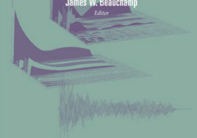
Analysis, Synthesis, and Perception of Musical Sounds

The Sound of Music

James W. Beauchamp
Editor



Modern Acoustics and Signal Processing

BEAUCHAMP
Editor

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

Springer

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 McAulay-Quatieri 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 envelope, 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: "Sines Plus Transients Plus Noise" and "Spectrotemporal Reassignment" are covered. Methods for timbre morphing are given. The last two chapters discuss the perception of musical sounds based on discrimination and multidimensional scaling timbre 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."
—Joel Chabade, Electronic Music Foundation

Physics
ISBN 0-387-32496-8
9 780387 324968

Springer

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.

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 ith 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 ith 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

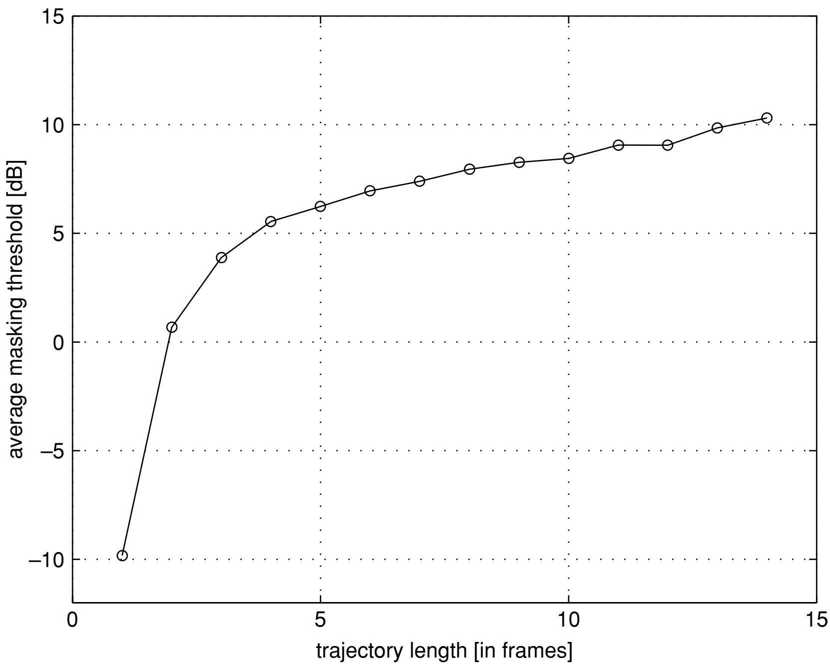


FIGURE 4.4. Average maximum signal-to-masking threshold (in decibels) vs sinusoidal trajectory length. Note that the longer a trajectory lasts, the higher its signal-to-masking threshold. These data were derived from the top frequency range of 8 s of pop music, where each frame length is approximately 6 ms.

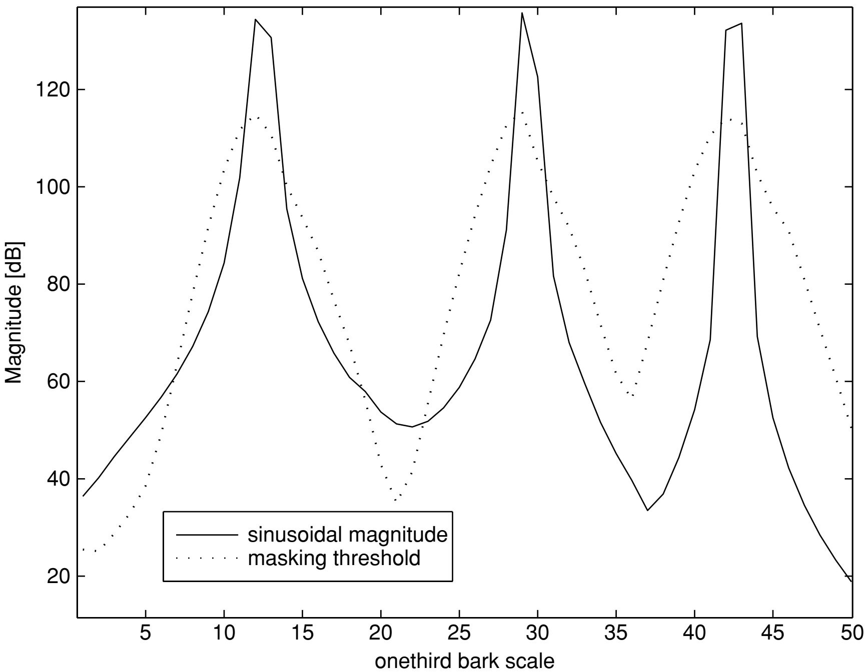


FIGURE 4.5. The original spectral energy vs the masking threshold for three pure sinusoids at frequencies 500, 1500, and 3200 Hz. Note that the masking threshold is approximately 18 dB below each sinusoidal peak.

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

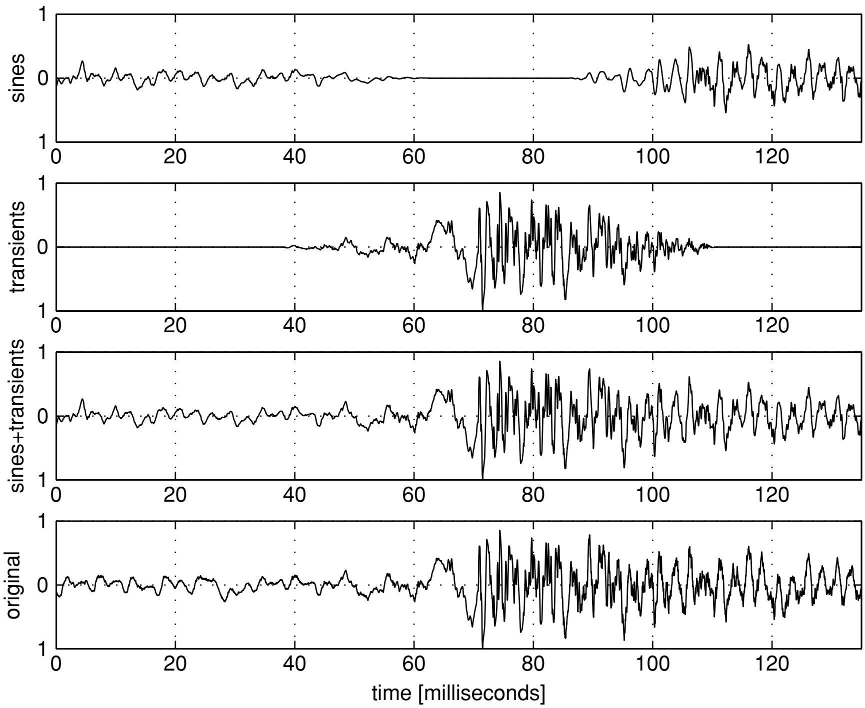


FIGURE 4.6. How sines and transients are combined: The top plot shows the multiresolution sinusoidal modeling component of the original signal. The sinusoids are faded-out during the transient region. The second plot shows a transform-coded transient. The third plot shows the sum of the sines plus the 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 $t \cong 60$ 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 lth 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 $\{\phi_r[l], \phi_\mathrm{r}[l-1]\}$. 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

$$s(m+lS) = \sum_{r=1}^{R[l]} A_r[l, m]\cos(\theta_r[l, m]), m = 0, \ldots, S-1 \tag{4.3}$$

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:

$$\omega_r[l, m] = \omega_r[l] + \frac{(\omega_r[l+1] - \omega_r[l])}{S} m, m = 0, \ldots, S-1 \tag{4.4}$$

Then, the instantaneous phase for the rth trajectory in the lth frame is

$$\theta_r[l, m] = \theta[l, m-1] + \omega_r[l, m], m = 0, \ldots, S-1, \tag{4.5}$$

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 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* (Klumer 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 modications," *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).

5

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