TIME-FREQUENCY SIGNAL MODELS FOR MUSIC ANALYSIS, TRANSFORMATION, AND SYNTHESIS

Michael Goodwin[†] *Martin Vetterli*

Department of Electrical Engineering and Computer Science &
[†]Center for New Music and Audio Technologies
University of California at Berkeley

ABSTRACT

In signal analysis-synthesis, the analysis derives a set of parameters that the synthesis uses to reconstruct the original signal. In musical applications, this reconstruction should be perceptually accurate, and the parameterization should allow for such desirable signal modifications as time-scaling, pitch-shifting, and cross-synthesis; the analysis parameters should correspond to a signal model that is flexible enough to allow these transformations. Sinusoidal modeling meets this flexibility requirement, but has difficulty representing some salient features of musical signals such as attack transients and noiselike processes. In this paper, sinusoidal modeling is reviewed and some variations are proposed to account for its shortcomings; also, wavelet-based representations of musical signals are considered.

1. SIGNAL MODELING FOR MUSIC

A signal model provides a mathematical representation of a signal in terms of salient parameters. For a given signal, these parameters can be determined by a suitable analysis, and can be used in a synthesis process to construct an estimate of the original signal; note that the analysis-synthesis method and the signal model are inherently coupled. In music analysis-synthesis, it is often desirable to effect transformations such as time-scaling, pitch-shifting, and cross-synthesis; this is made possible by using a flexible signal model whose parameters can be modified before synthesis to produce the desired transformation.

In some analysis-synthesis scenarios, there is a difference between the synthesized signal and the original. This is termed the analysis-synthesis *residual*; it exists when the signal model does not account for all of the features of the original signal, or if the accompanying analysis-synthesis is inaccurate. Shortcomings of the analysis-synthesis and the underlying signal model are thus manifested in the residual.

Perfect reconstruction is frequently desired in analysis-synthesis applications; in that case, there is no energy in the residual and the synthesized signal is identical to the original. In audio applications, on the other hand, it is generally sufficient to achieve *perceptually lossless* reconstruction, in which the synthesized sound is perceptually equivalent to the original. To achieve this, either the residual must contain only components that would be perceptually masked in the synthesis, or the residual must be separately modeled and reinjected into the reconstruction. Note that perfect reconstruction is perceptually lossless, and that a perceptually lossless system can invoke psychophysical phenomena such as masking to effect data reduction, *i.e.* its intermediate parameterization may be more efficient than that of a perfect reconstruction system, and may also be more readily transformable since the parameters are based on perception.

Analysis-synthesis can be generally viewed as follows:

$$x[n] \implies \{a_i, f_i\} \implies \hat{x}[n] = \sum_{k=1}^{K} a_k f_k[n] \qquad (1)$$

The synthesis estimates the original signal using a set of functions based on the signal model and a set of coefficients provided by the analysis; in some cases the expansion functions are signal-dependent, and are derived by the analysis as well. The functions and coefficients comprise the intermediate parameterization of the signal. If coding efficiency is desired, this parameterization should be as sparse as possible. In musical applications, however, the ability to transform the parameters to achieve a certain effect often takes some priority over data reduction, so in this case the parameterization should relate to natural musical features such as pitch, harmonic structure, modulation (*e.g.* vibrato), noise (*e.g.* breathiness), and dynamics such as attacks. If these features are apparent in the parameterization, they can be appropriately handled by the modification process. For instance, in time-scaling it is generally undesirable to alter the time structure of sharp attacks, so it is necessary to identify the attacks at the parametric level in order to preserve them for accurate synthesis. In this framework, modifications can be carried out by altering the coefficients or the synthesis functions or both.

2. SINUSOIDAL MODELING

In sinusoidal modeling, the signal is modeled as a sum of evolving sinusoids called *partials*:

$$d(t) = \sum_{q=1}^{Q(t)} A_q(t) cos\Theta_q(t) \qquad (2)$$

Here, $Q(t)$ is the number of partials at time t; $A_q(t)$ is the time-varying amplitude of the q-th partial and the total phase $\Theta_q(t)$ describes its frequency evolution and phase offset. Since this sum-of-partials model is not well-suited for representing broadband noise processes, an additive noise component $s(t)$ is often included in the signal model, resulting in a *deterministic plus stochastic decomposition* [1]: $x(t) = d(t) + s(t)$. For musical signals, the additive noise accounts for such inherently stochastic musical features as breath noise. Because these processes are an integral part of music, the noise component must be modeled independently of the partials and reinjected at the synthesis stage to insure the realism of the output music [1, 2]; this inclusion of noise is a prerequisite for perceptual losslessness.

Analysis methods for the sinusoidal model are generally frame-by-frame approaches; the analysis parameters

are then frame-rate representations of the time-varying amplitude and frequency *tracks* of each partial. In [1, 3, 4], the analysis uses the short-time Fourier transform (STFT); the parameters of the partials in a given frame are found by estimating the amplitude, frequency, and phase of the peaks in the magnitude spectrum of the oversampled discrete Fourier transform (DFT) of that frame. This contrasts with the time-domain analysis-by-synthesis proposed in [5], where the partials are estimated by exhaustively searching for sinusoids that correlate strongly with the signal frame; when the maximally correlated sinusoid is found, its contribution is subtracted from the frame and the process is iterated as in a matching pursuit. In either approach, the analysis examines frames of length N and uses a stride L (often $N/2$) to advance through the signal; N and L are chosen to give a reasonable tradeoff between the efficiency of the parameterization and its accuracy in representing the time-domain variations of the signal. The result of the analysis is a set of amplitude, frequency, and phase parameters $\{A_{q,i}, \omega_{q,i}, \phi_{i,q}\}$ for each partial q in each frame i.

Synthesis for the sinusoidal model generally involves accumulating the time-domain outputs of a bank of sinusoidal oscillators as in the signal model of equation 2. Synthesis can also be done in the frequency domain by accumulating the spectral contributions of the partials for each frame and then using an inverse DFT and overlap-add to construct the output from the frame-by-frame spectra [6]. This approach will not be discussed further, however, since time-domain synthesis more directly illustrates the signal representation issues this paper is concerned with.

In time-domain additive synthesis, the output of the q-th oscillator is $A_q[n]\cos\Theta_q[n]$; it is dictated by amplitude and total phase control functions that must be calculated in the synthesis process. This involves two difficulties, *line tracking* and parameter interpolation, both of which arise because of the time-evolution of the partials and the resultant analysis parameter differences from frame to frame. Since the analysis does not track the partials, but instead merely derives sets of parameters for the partials that it finds in the signal frames, the synthesis must establish continuity by relating the parameter sets in adjacent frames to form partials that endure appropriately in time. This line tracking is generally done by associating the q-th partial in frame i to the partial in frame $i+1$ with frequency closest to $\omega_{q,i}$; this procedure is carried out until each of the partials in adjacent frames are either coupled or accounted for as a birth or a death, *i.e.* a partial that is newly entering or leaving the signal. For signals such as music with dynamic spectral content, line tracking is a difficult problem, and a variety of algorithms have been proposed [1, 3].

After partial continuity is established by line tracking, it is necessary to interpolate the frame-rate partial parameters to determine the sample-rate oscillator control functions. Typically, interpolation is done using low-order polynomial models such as linear amplitude and cubic phase. The partial amplitude interpolation in synthesis frame i is a linear progression from the amplitude in analysis frame i to that in frame $i+1$ and is given by

$$\hat{A}_{q,i}[n] = A_{q,i} + \frac{A_{q,i+1} - A_{q,i}}{S} n \qquad (3)$$

where $n = 0, 1, \ldots, S-1$ is the time sample index, and S is the length of the synthesis frame. Unless the analysis parameters are intermediately interpolated to a different time resolution, $S = L$, the analysis stride. The phase interpolation is given by

$$\hat{\Theta}_{q,i}[n] = \Theta_{q,i} + \omega_{q,i} n + \alpha_{q,i} n^2 + \beta_{q,i} n^3 \qquad (4)$$

where Θ and ω enforce phase and frequency matching constraints at the frame boundaries, and α and β are chosen the make the total phase progression maximally smooth [3].

Sinusoidal modeling has found many applications in speech and audio coding and analysis-synthesis. This is primarily because the representation in terms of sinusoidal parameters is efficient and readily allows for such desirable transformations as time-scaling, pitch-shifting, and cross-synthesis [3, 4, 6]; it also provides for novel modifications based on a musical timbre space, such as arbitrary interpolation between disparate sounds. However, sinusoidal modeling only derives an accurate reconstruction for signals that vary smoothly on a time scale comparable to the frame rate. Substantial changes that occur from frame to frame are not well-represented by the analysis parameters since the STFT is not adequately time-localized; also, the low-order interpolation used by the synthesis will not necessarily match the behavior of the original signal. Thus, the dynamics of the original signal are not accurately reconstructed in the synthesis. As a result, the residual tends to contain rapid transients related to note attacks and extraneous partials brought about by the interpolation mismatch. These signal features are not meant to appear in the residual, which ideally should only contain the broadband stochastic component of the signal model; models of the residual are generally simple and can not specifically account for these features, so they will not be preserved by analysis-synthesis of the residual and thus will not appear accurately in the final deterministic plus stochastic reconstruction [2], which will then be perceptually lossy. Note that narrowband noise is represented by the partials since it can be expressed as a sinusoid modulated by a lowpass random process, namely a sinusoid varying at the frame-rate time scale.

The sinusoidal synthesis can be expressed as a sum of non-overlapping frames, each of which is a sum of partials:

$$\hat{x}[n] = \sum_i \hat{x}_i[n] = \sum_i \sum_{q=1}^{Q_i} A_{q,i}[n]\cos\Theta_{q,i}[n] \qquad (5)$$

Each of the modulated sinusoids in this expression spans a frame, implying a constant time resolution. A transient event that occurs on a time scale shorter than an analysis frame is inherently spread out across the synthesis frame because of the fixed time resolution of these sinusoidal expansion functions. Furthermore, the parameter interpolation across frames results in additional smoothing of transients. The combination of these effects creates a *pre-echo* before an attack; Figure 1a depicts a rapid onset of a single sinusoid and Figure 1b shows how the attack is spread across several frames in the synthesis. In Figure 1c, a more accurate reconstruction of the attack is obtained by shortening the frames near the onset; the accuracy could be improved by decreasing the frame widths further. In this adaptive frame-rate approach, the same parameter interpolation models are used from frame to frame; the reconstruction is improved because the time-varying frame rate results in synthesis functions with varying time support that are better suited for representing dynamic signals. The improvement in accuracy, however, comes at the cost of additional analysis computation to locate transient behavior, additional parameters for encoding the frame rate, an increased data rate for decreased frame sizes, and possible limitations on real-time synthesis for short frames. For applications where these drawbacks are not prohibitive, this framework allows a combination of the modification flexibility of sinusoidal modeling with the representational accuracy of methods based on appropriate time-frequency resolution tradeoffs.

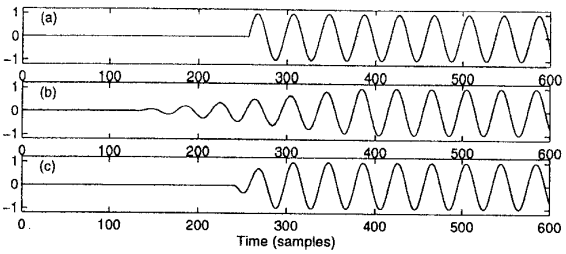

Figure 1. *(a) The onset of a partial, (b) its delocalized reconstruction for $N = 256$, and (c) a more accurate synthesis obtained by shortening the frames near the onset time.*

3. WAVELETS

The above consideration of time-frequency resolution naturally leads to a discussion of wavelets and their underlying signal model. Though some musical applications of the continuous wavelet transform have been developed [7], the focus here will be limited to the discrete wavelet transform and its implementation using a tree-structured filterbank.

The discrete wavelet transform is based on critically sampled two-channel perfect reconstruction filterbanks with halfband highpass and lowpass branches. If such a filterbank is iteratively applied to its own lowpass branch, the discrete wavelet transform results; it is a decomposition of the signal into octave bands (the highpass channels of the iterated two-channel filterbanks) plus the lowpass channel where the iteration is stopped. The lowpass channel signal is a coarse estimate of the original signal. The reconstruction can successively approximate the original signal as a sum of this lowpass version and the successively finer details of the highpass channels; if all the details are added, the reconstruction is perfect. In this light, the wavelet transform supports a model in which the signal is a a sum of a lowpass approximation and several levels of detail.

The wavelet analysis filterbank derives the coefficients for the linear expansion of the signal with respect to the basis functions corresponding to the impulse responses of the synthesis filterbank. At each stage of the filterbank, the basis functions have a different support in time: the highpass channel of the first stage has the shortest impulse response, so this channel signal contains short-time features of the original signal; deeper stages of of the filterbank have longer impulse responses (basis functions). The channel signals thus represent signal behaviors on different time scales; this suggests that the wavelet transform is suitable for separating rapidly varying components like noise and transients from long-term components like enduring low-frequency partials. Note that if multiple partials fall in a single band, they are not separated by the wavelet analysis.

Unlike sinusoidal modeling, the wavelet transform allows for perfect reconstruction; also, depending on the specific filters used, it can be more efficient to implement than sinusoidal analysis-synthesis. These advantages, however, are accompanied by a loss of modification capabilities: though the aforementioned separation of transients can be useful for accurate time-scaling based on modifying the transform coefficients, frequency-based modifications are difficult to formalize. To get both the desirable time-frequency tradeoff of the wavelet transform and the musically meaningful modifications of the sinusoidal model, the two approaches can be merged by applying a sinusoidal analysis in each of the wavelet transform channels [8]. This amounts to using a different frame rate for sinusoids in different frequency bands, and supplements the wavelet transform with the ability to separate and estimate multiple sinusoids in a single band. Synthesis for this multi-rate sinusoidal model can be done using oscillators in the channels followed by the synthesis filterbank, or by using a bank of oscillators with multi-rate interpolation models. This approach is basically an extension of the classic phase vocoder [9] that allows more than one partial in each frequency band and achieves a time-frequency tradeoff that represents partial behavior more accurately than fixed resolution methods.

4. PITCH-SYNCHRONOUS WAVELETS

For pseudo-periodic signals such as music, an appealing alternative to the lowpass-plus-details wavelet model is the approach presented by Evangelista in [10], in which a pseudo-periodic signal is modeled as a sum of a periodic signal plus deviations from periodicity. The coarse signal estimate, namely the periodic signal, is derived by bandpass filtering the signal around a set of harmonic frequencies determined by estimating the signal's pitch period; the deviation details correspond to frequency bands that get wider as they get farther from the harmonic frequencies.

The decomposition of the signal into harmonics and modulation details around the harmonics is done by a pitch-synchronous wavelet transform (PSWT), which relies on an accurate estimate of the possibly time-varying signal pitch. If the estimated pitch is P, the signal is demultiplexed into P channels; the p-th channel signal consists of the p-th sample of each pitch period. Then, each of these P pitch-rate signals undergoes a separate wavelet transform. The lowpass estimate of the p-th channel by the p-th wavelet transform is then used to construct the p-th sample of the periodic estimate; the highpass details of the P wavelet transforms correspond to the various levels of deviation from periodicity. An example of this model is given in Figure 2.

The PSWT is useful for obtaining musically important modulation information about the signal, and provides a more appropriate successive approximation method for pseudo-periodic signals than the lowpass-plus-details model of the wavelet transform. Time-scale and pitch modifications can be achieved by interpolating or decimating the channel signals in time or across channels, respectively, and interesting cross-synthesis results can be obtained by applying the modulation of one musical signal to the periodic estimate of another [10]; these modifications are facilitated by the pseudo-sinusoidal nature of the representation.

In this model, transients and noise are not part of the fundamental signal estimate, and are primarily represented by the wide bands away from the harmonics. In some cases, robust pitch detection may improve this representation; if no pitch is detected, the PSWT can revert to the standard wavelet transform, which is more suitable for non-periodic signal components [10].

5. WAVELET PACKETS

Wavelet packets, like wavelets, are based on iterating two-channel perfect reconstruction filterbanks. In this case, however, the filterbank can be iterated on either branch. A tree-structured filterbank is grown by applying a metric at each node to determine if iterating the filterbank at that node improves the signal representation; this process is analogous to pruning the full tree-structured filterbank. The resultant transform creates a division of the frequency domain that represents the signal optimally with respect to the applied metric; this frequency division can be time-

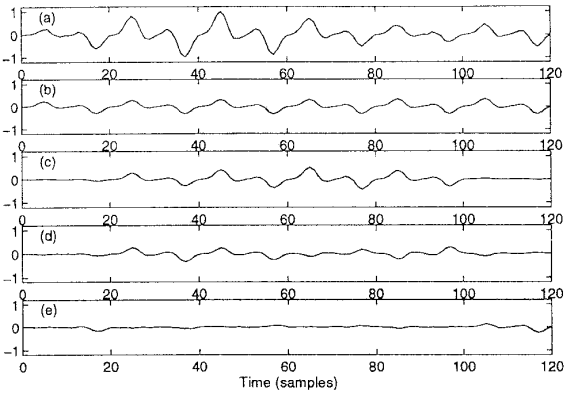

Figure 2. *(a) A pseudo-periodic signal, (b) its periodic estimate by the PSWT, and (c,d,e) three levels of details.*

varied to create a wavelet packet that adapts to optimally represent a signal whose characteristics change in time [11].

Pruning the filterbank tree corresponds to choosing a basis for expanding the input signal. One proposed metric for pruning is the entropy of the expansion coefficients for the basis that corresponds to the particular pruning. Entropy measures the energy spread of the coefficients; the lower the entropy, the fewer large-valued coefficients. Thus, minimizing the entropy results in a "best basis" where the basis functions with large expansion coefficients are similar to the principal components of the input signal. Small-valued coefficients can be considered noise, and a thresholding scheme can be applied so that these noise contributions are excluded from the synthesis. This approach was proposed as a technique for removing unwanted noise from music recordings [12], but can also serve to derive the deterministic-plus-stochastic signal decomposition discussed earlier; because this method does not involve interpolation, it results in a possibly cleaner stochastic component than the sinusoidal analysis-synthesis residual, and allows for separate processing of noiselike components as in the sinusoidal model.

As in the wavelet transform, modifications based strictly on wavelet packet coefficients are not easily tractable. This difficulty can be circumvented by using the metric proposed in [13], which measures the number of partials in the frequency band corresponding to a node. The two-channel filterbank is iterated at a given node if the number of partials found in the prospective children nodes is greater than or equal to the number of partials at the given parent node. This process continues until each frequency band contains at most one partial; the partial parameters are then estimated based on the Tufts-Kumaresan method, which improves in accuracy when applied in subbands as in this approach [13]. This analysis, which can be viewed as a phase vocoder based on a wavelet packet filterbank, provides an optimal sinusoidal decomposition of the signal. If time segmentation is introduced to account for attacks and signal changes, and if line-tracking is introduced between adjacent segments, the synthesis can be phrased in terms of oscillators instead of a filterbank as in regular wavelet packet approaches. This enables the wide class of modifications achievable in the general sinusoidal model.

6. CONCLUSION

A variety of musical signal models have been discussed. The sinusoidal model provides perceptually meaningful parameters relating to loudness and pitch, and is suitable for performing the wide range of desired modifications, but has difficulty representing some signal features because of its fixed resolution. In perfect reconstruction approaches such as wavelets and wavelet packets, however, modifications based on transforming the coefficients or filterbanks are difficult to generalize, and in cases where real-time synthesis is desired, any complicated modification models are inapplicable. Thus, the trend in the signal models presented is to form hybrid representations that combine the flexibility of sinusoidal parameters with the representational accuracy of methods with appropriate time-frequency tradeoffs.

7. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Gianpaolo Evangelista for his helpful software contributions and Prof. Edward Lee for his continuous support.

REFERENCES

[1] X. Serra and J. Smith. Spectral modeling synthesis: A sound analysis / synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24, Winter 1990.

[2] M. Goodwin. Residual modeling in music analysis-synthesis. *ICASSP-1996.*

[3] R. McAulay and T. Quatieri. Speech analysis / synthesis based on a sinusoidal representation. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Aug. 1986.

[4] T. Quatieri and R. McAulay. Speech transformations based on a sinusoidal representation. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Dec. 1986.

[5] E. George and M. Smith. Analysis-by-synthesis / overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones. *Journal of the Audio Engineering Society*, 40(6):497–516, June 1992.

[6] X. Rodet and P. Depalle. Spectral envelopes and inverse FFT synthesis. *Proceedings of the 93rd Audio Engineering Society Convention*, Oct. 1992.

[7] R. Kronland-Martinet. The wavelet transform for analysis, synthesis, and processing of speech and music sounds. *Computer Music Journal*, 12(4):11–20, Winter 1988.

[8] M. Rodriguez-Hernandez and F. Casajus-Quiros. Improving time-scale modification of audio signals using wavelets. *ICSPAT-1994*, 2:1573–1577.

[9] M. Dolson. The phase vocoder: A tutorial. *Computer Music Journal*, 10(4):14–27, Winter 1986.

[10] G. Evangelista. Pitch-synchronous wavelet representations of speech and music signals. *IEEE Trans. on Signal Processing*, 41(12):3313–3330, Dec. 1993.

[11] C. Herley et al. Time-varying orthonormal tilings of the time-frequency plane. *ICASSP-1993*, 3:205–208.

[12] J. Berger, R. Coifman, and M. Goldberg. Removing noise from music using local trigonometric bases and wavelet packets. *Journal of the Audio Engineering Society*, 42(10):808–818, October 1994.

[13] C. van den Branden Lambrecht and M. Karrakchou. Wavelet packets-based high-resolution spectral estimation. *Signal Processing*, 47(2):135–144, Nov. 1995.