# MULTIRESOLUTION SINUSOIDAL MODELING USING ADAPTIVE SEGMENTATION

Michael Goodwin

Department of Electrical Engineering and Computer Science
University of California at Berkeley
michaelg@eecs.berkeley.edu

## ABSTRACT

The sinusoidal model has proven useful for representation and modification of speech and audio. One drawback, however, is that a sinusoidal signal model is typically derived using a fixed frame size, which corresponds to a rigid signal segmentation. For nonstationary signals, the resolution limitations that result from this rigidity lead to reconstruction artifacts. It is shown in this paper that such artifacts can be significantly reduced by using a signal-adaptive segmentation derived by a dynamic program. An atomic interpretation of the sinusoidal model is given; this perspective suggests that algorithms for adaptive segmentation can be viewed as methods for adapting the time scales of the constituent atoms so as to improve the model by employing appropriate time-frequency tradeoffs.

## 1. ADAPTIVE SIGNAL MODELS

Compact signal models are useful for analysis, compression, enhancement, and modification [1]. To achieve compaction for arbitrary signals, models must be constructed in a signal-adaptive manner. Such signal adaptivity is the central principle in methods such as best bases [2], adaptive wavelet packets [3], and various atomic decomposition approaches such as matching pursuit [4, 5]; these models can be interpreted as signal expansions in which the expansion functions are chosen in a signal-adaptive fashion from an overcomplete set [1]. Signal adaptivity can also be achieved in parametric methods such as the sinusoidal model, in which the sinusoidal expansion functions are constructed using parameters extracted from the signal [1].

This paper is concerned with the sinusoidal model. The basic problem is that the sinusoidal model is typically carried out with a fixed frame size which may not be appropriate for all regions of a nonstationary signal. It is demonstrated that a fixed signal segmentation leads to resolution limitations that result in artifacts such as pre-echo, which is a well-known difficulty in audio coding [6]. An atomic interpretation of the sinusoidal model suggests that these reconstruction artifacts can be reduced by adapting the time scales of the atoms according to the signal behavior, i.e. using long scales for stationary behavior and short scales for transients. It is demonstrated that such adaptation can be carried out effectively using a segmentation algorithm based on a dynamic program.

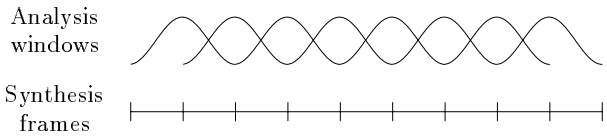


Figure 1: Analysis windows and synthesis frames in the sinusoidal model; the synthesis frames are defined by the window centers.

## 2. SINUSOIDAL MODELING

The sinusoidal model has been widely used for speech and audio processing [7, 8, 9]. This section reviews sinusoidal analysis-synthesis, discusses reconstruction artifacts, and suggests an atomic interpretation of the model.

### 2.1. Analysis-Synthesis

In sinusoidal modeling, a signal is represented as a sum of slowly evolving sinusoids:

$$x(t) \;\approx\; \hat{x}(t) \;=\; \sum_{q=1}^{Q} A_q(t) \cos \Phi_q(t). \qquad (1)$$

Analysis for this model corresponds to finding moving estimates of the amplitude, frequency, and phase of the constituent partials [7, 8, 9]. This estimation is typically carried out by peak picking in the short-time Fourier domain. The analysis yields partial parameters for each analysis frame, and the data rate of the parameterization is given by the analysis stride and the order of the model, i.e. the number of partials $Q$ in Eq. (1). In the synthesis stage, the stride-rate model parameters are connected from frame to frame by a line tracking process and then interpolated using low-order polynomial models to derive sample-rate control functions for a bank of oscillators; the interpolation is carried out based on underlying synthesis frames, which are implicitly established by the analysis stride. Analysis and synthesis frames are depicted in Fig. 1.

The sum-of-partials model in Eq. (1) has difficulty representing broadband processes; these appear in the residual $r(t) = x(t) - \hat{x}(t)$. In applications involving musical signals, where processes such as breath or bow noise are important for synthesis realism, the residual can be independently modeled to account for these features [8, 10]. As discussed below, however, the residual also contains artifacts related to time-localized events in the original signal.

Noise-based models of the residual are not well-suited for representing such features. To achieve perceptually accurate modeling in a partials-plus-residual framework, then, it is necessary to modify the sinusoidal model so as to reduce these reconstruction artifacts.

### 2.2. Reconstruction Artifacts

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, on the other hand, the frequency resolution is degraded such that estimation of sinusoidal components becomes difficult. Thus, the sinusoidal model is governed by the same basic tradeoffs as any time-frequency representation.

In compact models, limitations in time-frequency resolution tend to result in artifacts in the signal reconstruction; in turn, the analysis-synthesis process yields a nonzero residual. The components of this residual arise both due to errors made by the analysis as well as shortcomings of the particular model. In the sinusoidal model, such errors occur if the original signal does not behave in the manner assumed by the line tracking and parameter interpolation used in the synthesis. Then, the residual contains such model artifacts in addition to the noiselike processes discussed above. While the algorithm to be discussed is useful for reducing a variety of artifacts, the specific artifact that will be considered here is pre-echo in the reconstruction of signal onsets; this is of interest since high-quality music synthesis requires preservation of note attacks [8].

Pre-echo in the sinusoidal model is caused by the following mechanism. Before the signal onset, there is an analysis frame in which the signal is not present and no partials are found. Thereafter, various partials are identified in the frame in which the onset occurs; the line tracking interprets these as new partials and connects them to zero-amplitude partials in the previous frame using the interpolation models [7]. This results in a smooth amplitude envelope for each partial instead of a sharp onset. The synthesis pre-echo in the case of linear amplitude interpolation is shown in Fig. 2 for two simple signals, and in Fig. 5 for a saxophone note; note the artifacts in the residuals.

### 2.3. Atomic Interpretation

In the sinusoidal model, the reconstruction process corresponds to a concatenation of nonoverlapping synthesis frames. Each frame consists of a sum of time-limited partials which can be interpreted as time-frequency atoms constructed from the sinusoidal parameters derived by the analysis. The sinusoidal model can thus be written as

$$x[n] \;\approx\; \sum_j \sum_q g_{q,j}(t) \;=\; \sum_j \sum_q A_{q,j}(t) \cos \Phi_{q,j}(t), \quad (2)$$

where $j$ is a synthesis frame index and the functions $A_{q,j}(t)$ and $\Phi_{q,j}(t)$ are time-limited to the $j$-th frame. The atomic amplitude and phase functions are dictated by the parameter interpolation; in the typical case of first order amplitude interpolation and third order phase, the sinusoidal model computes a signal decomposition in terms of atoms with linear amplitude and cubic phase.

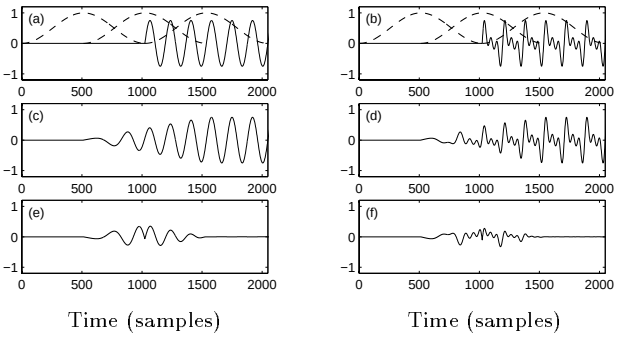


Figure 2: Pre-echo in the sinusoidal model. Modeling the onsets of (a) a sinusoid and (b) a simple harmonic signal using the analysis windows depicted leads to (c,d) delocalized reconstruction and (e,f) residuals with artifacts.

The atomic interpretation of the sinusoidal model indicates why it has difficulties representing transients. Each atom in the decomposition spans a fixed-length synthesis frame; this fixed resolution results in dispersion of events that occur on short time scales, where the spreading is caused both by the use of a long analysis window and the accompanying long stride. To model nonstationary signals effectively, it is necessary to admit atoms with a variety of time supports into the decomposition. Such multiresolution sinusoidal modeling can be achieved by two methods: filter bank approaches, wherein subband filtering is followed by sinusoidal modeling of the channel signals with long frames for low-frequency bands and short frames for high-frequency bands [1, 11]; or, segmentation methods in which the frame size is varied based on the signal characteristics. This paper focuses on the latter, 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.

## 3. ADAPTIVE SEGMENTATION

Since the sinusoidal model is inherently parametric and approximate, the analysis windows do not have to satisfy an overlap-add property as in the perfect reconstruction STFT [1]. This allows for flexibility in the window design and furthermore justifies the use of time-varying windows such as those depicted in Fig. 3; note that the synthesis frames are defined by the centers of the analysis windows. In this section, algorithms for signal-adaptive derivation of such multiresolution segmentations are discussed.

In the sinusoidal model, the number of partials is a main factor in the rate-distortion tradeoff. In a frame-wise sense, this model order imposes a constraint on the maximum number of partials incorporated in a given frame. In this consideration, a fixed model order will be assumed, meaning that the analysis attempts to identify the same number of partials in short segments as in long segments. This constraint simplifies the line tracking process.

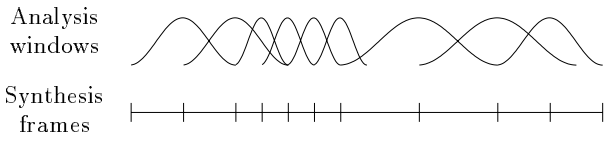


Figure 3: Analysis windows and synthesis frames in a multiresolution sinusoidal model with adaptive segmentation.

### 3.1. Global Search

Given a fixed model order, the problem at hand is that of finding a segmentation of the signal that minimizes some metric such as the mean-squared error between the original and the reconstruction. This optimization procedure can be carried out by an exhaustive global search in which each possible segmentation is considered in turn. Since a model must be evaluated on each segment in each segmentation, a simple estimate of the computational cost of a global search can be arrived at by counting the total number of segments in all of the possible segmentations. Denoting the length of the signal by $N\epsilon$ and using the set of segment sizes $\lambda = \{\epsilon, 2\epsilon, 3\epsilon, \ldots, L\epsilon\}$, i.e. integer multiples of a cell size $\epsilon$, this enumeration of segments is governed by an exponential dependence on the signal length:

$$\mathcal{C}_{\text{global}} \propto 2^N. \tag{3}$$

While this is unrealistic as a measure of computational cost in that it assumes an equal cost of model evaluation on segments of different length, such an enumeration does provide a basic indication that a global search is computationally prohibitive for long signals.

### 3.2. Dynamic Segmentation

If the optimization metric is additive and independent on disjoint segments, an exhaustive evaluation of all possible segmentations involves redundant computation since a given segment appears in many different segmentations. This redundancy can be removed by formulating the computation as a dynamic program, which is is based on treating the time span of the signal as a concatenation of cells [12]. The boundaries between these cells will be referred to as markers; these markers serve as nodes in the dynamic program. Because of the integer construction of the allowable lengths in the set $\lambda$, the segment boundaries in any candidate segmentation align with some of these markers.

The operation of the dynamic algorithm for the case $L = 3$ is depicted in Fig. 4; the expression $D_{ab}$ is used in the figure to represent 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. 3). When the end of the signal is reached, the globally 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

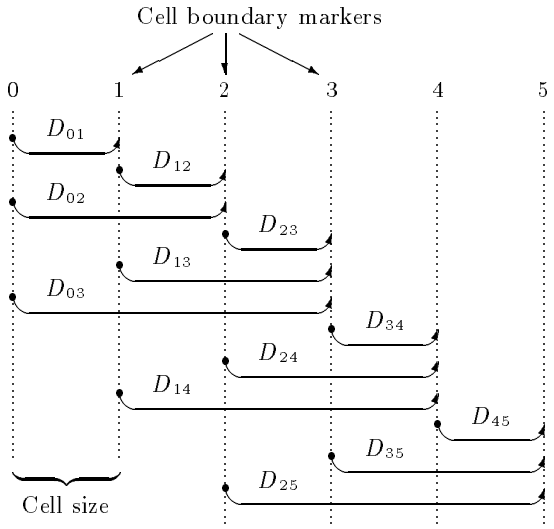


Figure 4: Depiction of a dynamic algorithm for signal segmentation. Note the regularity of the computation after the startup; the cost grows linearly with the signal length.

marker; this is done by analyzing the signal with a window based on the segment length (as in Fig. 3), synthesizing the model based on 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. As suggested in Fig. 4, the total number of segments on which models are computed has a linear dependence on the length of the signal:

$$\mathcal{C}_{\text{dynamic}} \propto LN. \tag{4}$$

For more details on dynamic programs, the reader is referred to [13] or other texts on digital communication; the widely used Viterbi algorithm for sequence detection is an example of a dynamic program.

In signal modeling, it is generally important to achieve continuity at synthesis frame boundaries. An example of this is the use of lapped transforms in image processing to reduce blocking artifacts caused by quantization. In the sinusoidal model, the continuity problem is resolved by using overlapping analysis frames and parameter interpolation functions that match the frequency and phase of the reconstructed partials at the synthesis frame boundaries. The caveat here is that if such overlap methods are used in the dynamic segmentation algorithm, the modeling metrics for adjacent segments are not strictly independent. As a result, the algorithm is not guaranteed to arrive at the absolute optimal segmentation that a global search would find. In practice, however, this dependency does not degrade the performance of the algorithm [1, 3]; an exemplifying model of a saxophone onset is given in Fig. 5.

### 3.3. Heuristic Segmentation

As an alternative to global optimization via dynamic segmentation, a signal-adaptive segmentation can be derived in a forward manner by the following heuristic approach.

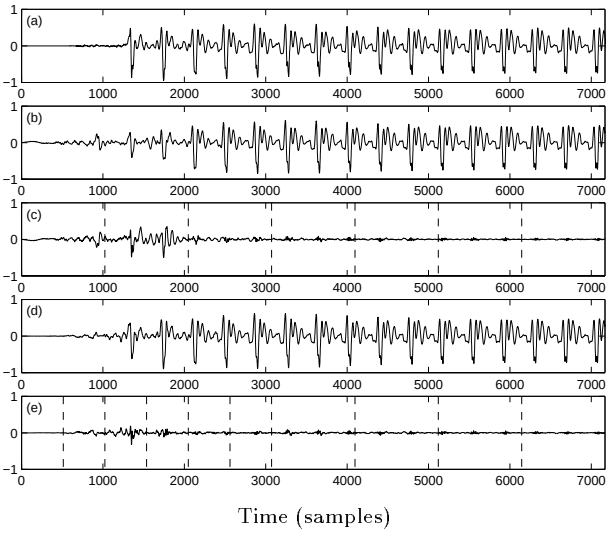


Figure 5: Modeling example: (a) the onset of a saxophone note, (b) a fixed resolution reconstruction and (c) the corresponding residual, where the dashed lines indicate the synthesis frames; and, (d) a multiresolution model derived by dynamic segmentation and (e) its residual, which exhibits fewer artifacts.

At marker $a$ in the signal, the weighted metrics $D_{ab}/(b-a)$ are evaluated for $b \in \{a+1, a+2, \ldots, a+L\}$; the segment length is then chosen according to the value $\hat{b}$ which minimizes the weighted metric, and the algorithm is continued from the new starting point $a = \hat{b}$. For such an approach, the number of segments considered is signal-dependent, so the computational cost can only be formulated in an average sense; assuming that the average size of the chosen segments is the mean of the set $\lambda$, the segment enumeration is governed by a linear dependence on the signal length:

$$\mathcal{C}_{\text{forward}} \propto 2N. \tag{5}$$

This heuristic algorithm can achieve similar results as the dynamic approach, but it is not as robust since the segmentation decisions are based on greedy local optimization of the modeling metric [1]. In time-critical applications, the reduced cost with respect to dynamic segmentation may merit the accompanying decrease in model accuracy.

## 4. CONCLUSION

It has been demonstrated that using an adaptive segmentation in the sinusoidal model reduces reconstruction artifacts such as pre-echo. In audio signal modeling, this leads to an analysis-synthesis residual that can be more effectively described using critical-band noise shaping as in [10]. It should be noted that similar localization of onsets can be achieved by independently modeling the signal envelope [9]; in this adaptive segmentation approach, however, a uniform sinusoidal parameterization is maintained, which is advantageous for synthesis computation. The dynamic segmentation algorithm discussed herein can be used to derive signal-adaptive models that are nearly optimal with respect to segment-wise additive cost measures such as rate-distortion.

## 5. ACKNOWLEDGMENTS

The author would like to thank Martin Vetterli for sparking this work and Paolo Prandoni for various insights. Also, the author is indebted to both Martin and Edward Lee for their support during the course of this research.

## 6. REFERENCES

[1] M. Goodwin. Adaptive Signal Models: Theory, Algorithms, and Audio Applications. PhD thesis, University of California at Berkeley, Fall 1997.

[2] R. Coifman and M. Wickerhauser. Entropy-based algorithms for best basis selection. IEEE Transactions on Information Theory, 38(2):713–718, March 1992.

[3] K. Ramchandran and M. Vetterli. Best wavelet packet bases in a rate-distortion sense. IEEE Transactions on Image Processing, 2(2):160–75, April 1993.

[4] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. IEEE Transactions on Signal Processing, 41(12):3397–3415, December 1993.

[5] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit, February 1996. Available at ftp.playfair.stanford.edu.

[6] K. Brandenburg and G. Stoll. ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio. Journal of the Audio Engineering Society, 42(10):780–791, October 1994.

[7] 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.

[8] 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.

[9] E. B. George and M. J. T. Smith. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. IEEE Transactions on Speech and Audio Processing, 5(5):389–406, September 1997.

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

[11] 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.

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

[13] E. A. Lee and D. G. Messerschmitt. Digital Communication. Kluwer Academic Publishers, Boston, 1988.