View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Springer - Publisher Connector

EURASIP Journal on Applied Signal Processing 2003:1, 15–20
© 2003 Hindawi Publishing Corporation

# Sinusoidal Analysis-Synthesis of Audio Using Perceptual Criteria

**Ted Painter**
*Intel Corporation HD2-230, Handheld Computing Division, 77 Reed Road, Hudson, MA 01749, USA*
*Email: ted.painter@intel.com*

**Andreas Spanias**
*Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287-7206, USA*
*Email: spanias@asu.edu*

*Received 23 May 2002 and in revised form 4 November 2002*

This paper presents a new method for the selection of sinusoidal components for use in compact representations of narrowband audio. The method consists of ranking and selecting the most perceptually relevant sinusoids. The idea behind the method is to maximize the matching between the auditory excitation pattern associated with the original signal and the corresponding auditory excitation pattern associated with the modeled signal that is being represented by a small set of sinusoidal parameters. The proposed component-selection methodology is shown to outperform the maximum signal-to-mask ratio selection strategy in terms of subjective quality.

**Keywords and phrases:** audio-coding, sinusoidal synthesis, audio coders.

## 1. INTRODUCTION

Sinusoidal modeling of speech and audio has been successfully used in several speech-coding applications such as the sinusoidal transform coder [1], the multiband excitation coder by [2], as well as in some of the recent wideband multiresolution audio applications [3]. One of the most recent enhancements of the sinusoidal model is the introduction of a new method that handles not only the harmonic aspects of the signal but also its broadband and transient components. This new form of adaptive signal representation is called the sines + transients + noise (STN) model [4].

The paper presents a new method for the selection of sinusoids in hybrid (STN) sinusoidal modeling of audio. This consists of ranking and selecting the most perceptually relevant sinusoids. The method maximizes the matching between the excitation pattern associated with the signal and the corresponding pattern associated with the sinusoidal model. The new method is based on excitation similarity weighting (ESW). The reconstruction quality provided by ESW is compared against a quality benchmark established with the maximum signal-to-mask ratio (maximum SMR) methodology. The ESW component-selection methodology is shown to outperform the maximum SMR selection strategy in terms of both objective and subjective quality.

This method is inherently different than previously proposed methods that select components by either peak picking [5] or by harmonic constraints [1, 2]. In fact, the sinusoids chosen by ESW are generally neither harmonic nor maximum amplitude. The paper is organized as follows. In Section 2, the classical sinusoidal model is presented along with the STN extensions. Section 3 describes the ESW selection process and gives sample results. Section 4 gives our concluding remarks.

## 2. SINUSOIDAL ANALYSIS-SYNTHESIS

The classical sinusoidal model comprises an analysis-synthesis framework [5] that represents a signal $s(n)$ as the sum of a collection of $K$ sinusoids (*partials*) with time-varying frequencies, phases, and amplitudes, that is,

$$s(n) \approx \hat{s}(n) = \sum_{k=1}^{K} A_k(n) \cos\left(\omega_k(n)n + \phi_k(n)\right), \qquad (1)$$

where $A_k(n)$ represents the amplitude, $\omega_k(n)$ represents the instantaneous frequency, and $\phi_k(n)$ represents the instantaneous phase of the $k$th sinusoid. Estimation of parameters is typically accomplished by peak picking the short-time Fourier transform (STFT) [5]. In the synthesis stage, the

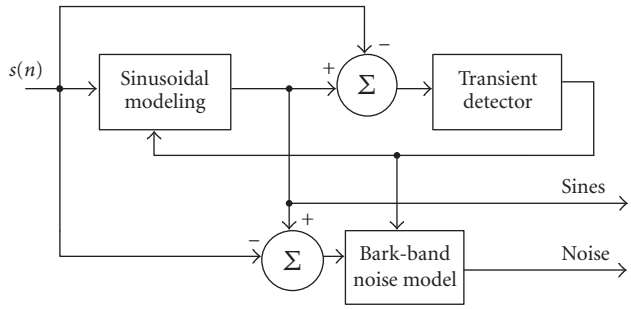



Figure 1: STN model.

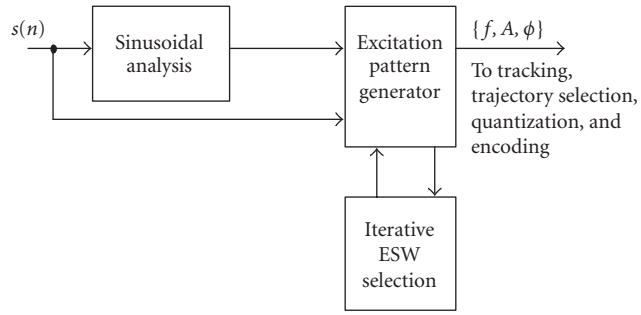



Figure 2: The ESW scheme.

model parameters are subjected to spectral line tracking and frame-to-frame amplitude and phase interpolation.

Although the basic sinusoidal model achieves efficient representation of harmonically structured signals, extensions to the basic model have also been proposed for signals containing nontonal energy [6]. The spectral modeling and synthesis system treats audio as the sum of $K$ sinusoids along with a stochastic component ($e_n$), that is,

$$s(n) \approx \hat{s}(n) = \sum_{k=1}^{K} A_k(n) \cos\left(\omega_k(n)n + \phi_k(n)\right) + e(n). \quad (2)$$

Although the sines + noise signal model gave improved performance, the addition of transient components giving rise to a three-part model consisting of STN [4, 7] (Figure 1) provides additional enhancements. In STN, sinusoidal modeling is applied to the input. Then, transients are detected via an energy threshold combined with a partial loudness edge detection scheme that operates on the sinusoidal modeling residual. The idea behind this system is to identify unmasked transients, while, at the same time, disregarding masked transients. Both masked and unmasked transients have the potential to trip the energy threshold detector, but masked transients will have a significantly lower impact on residual noise loudness than will unmasked transients. Standard time resolution is adequate for masked transients, at least in the low-rate coding scenario. Once the tonal and transient components have been analyzed, the residual of the sines + transients modeling procedure is captured by the Bark-band noise model [8, 9]. Although the methods proposed in this paper are concerned with sinusoidal model estimation, ultimately they can also be applied to optimize the STN model for a scalable audio-coding application.

## 3. COMPACT REPRESENTATION OF STN PARAMETERS

This section is concerned with the ranking and selection of perceptually relevant sinusoids on a compact set. We call this the ESW ranking and selection procedure. Whereas some of the current audio coders tend to choose maximum SMR components and therefore base the selection decision on the masked threshold, the ESW methodology seeks to maximize the matching between the excitation patterns evoked by the coded and original signals on a short-time basis.

In contrast to ESW, the maximum SMR selection criterion does not guarantee maximal matching between the modeled and the original excitation patterns [8]. The idea behind the ESW technique is to select sinusoids such that each new sinusoid added will provide a maximum incremental gain in matching between the auditory excitation pattern associated with the original signal and the auditory excitation pattern associated with the modeled signal. In order to accomplish this goal, an iterative process is proposed in which each sinusoid extracted during conventional analysis is assigned an excitation similarity weight. During each iteration, the sinusoid having the largest weight is added to the modeled representation. New sinusoids are accumulated until some constrain is exhausted, for example, a bit budget. The algorithm tends to converge as the number of modeled sinusoids increases. The ESW sinusoidal component-selection strategy (Figure 2) works as follows. First, a complete set of sinusoids is estimated using the STFT. Then, a reference excitation pattern is computed for the original signal in a manner similar to the method outlined in the description of PERCEVAL [10]. PERCEVAL is a software that was developed to evaluate audio signals corrupted by noise. This is based on a frequency-domain model that computes a basilar energy distribution in terms of *Mel* from a high-fidelity energy spectrum (0–20 kHz). This pattern may contain up to 2500 discrete excitation levels (0–2500 Mel) that correspond to assumed discrete detectors along the basilar membrane. A logarithmic function is applied to these energy values and a 2500-component basilar sensation vector (reference excitation pattern) is obtained. This reference excitation pattern is then used in conjunction with an iterative ranking procedure to select the sinusoids. The objective of the $k$th iteration is to extract from the candidate set the most perceptually salient sinusoid, given the previous $k-1$ selections. The method assumes that maximum perceptual salience is associated with the component able to affect the greatest improvement in matching between the excitation pattern associated with the original signal and the excitation pattern that is associated with the modeled signal. To select from the candidates during the $k$th iteration, a complete set

of candidate excitation patterns is computed, one for each of the patterns associated with the modeled signal containing the first $k-1$ selected sinusoids, as well as each of the candidates currently available. The candidate that minimizes the difference between the reference and the modeled excitation patterns is selected for the $k$th iteration. The resulting sinusoidal parameters of the best candidate are passed to the trajectory tracking and model pruning components. The core ESW calculation comprises an average difference calculation that operates on the reference and test excitation patterns. In particular, the average difference $\Delta_k$ between the original (reference) and the test patterns on the $k$th iteration is given by

$$\Delta_k = \frac{1}{D}\sum_{i=1}^{D}\left[E(i) - X_k(i)\right], \tag{3}$$

where $E(i)$ is the reference excitation pattern level (in dB), $X_k(i)$ is the level (in dB) of any of the candidate test excitation patterns on the $k$th iteration, and $D$ is the number of detectors. Therefore, for each pattern, the improvement in matching on the $k$th iteration for each candidate pattern $X_k(i)$ is given by

$$\Delta_k - \Delta_{k+1} = \frac{1}{D}\sum_{i=1}^{D}\left[X_{k+1}(i) - X_k(i)\right]. \tag{4}$$

The ESW technique computes the matching improvement for all candidate patterns during the $k$th iteration and selects the component that maximizes (4). Once the best candidate pattern $X_k^*(i)$ has been identified on the $k$th iteration (in the sense of maximizing (4)), an excitation similarity weight is assigned to the sinusoidal component that provided the maximum incremental matching improvement. The ESW assigned to the $k$th component is

$$\mathrm{ESW}_k = \Delta_{k-1} - \Delta_k. \tag{5}$$

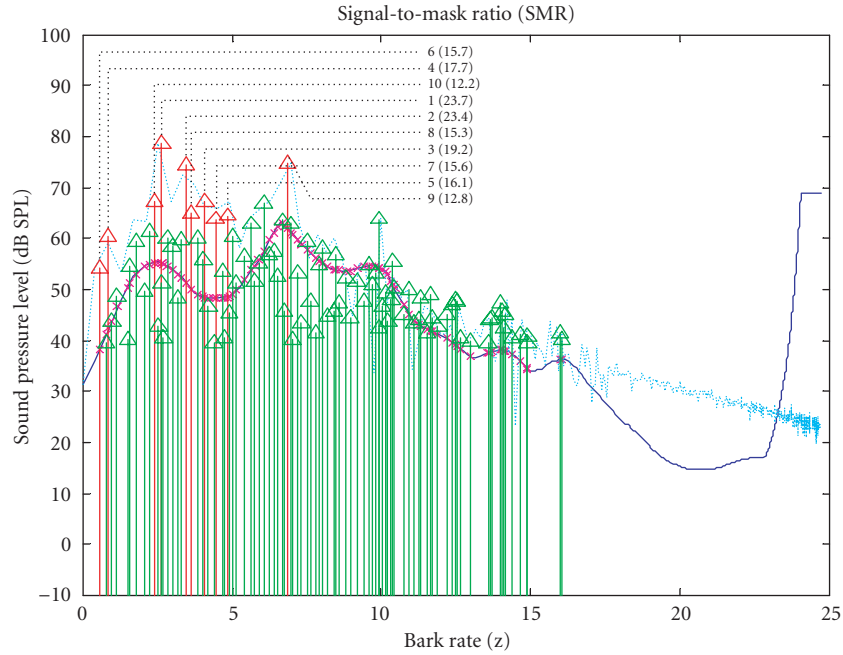
### 3.1. Comparison of ESW versus maximum SMR

For validation, the ESW component-selection and ranking scheme was compared against a reference maximum-SMR selection scheme over a diverse collection of audio program material. The ESW-based output samples generated from STN model parameters consistently outperformed the SMR-based audio samples in terms of both subjective informal listening tests and objective evaluations using the partial loudness model described earlier. We give here sample comparative results in graphical format for a selection of rock music that was judged to be spectrally complex and therefore challenging for a low-rate coding application. Figure 3 provides insight on how the ESW methodology selects components in contrast to the maximum SMR methodology. These comparative results (Figure 3) show a spectral view corresponding to 23 milliseconds of audio. The vertical arrows in both figure panels correspond to the complete set of sinusoids returned by classical sinusoidal analysis. The dashed line corresponds to a short-time spectral estimate (magnitude FFT) mapped to SPL, and the solid line corresponds to an estimate of the masked threshold generated by the MPEG-1 Psychoacoustic model 2.

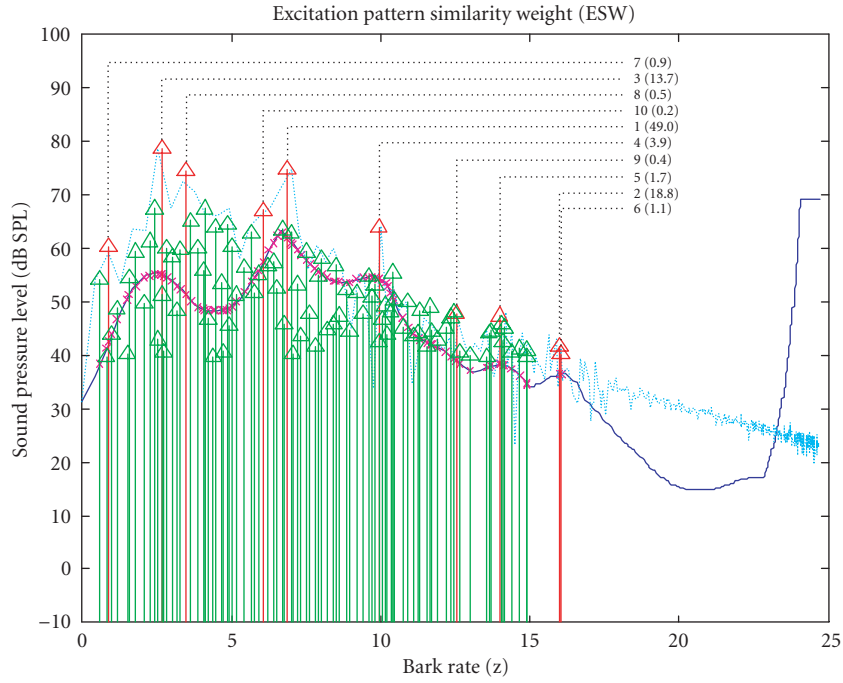
Sinusoids labeled in panel (a) of Figure 3 were selected on the basis of maximum SMR. Each of the selected sinusoids is labeled with its rank, one through ten, and its SMR in dB. It is clear from the figure that the ranking is in terms of descending SMR. This ranking directly corresponds to the currently popular method of sinusoid selection. Panel (b) of Figure 3 shows the selection process for the ESW methodology. In this figure, each of the ten selected sinusoids is labeled with its rank and ESW score (5).

A comparison of the figures reveals that the ESW method tends to choose sinusoids across the spectrum, whereas the maximum SMR method tends to choose sinusoids of higher energy that are clustered at lower frequencies. This trend was manifested across time in the given example and also across many musical selections. The second set of comparative results (Figure 4) shows the convergence trends for each selection methodology. In both panels of Figure 4, the reference excitation pattern (same in both) is labeled with an arrow. The reference pattern corresponds to the internal representation that is associated with the original short-time spectral slice shown in Figure 3.

The second solid line labeled in each panel of Figure 4 shows the final modeled excitation pattern, that is, the pattern generated by the subset of sinusoids selected during the SMR and ESW pruning processes illustrated in Figure 3. Finally, the set of dashed lines in each figure (Figure 4) illustrate the best excitation patterns generated by the sets of sinusoids selected during iterations 1 through 10. In addition, each panel is labeled with the average detector difference in dB that is present at the conclusion of the selection process. Panel (a) of Figure 4 clearly shows that the SMR method tends to cluster its estimates of the most important sinusoids in the low-frequency regions. Inspection of the final maximum SMR modeled pattern demonstrates how this strategy handicaps the excitation pattern matching. Substantial gaps in excitation pattern matching occur at high frequencies, where the SMRs tend to be quite small. As a result, after choosing ten sinusoids, the average dB difference between the reference and modeled patterns after 10 iterations exceeds 30 dB. Given Zwicker's 1 dB difference detection criterion, it is likely that this short-time segment will not resemble the original sound very closely. In contrast, panel (b) shows that the ESW method tends to push the modeled excitation pattern very close to the reference pattern across the entire spectrum (Bark rate shown), such that the final ESW pattern creates an average detector difference of only 7.7 dB. The demonstrated trend of dramatically improved matching achieved by ESW relative to maximum SMR in this example for very few sinusoidal components generalizes across time for this selection and across musical selections to other samples as well. The significant improvement in pattern matching was observed for a diverse set of music samples and, perhaps most importantly, informal subjective quality evaluations confirmed the expected improvements in output quality associated with the ESW selection scheme.




(a)




(b)

Figure 3: (a) Comparison of sinusoidal pruning methodologies for the maximum SMR method. (b) Comparison of sinusoidal pruning methodologies for the maximum ESW method.

The final set of comparative results (Figure 5) shows the time-domain residuals associated with each component-selection strategy, and then provides a view of the partial loudness measured in sones for each residual across time. The results are for a compact set of 10 out of more than 200 sinusoids on each frame. A dashed line on each of the loudness plots represents the time-averaged loudness over the entire record. Although it is difficult to detect significant

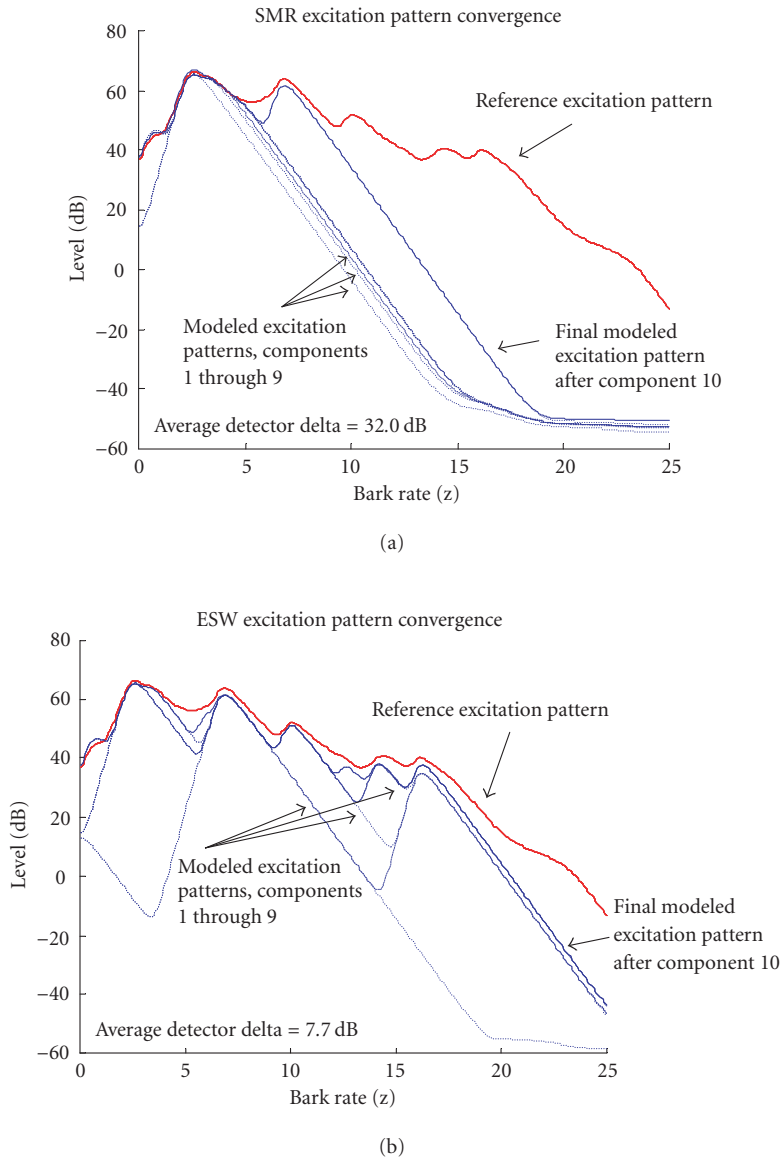



FIGURE 4: (a) Excitation pattern convergence for the spectral slices shown in Figure 3 for the maximum SMR method. (b) Excitation pattern convergence for the spectral slices shown in Figure 3 for the maximum ESW method.

differences in the time-domain residuals, comparison of the partial loudness results shows a significant difference. Note that the SMR method creates a residual with an average partial loudness of 5.3 sones, with maxima in the vicinity of 7 to 8 sones. In contrast, the ESW method is characterized by an average partial loudness of only 3.5 sones, with worst-case values in the vicinity of only 5 sones.

## 4. CONCLUDING REMARKS

The results presented in Figures 4 and 5 clearly suggest that the ESW sinusoidal component-selection strategy tends to outperform the now popular maximum SMR method on compact sets of sinusoidal parameters. This implied result was verified through extensive informal subjective listening tests across a diverse set of program material. The results suggest that the realized enhancements in sinusoidal selection lead to several methods for achieving compact representations of ESW-ranked sinusoidal components. Perhaps the most intuitive is that of thresholding on the basis of a minimum ESW. All sinusoids below the minimum ESW can be discarded. We note that the ESW method provided improvements in cases where the number of sinusoids selected was small. For large sets of sinusoids, we anticipate that a combined ESW/SMR-selection process will have to be developed.

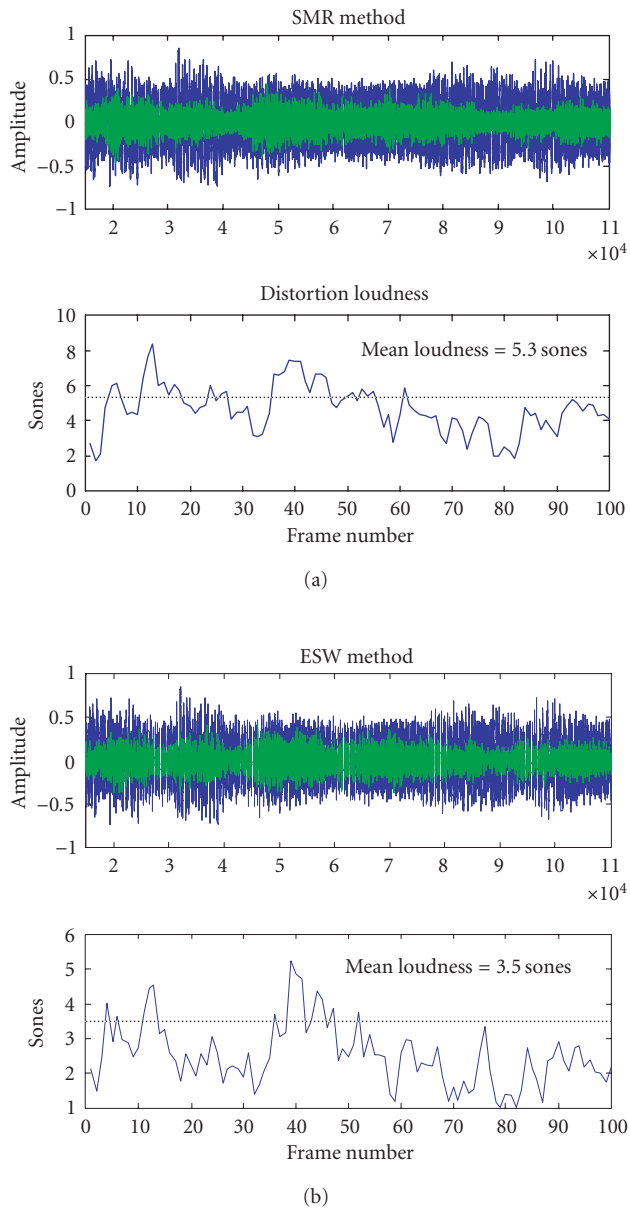



FIGURE 5: (a) Time-domain residuals and their partial loudness for the maximum SMR method. (b) Time-domain residuals and their partial loudness for the maximum ESW method.

## ACKNOWLEDGMENTS

Part of this work was presented at MMSP-02. Research performed at Arizona State University as part of a Ph.D. thesis of Dr. Painter. Paper was invited by J. Dugaley.

## REFERENCES

[1] R. McAulay and T. Quateri, "The sinusoidal transform coder at 2400 b/s," in *Military Communications Conference*, San Diego, Calif, USA, October 1992.

[2] D. Griffin and J. Lim, "Multiband excitation vocoder," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, no. 8, pp. 1223–1235, 1988.

[3] D. V. Anderson, "Speech analysis and coding using a multiresolution sinusoidal transform," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 1045–1048, Salt Lake City, Utah, USA, May 1996.

[4] S. Levine and J. Smith, "A Sines+Transients+Noise Audio representation for data compression and time/pitch scale modifications," in *Proc. Audio Engineering Society 105th Int. Conv*, San Francisco, Calif, USA, preprint #4781, September 1998.

[5] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[6] X. Serra, *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*, Ph.D. thesis, Stanford University, Stanford, Calif, USA, 1989.

[7] T. Verma, S. Levine, and T. Meng, "Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals," in *International Computer Music Conference*, Thessaloniki, Greece, September 1997.

[8] T. Painter, *Scalable perceptual audio coding with a hybrid adaptive sinusoidal signal model*, Ph.D. thesis, Arizona State University, Tempe, Ariz, USA, June 2000.

[9] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–513, 2000.

[10] B. Paillard, P. Mabilleau, S. Morisette, and J. Soumagne, "PERCEVAL: Perceptual evaluation of the quality of audio signals," *J. Audio Eng. Soc.*, vol. 40, no. 1/2, pp. 21–31, 1992.

**Ted Painter** received his Ph.D. in electrical engineering from Arizona State University in August 2000. He is currently a software Architect in the handheld Computing Division of Intel Corporation in Hudson, Mass, USA. He specializes in the development of high-performance multimedia software for next generation portable and wireless computing devices. His primary interests are in speech and audio signal processing, perceptual coding, and psychoacoustics. Ted Painter is corecipient of the 2002 IEEE Donald G. Fink Best Paper Award along with Andreas Spanias.

**Andreas Spanias** is a Professor in the Department of Electrical Engineering at Arizona State University (ASU). His research interests are in the areas of adaptive signal processing, speech/audio and multimedia signal processing. While at ASU, he developed and taught courses in digital signal processing (DSP), adaptive signal processing, and speech-coding. He is a senior member of the IEEE and has served as a member in the Technical Committee on Statistical Signal and Array Processing of the IEEE Signal Processing society. He has also served as an Associate Editor of the IEEE Transactions on Signal Processing and as a General Cochair of the 1999 International Conference on Acoustics Speech and Signal Processing (ICASSP-99) in Phoenix. He is currently the IEEE Signal Processing Vice-President for Conferences and the Chair of the Conference Board. He is also a member of the IEEE Signal Processing Executive Committee and an Associate Editor of the IEEE Signal Processing Letters. Andreas Spanias has served as the Chair (for four years) of the IEEE Communications and Signal Processing Chapter in Phoenix, and is a member of Eta Kappa Nu and Sigma Xi. Andreas Spanias is a corecipient of the 2002 IEEE Donald G. Fink Best Paper Award along with Ted Painter.