Analysis, Synthesis, and Perception of Musical Sounds

Modern Acoustics and Signal Processing

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.

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 indispensible 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 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 Interharmonic 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

1

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 acccuracy 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

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

where

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

The various parameters are defined as follows:

t = time.
$A_k(t)$ = amplitude of the kth 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

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

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 $1 \le k \le K$. 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.

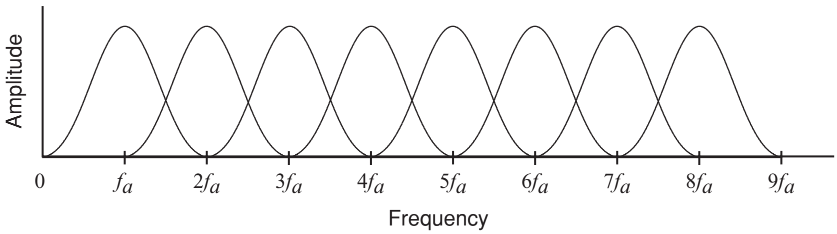

FIGURE 1.1. Overlapping band-pass analysis filter responses centered at harmonics of f_a.

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 $k = 1, \ldots, K$, 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.,

$$s_k(t) = A_k \cos(2\pi k f_a t + \theta_{k_0}). \tag{1.3}$$

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 kth 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

$$f_k(t) = kf_a + \Delta f_k(t), \tag{1.4a}$$

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

The frequency deviation can be written as

$$\Delta f_k(t) = f_k(t) - kf_a, \tag{1.4b}$$

and the relative frequency deviation as

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

Also useful is the normalized frequency deviation

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

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 kth 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

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

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

$$\Delta f_k(t) = k \Delta f_1(t), \tag{1.5a}$$

which leads to a definition of *inharmonicity*:

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

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

$$\Delta f_{c_1}(t) = \frac{\sum_{k=1}^{5} A_k(t) \Delta f_k(t)/k}{\sum_{k=1}^{5} A_k(t)}, \tag{1.5c}$$

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 formlula. Thus, for cases where A_1 is weak, $\Delta \mathrm{f}_{c1}$ 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 kth harmonic analysis filter, also called the kth *bin*. When $k\Delta f_1 \geq 0.5\ f_a$, the kth frequency component is reported from the $(k+1)$st bin with as much or greater amplitude than from the kth bin. Meanwhile, the kth 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 kth harmonic of $s(t)$ is given by

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

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 $e^{-j2\pi k f_a t}$ [which can also be written as $\cos(2\pi k f_a t) - j\,\sin(2\pi k f_a t)$], 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

$$s'_k(t) = e^{-j2\pi k f_a t} s(t) \tag{1.6b}$$

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

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

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

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

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 $(-0.5f_a, 0.5f_a)$ 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 $\pm 0.5f_a$ 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 $-T \le t \le T$. 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

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

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:

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

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:

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

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:

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

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 Nuttall (1981). Figure 1.2a compares the four window functions given above (normalized by f_a). They can be generalized to the form:

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

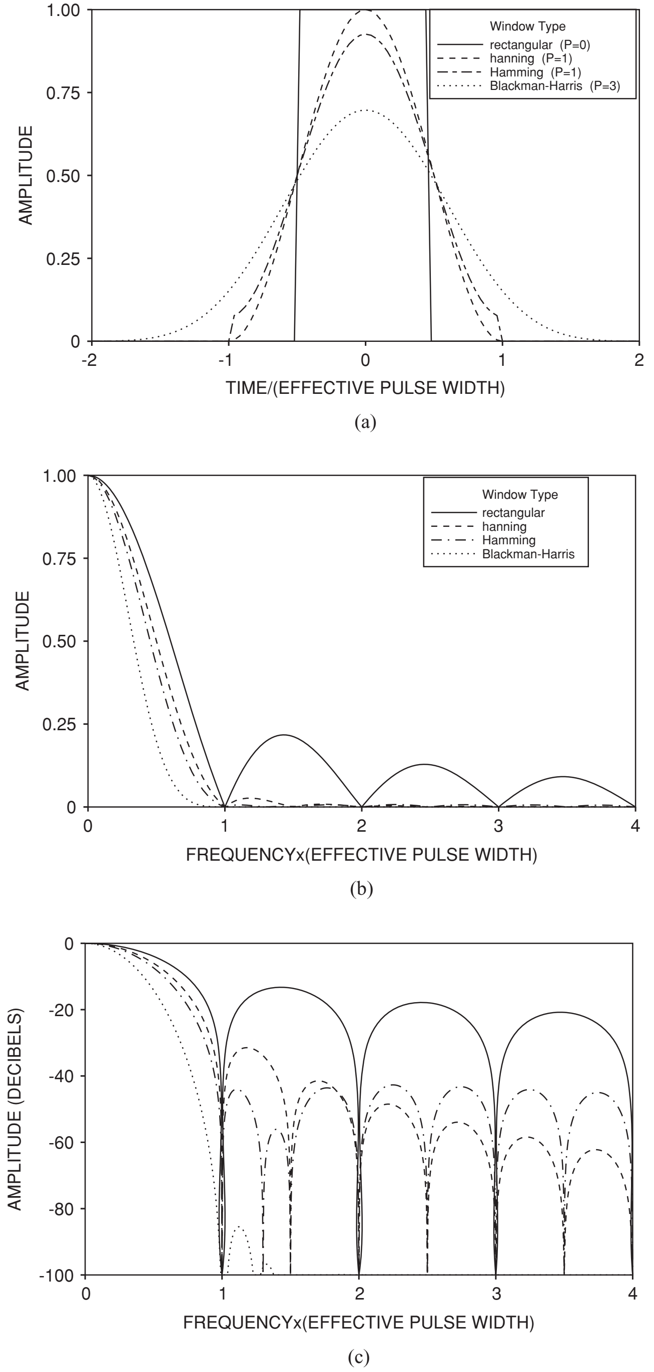

FIGURE 1.2. Comparison of four window types: rectangular, hanning, Hamming, and Blackman–Harris. (a) Normalized window functions, $w(tf_a)/f_a$; (b) Normalized window frequency responses,$W(f/f_a)$; (c) Window responses in decibels, $20\log(W(f/f_a))$.

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

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

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

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

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

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

where f is the frequency.

Knowing that

$$\int_{-T}^{T} \cos(\beta\tau) e^{-j\omega\tau} d\tau = \frac{\sin((\omega+\beta)T)}{\omega+\beta} + \frac{\sin((\omega-\beta)T)}{\omega-\beta}$$
$$= T\left[\mathrm{sinc}((\omega+\beta)T) + \mathrm{sinc}((\omega-\beta)T)\right], \tag{1.11}$$

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

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

From Eq. (1.12a), considering that $\alpha_0 = 1/P$ and $f = 0$, it follows that $H(0) = P\alpha_o = 1.0$, the maximum value of the response. Also, if the frequency is a harmonic of f_a, i.e., $f = kf_a$, $k = 1, 2, 3, \ldots$, 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 $f = qf_a/P$, $q = 1, 2, \ldots, P-1$, $W(qf_a/P) = 0.5P\alpha_q$. 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, $W_{\mathrm{db}} = 20\log_{10}(W)$, is zero for $f = 0$ and less than zero for $f > 0$. The "half-way" pass-band values at $W_{\mathrm{db}}(f_a/2)$ 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

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

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 $q = P, P+1, P+2, \ldots$. 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 $W_{\mathrm{db}}(f)$ 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

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

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

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

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 $\pi T f/\alpha = 1$, 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:

$$W(f) = \frac{\alpha}{\sinh(\alpha)}\mathrm{sinc}\left(\alpha\sqrt{(\pi T f/\alpha)^2-1}\right) \approx \frac{\alpha}{\sinh(\alpha)}\mathrm{sinc}\,(\pi T f). \tag{1.13c}$$

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,