Sound Representation and Modification With Multiresolution Sinusoidal Models

Michael Goodwin*, Paolo Prandoni†, and Martin Vetterli†

*Silicon Graphics, Inc., Mountain View, CA and †EPFL, Lausanne, Switzerland

Abstract: In the standard sinusoidal model, the analysis and synthesis are typically carried out at fixed frame rates. This can result in representation error for signals with dynamic behavior. To improve the modeling performance for nonstationary signals, the sinusoidal analysis-synthesis can be carried out in a multiresolution framework based on adaptive time segmentation; the adaptation is guided by a metric such as rate-distortion. This approach improves the model accuracy while preserving the ability to carry out the rich class of modifications afforded by the standard sinusoidal model.

SINUSOIDAL MODELING

In sinusoidal modeling, a signal is represented as a sum of evolving sinusoidal partials. Analysis involves finding frame-by-frame estimates of the amplitude, frequency, and phase of the constituent partials [1, 2, 3]. In the synthesis stage, these parameters are connected from frame to frame by a line tracking process and then interpolated to derive sample-rate control functions for a bank of oscillators; the interpolation is carried out based on underlying synthesis frames, which are established by the analysis stride. Typically, the frames are of fixed size as depicted in Fig. 1(a). While such models have proven useful for speech and audio processing, using a fixed frame size has various drawbacks which motivate the use of adaptive frame sizes such as those shown in Fig. 1(b); these drawbacks are discussed below.

The resolution of the sinusoidal model is limited by the choice of the analysis frame size and stride. For long frames, the time resolution is inadequate for capturing signal dynamics such as attack transients. For short frames, the frequency resolution is degraded such that estimation of sinusoidal components becomes difficult. Furthermore, the original signal may not behave in the manner assumed by the line tracking and parameter interpolation used in the synthesis. Finally, the sum-of-partials model simply has difficulty representing broadband noiselike processes such as breath noise. As a result of these various limitations, the sinusoidal analysis-synthesis process yields a nonzero residual. Noise-based models of the residual have been developed, but such approaches are primarily useful for modeling the aforementioned noiselike processes [2, 4]; while these methods do improve the synthesis realism, they are limited in that noise-based models are not effective for representing coherent features. For noise-based residual models to be effective, then, it is necessary to modify the sinusoidal model so as to reduce coherent artifacts, for instance *pre-echo* in the reconstruction of signal onsets. Pre-echo, a well-known difficulty in audio coding, is of interest here since high-quality music synthesis requires preservation of note attacks [2, 3].

Pre-echo is introduced in the sinusoidal model in the following way. Before a signal onset, there is an analysis frame in which the signal is not present and no partials are found. Thereafter, partials are identified in the frame in which the onset occurs; the line tracking interprets these as new partials and connects them via interpolation to zero-amplitude partials in the previous frame [1]. This results in a smooth amplitude envelope for each partial instead of a sharp onset. The pre-echo caused by linear amplitude interpolation is shown in Fig. 1 for two simple signals, and in Fig. 2 for a saxophone note.

The pre-echo problem in the sinusoidal model can be alleviated by casting the modeling into a multiresolution framework [3, 5, 6]. This paper is concerned with multiresolution based on time-varying segmentation, in which short frames are used near transients, which improves time localization, and long frames are used for regions with stationary behavior, which improves frequency resolution and allows for coding gain. Such adaptivity is of great importance in signal modeling [3].

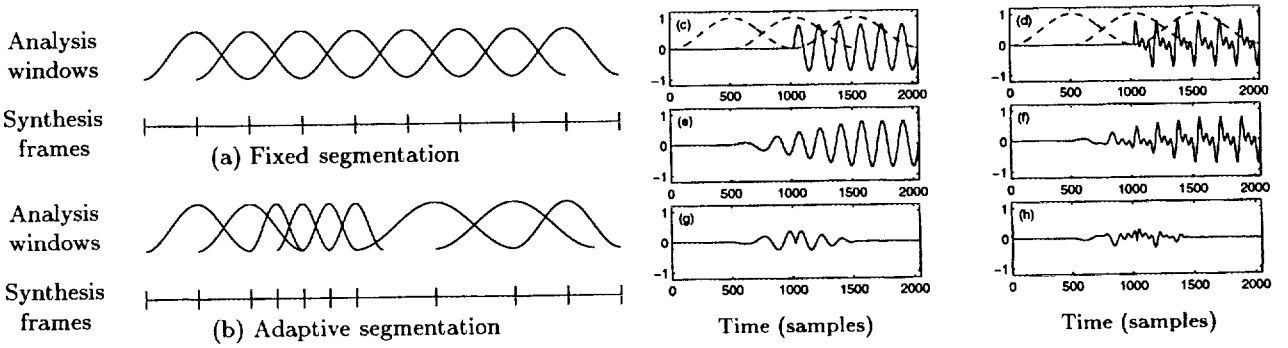

Figure 1: Analysis windows and synthesis frames in sinusoidal models with (a) fixed and (b) adaptive resolution. Using a fixed frame size introduces pre-echo; modeling the onsets of (c) a sinusoid and (d) a simple harmonic signal using the analysis windows depicted leads to (e,f) delocalized reconstruction and (g,h) residuals with artifacts.

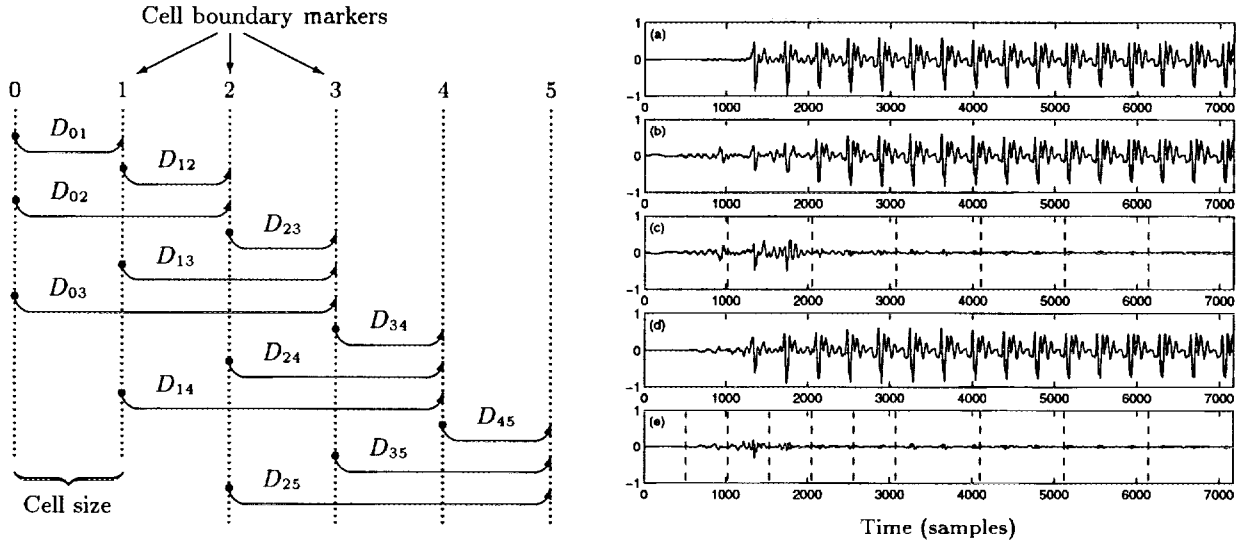

Figure 2: Depiction of a dynamic segmentation algorithm and a modeling example: (a) the onset of a saxophone note, (b) a fixed resolution reconstruction and (c) the corresponding residual, where the dashed lines indicate the synthesis frames; and, (d) a dynamic multiresolution model and (e) its residual, which exhibits fewer artifacts.

ADAPTIVE SEGMENTATION

In the sinusoidal model, the analysis windows do not have to satisfy an overlap-add property as in the perfect reconstruction STFT [3]. Thus, time-varying windows such as those depicted in Fig. 1(b) can be readily used. Such windows correspond to a time-varying synthesis segmentation; the goal, then, is to derive such a segmentation which optimizes some metric such as the mean-squared reconstruction error. This optimization can be carried out by an exhaustive global search in which each possible segmentation is considered in turn. If the optimization metric is additive and independent on disjoint segments, however, an exhaustive evaluation involves redundant computation. This redundancy can be removed by formulating the computation as a dynamic program. This approach is based on treating the time span of the signal as a concatenation of *cells*. The boundaries between these cells, which will be called *markers*, serve as nodes in the dynamic program; allowable segment lengths correspond to integer multiples of the cell size [7].

The operation of a dynamic segmentation algorithm is depicted in Fig. 2; the expression D_{ab} represents the metric associated with the signal model on the segment between markers a and b. At each marker, the algorithm computes and records the minimum modeling metric to reach that marker; it also records the length of the last segment in the corresponding segmentation, which is the optimal segmentation up to that point in the signal, and the sinusoidal parameters computed for that particular segment using an analysis window of a corresponding scale (see Fig. 1(b)). When the end of the signal is reached, the optimal segmentation can be recovered by backtracking through the recorded lengths. The computation at a given marker thus amounts to evaluating the modeling metric on each segment that leads to that marker; this is done by analyzing the signal with a window based on the segment length, synthesizing the model by tracking and interpolating the analysis data back to the data recorded for the marker at the start of the segment, and computing the reconstruction error; the tracking is simplified by constraining the maximum number of partials to be the same in short segments as in long segments. Fig. 2 depicts a fixed model of a saxophone onset and an improved model derived by a dynamic segmentation algorithm. Additional details on signal-adaptive segmentation for sinusoidal modeling can be found in [3, 6]. Such models allow the same extensive modification capabilities as the standard sinusoidal model while providing improved accuracy; the analysis-synthesis residuals for such multiresolution models can be effectively described using noise-based methods.

REFERENCES

[1] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):744 - 754, August 1986.

[2] 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.

[3] M. Goodwin. *Adaptive Signal Models: Theory, Algorithms, and Audio Applications.* PhD thesis, Berkeley, 1997.

[4] M. Goodwin. Residual modeling in music analysis-synthesis. *ICASSP-1996*, 2:1005–1008.

[5] S. Levine, T. Verma, and J. Smith. Alias-free, multiresolution sinusoidal modeling for polyphonic wideband audio. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 1997.

[6] M. Goodwin. Multiresolution sinusoidal modeling using adaptive segmentation. *ICASSP-1998.*

[7] P. Prandoni, M. Goodwin, and M. Vetterli. Optimal segmentation for signal modeling and compression. *ICASSP-1997*, 3:2029–2032.