowing control of a digital processor. One dimension of the space was used to manipulate the shape of the spectral-energy distribution. This was accomplished by appropriately scaling line-segment spectral envelopes according to a shaping function. The other axis of the space was used to control either the attack rate or the extent of synchronicity among the various components. Overall, the timbral trajectories in these spaces were reported by the author to be smooth and otherwise perceptually well-behaved.

All of these results and observations suggest that some intermediate regions of the timbre space could be filled in and that regular, finely graded transitions are conceivable, thus supporting the hypothesis that timbre perception can be modeled by continuous physical dimensions that underlie a small number of perceptual dimensions.

2.4.2 Perception of Timbral Intervals

Classical musical structures are based on the separation and the grouping of sound events according their relative differences in pitch (melody), intensity (dynamics),

duration (rhythm), and timbre (instrument). Research on timbre tries to expand this conception of the organization of musical sequences. Indeed, transposing timbral sequences may be heard by listeners and used consciously by composers. The aim of the following studies was to test the idea of the composer Arnold Schoenberg (1911) that musical phrases can be formed by notes which differ only in timbre. Once a timbre space has been quantified, one might ask whether the structure of the common dimensions is useful as a tool for predicting listeners' abilities to compare relations among the different timbres.

Ehresman and Wessel (1978) were among the first to apply Rumelhart and Abrahamson's (1973) parallelogram model of analogical reasoning with a two-dimensional space composed of traditional musical instrument sounds (Grey, 1977). This model predicts that if the relation between two objects A and B is represented as the vector A-B in the space, another vector C-D will be perceived as analogous if it has the same magnitude and orientation as A-B. In the analogy task, vector A-B is presented and a series of vectors C-D_i are presented. According to the model, the subjects will choose the D_i that is closest to the end point of a vector starting at C and having the same magnitude and direction as A-B. This ideal point is called I and the vectors A-B and C-I thus form a parallelogram in the space. Analogies of the form A, B, C (D₁, D₂, D₃, D₄), where D_i was varied according to its distance from I, were constructed. The probability of choosing D_i as the best solution was found to be a monotonically decreasing function of the absolute distance of D_i from I, thus supporting the parallelogram model. Ehresman and Wessel proceeded in analogous fashion with musical instrument tones. The two perceptual dimensions of their space corresponded to (1) "spectral energy distribution" of the tones and (2) "nature of the onset transients." The results were better predicted by this model than a number of other models. In addition, timbral vectors were computed from a two-dimensional solution and only relative vector magnitude (corresponding to the estimated perceived dissimilarity) was tested, ignoring the direction components.

McAdams and Cunibile (1992) tested a similar geometric model for the three-dimensional space observed by Krumhansl (1989) taking into account separately the magnitude and orientation of the different timbral vectors. Sequences of four timbres of the perceptual space (five different sets for each experimental condition) were constructed according to four experimental conditions differing in the degree to which they corresponded to the "good" analogy defined by the model: (1) good magnitude, good orientation; (2) good magnitude, bad orientation; (3) bad magnitude, good orientation; and (4) bad magnitude, bad orientation (see Fig. 8.10). Two sequences of four timbres, where only the last varied between the two sequences, were presented to listeners (musicians and non-musicians). The task was to choose the sequence that best corresponded to an analogy of the form: timbre A is to timbre B as timbre C is to timbre D. The hypotheses were: (1) sequences in which the A-B and C-D vectors formed a parallelogram would be preferred; (2) sequences in which the C-D vector had a good magnitude but a bad orientation would be preferred over those with bad magnitude and orientation; (3) sequences in which the C-D vector had a bad magnitude but a good orientation would be

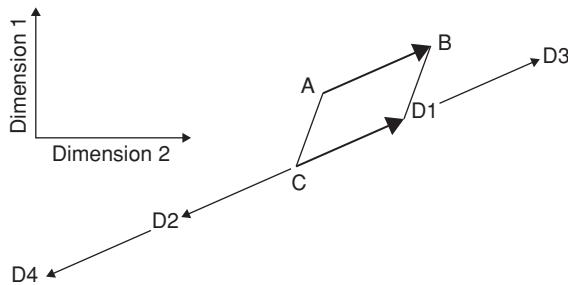


FIGURE 8.10. Parallelogram model of timbre analogies. (The two-dimensional case is shown.) A to B is a given change in timbre; C to D is a desired timbral analogy, with C given. D₁, D₂, D₃, D₄ are the different analogies offered to the listeners with D₁ corresponding to the ideal point according to the model.

preferred over those with bad magnitude and orientation; and (4) there would be no differences among the different versions of each comparison type because the analogy judgment is based on a perception of abstract relations among the timbres of the stimulus tones. The results showed that: (1) the listeners preferred sequences with good magnitude and orientation; (2) sequences with either good magnitude and bad orientation or bad magnitude and good orientation were preferred significantly more often than those with bad magnitude and bad orientation; (3) however, the judgments for the different versions of each comparison differed significantly from one another. According to the authors, these latter differences may have been due to the presence of specificities that were not taken into account in computing the vectors in this experiment. Indeed, if we consider that certain timbres had specificities, this would distort the vector established on the basis of the common dimensions alone.

Overall the results were encouraging, indicating an ability to perceive timbral analogies on the basis of a timbre space describing the dissimilarity relations among different timbres. However, these studies lacked control for specificities, which can influence the dissimilarity between timbres and thus the distances separating them in the space. They also needed to control for the positions of vector pairs in the perceptual space by using a synthetic space in which the timbres are distributed in a homogeneous fashion.

2.4.3 The Role of Timbre in Auditory Streaming

We know now that many aspects of sound are important in auditory streaming: intensity (Van Noorden, 1975), fundamental frequency (Bregman, 1990; Bregman et al., 1990; Miller and Heise, 1950; Singh, 1987; Van Noorden, 1975), spectral factors (Hartmann and Johnson, 1991; McAdams and Bregman, 1979), and temporal factors (Hartmann and Johnson, 1991). Even if the majority of researchers consider timbre to be an attribute of sound processed after auditory grouping, it

seems that the spectral, temporal, or spectrotemporal properties giving rise to timbral attributes may also contribute to auditory stream segregation. A hypothesis may be made that sequential groupings of complex sounds are based on the spectral or temporal similarity of the sounds. In these cases, the auditory system would organize sound events in the same stream when they are sufficiently similar.

Several researchers (McAdams and Bregman, 1979; Wessel, 1979; Iverson, 1993; Gregory, 1994; Bey and McAdams, 2003) have studied streaming by musical timbre. Wessel (1979) conducted an early demonstration of streaming employing 16 synthetic instrument tones. In a previous experiment he had subjects rate the similarity of these tones and used MDS to fit the judgments to a two-dimensional space with one dimension corresponding to spectra and the other to onset transients. To test the relationship between similarity judgments and streaming, he constructed repeated sequences of three ascending notes with alternate notes differing in timbre, but otherwise the pitch sequence and rhythmic timing remained fixed. When the timbral distance between the adjacent notes was small along the spectral dimension, a repeating, ascending pitch line was heard. However, when the timbral distance was enlarged along this same dimension, listeners heard two streams with one stream comprised of timbre A and the other of timbre B. This phenomenon is called "melodic fission" or "auditory stream segregation." On the other hand, a different effect was obtained when the note timings were modified. In this case, a single stream with perceptually irregular rhythm was perceived regardless of the timbral distance separating the different notes. This result suggested that the spectral dimension influenced auditory streaming but the temporal dimension did not.

Iverson (1993) also conducted a series of experiments to test the relationship between similarity judgments and auditory streaming. In a previous study, the author examined 16 tones using a standard similarity-scaling technique and found a two-dimensional MDS space where the 16 tones were represented. The second experiment assessed the relationship between similarity judgments and streaming. Pairs of sequences, constructed with the same 16 tones used in the first experiment, were presented to listeners. The task was to rate the degree of streaming of each sequence on a continuous scale, resulting in a triangular matrix giving the relative streaming of each pair of tones. The streaming ratings were used as a similarity metric for MDS, so tones that formed one stream were closer in the space than tones that formed two streams. A two-dimensional space was obtained similar to those observed with the similarity ratings on single tones. The first dimension corresponded to attack quality and the second to the perceived brightness of the sounds. Acoustical attributes were identified and correlated with the judgments. Iverson showed that sounds with similar spectral or temporal envelopes were integrated into one stream and sounds with different spectral or temporal envelopes were segregated into different streams. This result showed the importance of temporal factors in auditory streaming, contrary to the results observed by Wessel (1979) and Hartmann and Johnson (1991).

Gregory (1994) tested the influence of each perceptual dimension of timbre in auditory streaming. The three dimensions of his MDS space were "relative

percentage of energy in the first three partials,” “decay duration,” and “relative strengths of odd to even partials.” Listeners were tested to determine their abilities to separate streams according to the perceptual distances of timbres observed in the timbre space. When the timbral difference was increased, auditory streaming was not based on the pitch difference but on timbral difference. Moreover, the temporal dimension seemed more important than the two other spectral dimensions in auditory streaming.

A study conducted by Bey and McAdams (2003) confirmed the role of temporal factors. The subjects’ task corresponded to a recognition of interleaved melodies. Sequences were composed of two melodies with timbres that were more or less distant in Krumhansl’s (1989) perceptual space. Results showed that differences along the spectral and spectrotemporal dimensions were not sufficient to separate the two melodies, and recognition of the embedded melodies was thus not possible. However, if sounds also varied on the temporal dimension, listeners could separate the two melodies, and recognition performance was improved. Furthermore, the authors showed that a timbre difference combined with a pitch difference led listeners to separate the two melodies even more than if only a timbre or pitch difference distinguished them.

The studies conducted by Gregory (1994), Iverson (1993), and Bey and McAdams (2003) illustrate the contribution of temporal factors to listeners’ abilities to separate sound streams and counter the idea that only spectral factors are significant in auditory streaming (Bregman, 1990; Bregman et al., 1990; McAdams and Bregman, 1979; Miller and Heise, 1950; Singh, 1987; Van Noorden, 1975; Wessel, 1979).

2.4.4 Context Effects

While spectral factors seem to systematically influence timbre perception, depending on context, temporal factors are not always as salient. Wedin and Goude (1972) observed that the presence or absence of attack transients did not influence the perceptual representation of a set of musical timbres. The mean dissimilarity of the two tested conditions was highly correlated ($r = 0.92$).

Miller and Carterette (1975) attempted to demonstrate the perceptual importance of temporal parameters of timbre. Their stimuli were synthetic tones with variable harmonic spectra, variable amplitude-vs-time envelope, and variable onset delays for the harmonics (temporal properties). They obtained a three-dimensional MDS solution that accounted only for harmonic structure (in dimensions 1 and 2) and the amplitude-vs-time envelopes (in dimension 3), so that the contribution of harmonic onset delay pattern on timbre perception remained in doubt. However, their temporal properties were indeed organized and combined along the third dimension.

According to Iverson and Krumhansl (1991), when sounds are isolated, the attack seems essential to their recognition but does not seem to be the determining factor in similarity judgments. Iverson and Krumhansl (1993) confirmed these results showing that attributes on which listeners based their dissimilarity judgments

among different timbres were present in the duration of the sound. Indeed, they observed similar multidimensional spaces for sounds in which only the first 80 ms were presented, sounds where only the first 80 ms were removed, as well as original sounds.

Many studies have shown that temporal aspects of sounds are perceptually less pertinent when situated in a musical context. Grey (1978) used simplified sounds of three instruments: bassoon, trumpet, and clarinet. He first created notes of different pitches by transposing each instrument's spectrum to higher or lower frequencies. He then asked listeners to distinguish simplifications applied for isolated instrument sounds or for the same sounds placed in different musical configurations, differing in the number of simultaneous melodic lines, rhythm variety, and temporal density. The musical context effect was measured by noting the difference in discrimination ability for the various conditions. While for the bassoon no effect of musical context was observed on discrimination between the original and modified versions, discrimination performance was found to decrease with musical context for the clarinet and trumpet. An acoustical analysis of the original and modified bassoon sounds showed that the simplification involved changes in the spectral envelope, which was not the case for the other instruments. For the bassoon, the changes were described by listeners as brightness differences, which corresponded to spectral envelope changes. On the other hand, changes described for the trumpet and clarinet were located in the "attack" or in the articulation. Small spectral differences were thus slightly enhanced in single-voice contexts compared with isolated tones and multivoiced contexts, although discrimination remained high. Articulation differences, on the other hand, were increasingly disregarded as the complexity and density of the context increased.

Similarly, Kendall (1986) conducted an experiment in which tone modifications were made by time-domain editing. Two different note sequences were presented to listeners whose task was to decide which instrument in the second sequence corresponded to the instrument sounded in the first sequence. The first sequence was an edited version of the same melody played by one of the three instruments used: clarinet, trumpet, or violin. The second sequence consisted of the melody played in unedited form in random order by each of the three instruments under the following conditions: (1) normal tones, (2) sustain portion only (cut attacks and decays), or (3) transients only (with either a silent gap in the sustain portion or an artificially stabilized sustain portion). The results suggested that transients in isolated notes enhance instrument recognition when they were alone or coupled with a natural (time-varying) sustain portion but were of little value when coupled with a static sustain part. They were also of less value in continuous musical phrases where the information present in the sustain portion (probably related to the spectral envelope) was more important. This conclusion confirmed Grey's (1978) discrimination study and was verified by McAdams's (1993) study, which utilized stimuli with more realistic variations.

In comparison to studies on isolated sounds (Berger, 1964; Charbonneau, 1981; Grey and Moorer, 1977; Saldanha and Corso, 1964), these results suggest that

attack transients play a less important perceptual role for musical phrases than they do for isolated tones.

2.5 *Verbal Attributes of Timbre*

2.5.1 Semantic Differential Analyses

One approach to the study of timbre perception of complex sounds is the analysis of verbal attributes used to describe them. Some authors (Lichte, 1941; Solomon, 1959; Terhardt, 1974; Vogel, 1974; von Bismarck, 1974) have hypothesized that timbre can be described by semantic scales. For example, scales can be presented to listeners in which the extremities are two opposing verbal attributes such as “smooth–rough” or “light–dark.” They are asked to rate each timbre on each scale. A factor analysis is used to identify a number of factors or scales contributing to explaining variance in the judgments. The remaining scales are considered to describe the different timbres used.

Semantic studies began with Lichte (1941) study of the “bright/dull” and “thin/full” scales using synthetic harmonic tones. Solomon (1959) investigated seven timbral attributes of sonar recordings and the contribution of each spectral region made to each attribute. Terhardt (1974) and Vogel (1974) both examined the notion of “roughness” for the steady-state portion of synthetic sounds.

One of the most complete psychophysical timbre studies was performed by von Bismarck (1974) in which subjects had to rate 35 speech sounds (having equal loudness but different spectral envelopes) on 30 verbal scales such as “brilliant-dull” or “wide-narrow.” A factor analysis showed that four orthogonal factors were sufficient to account for 90% of the variance. Timbre would have, according to this study, four dimensions: (1) thick/thin; (2) compact/diffuse; (3) colorful/colorless; and (4) full/empty. A major problem with this type of study is that the choice of the verbal attributes characterizing the scales does not always correspond to scales that subjects would choose spontaneously. A timbral dimension correlating with a specific acoustic parameter such as spectral fine structure cannot be revealed by such a study. Moreover, the meaning of certain terms is likely to vary according to the musical culture of the subject.

2.5.2 Relations between Verbal and Perceptual Attributes or Analyses of Verbal Protocols

To eliminate some problems posed by semantic differential studies, Faure et al. (1996) and Faure (2000) used subjects’ free verbalizations analyzed by a paradigm developed by Samoylenko et al. (1996). Free verbalization does not impose a vocabulary on the listener. The aim was to define the verbal correlates of the different perceptual dimensions of timbre. The listeners (musicians, non-musicians, and amateurs) were asked to judge the degree of dissimilarity of pairs of timbres [a subset of Krumhansl’s sounds (1989)] and then to describe all the dissimilarities and similarities between the timbres. The listeners could modify their dissimilarity judgment after their verbalization.

Two different multidimensional analyses were performed on the ratings given before and after the verbalization. The two resulting timbre spaces were similar suggesting that the verbalization process did not affect the mental structure of timbre. This result allowed a comparison of the dissimilarity judgments to the verbalization.

To find verbal correlates, 22 descriptors were extracted from expressions of the form: "sound 1 is more (or less) X than sound 2". These descriptors were /high/-/sharp/-/shril/, /low/-/deep/, /long/, /clean/-/distinct/, /mussed/-/dull/, /round/, /clear/-/light/, /resonant/, /nasal/, /metallic/, /vibrated/, /strong/-/loud/, /dry/, /soft/, /rich/, /high/, /low/, /wide/, /diffuse/, /brilliant/-/bright/, /plucked/ and /blown/. Some descriptors were correlated with one MDS dimension while others were correlated with more than one dimension.

Coefficients from multiple regressions were used to project verbal vectors in the multidimensional timbre space. If a descriptor's vector was correlated to only one dimension, it was aligned along the axis of this dimension. If a descriptor was partially correlated to two different dimensions, the vector formed an angle with the two dimension axes, the slope reflecting the ratio between the regression coefficients. Only a few descriptors were correlated to only one dimension. These descriptors were /dry/—correlated with the log of the attack time dimension; /round/—correlated to the spectral centroid; and /brilliant/-/bright/—correlated to the spectral flux. The other descriptors were correlated to more than one dimension: /metallic/, for example, was correlated with three perceptual dimensions. The authors explained these multiple correlations by the fact that sounds characterized as /metallic/ generally have a fast attack, a resonance with much energy in the high frequencies, and a spectral evolution that reflects more rapid damping of high frequencies. On the other hand, the descriptor /mussed/-/dull/ was very often the antonym of /metallic/. Indeed, its vector formed a 180° angle with that descriptor. This result suggests that Faure's approach may be very useful for research to determine verbal antonyms and synonyms describing the timbre of complex sounds.

3 Categories of Timbre

A different view of perceptual activity will now be presented. According to this view, when we are subjected to multiple physical stimulation coming from the environment, we experience multiple sensations. In order to behave coherently when faced with this environment, we need to classify the stimuli. How is this done and what is the structure of our mental representation when this classification is accomplished? Numerous authors have proposed the existence of categorization processes which could be at the origin of a categorical structure of the perceptual representation of most stimuli. An example is the categorical perceptual phenomenon of speech, in which the capacity to discriminate differences between speech sounds is determined by the capacity to differently categorize these kinds of sounds. In this case, we seem to transform initial continuous information into a discrete form. Some authors postulate that the conversion from a continuous variation of a stimulus to a discrete form is based on a late stage of the recognition

process, while others postulate that this conversion occurs during low-level stages of the perceptual process.

Besides spatial models used to determine the mental structure of the timbre of complex tones presented in the last section, there are non-spatial models in which each object is described in terms of its common and distinctive features and represented by discrete clusters. Tversky (1977) proposed a “feature matching model” based on the idea that when faced with a set of objects, subjects often sort them into clusters to reduce information load and facilitate further processing. This model is based on a similarity relation that is very different from that of the geometric models. According to Tversky, each object is represented by a set of features or attributes. Thus, the degree of similarity $s(A, B)$ between objects A and B, for all distinct A and B, is defined by a matching function between the common and distinctive features of the two compared objects. This function is composed of three arguments: (1) A intersect B: the features that are common to the two compared objects A and B; (2) A minus B: the features belonging to A but not B; and (3) B minus A: the features belonging to B but not A.

This approach is formalized by cluster analysis. In a cluster analysis, objects that are similar belong to the same cluster and objects that are dissimilar belong to different clusters. Clustering of objects can be hierarchical or nonhierarchical. In the case of nonhierarchical clustering, objects can belong to one and only one cluster. However, with hierarchical clustering, objects can belong to more than one cluster as long as they are hierarchically nested; i.e., all members of a lower-level cluster belong to a higher-level cluster. One way to represent hierarchical clustering is with a tree, a graph in which the similarity between two objects is represented by the length or the height of the link joining the two objects. In a hierarchical representation obtained by the application of the HICLUS model proposed by Johnson (1967), objects that are most similar are joined at lower levels in the tree, whereas dissimilar objects are joined together only at higher levels in the tree. Also, the ADCLUS model proposed by Shepard and Arabie (1979) provides a representation that allows partial overlapping of clusters. Finally, there are additive trees in which similarity between objects is given by the lengths of links between nodes in the trees.

Other authors (Gibson, 1966, 1979; Rosch, 1973a,b), have postulated that the physical world that surrounds us has discontinuities, which eliminates the problem of determining the level at which such a categorization occurs. According to Gibson (1966, 1979), all information necessary for visual perception is present in the environment and the perceiving subject has only to pick it up. This conception leads us to consider only natural situations (from which the term “ecological” is derived) and to reject the general validity of laboratory experiments. Gibson’s theory is opposed to all constructivist positions according to which information is extracted from sense systems (visual, auditory, and the like) by computational procedures and processes. All these processes are judged to be useless because information given by the physical environment to the perceiving subject is already structured and organized in a coherent manner. Perception is thus direct because the information is presorted and does not need to be processed.

For auditory nonverbal perception, this conception suggests that the physical nature of the sound object, the means by which it is set into vibration, and its function for the listener are perceived directly, without intermediate processes. In other words, there is no analysis of the individual elements that comprise a sound event; nor is there a reconstruction of an auditory image that is compared to a representation in memory (McAdams, 1993). Thus, the approach for ecological psychologists is to describe the structure of the physical world in order to understand perceived properties as invariants. Note that we can usually recognize a saxophone or a piano played on the radio even if the signal is modified by bad transmission. If invariants can be isolated, then the task of the psychologist is to determine how the listeners detect these properties. This approach allows us to evoke a mechanism of “causality inference”: Received data are indices considered as effects of a causality, which is the perceived object. This conception thus suggests a strong relation between the mental representation of a sound event, its production mode, and its perceptual identity.

The question whether perception of timbre is categorical is not neutral with respect to causality. Historically, the relation between the physical production of a sound event and its auditory result has been obvious. Indeed, at one time the term “timbre” designated a particular instrument, a sort of drum with stretched strings that gave a characteristic “color” to its sound (Dictionnaire de l’Academie Francaise, 1835). But the predominance of pitch in most musical cultures has relegated timbre to a secondary role. Classical instruments, excluding some percussion instruments, were constructed so that anything that disturbed pitch recognition was eliminated. In the absence of an explicit musical function, it is natural that “timbre” tends to no longer refer to a particular sound source or instrument. However, even today, classical instruments are categorized, and if categorization appears at a perceptual level, it is likely to be due to the type of sound source. While it is difficult to physically construct an intermediate instrument between a percussive instrument and a sustained instrument, electronic synthesis allows us to create hybrid timbres and place perception outside of the mechano-acoustical instrument categories. Even so, the perception of timbre as revealed by multidimensional space analysis, where continua of timbre are theoretically possible, seems partially categorical. According to Grey (1975) “the scaling for sets of naturalistic tones suggests a hybrid space, where some dimensions are based on low-level perceptual distinctions made with respect to obvious physical properties of tones, while other dimensions can be explained only on the basis of a higher level distinction, like musical instrument families” [cited by Risset and Wessel (1982, p. 48)]. The intervention of cognitive processes, such as familiarity with or recognition of an instrument, shows that it is perhaps impossible to obtain a totally continuous timbre space.

3.1 Studies of the Perception of Causality of Sound Events

An alternative approach for studying timbre perception is to consider that musical instruments are often grouped on the basis of their belonging to resonator and/or exciter categories and that the mechanical properties of sound sources could

influence dissimilarity judgments among different timbres. Indeed, some categorization processes may be likely to influence listeners' dissimilarity judgments on which the notion of timbre space is based. The aim of an experiment performed by Donnadieu (1997) was to examine such categorization processes by a classification task and to specify the relation that could exist between a multidimensional representation of timbre and a categorical representation. In other words, the study's goal was to determine the perceptual categories underlying 36 digitally recorded musical instrument sounds selected from the McGill University Master Samples compact disk (Opolko & Wapnick, 1987). These included tones produced by traditional pitched sustained instruments (e.g., flute, trumpet, piano), tones of strongly pitched percussion instruments (e.g., celesta, marimba, vibraphone bowed, vibraphone struck, tympani), weakly pitched (e.g., bowed cymbal, log drum), and unpitched (e.g., tam-tam, bamboo chimes), representing most of the types of excitors and resonators used in the orchestra. The objective was to determine whether listeners based their classifications on instrument families or on certain physical attributes of sound objects. A multidimensional representation of the categorical structure was used in order to define how timbre categories are partitioned in a timbre space and to evaluate the influence of the physical functioning of instruments on perceptual categorical structure.

Sixty subjects were asked to perform a free classification task. Two advantages of this type of task are that it is easily performed by listeners and it can help to determine the kinds of sound properties that are worth investigating more systematically. All stimuli were first presented to the subjects, and they were asked to create their own categories and to assign similar stimuli to the same category and dissimilar stimuli to different categories. In a free classification task, subjects can create as many categories as they want and can assign as many stimuli as they wish in each category. To determine the categorical structure of this set of stimuli, an ADTREE analysis (Barthélemy and Guénoche, 1988) was used, which allowed the development of an additive tree, a graph in which the similarity between any two nodes, corresponding to the objects, is given by the length of the link between those nodes. The observed tree for the 36 orchestral instruments is represented in Fig. 8.11. According to Tversky's model (1977), the nodes can be interpreted as the prototype of a category which corresponds to the object that shares common features with the objects belonging to this category, while the length of the link between two nodes corresponds to the weight of the features belonging to class A but not to class B, for example. This last model was used because it was particularly easy to interpret this type of representation according to the model proposed by Tversky. Trees were established for all the subjects. An attempt was made to establish a relation between different perceptual categories and the stimulus properties, most of the time by seeking structural similarities among stimuli classed together and differences between stimuli classed in different categories. Such a classification was observed with all impulsive excitation (percussion) instruments in one category and all sustained excitation instruments in another category. Classifications were also observed according to each instrument's resonator type, with strings, plates, and bowed membranes placed in different categories. Influence of resonator type was particularly evident when two types of vibraphone sounds were examined:

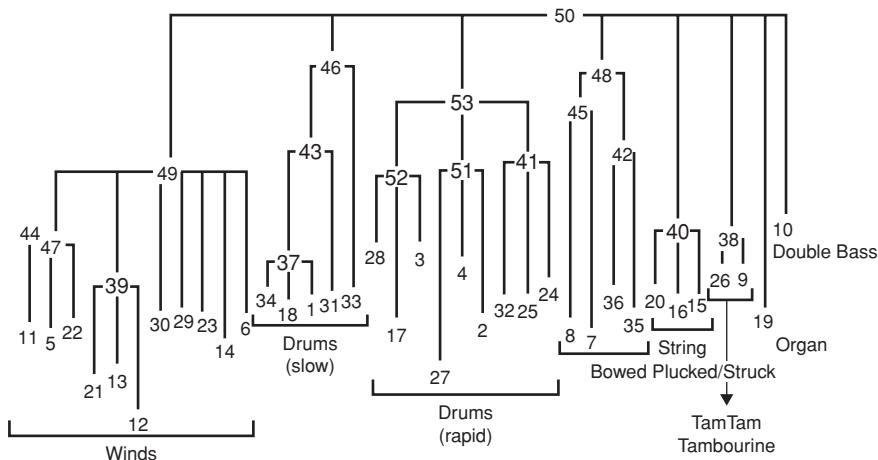


FIGURE 8.11. Hierarchical representation of the timbral classification tree structure for 36 acoustic sounds derived from an ADTREE analysis of free classifications by 60 subjects. Numbers from 1 to 36 correspond to the 36 sounds used and numbers from 37 to 53 are nodes computed by the algorithm.

They fell in the same category even if the type of excitation was very different [e.g., see Fig. 8.11, numbers 33 (vibraphone bowed) and 34 (vibraphone struck)].

In summary, we can conclude that the notion of categories of timbre corresponds to a perceptual reality and that these categories seem in most cases to be based on the physical functioning of the different instruments. This result suggests that an implicit knowledge of the physical functioning of instruments has a strong correlation with perceptual classification of their corresponding sounds. An idea worth exploring would be to systematically define these perceptual categories within a timbre space. This space could probably be continuous in that the boundaries between categories could be fuzzy. Still, discrimination within and between those boundaries would be possible.

3.2 Categorical Perception: A Speech-Specific Phenomenon

3.2.1 Definition of the Categorical Perception Phenomenon

The *categorical phenomenon*, first described by Liberman (1957), refers to a situation where it is possible to identify and discriminate two objects belonging to two distinctive categories, but not possible to discriminate two objects belonging to the same category. A procedure generally employed to demonstrate such a phenomenon is the following: A continuum of N stimuli is constructed (N is typically equal to 10) by variation of a control parameter. This stimuli continuum is composed of two distinct end-point stimuli (e.g., two distinct phonemes) at the extremes of the continuum and different intermediate synthetic stimuli between the extremes. Listeners are asked to perform two tasks: (1) an *identification* task for them to identify each stimulus according to two contrasted categories; and

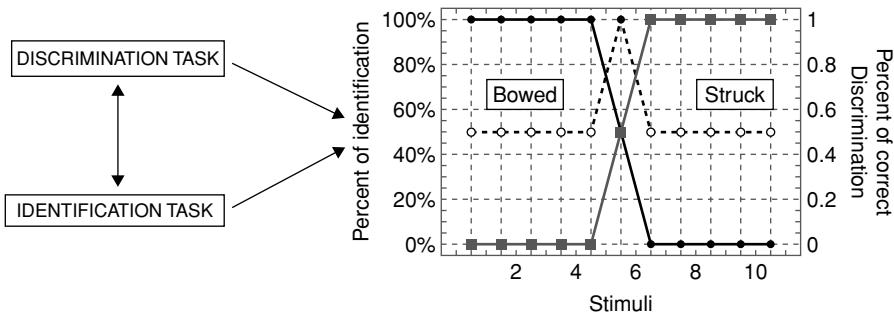


FIGURE 8.12. Theoretical discrimination and identification functions are shown for stimuli gradually changing from “bowed” (left) to “struck” (right). Stimuli 1–5 are categorically identified as “bowed,” whereas stimuli 7–11 are categorically identified as “struck.” However, percent correct discrimination between adjacent stimuli is 50% (guessing) except at the category boundary corresponding to stimulus 6, where discrimination is 100%.

(2) a *discrimination task* where stimuli are presented in pairs and subjects are asked to respond whether the stimuli are identical or not.

According to Studdert-Kennedy et al. (1970), three criteria are necessary to conclude that categorical perception exists in a continuum: (1) “peaks,” regions of high discriminability in the discrimination function; (2) “troughs,” regions where discrimination is near the chance level; and (3) a correspondence between the peaks and troughs and the shape of the identification function, with peaks occurring at the identification boundaries and troughs within each category. In other words, in contrast to continuous perception, categorical perception occurs when intracategorical discrimination is absent: Subjects discriminate two neighboring stimuli only if they (or their control parameters) are situated on either side of the boundary separating the two categories. Fig. 8.12 represents hypothetical results of such a categorical perception experiment.

3.2.2 Musical Categories: Plucking and Striking vs Bowing

It has been asserted that one of the most important differences between speech stimuli and non-speech stimuli is that the former are categorically perceived whereas the latter are not. However, it seems possible to observe this categorical perception phenomenon for non-speech stimuli. Miller et al. (1976) used noise and buzz sounds, with the onset of the noise varying from -10 to $+80$ ms with respect to the onset of the buzz. Discrimination was best when the noise led the buzz by about 16 ms, which was about the same amount of delay as the category boundary in a labeling task. Pisoni (1977) and Pastore (1976) also observed such a phenomenon for two-tone stimuli and critical-flicker fusion, respectively. Locke and Kellar (1973) and Siegel and Siegel (1977), as well as others, have observed categorical perception for musical intervals.

Cutting and Rosner (1974, 1976) used sawtooth waveforms varying in rise time from 0 to 80 ms in increments of 10 ms and found that best discrimination occurred between tones whose rise times straddled 40 ms, the position of the plucked/bowed perception category boundary between where subjects had identified rapid (0 to 30 ms) rise-time stimuli as plucked strings and slower (50 to 80 ms) rise-time stimuli as bowed strings. While discriminations between the plucked and bowed category regions were easy to make, subjects were not able to discriminate rise time differences very well within the bowed and plucked category regions. Also, Remez (1978) created a plucked-to-bowed continuum by tailoring natural tokens of musical sounds played on a bass viol. These, too, were perceived categorically. However, his continuum was a rise-time-by-amplitude-at-onset continuum rather than simply a rise-time continuum. Macmillan (1979), using analog-generated stimuli of considerably lower fundamental frequency, also found categorical perception. However, the boundaries fell at 25 ms rather than 40 ms for the discrimination and identification functions.

Cutting et al. (1976) extended their previous findings (Cutting and Rosner, 1974, 1976) by demonstrating selective adaptation effects with the same stimuli. Subjects had to categorize stimuli before and after repeated exposure to an *adaptor*, which corresponded to either a stimulus with the same spectral envelope, a stimulus with the same frequency, or a stimulus with the same spectral envelope and frequency. The boundary between the two categories shifted as expected after the exposure to the adaptor, and the greatest shift was observed when the adapting stimulus shared all dimensions with the test continuum. Remez et al. (1980) found reliable adaptation by using end-point adaptors on a plucked-bowed continuum. More recently, Pitt (1995) found such an effect on identification and reaction-time performance using a trumpet-to-piano continuum of acoustic sounds. He showed that after the exposure to adaptors corresponding to the end-points of the continuum, the categorization boundary indicated by the identification function shifted significantly, and reaction times were significantly faster for stimuli situated near the end-points of the continuum. Direct comparison of the identification results with those from previous timbre adaptation studies is not possible because different measures of adaptation magnitude were used. However, visual comparison of identification functions suggests that the trumpet-to-piano continuum produced a larger boundary shift than the pluck-to-bow continuum of Cutting et al. (1976).

Such results suggest that categorical perception of music-like sounds may be explained by a theory based on feature detection. Indeed, according to such a theory, the repetitive presentation of a stimulus belonging to a perceptual category would lead to a decrease in the rate response of the detector for other stimuli in the same category.

3.2.2.1 Are the Same Feature Detectors Used for Speech and Nonspeech Sounds?

In the light of these results, one might ask whether specific detectors involved in the processing of speech sounds and nonverbal sounds are the same. Some authors

have indeed demonstrated a similar adaptation for verbal stimuli using nonverbal stimuli as adaptors (Diehl, 1976; Kat and Samuel, 1984; Samuel, 1988; Samuel and Newport, 1979), although other authors (Remez et al., 1980) have not observed such a phenomenon. For example, Diehl (1976) showed that the spectrum of a plucked string could influence the perception of a continuum from /ba/ to /wa/, but that the spectrum of a bowed string did not influence the result. Samuel and Newport (1979) conducted this experiment with continua from /ba/ to /wa/ and from /tʃ a/ to /ʃ a/. They used four types of nonverbal adaptors: two periodic sounds (where the fundamental frequency was different from that of the verbal sounds) both imitating either a plucked or a bowed string. Results showed that periodic sounds with rapid attack times had an influence if they shared properties with the /ba/ sound but not with the /tʃ a/ sound, while the sounds with slow attack times had an effect if they shared a property with the /ʃ a/ sound but not with the /wa/ sound.

Nonetheless, results observed by Remez et al. (1980) argue against the hypothesis of common specific detectors for nonspeech and speech stimuli. In their study, the authors crossed adaptor stimuli, which could be either verbal or nonverbal, with test stimuli that were either verbal or not. Adaptor stimuli were either extreme stimuli of the two types of tested continua or difference stimuli according to their acoustical properties. Adaptor stimuli differed from the continuum neither in attack time, nor in fundamental frequency, but only in terms of their spectral envelopes. This difference gave rise to a difference in source identity. Results showed adaptation only when the type of the adaptor (e.g., verbal vs nonverbal) corresponded to that of the test stimuli; i.e., adaptation occurred for verbal test stimuli with a verbal adaptor and for nonverbal test stimuli with a nonverbal adapter, while nonverbal and verbal adaptor stimuli did not influence the verbal and non-verbal test stimuli, respectively. These results thus suggest that specific detectors involved in the processing of verbal sounds and those involved with nonverbal sounds are different in nature, confirming the idea formulated by Cutting et al. (1976) that the importance of the adaptation effect is a function of the number of auditory attributes shared by the continuum and the adaptor.

These results suggest that nonspeech stimuli could be categorically perceived and could be explained by a feature-detector theory (Cutting et al., 1976; Pitt, 1995; Remez et al., 1980). However, it is difficult to make any conclusions about existence of common feature detectors for speech and nonspeech stimuli (Diehl, 1976; Remez et al., 1980; Samuel and Newport, 1979).

3.2.2.2 Categorical Perception in Young Infants

Infants, like adults, seem to perceive nonspeech stimuli in a categorical manner. Jusczyk et al. (1977) used a high-amplitude sucking technique to explore 2-month-olds' perception of rise-time differences for the same stimuli used by Cutting and Rosner (1974, 1976). The authors observed that the sucking rate did not vary if the change was within one of the two categories, but that it was significantly higher when the change was across the two categories: "bowed" vs "plucked." More

specifically, infants seemed to perceive a difference between stimuli with 30-ms to 60-ms rise times, which corresponded to the boundaries observed by Cutting and Rosner (1974) for adults, but not for stimuli between 0 to 30 ms and 60–90 ms rise times. Like adults, infants discriminated rise-time differences between the two category boundaries but not equal differences within either category. The presence of such categorical perception in 2-month-old infants suggests that it is relatively independent of auditory experience. To account for similar results in infants for verbal stimuli, Eimas (1975) proposed the hypothesis that newborns are equipped with specific detectors which respond to relevant acoustical properties of verbal sounds. Results observed for nonverbal sounds lead to a similar interpretation. However, research on prenatal audition (Granier-Deferre & Busnel, 1981; Granier-Deferre & Lecanuet, 1987; Lecanuet et al., 1988; Lecanuet et al., 1992) has shown that newborn infants do not begin their auditory experience at birth, but actually several months before, therefore providing several months of auditory experience during which perceptual learning can take place.

3.2.2.3 *The McGurk Effect for Timbre*

McGurk and MacDonald (1976) and MacDonald and McGurk (1978) showed that the perception of an acoustic syllable could be affected by the simultaneous presentation of visual information specifying a speaker's articulatory movement of a different syllable. For example, if the auditory syllable is a /ba/ and if the subjects see a video tape of a speaker producing a /ga/, they report having heard a /da/. This /da/ syllable is an intermediate syllable, the place of articulation of which is between those of /ba/ and /ga/. This effect, called the "McGurk effect," clearly shows that visual and auditory information can be integrated by subjects, the response being a compromise between normal responses to two opposing stimuli. Moreover, Kuhl and Meltzoff (1982) observed that young infants show a preference for pairs of stimuli in which auditory and visual information are matched. Infants look longer at a mouth which presents the articulatory movement of the heard syllable than at one whose articulatory movement does not correspond to the sound. This result suggests a predisposed functional relationship between the perception and the production of language.

For nonverbal sounds, Rosenblum and Fowler (1991) observed that visual information could have an influence on auditory judgment. They showed, for example, using the McGurk paradigm, that loudness judgments of syllables or hand-clapping could be influenced by visual information. More recently, Saldana and Rosenblum (1993) observed the same type of effect for plucked and bowed string sounds. In their first experiment, they presented each sound along a continuum between a plucked and a bowed string. At each presentation of a sound, the subject had to estimate whether the sound was plucked or bowed on a continuous scale. The instructions were to use the middle of this scale if the sound was ambiguous. In one condition, the sound was presented simultaneously with a video tape showing a player plucking or bowing a string. Results showed that subjects' responses were greatly influenced by the visual information. Indeed, the identification function for

judgments based only on the auditory presentation of the sound was significantly different from that based on the audiovisual presentation. In fact, the authors observed that the identification function corresponding to the audiovisual condition shifted to the plucked response scale and inversely for the condition where the video tape presented a bowed string. However, this study did not include the opposite possibility of allowing the subjects to identify a plucked string as a bowed string when the visual information described a bowed string. The hypothesis of the authors was that the effect could be explained by the ecological theory according to which the influence of the visual information would be in direct relation with the production mode of the sound event. To test this last hypothesis they replaced the visual information by a visual presentation of the two words “plucked” and “bowed.” In this last case no effect was observed.

3.2.3 Is There a Perceptual Categorization of Timbre?

In contrast to the above discussion, some researchers argue that nonspeech sounds are not categorically perceived. Van Heuven and van den Broecke (1979) measured the variability of settings in a rise-time reproduction task. They found that the standard deviation of adjustments was an increasing linear function of rise time. They felt that the differences between their results and Cutting and Rosner's could be attributed to differences in stimulus generation techniques. Rosen and Howell (1981) synthesized a new continuum of sawtooth waves differing in linear increments of rise time, analogous to the array reported by Cutting and Rosner (1974). In order to test the hypothesis that a different generation technique could produce different results, Van Heuven and van den Broecke (1979) included a condition in which stimuli were recorded before presentation to the subjects. They did not obtain results consistent with categorical perception. Although they obtained a similar identification function, they did not observe a peak in the discrimination function. Instead, they found a discrimination function that might be predicted better on the basis of a Weber fraction for rise time. So, the method of stimulus generation and presentation was not responsible for the discrepancies between the results. Using the original tapes of Cutting and Rosner (1974) to replicate their results, they found categorical perception of these stimuli. To reconcile the difference in the two findings, they measured the original stimuli and found that the rise times differed from those reported in the Cutting and Rosner paper. Moreover, the discrepancies were such that they predicted nonlinearities in the discrimination results. Thus, they concluded that plucked and bowed music-like sounds are not categorically perceived.

According to Hary and Massaro (1982), categorical perception results do not necessarily imply categorical perception. Indeed, they showed that a bipolar continuum of increasing and decreasing onset times yielded traditional categorical results but that when only half of this continuum was tested, the same sounds were perceived continuously. On the other hand, contradictory results for the identification function have also been found. Smurzynski (1985) asked subjects to learn envelopes by rise-time value and later to identify them. Analysis of responses

showed that trained subjects did not classify a continuum of sawtooth waveforms varying in rise time into two sharply defined categories, but were able to resolve rise-time values with much greater accuracy than would be achieved by simply dividing the continuum into two categories such as "plucked" and "bowed." To conclude, Cutting (1982) found that stimuli with equal linear increments of rise time were not categorically perceived, but stimuli with logarithmic increments of rise time were categorically perceived. The stimuli and the results observed by Rosen and Howell (1981) are shown in Figs. 8.13a and 8.13b, and the results found by Cutting and Rosner (1974) are represented in Fig. 8.13c.

Donnadieu and McAdams (1996), Donnadieu et al. (1996), and Donnadieu (1997) confirmed the idea of noncategorical perception of rise time using two continua of attack time constructed on the basis of two original vibraphone sounds. One sound resulted from a vibraphone struck by a hard mallet, and the other came from the vibraphone bowed with a violin bow on its edge. (Both sounds were taken from a McGill University Master Samples compact disk.) Most studies on categorical perception of rise time have utilized synthesized sounds. We chose to use acoustic sounds even though in this case the definition of attack time is somewhat arbitrary. For both sounds, 10 stimuli were constructed in which only the rise time of the amplitude-vs-time envelope differed. Utilizing a phase-vocoder analysis/resynthesis program (Beauchamp, 1993), a "struck" continuum was constructed by successively modifying the rise time of the struck vibraphone sound so that it started at 0.13 s, increasing from step to step by a factor of 1.29, and ending at 1.30 s. A corresponding "bowed" continuum was constructed by decreasing the rise time of a bowed vibraphone sound, starting at 0.35 s, decreasing by factors of 0.76, and ending at 0.03 s. For each continuum, the rapid-onset stimuli tended to sound like a struck instrument and the slower onset stimuli like a bowed instrument.

Subjects were asked to perform three tasks: (1) discriminate pairs of stimuli along the two continua, (2) identify (or categorize) them as one of the end-points ("struck" or "bowed"), and (3) judge the perceptual dissimilarity of the stimulus pairs. The stimulus pairs were separated by two steps on each continuum. The three tasks were performed separately for each of the two continua ("reduced contexts") and for the combined set of these two continua stimuli ("extended context"). The discrimination task was of type AX. For this task, the subjects heard each stimulus pair with a 1-s interstimulus interval (ISI) and were asked to judge if the stimuli were "same" or "different." Raw discrimination scores were adjusted by subtracting "false alarm" scores (responding "different" when the sounds were identical). For the identification task, subjects were asked to label the stimuli as "struck" or "bowed." For the dissimilarity task, subjects judged the degree of dissimilarity (on a scale varying from very similar to very dissimilar) of stimuli pairs (1-3, 2-4, etc.). A scale was presented on a computer screen and listeners had to push the button at the desired position. For each task, subjects participated in three sessions corresponding to the three contexts: (1) stimuli from the "bowed" continuum ("reduced context"), (2) stimuli from the "struck" continuum ("reduced context"), and (3) stimuli from the union of these two continua ("extended context"). Subjects completed all three types of tasks (discrimination, identification, and dissimilarity)

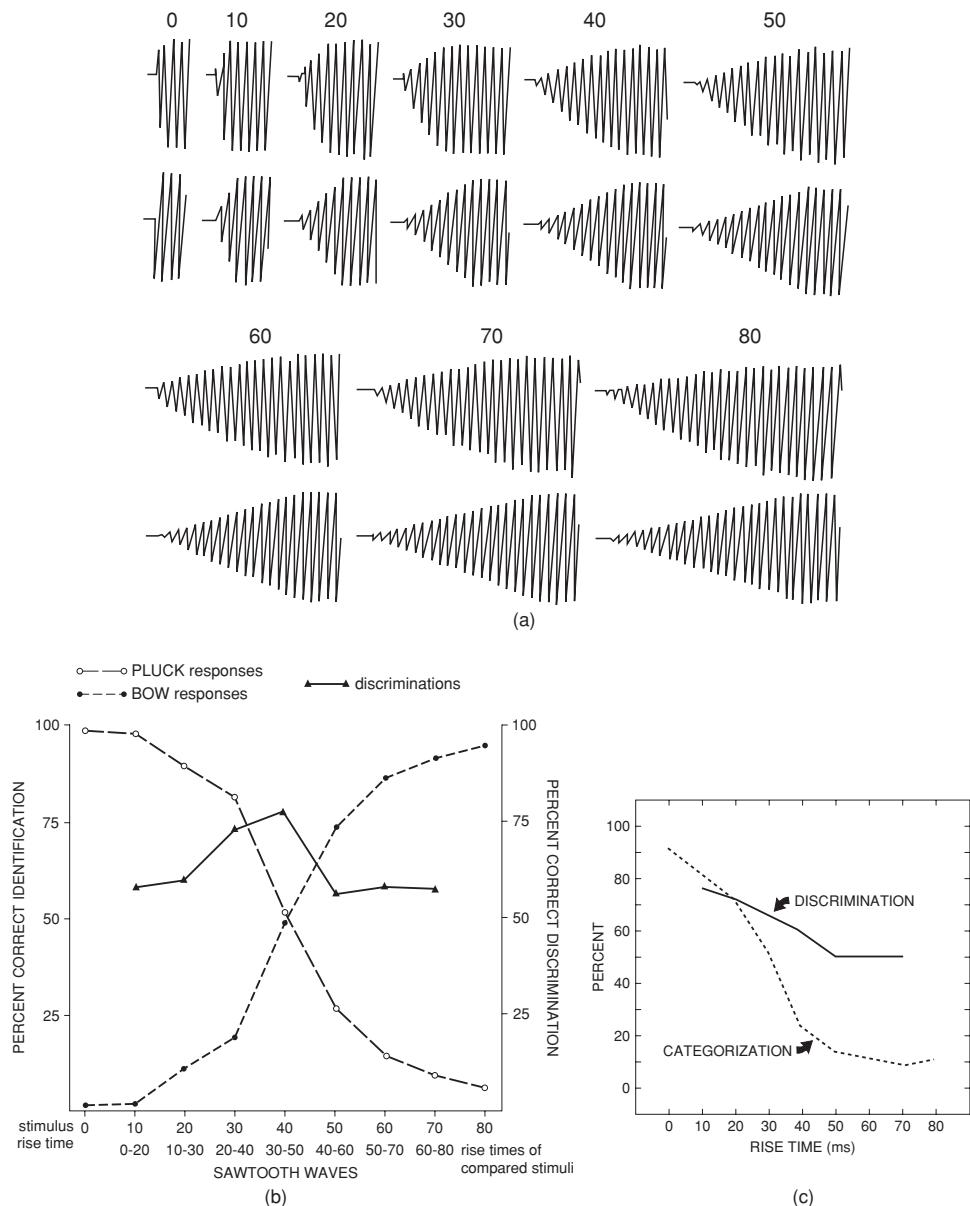


FIGURE 8.13. Categorical perception of a rise-time continuum. (a) Oscillograms for a nine sawtooth-wave stimuli continuum used by Cutting and Rosner (1974) and by Rosen and Howell (1981). (b) Identification and discrimination functions observed by Cutting and Rosner (1974). (c) Results observed by Rosen and Howell (1981) on the Cutting and Rosner stimuli. [From Cutting and Rosner (1974), Fig. 2 and Rosen and Howell (1981), Figs. 3 and 4, adapted by permission of the Psychometric Society.]

in each session. From the results, it was clear that although subjects on average gradually changed their classification from bowed vibraphone to struck vibraphone and vice versa along the two continua, discrimination performance was fairly constant along the continua. Figures 8.14 and 8.15 give the results of the discrimination and identification experiments, respectively. Note that two sets of data were extracted from the extended context session: one which focused on the listener's ability to correctly identify or discriminate the "struck" continuum data in the presence of the "bowed" continuum data, and vice versa.

These results were not consistent with the numerous studies that have shown categorical perception of rise time (Cutting, 1982; Cutting and Rosner, 1974, 1976; Cutting et al., 1976; Jusczyk et al., 1977; Macmillan, 1979; Miller et al., 1976; Pitt, 1995; Remez, 1978; Remez et al., 1980). However, Rosen and Howell (1981) observed that discrimination performance for equally spaced stimuli is always best for shortest rise times. Our results were partially consistent with these results because, although we did not observe a categorical perception of the rise time of acoustic struck or bowed vibraphones, we did observe that discrimination performance was relatively constant across the two continua tested. The difference in our results could be due to the fact that our continua were constructed by logarithmic rather than linear rise time increments. We chose logarithmic increments first because Cutting's last results showed that only in this case is categorical perception observed for the attack time of nonspeech sounds and second because the first dimension of timbre is generally more correlated with a measure of the logarithm of the attack time than with linear rise time of the temporal envelope (McAdams et al., 1995). On the other hand, category boundaries observed in previous studies were very different from our category boundaries. This difference could be due to the fact that our stimuli corresponded to resynthesized transformations of recorded acoustic sounds. Moreover, in this experiment we used a bowed bar (a vibraphone) rather than a bowed string. The modes of vibration of a metal bar are very different than that of a string and take more time to be set into vibration when bowed.

We also noted during the construction of the stimuli that there was a large difference between the attack times of the two continua endpoints. Indeed, to induce a perception of the bowed vibraphone we had to considerably augment the rise time of the temporal envelope of the original struck vibraphone beyond that of the rise time of the original bowed vibraphone. Moreover, the boundary between the "struck" and "bowed" categories was quite different for the struck and bowed continua. This calls into question the definition of the attack as being characterized uniquely by rise time and suggests that other factors in addition to the logarithm of the attack time of the sounds contributed to the identity of the type of excitation. Indeed, the attack epochs of these sounds may include many characteristics, such as the presence of a high-frequency component or the presence of noise produced by the contact between mallet and bar. These characteristics could be used by the auditory system to identify an instrument's resonator (e.g., as a bar or a string) or the type of excitors used. These aspects would correspond to the structural invariants of the ecological approach.

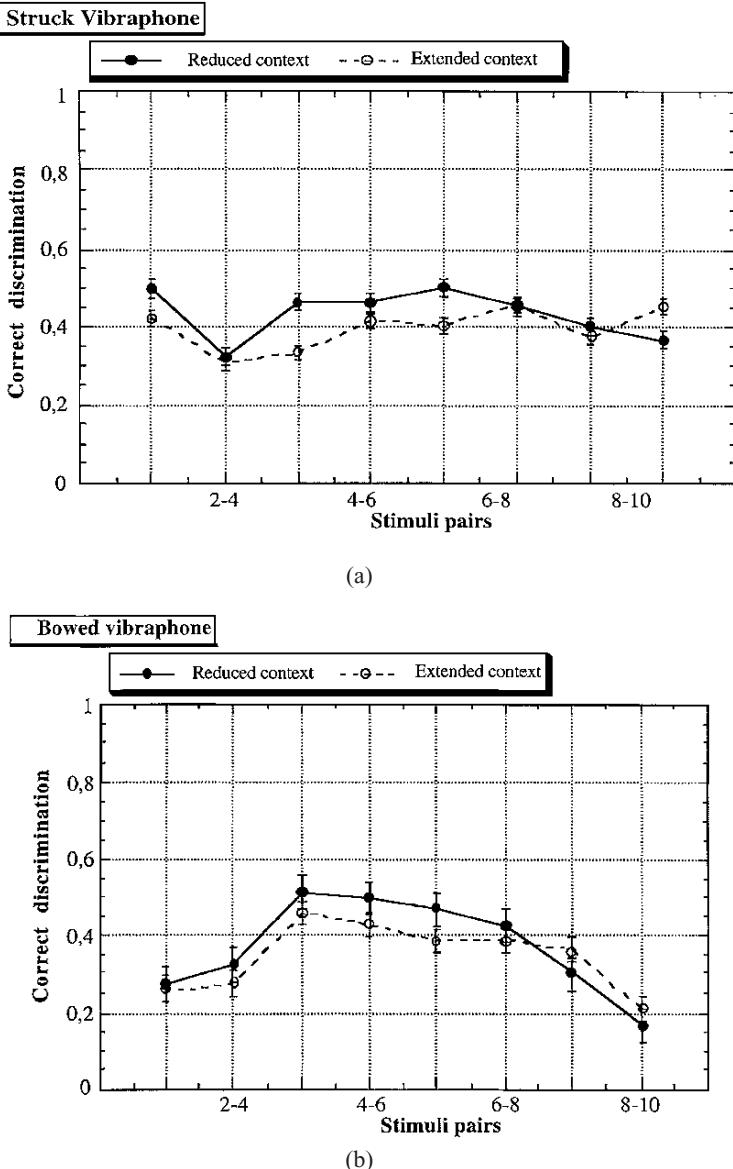


FIGURE 8.14. Mean discrimination functions for “struck” and “bowed” vibraphone continua stimuli presented in “reduced context” (each continuum alone) and “extended context” (continua stimuli combined). (a) Discrimination of stimuli pairs (1 and 3, 2 and 4, etc.) along continua of gradually increasing rise time of struck vibraphone sounds. (b) Discrimination of stimuli pairs along continua of gradually decreasing rise time of bowed vibraphone sounds. 0% discrimination corresponds to the guessing level. Note that in the “extended context” responses to the same stimuli are scored as in the “reduced context,” but in the former case the stimuli are intermixed with stimuli from the other continuum.

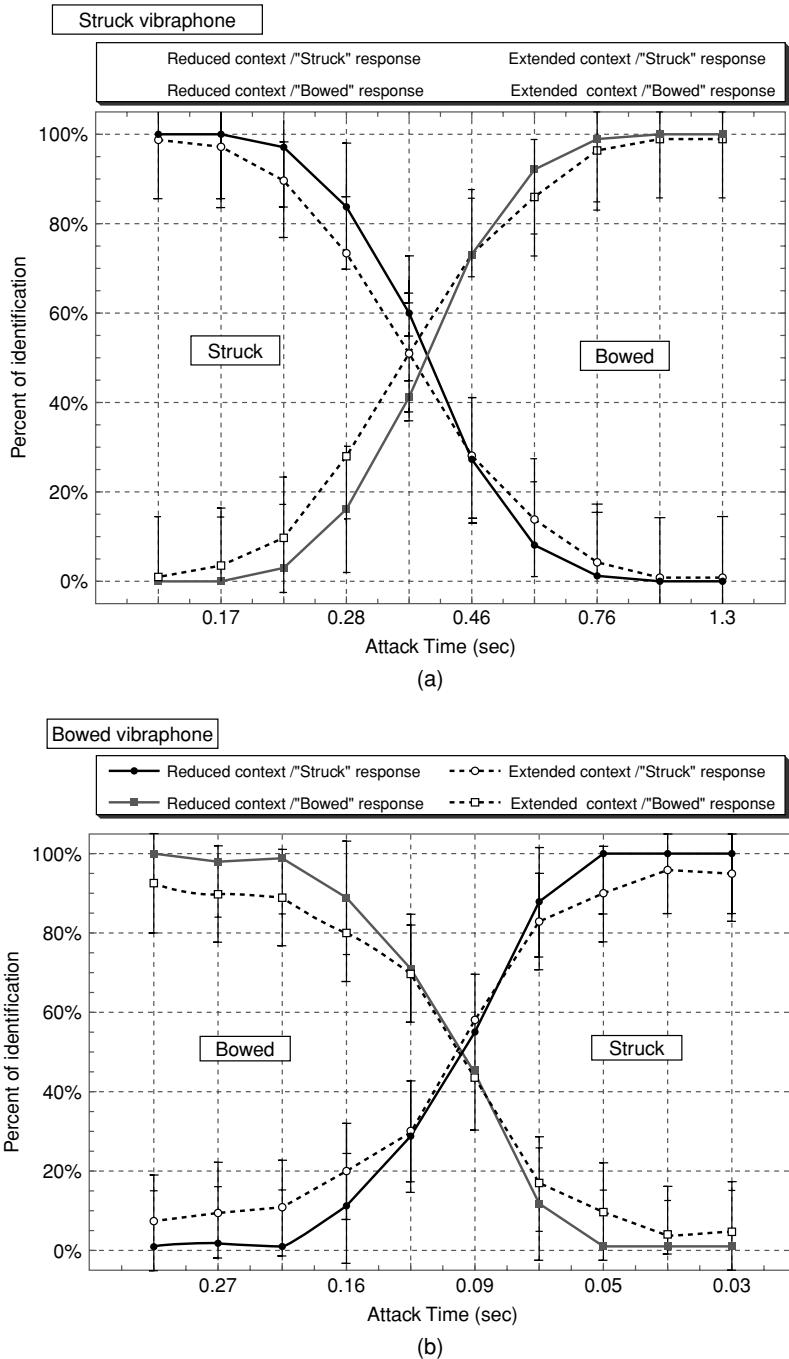


FIGURE 8.15. Mean identification functions for identifying the same stimuli as given in Fig. 8.14 as either "struck" or "bowed" vs attack time of the stimuli for the "reduced" and "extended" contexts. (a) Identification of struck vibraphone continua stimuli. (b) Identification of bowed vibraphone continua stimuli.

The fact that attack quality of sounds seems to be perceived along a continuum suggests that the mental structure of musical timbre could be represented in terms of several perceptually continuous dimensions. However, the fact that we can classify the timbres of vibraphone as struck or bowed, as the identification function shows, suggests that this type of representation could be influenced by a higher-level process of categorization.

4 Conclusions

This chapter describes studies on the perception of timbre of complex tones. Two approaches were presented corresponding to two ways of understanding the perceptual representation of musical timbre.

The first approach describes different perceptual dimensions of timbre in terms of abstract properties. It seeks to determine which acoustical parameters of the complex signal are processed by the auditory system and in the end contribute to the perception of timbre. Multidimensional scaling has been particularly fruitful for this type of study. Results suggest that essentially three dimensions can be used to describe the timbres of a given set of musical complex tones. The physical correlates of the different dimensions seem to be well identified, corresponding to spectral, temporal, and spectrotemporal aspects of the acoustical signal (Grey, 1975, 1977; Krumhansl, 1989; Miller and Carterette, 1975; Plomp, 1970, 1976; Samson et al., 1996; Wedin and Goude, 1972; Wessel, 1979). However, if the contribution of spectral aspects of the sounds is clear today, the influence of temporal aspects is less clear. Indeed, some multidimensional studies call into question the perceptual importance of such temporal aspects in the perception of timbre. Wedin and Goude's (1972), Miller and Carterette's (1975), and Iverson and Krumhansl's (1991) results suggest a predominance of spectral factors. Moreover, it seems that the perceptual salience of temporal judgments depends largely on context, and, in particular, on the musical context in which the sounds are presented (Grey, 1978; Kendall, 1986). Nevertheless, for the case where sounds are presented in isolation, we cannot doubt the importance of such factors. Results from experiments based on deletion of parts of sounds (Berger, 1964; Saldanha and Corso, 1964), those based on spectral modifications associated with discrimination tasks (Charbonneau, 1981; Grey and Moorer, 1977), and observations from some multidimensional studies (Grey, 1977; Krumhansl, 1989; Wessel, 1979) support the importance of the influence of temporal aspects on timbre perception.

Results from multidimensional studies suggest a continuous representation of the timbre of complex sounds. In the same way, studies based on sound modifications (Saldanha and Corso, 1964; Grey and Moorer, 1977; Grey, 1978; Kendall, 1986) have shown that the capacity of listeners to identify sounds diminishes when acoustical parameters are manipulated. This degradation of identification efficacy may depend on whether an auditory stimulus varies continuously along dimensions related to specific acoustical parameters and whether the categories involved have fuzzy boundaries. Moreover, it appears from studies of timbral analogies

(Ehresman and Wessel, 1978; McAdams and Cunibile, 1992) and studies involving modification of sounds to examine their consequences on timbre space (Grey and Gordon, 1978; Wessel, 1979, 1983), that intermediate areas of a timbre space can be filled in and that regular perceptual transitions based on a few physical dimensions are possible. In the same way, studies on the role played by timbre in auditory organization (Bey and McAdams, 2003; Gregory, 1994; Hartmann and Johnson, 1991; Iverson, 1993; McAdams and Bregman, 1979; Wessel, 1979) indicate that the auditory-stream-segregation process is based on the same perceptual attributes as those used by listeners when they were asked to do dissimilarity judgments between different timbres. These results suggest that MDS timbre spaces can account for the similarity relations between different timbres. Fusion and segregation processes could be based on the metric distance separating timbres in a geometric space, and the perceptual dimensions of timbres may be related to the representation upon which such processes operate. Finally, the development of verbal attributes of timbre (Faure et al., 1996; Samoylenko et al., 1996) allow us to complete our knowledge concerning timbre space and to establish the relation between perceptual representations and semantic representations.

The second approach is related to ecological considerations (Gibson, 1966) and the notion of perceptual categories of timbre. According to this approach, timbre perception is a direct function of the physical properties of the sound source. In this case, the aim of various studies (e.g., Donnadieu, 1997; Lakatos, 2000) has been to describe perceptually relevant physical properties of sound objects and their relative roles in the perception of musical instrument sounds, in addition to the major roles played by perceptual attributes such as pitch salience, spectral envelope, and roughness. The idea is that the auditory system can code the timbre of complex sounds in terms of the details of physical source sound production. Indeed, this research suggests that the relation between timbre and physical causality could be a fundamental aspect of our perception and of the categorical organization of the perceptual structure of timbre. However, studies that investigated whether the attack quality of complex sounds is categorically perceived gave contradictory results. In summary, these results suggest that attack qualities can be continuously perceived and support a model of perceptually continuous timbre space, but they do not exclude the possibility that higher-level classification organizations could be present and that timbre categories could be organized in such a timbre space.

References

- American National Standards Institute (1973). *Psychoacoustical Terminology, S3.20-1973* (American National Standards Institute, New York).
- Barthélemy, J.-P. and Guénoche, A. (1988). *Arbres et les représentation des proximités* [Trees and proximity representations]. (Masson, Paris).
- Beauchamp, J. W. (1993). "Unix workstation software for analysis, graphics, modifications, and synthesis of musical sounds," *94th Convention of the Audio Engineering Society*, Berlin, (Audio Eng. Soc., New York), Audio Eng. Soc. Preprint 3479.

- Berger, K. W. (1964). "Some factors in the recognition of timbre," *J. Acoust. Soc. Am.* **36**(10), 1888–1891.
- Bey, C. and McAdams, S. (2003). "Postrecognition of interleaved melodies as an indirect measure of auditory stream formation," *J. Exp. Psychol.: Human Percept. Perform.* **29**, 267–279.
- Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).
- Bregman, A. S., Liao, C., and Levitan, R. (1990). "Auditory grouping based on fundamental frequency and formant peak frequency," *Canadian J. Psychol.* **44**, 400–413.
- Cadoz, C. (1991). "Timbre et causalité," in *Le timbre: Métaphore pour la composition*, J.-B. Barrière, ed. (Christian Bourgois, Paris), pp. 17–46.
- Carroll, J. D. and Chang, J. J. (1970). "Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition," *Psychometrika* **35**, 283–319.
- Charbonneau, G. R. (1981). "Timbre and the perceptual effects of three types of data reduction," *Computer Music J.* **5**(2), 10–19.
- Cutting, J. E. and Rosner, B. S. (1974). "Categories and boundaries in speech and music," *Perception and Psychophysics* **16**(3), 564–570.
- Cutting, J. E. and Rosner, B. S. (1976). "Discrimination functions predicted from categories in speech and music," *Perception and Psychophysics* **20**, 87–88.
- Cutting, J. E., Rosner, B. S., and Foard, C. F. (1976). "Perceptual categories for musiclike sounds: Implications for theories of speech perception," *Quarterly J. Exp. Psychol.* **28**, 361–378.
- Cutting, J. E. (1982). "Plucks and bows are categorically perceived, sometimes," *Perception and Psychophysics* **31**, 462–476.
- De Brujin, A. (1978). "Timbre classification of complex tones," *Acustica* **40**, 108–114.
- Dictionnaire de l'Academie Française, 1835.
- Diehl, R. (1976). "Feature analyzers for the phonetic dimension stop vs. continuant," *Perception and Psychophysics* **19**, 267–272.
- Donnadieu, S. and McAdams, S. (1996). "Effect of context change on dissimilarity, discrimination and categorization task on timbre perception," in *Proc. 12th Annual Meeting of the Int. Society for Psychophysics*, Padua, Italy, S. Masin, ed. (Univ. of Padua, Padua, Italy), pp. 239–244.
- Donnadieu, S., McAdams, S., and Winsberg, S. (1996). "Categorization, discrimination and context effects in the perception of natural and interpolated timbres," in *Proc. 4th Int. Conf. on Music Perception and Cognition* (ICMPC4), Montréal, Canada, B. Pennycook and E. Costa-Gomi, eds. (McGill University, Montréal), pp. 73–78.
- Donnadieu, S. (1997). "Représentation mental du timbre des sons complexes et effets de contexte [Mental representation of timbre of complex sounds and the effects of context]," unpublished doctoral dissertation, Université Paris V.
- Ehresman, D. and Wessel, D. (1978). *Perception of Timbre Analogies*, IRCAM Technical Report 13/78 (Centre Georges Pompidou, Paris).
- Eimas, P. D. (1975). "Auditory and linguistic processing of cues for place of articulation by infants," *Perception and Psychophysics* **16**, 513–521.
- Faure, A., McAdams, S., and Nosulenka, V. (1996). "Verbal correlates of perceptual dimensions of timbre," in *Proc. 4th Int. Conf. on Music Perception and Cognition* (ICMPC4), B. Pennycook and E. Costa-Gomi, eds., McGill University, Montreal, Canada, pp. 79–84.

- Faure, A. (2000). "Des sons aux mots: Comment parle-t-on du timbre musical [From Sounds to Words: How Does One Speak of Musical Timbre?]", unpublished doctoral dissertation, Ecoles des Hautes Etudes en Sciences Sociales, Paris.
- George, W. H. (1954). "A sound reversal technique applied to the study of tone quality," *Acustica* **4**, 224–225.
- Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems* (Houghton-Mifflin, Boston).
- Gibson, J. J. (1979). *The Ecological Approach to Visual Perception* (Houghton-Mifflin, Boston).
- Granier-Deferre, C. and Busnel, M-C. (1981). "L'audition prénatale [Prenatal Hearing]", in *L'aube des sens, Cahiers du Nouveau-né [The dawn of the senses, Newborn Journal]*, E. Herbinet and M-C. Busnel, eds. (Stock, Paris), pp. 147–175.
- Granier-Deferre, C. and Lecanuet, J-P. (1987). "Influence de stimulations auditives précocees sur la maturation anatomique et fonctionnel du système auditif [Influence of early auditory stimulation on anatomical and functional maturation of the auditory system]," *Progrès en Néonatalogie* **7**, 236–249.
- Gregory, A. H. (1994). "Timbre and auditory streaming," *Music Perception* **12**(2), 161–174.
- Grey, J. M. (1975). "An Exploration of Musical Timbre," unpublished doctoral dissertation, Stanford University, Stanford, CA. Also available as Stanford Dept. of Music Report STAN-M-2.
- Grey, J. M. (1977). "Multidimensional perceptual scaling of musical timbres," *J. Acoust. Soc. Am.* **61**(5), 1270–1277.
- Grey, J. M. and Moorer, J. A. (1977). "Perceptual evaluations of synthesized musical instrument tones," *J. Acoust. Soc. Am.* **62**(2), 454–462.
- Grey, J. M. and Gordon, J. W. (1978). "Perceptual effects of spectral modifications on musical timbres," *J. Acoust. Soc. Am.* **63**(5), 1493–1500.
- Grey, J. M. (1978). "Timbre discrimination in musical patterns," *J. Acoust. Soc. Am.* **64**(2), 467–472.
- Guyot, F. (1992). "Etude de la pertinence de deux critères acoustiques pour caractériser la sonorité des sons à spectre réduit [Study of the relevance of two acoustic criteria for characterizing the sonorities of simplified sounds]," unpublished DEA thesis, Université du Maine, France.
- Hartmann, W. M. and Johnson, D. (1991). "Stream segregation and peripheral channeling," *Music Perception* **9**(2), 155–183.
- Hary, J. M. and Massaro, D. W. (1982). "Categorical results do not imply categorical perception," *Perception and Psychophysics* **32**(5), 409–418.
- Iverson, P. and Krumhansl, C. L. (1991). "Measuring similarity of musical timbres," *J. Acoust. Soc. Am.* **89**(4), Pt. 2, 1988 (abstract).
- Iverson, P. (1993). "Auditory segregation by musical timbre," doctoral dissertation, Cornell University, Ithaca, NY. *Dissertation Abstracts International*, 54 (4-B), 2249.
- Iverson, P. and Krumhansl, C. L. (1993). "Isolating the dynamic attributes of musical timbre," *J. Acoust. Soc. Am.* **94**(5), 2595–2603.
- Johnson, S. C. (1967). "Hierarchical clustering schemes," *Psychometrika* **32**, 241–254.
- Jusczyk, P. W., Rosner, B. S., Cutting, J., Foard, C. F., and Smith, L. B. (1977). "Categorical perception of nonspeech sounds by 2-month-old infants," *Perception and Psychophysics* **21**(1), 50–54.
- Kat, D. and Samuel, A. G. (1984). "More adaptation of speech by nonspeech," *J. Exp. Psych: Human Percept. Perform.* **10**, 512–525.

- Kendall, R. A. (1986). "The role of acoustic signal partitions in listener categorization of musical phrases," *Music Perception* **4**, 185–214.
- Krimphoff, J. (1993). "Analyse acoustique et perception du timbre," unpublished DEA thesis, Université du Maine, Le Mans, France.
- Krimphoff, J., McAdams, S., and Winsberg, S. (1994). "Caractérisation du timbre des sons complexes. II : Analyses acoustiques et quantification psychophysique. [Characterization of the timbre of complex sounds. 2. Acoustic analysis and psychophysical quantification]," *J. de Phys.* **4**(C5), 625–628.
- Krumhansl, C. L. (1989). "Why is musical timbre so hard to understand?" in *Structure and Perception of Electroacoustic Sound and Music: Proc. Marcus Wallenberg Symposium*, Lund, Sweden, August, 1988, S. Nielzén and O. Olsson, eds. (Excerpta Medica, Amsterdam), pp. 43–53.
- Kuhl, P. K. and Meltzoff, A. N. (1982). "The bimodal perception of speech in infancy," *Science* **218**, 1138–1144.
- Lakatos, S. (2000). "A common perceptual space for harmonic and percussive timbres," *Perception and Psychophysics* **62**(7), 1426–1439.
- Lecanuet, J-P., Granier-Deferre, C., and Busnel, M-C. (1988). "Fetal cardiac and motor responses to octave-band noises as a function of central frequency, intensity and heart rate variability," *Early Human Development* **18**, 81–93.
- Lecanuet, J-P., Granier-Deferre, C., Jacquet, A-Y., and Busnel, M-C. (1992). "Decelerative cardiac responsiveness to acoustical stimulation in the near-term fetus," *Quarterly J. Exp. Psychol.* **44B**, 279–303.
- Liberman, A. M. (1957). "Some results of research on speech perception," *J. Acoust. Soc. Am.* **29**, 117–123.
- Lichte, W. H. (1941). "Attributes of complex tones," *J. Exp. Psychol.* **28**, 455–480.
- Lindsay, P. H. and Norman, D. A. (1977). *Human Information Processing: An Introduction to Psychology*, 2nd ed. (Academic Press, New York).
- Locke, S. and Kellar, L. (1973). "Categorical perception in a nonlinguistic mode," *Cortex* **9**(4), 355–369.
- MacDonald, J. and McGurk, H. (1978). "Visual influences on speech perception processes," *Perception and Psychophysics* **24**, 253–257.
- Macmillan, N. A. (1979). "Categorical perception of musical sounds: The psychophysics of plucks and bows," *Bull. Psychonomic Soc.* **14**, 241 (abstract).
- Manoury, P. (1991). "Les limites de la notion de 'timbre,'" in *Le timbre: Métaphore pour la composition*, J.-B. Barriere, ed. (Christian Bourgois, Paris), pp. 293–299.
- Mathews, M. V., Miller, J. E., Pierce, J. R., and Tenney, J. (1965). "Computer study of violin tones," *J. Acoust. Soc. Am.* **38**, p. 912 (abstract).
- McAdams, S. and Bregman, A. (1979). "Hearing musical streams," *Computer Music J.* **3**(4), 26–43.
- McAdams, S. and Cunibile, J.-C. (1992). "Perception of timbral analogies," *Philosophical Transactions of the Royal Society, London, series B*, **336**, 383–389.
- McAdams, S. (1993). "Recognition of sound sources and events," in *Thinking in Sound: The Cognitive Psychology of Human Audition*, S. McAdams and E. Bigand, eds. (Oxford University Press, Oxford), pp. 146–198.
- McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). "Perceptual scaling of synthesized musical timbres : Common dimensions, specificities, and latent subject classes," *Psychol. Res.* **58**, 177–192.

- McAdams, S., Beauchamp, J. W., and Meneguzzi, S. (1999). "Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters," *J. Acoust. Soc. Am.* **105**(2), 882–897.
- McGurk, H. and MacDonald, J. (1976). "Hearing lips and seeing voices," *Nature* **264**, 746–748.
- Miller, G. A. and Heise, G. A. (1950). "The trill threshold," *J. Acoust. Soc. Am.* **22**, 637–638.
- Miller, J. R. and Carterette, E. C. (1975). "Perceptual space for musical structures," *J. Acoust. Soc. Am.* **58**(3), 711–720.
- Miller, J. D., Wier, C. C., Pastore, R. E., Kelly, W. J., and Dooling, R. J. (1976). "Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception," *J. Acoust. Soc. Am.* **60**, 410–417.
- Opolko, F. and Wapnick, J. (1987). *McGill University master samples* [CD-ROM] (McGill University, Montreal).
- Pastore, R. E. (1976). "Categorical perception: A critical re-evaluation," in *Hearing and Davis: Essays Honoring Hallowell Davis (contributed by present and former colleagues on the occasion of his 80th birthday)*, S. K. Hirsh, D. H. Eldredge, I. J. Hirsh, and S. R. Silverman, eds. (Washington University Press, St. Louis), pp. 253–264.
- Pisoni, D. B. (1977). "Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops," *J. Acoust. Soc. Am.* **61**, 1352–1361.
- Pitt, M. A. (1995). "Evidence for a central representation of instrument timbre," *Perception and Psychophysics* **57**(1), 43–55.
- Plomp R. (1970). "Timbre as a multidimensional attribute of complex tones," in *Frequency Analysis and Periodicity Detection in Hearing*, R. R. Plomp and G. F. Smoorenburg, eds. (Sijthoff, Leiden), pp. 397–414.
- Plomp, R. (1976). "Timbre of complex tones," in *Aspects of Tone Sensation: A Psychophysical Study*, R. Plomp, ed. (Academic Press, London), pp. 85–110.
- Preis, A. (1984). "An attempt to describe the parameters determining the timbre of steady-state harmonic complex tones," *Acustica* **55**(1), 1–13.
- Remez, R. E. (1978). "An hypothesis of event-sensitivity in the perception of speech and bass violins." *Dissertation Abstracts International*, **39** (11-B), 5618-B (University Microfilms No. 7911404).
- Remez, R. E., Cutting, J. E., and Studdert-Kennedy, M. (1980). "Cross-series adaptation using song and string," *Perception and Psychophysics* **27**, 524–530.
- Risset, J.-C. and Mathews, M. V. (1969). "Analysis of musical-instrument tones," *Physics Today* **22**(2), 23–30.
- Risset, J-C. and Wessel, D. (1982). "Exploration of timbre by analysis and synthesis," in *The Psychology of Music*, D. Deutsch, ed. (Academic Press, New York), pp. 25–58.
- Rosch, E. H. (1973a). "Natural categories," *Cognitive Psychology* **4**, 328–350.
- Rosch, E. H. (1973b). "On the internal structure of perceptual and semantic categories," in *Cognitive Development and the Acquisition of Language*, T. E. Moore, ed. (Academic Press, New York), pp. 111–144.
- Rosen, S. M. and Howell, P. (1981). "Plucks and bows are not categorically perceived," *Perception and Psychophysics* **30**(2), 156–168.
- Rosenblum, L. D. and Fowler, C. A. (1991). "Audiovisual investigation of the loudness-effort effect for speech and nonspeech events," *J. Exp. Psychol.: Human Percept. Perform.* **17**, 976–985.

- Rumelhart, D. E. and Abrahamson, A. A. (1973). "A model for analogical reasoning," *Cognitive Psych.* **5**, 1–28.
- Saldana, H. M. and Rosenblum, L. D. (1993). "Visual influences on auditory pluck and bow judgments," *Perception and Psychophysics* **54**(3), 406–416.
- Saldanha, E. L. and Corso, J. F. (1964). "Timbre cues and the identification of musical instruments," *J. Acoust. Soc. Am.* **36**, 2021–2026.
- Samoylenko, E., McAdams, S., and Nosulenka, V. (1996). "Systematic analysis of verbalizations produced in comparing musical timbres," *Intern. J. Psychol.* **31**, 255–278.
- Samson, S., Zatorre, R. J., and Ramsay, J. O. (1996). "Multidimensional scaling of synthetic musical timbre: Perception of spectral and temporal characteristics," *Canadian J. Psychol.* **51**, 307–315.
- Samuel, A. G. and Newport, E. L. (1979). "Adaptation of speech by nonspeech: Evidence for complex acoustic cue detectors," *J. Exp. Psychol.: Human Perception Perform.* **5**, 563–578.
- Samuel, A. G. (1988). "Central and peripheral representation of whispered and voiced speech," *J. Exp. Psychol.: Human Percept. Perform.* **14**, 379–388.
- Schaeffer, P. (1966). *Traité des objets musicaux* [Treatise on musical objects] (Seuil, Paris).
- Serafini, S. (1993). "Timbre Perception of Cultural Insiders: A Case Study with Javanese Gamelan Instruments," unpublished masters thesis, University of British Columbia, Vancouver, Canada.
- Schoenberg, A. (1911). *Harmonielehre* [Harmony] (Universal, Leipzig/Vienna) [French translation (1983), Lattes, Paris].
- Shepard, R. N. and Arabie, P. (1979). "Additive clustering: Representation of similarity as combinations of discrete overlapping properties," *Psychol. Rev.* **86**, 87–123.
- Shepard, R. N. (1982). "Structural representations of musical pitch," in *The Psychology of Music*, D. Deutsch, ed. (Academic Press, New York), pp. 343–390.
- Siegel, J. A. and Siegel, W. (1977). "Categorical perception of tonal intervals: Musicians can't tell sharp from flat," *Perception and Psychophysics* **21**, 399–407.
- Singh, P. G. (1987). "Perceptual organization of complex-tone sequences: A tradeoff between pitch and timbre?" *J. Acoust. Soc. Am.* **82**(3), 886–899.
- Smurzynski, J. (1985). "Noncategorical identification of rise time," *Perception and Psychophysics* **38**(6), 540–542.
- Solomon, L. N. (1959). "Search for physical correlates to psychological dimensions of sounds," *J. Acoust. Soc. Am.* **31**, 492–497.
- Strong, W. and Clark, M. (1967a). "Synthesis of wind-instrument tones," *J. Acoust. Soc. Am.* **41**, 39–52.
- Strong, W. and Clark, M. (1967b). "Perturbations of synthetic orchestral wind-instrument tones," *J. Acoust. Soc. Am.* **41**, 277–85.
- Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., and Cooper, F. S. (1970). "Motor theory of speech perception: A reply to Lane's critical review," *Psychol. Rev.* **77**, 234–249.
- Terhardt, E. (1974). "On the perception of periodic sound fluctuations (roughness)," *Acustica* **30**, 201–213.
- Tversky, A. (1977). "Features of similarity," *Psychol. Rev.* **84**, 327–352.
- Van Heuven, V. J. J. P. and van den Broecke, J. P. R. (1979). "Auditory discrimination of rise and decay time in tone and noise bursts," *J. Acoust. Soc. Am.* **66**, 1308–1315.

- Van Noorden, L. P. A. S. (1975). "Temporal Coherence in the Perception of Tone Sequences," unpublished doctoral dissertation, Eindhoven Univ. of Technology, Eindhoven, Pays-Bas, Germany.
- Vogel, A. (1974). "Roughness and its relation to the time-pattern of psychoacoustical excitation," in *Facts and Models in Hearing*, E. Zwicker and E. Terhardt, eds. (Springer-Verlag, Berlin), pp. 241–250.
- von Bismarck, G. (1974). "Sharpness as an attribute of the timbre of steady sounds," *Acustica* **30**, 159–172.
- Wedin, L. and Goude, G. (1972). "Dimension analysis of the perception of instrumental timbre," *Scandinavian J. Psychol.* **13**, 228–240.
- Wessel, D. L. (1979). "Timbre space as a musical control structure," *Computer Music J.* **3**(2), 45–52.
- Wessel, D. L. (1983). *Le concept de recherche en musique*, IRCAM, Paris, Communication.
- Wessel, D., Bristow, D., and Settel, Z. (1987). "Control of phrasing and articulation in synthesis," in *Proc. 1987 Int. Computer Music Conf.*, Urbana, IL (Computer Music Assoc., San Francisco), pp. 108–116.
- Winsberg, S. and Carroll, J. D. (1989). "A quasi-nonmetric method for multidimensional scaling of multiway data via a restricted case of an extended INDSCAL model," in *Multiway Data Analysis*, R. Coppi and S. Bolasco, eds. (North-Holland, Amsterdam), pp. 405–414.

Index

- acoustic signal models
 - ACT, 254–256
 - deterministic sine model, 128
 - multiresolution sinusoidal model, 152–153
 - noise modeling, 148, 150–151, 163
 - Bark-band noise, 149, 164–165, 169
 - Bark-band quantization, 165–166
 - third-Bark bands, 157
 - parametric models, 145
 - physical models, 149, 177
 - sines-plus-noise model, 148
 - sines-plus-transients-plus-noise model, 145–174
 - sinusoidal models, 145–146, 148, 176
 - source-filter model, 178–185, 194
 - stochastic noise model, 128
- acoustic signal properties, 1, 286, 290, 304, 305
 - amplitude, 2–10, 20, 138
 - attack, 1, 39, 92, 309, 312
 - fundamental frequency, 1, 3, 4, 5, 33–35, 90, 228
 - multiphonics, 179
 - periodicity, 33
- additive synthesis, 19, 37, 39–40, 85, 122–123, 128, 135, 139, 264, 279
- aftertouch, 139–140
- algorithms
 - Cholesky decomposition, 199
 - Durbin-Levinson, 192, 211
 - least-squares, 230
- amplitude envelope (amplitude-vs.-time), 260, 281
 - amplitude transient, 253, 254
 - coherence, 263
 - smoothing, 263–264
- amplitude averaging, 47, 58
- analysis data, 44
 - musicological data, 85
 - analysis data formats, 43
- analysis/synthesis software
 - CHANT, 208, 220
 - C-Sound, 96
 - DPHONE, 220
 - Loris, 126
 - SNDAN, viii, 1, 15, 42, 75, 247, 264–265
- attack transients, 128, 149, 151, 159, 171, 257, 280
- auditory nonverbal perception, 299
- auditory stream segregation, 293, 313
- auditory streaming, 292–294
- auditory system, 93, 203, 274, 285, 293, 309, 313
- autoregression spectral envelope, 190
- bandwidth, 16, 27, 41, 52, 62, 79, 85, 94, 147, 207, 209, 215
- band-pass filter, 3, 5, 12–13, 84, 146–147, 165
- Bark-band quantization, 165
- Bark-bandwidth, 165
- Britten, Benjamin, 257
- categorical perception, 301
 - continua, 304
 - plucked-to-bow, 303
 - rise time, 303
 - trumpet-to-piano, 303
 - definition, 301
 - extended context, 303
 - fuzzy boundaries, 312
 - infants, 303
 - reduced context, 303
 - causality inference, 299
- cepstrum method, 194, 196, 199–200, 221
 - cepstral coefficients, 194–195, 204, 206, 217

- cepstrum spectral envelope, 194, 197, 202
 - quefrency, 194–195
- cluster analysis, 53, 276, 298
- coding algorithms
 - autoregression (AR), 190
 - AR filter coefficients, 211
 - Huffman coding, 158, 164–165, 167
 - linear predictive coding (LPC), 40, 190
 - LPC coefficients, 191
 - MPEG-AAC method, 163
 - Sound Description Interchange Format (SDIF), 218
- coefficient of variation, 251, 262
- coherence, 262, 264, 284, 289
- constant-Q transform, 91–95, 101
- context effect(s), 294–295
 - musical context, 295
 - melodic context, 295
 - effect of transients, 294
 - effect of reverse playback, 295
- convolution, 5, 101, 147, 194, 216
- correlation, 67, 260–262, 274
 - autocorrelation, 33, 82, 192, 211
 - cross-correlation, 101–103
- data compression, 145–150, 171
- discrete cepstrum method, 199–200
 - regularization, 221
 - stochastic smoothing, 221
- Dudley, Homer, 146
- ecological theory, 273, 306
- envelope
 - noise, 123–126, 128, 133, 136, 138
 - spectral, 40, 50–52, 69, 176
 - temporal, 149, 170, 258, 260, 274, 277, 282
- filter
 - band-pass, 3, 5, 12–13, 84, 146–147, 165
 - high-pass, 58
 - low-pass, 5, 58, 194, 206
- filter parameters, 198–199, 206, 216, 221
- filter bank, 2–3, 5, 12–13, 17, 26, 31, 36, 84, 93, 147–148, 154–155
- filter synthesis, 216, 221
- formants
 - fuzzy, 209–210, 214, 216, 221
 - morphing, 215
 - shifting, 213
- Fourier transform, 1, 5–6, 8, 13, 15, 27, 41, 84, 90, 93–94
- frequency(ies)
 - cents deviation, 4
 - composite fundamental frequency, 4
- frequency deviation, 3–5, 12, 14, 16–17, 19, 21, 31, 37, 44, 50, 75, 111, 264
- frequency modulation (FM), x, 4, 93, 104, 228–232, 236–247
- frequency ratios, 105–106, 236–239
- frequency resolution, 29, 60, 90, 153, 203
- frequency spacing, 27, 29, 90, 97
- frequency tracking analysis, viii, 2, 26–33, 36–43, 60, 75, 78–81, 90
 - fundamental frequency, 90–93, 96, 99, 101–102, 104, 106–107, 109, 111–112, 118
 - harmonics, 281
 - high resolution, 103–105, 107, 111–112, 116
 - inharmonic (inharmonicity), 1, 3–4, 58, 60, 62, 111, 175, 264, 286
 - normalized frequency deviation, 3–4, 75
 - sampling frequency, 90
 - variations, 182, 184, 264, 279–280
- frequency (vs. time) envelope, 131, 281
 - coherence, 264
 - flatness, 264
 - smoothness, 264
- Hall-effect sensors, 141
- harmonics, 3–5, 10–14, 16–17, 19, 31, 33–34, 36–37, 43–56, 62–63, 67, 75, 78–85, 91–92, 97–103, 106–111, 115–116, 136, 167, 229, 231, 237–243, 262–265, 275, 278, 281, 284–285, 294
 - harmonic ratios, 107
- Helmholtz, Hermann, 250
- high-pass filter, 58
- histogram, 100–101
- inharmonicity, 1, 3–4, 58, 60, 62, 111, 175, 264, 286
- incoherence, 67, 86
- line-segment approximation, 166, 279
- listeners
 - experienced musical, 1, 33
 - graduate level, 300
 - nonprofessional performer, 113
 - professional, 114
- listening tests, 146, 164–165, 184
 - discrimination task, 302, 307
 - dissimilarity, 265, 274–275, 285–287, 291–292, 294, 296–297, 300, 307, 313
- identification task, 301, 307
- method of adjustment, 112
- two-interval/two-alternative forced choice, 112
- psychometric curve, 113–115

- loudness, 123, 131–133, 135, 141, 186, 251, 272, 274, 280, 287, 296, 305
 loudness function, 275
 low-pass filter, 5, 58, 147, 194, 196, 206
- matching algorithms, 149
 genetic algorithm (GA), 229
 pattern matching, 92, 101
- McGurk effect, 305
- Mozart, Wolfgang, 183
- multidimensional scaling (MDS) vii, 259–263, 273–274, 283, 312
- musical sounds, 1, 30, 39, 44, 50, 85, 92–93, 103, 105, 112, 115, 148, 176–178, 221, 228, 272, 274, 281–282, 303
- alto flute, 107, 110, 264–265
- bowed string, 39, 110–111, 250, 274–275, 303–306, 309
- cello, 107–108, 131–133, 184–185, 264, 275
- chime, 60, 62–64, 69, 71
- Chinese pipa, 240, 247
- clarinet, 35, 81, 107, 250, 255, 263, 265, 267, 278, 282, 285–286, 295
- cymbal, 60, 66–67, 69, 71, 75, 300
- percussion, 39, 145, 149, 179, 275, 282, 287, 299–300
- piano, 58, 60, 62, 107, 111, 123, 140–141, 147, 176, 183, 217, 250, 252, 255–256, 282
- plucked string, 275, 287, 303–304, 306
- singing voice, 75, 177, 179, 182, 207, 209, 212–213, 219–220
- sustained instruments, 300
- tenor voice, 31, 35, 75, 81, 240, 243
- timpani, 66, 69, 71, 75
- trumpet, 17, 24, 26, 44, 47, 49, 51–52, 58, 69, 75, 136, 185–186, 240–241, 245
- viola, 107, 112, 177
- violin, 90, 94, 96–97, 99, 102, 106–107, 112, 184, 255
- voice, 31, 75, 179, 182, 219
- winds, 86, 110, 284
- nonlinear frequency scaling, 202–204, 221
- non-sinusoidal components, 178, 181–182, 189
- oversampling, 150
- parameter streams, 123–126, 128, 142
- perceptual qualities
 dissimilarity, 265, 274–275, 285–287, 291–292, 294, 296–297, 300, 307, 313
- discriminability, 302, 307
 similarity, 92, 267
- perceptual discriminability, 307
- phase
 phase change, 104
 relative phase, 2, 24, 106, 185
 phase variations, 185
- phase interpolation, 19, 24, 39, 159–160, 170
 phase-matching, 149
 phase switching, 161
- phaseless reconstruction, 159–161
- piecewise constant method, 20, 24
- piecewise linear method, 20, 24
- piecewise cubic (polynomial) method, 23
- piecewise quadratic method, 21, 24
- phase vocoder, vii, 2–26, 31, 36–37, 42–82, 84–86, 93, 147–148, 229, 264–266, 307
- physical models, 149, 177
- pitch
 intonation, 111, 114
 just noticeable difference (JND), 112
 pitch center, 111, 114
 pitch perception, 92, 111, 115
 pitch tracking, 98
 vibrato, 31, 75, 85, 93
- pitch (fundamental frequency) detection
 (tracking)
 cepstrum method, 194, 196, 199–200, 221
 error function, 35, 85, 100, 230–231, 236
 pattern recognition method, 99, 103
 phase difference method, 17, 21, 103–104
 two-way mismatch method, 34, 40, 81, 85
- polyphonic audio, 153
- psychoacoustic masking threshold, 155
- psychophysical analysis, 274
- rise time (see attack transients), 303
- Russian “futurists”, 146
- Schaeffer, Pierre, 273
- Seashore, Carl, 250
- semantic differential analysis, 296
- sharpness, 258
- signal modifications
 cross synthesis, 122, 167
 envelope morphing, 215
 pitch-scale modifications, 149–150, 167
 sound manipulation, 275
 spectral envelope exchange, 184
 spectro-temporal simplifications, 86
 time dilation, 123, 129, 136
 time scaling, 38, 149, 170
 tone modifications, 295
 vibrato morphing, 137

- signal properties
 - amplitude, 20
 - breath noise, 39, 47, 110, 146
 - frequency, 90, 182, 184, 264, 279–280
 - harmonics, 281
 - musical signal, 90
 - noise, 123–126, 128, 133
 - noise envelope, 123–126, 128, 133, 136, 138
 - partials, 84
 - phase, 104
 - residual, 39
 - sampled signal, 13
 - time-varying, 36
 - tonality, 151, 156, 182
 - uncoupled phases, 185
 - vibrato, 137
- singing voice, 75, 177–182, 219
- sinusoidal oscillator, 175, 217
- sound production, 107, 112, 178, 313
- sound sources
 - bowed string, 39, 110–111, 274–275, 303–304
 - continuant tones, 251, 255
 - exciter categories, 299
 - impulse (percussive) tones, 255, 268
 - mechanical properties, 299
 - nonharmonic, 128
 - nonpercussive tones, 253, 257
 - percussion instruments, 39, 299–300
 - plucked string, 275, 287
 - polyphonic, 148
 - quasi-harmonic, 40, 125–126, 138
 - resonator categories, 299
 - sawtooth waveform, 303
 - staccato continuant tones, 256
 - sustained instruments, 300
 - vocal tract, 177, 179
- source-filter model, 178
 - exciter, 178
 - resonator, 178
- spectral analysis/synthesis
 - additive-plus-residual analysis/synthesis, 175, 178
 - additive synthesis/resynthesis, 122–123
 - constant-Q, 116
 - frequency reassignment, 130
 - frequency tracking analysis, 26–33
 - harmonic analysis, 3–5, 10–14, 16–17, 19, 31, 33–34, 36
 - harmonic filter bank, 3
 - heterodyne-filter analysis, 5
 - McAulay-Quatieri (MQ), viii, 2, 23, 26, 31, 37, 43, 125–127, 148–152, 159–162, 175, 229
 - multi-resolution sinusoidal model, 153
- noise enhancement, 123–126, 128, 133, 136, 138
- peak continuation, 39, 155
- peak-to-valley ratio, 30
- phase vocoder, 2, 31, 44, 85
- pruned phase vocoder, 148
- residual noise, 39
- short-time, 2
- sinusoidal trajectories, 156–158
- spectral reassignment, 128
- spectrum peaks, 29, 86
- time-frequency pruning, 164
- time-frequency segmentation, 151
- time reassignment, 128
- wavelets, 93
- spectral envelope, 45–57, 175–221, 243, 251, 263–264, 274–285, 290, 295, 303, 313
 - adaptation, 187, 303
 - manipulation of, 275
 - modifications
 - shifting formants, 213
 - shifting fuzzy formants, 214
 - spectral simplifications, 213
 - transcoding, 211
 - morphing, 215
 - manual input, 212
 - memory space, 212
 - precision, 213
 - stability, 214
 - synthesis speed, 206
 - parameters, 190
 - representation of
 - autoregression, 190
 - basic formants, 209
 - break-point functions, 207
 - formants, 207
 - formants wave functions, 207
 - frequency domain sampled representation, 206
 - fuzzy formants, 209
 - geometric representation, 207
 - splines, 207
 - requirements, 190, 207
 - exactness, 190
 - flexibility, 208
 - locality envelope fit, 207
 - robustness, 190
 - smoothness, 190
 - residual signal spectral matching, 204
 - genetic algorithm (GA), 229
 - parameter search space, 190
 - relative-amplitude spectral error, 230
 - spectral snapshots, 232, 236

- spectral properties (factors)
 fine structure, 178, 181
 inharmonicity, 1, 3–4, 58, 60, 62, 111, 175,
 264, 286
 inverse spectral density, 69, 86
 log magnitude spectrum, 194
 RMS amplitude, 47, 49, 52, 58, 63, 67, 69, 71
 spectral centroid, viii, x, 1, 45–49, 54, 58, 69,
 75–77, 85–86, 251, 254–257, 261–263,
 268, 275, 278–280, 283–290, 297
 spectral envelope, 45–57, 175–221, 243, 251,
 263–264, 274–285, 290, 295, 303, 313
 spectral fluctuation, 262, 284
 spectral flux, 262–263
 spectral irregularity, 55, 58, 85, 251, 263,
 278
 spectral tilt, 212, 219
 spectral variation (deviation), 122, 234, 262
 spectro-temporal incoherence, 86
 time-varying, 86
- statistical models
 ADCLUS, 298
 ADTREE, 300
 HICLUS model, 298
 hierarchical clustering, 298
 individual differences scaling model
 (INDSCAL), 287
 latent-class structure, 287
 multidimensional scaling (MDS), 259,
 273–274, 283
 INDSCAL model (individual differences
 scaling), 287
 predictive power of, 290
 dimensional interpretation, 259
 geometric configuration, 259
 Tversky's model, 298, 300
- String tone, 106, 109
 Bowed, 109
 Plucked, 109
 Struck, 106
- synthesis (resynthesis)
 additive-plus-residual analysis/synthesis,
 175
 additive synthesis, 123
 additive synthesis/resynthesis, 124
 band pass filter bank equivalent, 12
 basis spectra, 232
 CHANT, 208, 220
 computational efficiency, 95, 147, 241
 DIPHONE, 220
 direct additive sinusoidal, 220
 double FM, 228, 237
 filter synthesis, 181, 216
- formant FM, 228, 231
 frequency modulation (FM), 228
 double FM, 228, 231
 formant FM, 228, 231–232
 nested FM, 228, 231, 239
 frequency-tracking synthesis, 37
 from time-varying spectral data, 231
 group additive, 231
 identity resynthesis, 20
 inverse Fast Fourier Transform (FFT^{-1}), 176,
 218
 noise modeling, 148, 150–151, 163
 oscillator bank, 17, 19, 264
 overlap-add, 17–18, 37, 40, 218
 real-time, 23, 122–123, 125, 129
 sampling, 15, 26, 28, 49, 90, 94–93
 time-scale modification, 18–19, 145,
 167–168, 170
 wavetable, 136, 228
 indexing, 214, 228, 231, 247
 interpolation, 19–21, 23, 29, 104, 141–142,
 155
 matching, 21, 101, 227–229
 multiple wavetable, 136, 232
- synthesizers
 Continuum Fingerboard Synthesizer, 139
- temporal envelope (time-envelope), 251–257,
 260–261
 ACT model, 254–256
 amplitude-vs.-time envelope, 290
 attack, 260
 decay, 260
 morphing, 260
 spectral-centroid-vs.-time envelope, 290
 steady-state, 290
- Threshold Calculation Partition Domain, 156
- timbral features (attributes)
 attack quality, 293
 auditory sensation, 272
 brightness, 258
 clarity, 258
 multiplicity, 258
 nasality, 259
 perceptual, 259
 realism, 259
 rise time, 260
 sharpness, 258
 spectral, 258
 strangeness, 260
 temporal verbal, 281
 warmth, 260
- timbral vectors, 291

- timbre
 - categorization of, 131
 - classification, 132
 - control space, 131
 - definition, 131
 - dissimilarity of, 133
 - feature-matching model, 285
 - interpolation 145
 - morphing ix–xi, 122–123, 135–139, 142, 211–215, 219–221, 250
 - perceptual attribute, 296
 - recognition, 297
 - relational studies, 296
 - semantic labels, 296
 - sound source identity, 297
 - spatial models, 298
 - specificities, 298
 - timbral analogies, 292
 - timbral interval perception, 290
 - verbal attribute magnitude estimation, 296
- transforms
 - convolution, 5
 - correlation, 5
 - constant-Q spectral transform, 90
 - discrete Fourier transform (DFT), 90, 93
 - fast Fourier transform (FFT), 15, 95
 - modified discrete cosine transform (MDCT), 150, 163–164
 - Parseval’s equation, 116
- transients
 - attack, 128, 159
 - decay, 106
 - detection of, 159
 - onset asynchrony, 275
 - structure of, 275
 - transform-coded transients, 161
 - transform coder, 163
- verbal attributes, 251, 258
- vibrato
 - morphing, 137
 - rate, 137
- vocoder
 - phase vocoder, 2, 31, 44, 85
 - voicing estimation, 182
- wavetable
 - index matching, 136, 228
 - interpolation matching, 136, 228
- window functions
 - Blackman–Harris, 6–16, 20, 27–28, 84, 192
 - Hamming, 6–16, 20, 27–28, 84, 93, 117
 - hanning (Hann), 6–16, 27–28, 40–41, 84, 107, 117–118
 - Kaiser, 9–10, 29
 - rectangular, 6–16, 27–28, 95
 - sinc (rectangular response), 11, 206

(continued from page ii)

Seismic Wave Propagation and Scattering in the Heterogeneous Earth, by

 Haruo Sato and Michael C. Fehler

Architectural Acoustics, by Yoichi Ando

Active Noise Control Primer, by Scott D. Snyder

The Science and Applications of Acoustics, by Daniel R. Raichel

Random Signals for Engineers Using MATLAB® and Mathcad®, by Richard
 C. Jaffe

Fundamentals of Ocean Acoustics, 3rd ed., by L.M. Brekhovskikh and
 Yu.P. Lysanov

Cochlear Implants: Fundamentals and Applications, by Graeme Clark

Analysis, Synthesis, and Perception of Musical Sounds: The Sound of Music,
 edited by James W. Beauchamp