

Efficient Multi-resolution Sinusoidal Modeling

David V. Anderson and Mark A. Clements

Abstract— Multi-resolution sinusoidal modeling is a method of representing audio signals as a sum of discrete sinusoids having various durations, and time-varying frequencies, phases, and amplitudes. This paper discusses the motivation for the multi-resolution sinusoidal model (MRSM) and various structures used to compute it. A new filter-bank analysis structure is presented that is simpler than previous methods. Finally, a matching-pursuit analysis method is presented for more efficiently modeling time-varying signals.

I. INTRODUCTION

A system that modifies speech and audio signals for human listeners requires a suitable signal representation on which to work. Common representations for general speech or audio signals include the Fourier transform of the signal, auto-regressive moving-average models, sub-band signals or wavelet coefficients, and the time samples themselves. Each of these representations emphasizes certain signal characteristics and they all represent trade-offs in the time resolution and frequency resolution at which the signal is portrayed.

The *sinusoidal model* (SM), as developed by Quatieri and McAulay [1], for speech provides a way of representing an audio signal in a sparse manner as a sum of discrete sinusoids [1]. The sinusoids in the SM are continually varying in amplitude, frequency, and phase; unlike the fixed, unconnected sinusoidal basis functions used in short-time Fourier analysis. The SM is more flexible than most other signal models in that it is easy to represent the time evolution of each frequency component in a signal using the SM. Representing a signal using the SM is a lossy process, i.e., in most cases the signal cannot be reproduced exactly from the model representation; however, the SM can be used to reproduce speech and audio signals which are nearly indistinguishable from the originals.

The SM was originally designed for use in speech coding, and it can be used to achieve large coding gains. However, it was soon discovered that the parametric method of representing speech in terms of discrete sinusoids was useful in speech modification. The SM has since been used for time-scale modification of speech [2], [3], [4], [5], music modeling [6], [7], [8], [9], pitch modification of speech [2], [10], co-channel interference suppression [11], [12], background noise suppression [13], speech synthesis [14], [10], singing voice synthesis [10], music synthesis and modification [15], [9], hearing compensation [16], and peak-to-RMS ratio reduction [17].

One of the problems with the sinusoidal model is that it does not do well at modeling wideband audio. This is not seem to be due to a fundamental problem with the sinusoidal model itself, but with the analysis methods used.

This work reviews some recent innovations to the SM in extending it to use a multi-resolution analysis. Several multi-resolution analysis methods are reviewed and several new methods are presented.

Georgia Institute of Technology, Atlanta, Georgia, USA. E-mail: david.anderson@ece.gatech.edu, mark.clements@ece.gatech.edu

II. SINUSOIDAL MODEL

A. Sinusoidal Model for Speech

A great strength of the SM is that it is well matched to speech signals and can represent the important auditory characteristics with very few parameters. These parameters can then be manipulated to modify certain signal characteristics.

The SM has an added advantage over most other speech signal representations in that it more accurately models a broader class of signals. The SM has been successfully used with multi-talker waveforms, music, speech in a musical background, and other signals [1], [18].

A.1 Voiced Speech

The sinusoidal transform is based on a speech production model in which the excitation waveform is a sum of sinusoids of the form

$$e(t) = \sum_{\ell=1}^{L(t)} a_{\ell}(t) \cos [\Omega_{\ell}(t)]. \quad (1)$$

The instantaneous phase, $\Omega_{\ell}(t)$, of the ℓ^{th} component is related to its instantaneous frequency, $\omega_{\ell}(t)$, by

$$\Omega_{\ell}(t) = \int_0^t \omega_{\ell}(\xi) d\xi + \phi_{\ell}. \quad (2)$$

Note that the sinusoids vary in amplitude and frequency as a function of time and that the number of sinusoids used, $L(t)$, also varies with time.

The excitation is then subject to the combined glottal waveform and vocal tract response. This system response is also time-varying and may be represented by

$$H(\omega; t) = M(\omega; t) \exp [j\psi(\omega; t)]. \quad (3)$$

It is convenient to track each component of the excitation signal separately. Therefore, for the ℓ^{th} frequency component we express the associated vocal tract magnitude and phase as

$$M_{\ell}(t) = M[\omega_{\ell}(t); t] \quad (4)$$

and

$$\psi_{\ell}(t) = \psi[\omega_{\ell}(t); t] \quad (5)$$

respectively. The speech signal which results from passing the excitation signal through the vocal tract is then

$$s(t) = \sum_{\ell=1}^{L(t)} A_{\ell}(t) \exp \theta_{\ell}(t) \quad (6)$$

where

$$A_{\ell}(t) = M_{\ell}(t) a_{\ell}(t) \quad (7)$$

and

$$\theta_{\ell}(t) = \int_0^t \omega_{\ell}(\xi) d\xi + \phi_{\ell} + \psi_{\ell}(t). \quad (8)$$

The $\psi_{\ell}(t)$ and ϕ_{ℓ} are often lumped together and phase expression becomes

$$\theta_{\ell}(t) = \int_0^t \omega_{\ell}(\xi) d\xi + \hat{\psi}_{\ell}(t). \quad (9)$$

Each

A.2 Unvoiced Speech

One shortcoming of the sinusoidal model for speech is that it is best suited for voiced speech. However, if there are enough sinusoids, and if they are “close enough” in frequency, it is possible to represent the noise-like unvoiced speech adequately [1]. This is one of the issues addressed later in this work. It is also possible to represent other signals such as multi-talker waveforms and music by increasing the number of sinusoids used and/or by using one of the alternate methods mentioned below.

A.3 Analysis

In practice the DFT is used to estimate the parameters in equations 6 and 9. Parameters are determined by finding peaks in the DFT spectrum of the zero-padded signal segment and noting the corresponding frequency, amplitude, and phase. The estimates are typically updated every 5-20 msec. To improve the accuracy of the peak-picking process, the segment of interest is windowed with a window whose length is at least 2-3 periods of the frequencies of interest. For high fidelity reproduction and in situations where there is no single pitch (as in multiple talkers or music) or with additive noise an even longer window is desirable.

A.4 Synthesis

The individual sinusoids associated with each segment are called *partials*. Since the SM parameters are only estimated once per segment, it is necessary to match the partials in adjacent segment to form a continuous sinusoid. Interpolation is then used to obtain a smooth evolution of the sinusoid between segments. The amplitude is usually interpolated linearly and the frequency and phase are treated together and interpolated so that they are “maximally smooth” [18].

Segment k of the discrete-time output of the sinusoidal model is expressed as

$$\tilde{x}^{(k)}[n] = \sum_{\ell=0}^{L^{(k)}-1} A_{\ell}^{(k)}[n] \cos\left(\phi_{\ell}^{(k)}[n]\right) \quad (10)$$

where the following terms are defined:

- $L^{(k)}$ the number of sinusoidal partials,
- ℓ the index to the sinusoidal partial,
- $A_{\ell}^{(k)}[n]$ the interpolated amplitude of the partial, and
- $\phi_{\ell}^{(k)}[n]$ the interpolated instantaneous phase value of the partial.

The final SM output signal for segment k is then given by summing the appropriate signal blocks, $\tilde{x}^{(k)}[n]$, as

$$\tilde{x}[n] = \sum_k \tilde{x}^{(k)} \left[n - n_0^{(k)} \right] \quad (11)$$

where $n_0^{(k)}$ is the time index at the beginning of segment k .

It is also possible to perform the synthesis using an overlap/add technique. Overlap/add synthesis does not do any explicit matching of tracks between segments; the interpolation between segments is performed by windowing the synthesized time waveform with a triangular window that is at least twice the segment length and overlapping and adding with the previous segment(s). One problem with the overlap/add technique is that it does not preserve waveform onsets as well as the matched-partial synthesis. Another problem is that the overlap/add method of interpolation can result in phase cancellations. This method works best if the SM parameters are estimated fairly often, at least every 10 msec.

A.5 Variations

There are many variations on the basic sinusoidal transform as described above. These include coding a residual separately [15], [19], [20], doing an analysis-by-synthesis or matching pursuit analysis [6], [21], and using harmonic signal plus noise models [20], [22], [23].

B. Multi-Resolution Sinusoidal Model

B.1 Motivation—Perceptual Considerations

The primary perceptual characteristic of human audition utilized by the SM is simultaneous auditory *masking*. Simultaneous masking describes the tendency of intense sounds or peaks in the short-term spectrum to mask out or hide softer sounds that are close in frequency [24]. Thus, it is possible to retain only the spectral peaks while the less intense sounds that are adjacent to peaks may be ignored without any perceivable change in the audio signal. The SM does this by encoding only the peaks, reconstructing the signal by representing them as sinusoids with smoothly varying phase and amplitude [25].

In addition to auditory masking, there are other perceptual or psychoacoustic phenomena which may be exploited in the design of an audio coder. One of these is *frequency resolution* in the human auditory system. Resolution in perception is often measured in terms of *just noticeable differences* (JNDs). Frequency JNDs are approximately logarithmic in frequency—the ear can resolve lower frequencies more closely than it can resolve high frequencies.

Another perceptual phenomenon that often plays a role in audio signal processing is temporal masking. Temporal masking occurs when an intense sound hides or masks less intense sounds *preceding* or *following* it. An example commonly encountered in signal compression is the larger quantization noise preceding a transient in a segment. In sinusoidal modeling there is often a pre-echo generated when the amplitudes of the partials are linearly interpolated between segments. Ideally, the segments will be close enough together so that the pre-echo is masked.

C. MRSM Motivation

The *multi-resolution sinusoidal model* (MRSM) offers improvements over the regular SM by addressing the frequency resolution aspect of auditory perception.

D. A Multitude of MRSM Analysis Methods

The input to the MRSM analysis stage is an audio signal, the output is a list of frequencies, amplitudes, phases, and durations of sinusoids which model the signal.

There are various ways to perform the MRSM analysis which will be discussed below. The goal of each of these methods is to produce an estimate of the sinusoids which represent the signal, with the high frequency sinusoids being updated frequently and the low frequency sinusoids updated less often but with better frequency resolution as mentioned above.

The one method discussed estimates the sinusoids using only DFTs of various lengths. The other three of the methods discussed use filter-banks to break the signal into sub-bands which are then analyzed to estimate the sinusoids.

D.1 Direct DFT Based MRSM Analysis

One method of performing MRSM analysis is to directly use DFTs (using FFTs) of different lengths [26]. The idea here is to “manually” tile the time-frequency plane as shown in Figure 1. This method suffers from two difficulties. First, the computational complexity is very high. Second, it is desirable to use

windowed DFTs, but then the bases used don't completely cover the space [27].¹

The peaks are encoded by recording their calculated amplitudes, phases, frequencies, and durations. Duration is proportional to the length of the DFT used to estimate the peak; when wavelet based analysis is used it corresponds to sample rate of the sub-band in which the peak was detected. If N is the duration of a sinusoid in the highest frequency band and L is the length of the shortest DFT, then the longer DFTs will have lengths $L2^d$, $d = 0, 1, 2, \dots$, and the corresponding sinusoids will have durations of $N2^d$.

Fig. 1. Discrete wavelet tiling of the time–frequency plane for a four band case

D.2 Wavelet Based MRSB Analysis

Initial attempts at MRSB analysis were built around a DFT analysis on the outputs of wavelet-like filter banks [26], [28]. It has been suggested that wavelets are particularly good at modeling the frequency response of the human auditory system [29]. Wavelets closely approximate the higher resolution in frequency exhibited by the human auditory system at lower frequencies, and the higher resolution in time exhibited by the human auditory system at higher frequencies. This MRSB analysis method uses wavelet-like analysis to provide a coarse frequency separation, then fine spectral determination is accomplished with the DFT.

The wavelet based MRSB analysis introduces aliasing that would normally be cancelled in a synthesis bank; however, when calculating the MRSB the aliasing must be canceled “manually” [26]. This causes it also to have high computation requirements.

Fig. 2. An analysis bank for the MRSB as used by Levine [30] (shown with the synthesis bank). This filter bank is oversampled in the high frequency band and the low frequency cutoff is chosen to be below $\frac{\pi}{2}$ so that there is little or no aliasing.

D.3 Oversampled Filter Bank MRSB Analysis I

A better method for MRSB analysis is based on oversampled filter-banks [30]. Instead of using critically sampled filter banks and then canceling aliased components, as with the wavelet method, the oversampled filter-bank method can eliminate or severely reduce aliasing in the frequencies of interest [31]. Thus, with good filter design, the aliasing cancellation requirement can be dropped. The outputs of the filter-bank are oversampled by a factor of two; therefore, the complexity growth is not as severe as with the multiple DFT method.

Levine [30] proposed using a tree structured filter-bank with each branch as shown in Figure 2. This structure was originally proposed by Fliege [31] as a modification to the Laplacian pyramid often used in multi-rate image processing. Filters $H_d(z)$ and $H_i(z)$ are half-band low-pass filters and they may be chosen so that $H_d(z) = H_i(z)$. Filter $H_b(z)$ eliminates, or attenuates to an arbitrary degree, aliased components introduced by the decimation following $H_d(z)$. Elimination of the aliasing is accomplished by choosing $H_b(z)$ so that the aliased portion of the signal is in the stop band of $H_b(z)$ as shown in the top

¹In other words, selecting bins from different length windowed DFTs to “tile” the time–frequency does not permit perfect reconstruction.

of Figure 5. The synthesis bank consists only of an up-sampler and anti-imaging filter for the low-pass filter branch.

Note that the filter bank is a perfect reconstructing filter bank regardless of the choice for $H_b(z)$. However, when using the filter bank in sinusoidal modeling, the synthesis bank is not used, so the perfect reconstructing aspect of the filter bank is moot.

The filter bank is implemented in a tree fashion with the analysis filter outputs being fed into a sinusoidal model analysis blocks (see Figure 3).

Fig. 3. The general analysis structure used for the MRSB. The sinusoidal model parameters are extracted from the sub-band signals. The sinusoidal model analysis window length is the same for each Sin. Mod. block but the effective window length is longer for the lower frequency bands.

The oversampled filter bank analysis approach is useful because it eliminates the complexity growth of the multiple DFT method and the aliasing associated with critically sampled filter banks (wavelets). However, there are some disadvantages to the filter bank used by Levine. The filter cut-off frequency, f_c , of $H_b(z)$ must satisfy

$$f_c < \pi - (f_{\tau_b} + f_{\tau_d}) \quad (12)$$

where f_{τ_b} and f_{τ_d} are the transition bandwidths of $H_b(z)$ and $H_d(z)$ respectively. Another disadvantage of the filter bank structure introduced above is that it is designed for use as a perfect reconstructing filter bank and it is more complex than needed.

D.4 Oversampled Filter Bank MRSB Analysis II

Since the synthesis portion of the filter bank is not used and the idea of perfect reconstruction is obliterated by the operation of sinusoidal modeling, the filter bank structure may be simplified. A better structure is shown in Figure 4. In this filter bank, $H_0(z)$ is chosen to have a cut-off less than $\frac{\pi}{2}$ such that $H_0(z) \approx 0$ for $\frac{\pi}{2} \leq \Omega \leq \pi$, where $z = e^{j\Omega}$. This structure is much simpler but, after the decimation, it can no longer be part of a perfect reconstructing system.² However, as mentioned above, perfect reconstructing filter banks are not necessarily the most appropriate for use as preprocessors for a sinusoidal modeler. This new system also has the advantage that the low-pass signal can have a higher cut-off frequency for a given complexity as shown in Figure 5.

Fig. 4. An improved analysis bank for the MRSB. The stop-band of $H_0(z)$ is the range $\frac{\pi}{2} \leq \Omega \leq \pi$. This two-channel filter bank is used in a tree structure to provide (nearly) octave band analysis.

E. Multiple Bases Analysis

In performing sinusoidal model analysis, the goal is to find the time-varying sinusoids that best model a signal. However, stationary sinusoids are usually used as basis vectors for this analysis. Ideally, the analysis procedure would optimally choose a small number of time-varying sinusoids from a large set of sinusoidal bases that vary over time in both amplitude and frequency. In this way a more accurate and compact model of the signal could be produced. Such an optimization problem has enormous complexity so a suboptimal solution is sought.

²The filter structure in Figure 4 can be part of a near-perfect reconstructing filter bank, but not a perfect reconstructing filter bank.

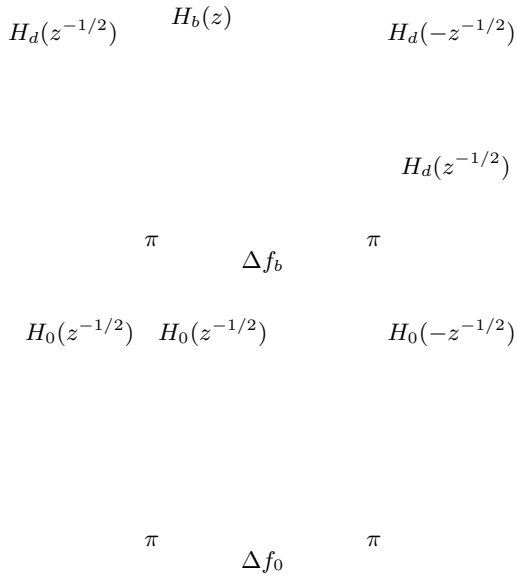


Fig. 5. The bandwidth of the low frequency bands in Levine's structure and the proposed structure are compared. The proposed structure (bottom) results in a nearly half-band split while the upper structure results in a one-third/two-third split for filters of reasonable complexity.

Fig. 6. The analysis structure used for the MRSM. The high frequency band is oversampled by a factor of two and the low frequency band has a cutoff below $\frac{\pi}{2}$ so that there is little or no aliasing.

E.1 Matching Pursuit

Matching pursuit is a sub-optimal method for finding a sparse representation of a signal given an over-complete set of bases [32], [33]. In a matching pursuit analysis, the signal is modeled with successive approximations, each approximation adding a single additional basis contribution. At each step, the residual energy is minimized. The iteration is stopped after the residual or error energy is brought below some threshold or some predetermined number of bases have been used. This method is similar to the analysis-by-synthesis procedure used by George [34], [6] with the exception that the basis functions are restricted to complex sinusoids.

E.2 Matching Pursuit and Perception

Matching pursuit is a greedy algorithm and, in general, the resulting model is not optimal in terms of the residual energy for number of basis vectors used. However, if the basis vectors are well chosen, it is well suited to audio signal representation. This is because the ear is not very sensitive to noise (modeling error) that is quiet and/or spectrally diffuse. In matching pursuit, the high energy spectral components tend to be modeled first, and small spectral components may not even be modeled. Thus, priority is given to those signal components to which the ear is most sensitive. Also, the overall energy in each critical band is very close to that of the original signal. If just peak picking is used, the energy may be noticeably different as the broad and narrow peaks may have the same maximum value but they contain different amounts of energy. In [27] the matching pursuit process is modified to explicitly take advantage of perception.

E.3 Procedure

Given some signal $x[n]$, the idea with matching pursuit is to iteratively create a representation based on the bases, \mathbf{b}_m , in

some dictionary of bases, \mathbf{D} , which minimizes a residual $r[n]$ at each iteration. The resulting approximation at iteration L is expressed as

$$\hat{x}_L[n] = \sum_{k=0}^{L-1} \alpha_k \mathbf{b}_{m(k)} \quad (13)$$

where $\mathbf{b}_{m(k)}$ is the basis chosen at the k^{th} iteration. The residual is given by

$$r_L[n] = x[n] - \hat{x}_L[n]. \quad (14)$$

Note that the residual can be computed recursively as

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k \mathbf{b}_{m(k)} \quad (15)$$

where, for notational convenience, $r_k[n]$ is represented as \mathbf{r}_k . After each iteration, k , the residual, $r_{k+1}[n]$, is only minimized if it is orthogonal to the chosen basis vector, $\mathbf{b}_{m(k)}$. If $\|\mathbf{b}_m\|^2 = 1$ then the α_k which minimizes $\|r_{k+1}\|$ is

$$\alpha_k = \langle \mathbf{r}_i, \mathbf{b}_{m(k)} \rangle. \quad (16)$$

The best \mathbf{b}_m for use at the k^{th} iteration is

$$\mathbf{b}_{m(k)} = \arg \max_{\mathbf{b}_m \in \mathbf{D}} |\langle \mathbf{b}_m, \mathbf{r}_k \rangle| \quad (17)$$

and it is found by minimizing $\|r_{k+1}\|^2$. In short, α_k is calculated for each possible $\mathbf{b}_{m(k)}$ and the basis function which produces the largest $|\alpha_k|$ is used.

E.4 Reducing Complexity

The matching pursuit algorithm as described above requires the calculation of a large number of inner products. However, the computation complexity may be reduced by iteratively updating the inner products:

$$\langle \mathbf{r}_{k+1}, \mathbf{b}_m \rangle = \langle \mathbf{r}_k, \mathbf{b}_m \rangle - \alpha_k \langle \mathbf{b}_{m(k)}, \mathbf{b}_m \rangle. \quad (18)$$

If enough memory is available, $\langle \mathbf{b}_n, \mathbf{b}_m \rangle$ may be precomputed and stored for use in Equation 18.

E.5 Selection of the Basis Vectors

Gabor dictionaries are often used for signal modeling because of the optimal time-frequency representation [32]. However, when modeling a specific type of signal it is often useful to choose basis vectors which better fit the signals to be modeled. Several different choices of dictionaries are discussed below.

E.6 Frames of Complex Exponentials

Frame is a mathematical term that refers to a basis set with some special properties [35]. Specifically,

Definition 1: A frame is a set of vectors ϕ_n such that

$$A \|\mathbf{x}\|^2 \leq \sum |\langle \phi_n, \mathbf{x} \rangle|^2 \leq B \|\mathbf{x}\|^2 \quad \text{for all } \mathbf{x} \quad (19)$$

$A > 0$ and $B > 0$ are the frame bounds.

Definition 2: A tight frame is a frame that has $A = B$.

Unlike the orthonormal bases usually used in signal decomposition, frames do not have the requirement of linear independence. Thus, frames can be thought of as over-sampling in the transform domain. Figure 7 illustrates the difference between a basis set and a frame set in \mathfrak{R}^2 .

Fig. 7. Examples of an orthogonal basis set (a), and a frame set (b) in \mathfrak{R}^2 .

E.6.a Complex Exponential Frames. In sinusoidal modeling, complex exponentials are usually used in the analysis. In order to get better frequency resolution, the DFT is typically padded with zeros to provide more frequency terms. Padding with zeros and taking the orthogonal projection onto the complex exponential basis vectors is equivalent to simply projecting the signal segment onto a tight frame of complex exponential vectors of the same length as the signal segment (see [36, pp 56-63]). Analysis with complex exponential basis functions is often performed by simply picking the peaks in the DFT spectrum. Peak picking is likely to result in poor signal modeling when sinusoids are clustered in the signal since they may meld into a single large spectral peak. Using matching pursuit analysis with the typical complex exponential basis vectors tends to reduce modeling error by better modeling close sinusoids and by reducing the explicit modeling of spurious peaks due to side-lobes.

E.7 Damped Sinusoids

Goodwin [21], [33] has expanded the analysis basis set to a dictionary of damped complex exponential basis vectors. This is particularly useful in modeling music since there are many sharp attacks with gradual decays. There are also some efficient ways of implementing matching pursuit analysis with damped complex exponential basis vectors that make it an appealing choice.

E.8 Chirped Sinusoids

Sinusoidal modeling assumes that the sinusoids are continually varying in amplitude and frequency. By using a dictionary of chirped sinusoids it is possible to directly account for frequency varying signal components in the analysis process. Note that the chirped sinusoids form a complete basis [37] so for each chirp rate used, a complete basis is added to the dictionary of basis vectors. The chirped basis vectors used in this work were of the form

$$e_{\Delta k}(k) = w_s[n] e^{j2\pi \frac{kn}{N_k}} e^{j2\pi \left(\frac{-\Delta k}{2} n + \frac{\Delta k}{2N-2} n^2 \right)} \quad (20)$$

$$\text{where } 0 \leq k < N_k, \quad (21)$$

$$0 \leq n < N, \quad (22)$$

$$\Delta k \in \{-2\beta, -\beta, 0, \beta, 2\beta\}, \quad (23)$$

$$\text{and } w_s[n] = \begin{cases} \frac{1}{\sqrt{N}} & 0 \leq n < N, \\ 0 & N \leq n < N_k \end{cases}. \quad (24)$$

The amount of frequency modulation in the chirp is determined by β ; when $\beta = 1$, $\Delta k = \beta$ produces a chirp whose instantaneous frequency range spans a single DFT bin width. Note, each value of β yields a frame or orthogonal basis (depending on the value of N_k).

The chirped sinusoid dictionary is convenient because it is easy to compute the basis projections and the inner products between basis vectors. The initial projection of the signal onto each basis vector of the basis vectors is computed using the FFT as shown in

$$\langle \mathbf{r}_0, e_{\Delta k}(k) \rangle = \frac{1}{N} \sum_{n=0}^{N_k-1} w_s[n] x[n] e^{-j2\pi \left(kn - \frac{\Delta k}{2} n + \frac{\Delta k}{2N-2} n^2 \right)}. \quad (25)$$

The matrix of basis vector inner products, $\langle e_{\Delta \kappa}(k), e_{\Delta k}(k) \rangle$, can be efficiently computed and stored because it is highly structured. The following values are defined to simplify the result:

$$\text{Let } d_{\Delta k} = \Delta k - \Delta \kappa, \quad (26)$$

$$d_k = k - \kappa, \text{ and} \quad (27)$$

$$\Phi(d_{\Delta k}, d_k) = \langle e_{\Delta k}(k), e_{\Delta \kappa}(k) \rangle. \quad (28)$$

Then

$$\Phi(d_{\Delta k}, d_k) = \frac{1}{N} \sum_{n=0}^{N-1} w_s^2[n] e^{j \frac{\pi n}{N_k} \left(2d_k - d_{\Delta k} + \frac{d_{\Delta k}}{N-1} n \right)}. \quad (29)$$

From this we see that

$$\Phi(d_{\Delta k}, d_k) = \Phi^*(-d_{\Delta k}, -d_k) \quad (30)$$

and that $\Phi(d_{\Delta k}, d_k)$ is only a function of the differences $\Delta k - \Delta \kappa$ and $k - \kappa$.

The sinusoidal model parameters are then chosen as $A_k = \frac{2}{\sqrt{N}} |\alpha_k|$, $\Phi_k = \angle \alpha_k$, and ω_k is taken from the frequency of $\mathbf{b}_{m(k)}$.

F. Synthesis

Synthesis is performed in a manner very similar to that used with the regular sinusoidal model. The partials in each sub-band are used to create sub-band signals that are then combined into the full-band signal. The final MRSMS output signal for segment k is then given by summing the appropriate signal segments, $\tilde{x}^{(k,b)}[n]$, as

$$\tilde{x}[n] = \sum_b \sum_k \tilde{x}^{(k,b)} \left[n - n_0^{(k,b)} \right] \quad (31)$$

where $n_0^{(k,b)}$ is the time index at the beginning of frame k in sub-band b .

III. CONCLUSIONS AND COMMENTS

We have discussed several methods of multi-resolution sinusoidal modeling and have presented two innovations improving upon current methods. The new MRSMS has been successfully used in various signal enhancement tasks including fast-acting dynamic range control for hearing compensation, peak-to-RMS ratio reduction of speech for communications, and background noise suppression [27]. In each of these cases the MRSMS outperforms traditional sinusoidal model and Fourier based techniques.

REFERENCES

- [1] Robert J. McAulay and Thomas F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744-754, August 1986.
- [2] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, no. 3, pp. 497-510, March 1992.
- [3] T. F. Quatieri, R. B. Dunn, and T. E. Hanna, "Time-scale modification of complex acoustic signals," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1213-216, April 1993.
- [4] 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*, 1996, In review.
- [5] Michael W. Macon and Mark A. Clements, "Sinusoidal modeling and modification of unvoiced speech," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 6, pp. 557-560, Nov. 1997.
- [6] E. B. George, *An Analysis-by-Synthesis Approach to Sinusoidal Modeling Applied to Speech and Music Signal Processing*, Ph.D. thesis, Georgia Institute of Technology, November 1991.
- [7] Xavier Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, C. Roads, S. Pope, A. Picialli, and G. De Poli, Eds. Swets and Zeitlinger Publishers, 1997.
- [8] Scott N. Levine, Tony S. Verma, and Julius O. Smith III, "Multiresolution sinusoidal modeling for wideband audio with modifications," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 1998, pp. IV:3585-3589.
- [9] Scott N. Levine and Julius O. Smith III, "A switched parametric & transform audio coder," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 1999.
- [10] M. W. Macon, *Speech and Voice Synthesis Based on Sinusoidal Modeling*, Ph.D. thesis, Georgia Institute of Technology, October 1996.

- [11] T. F. Quatieri and R. G. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 565–568, April 1988.
- [12] T. F. Quatieri and R. G. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 1, pp. 56–69, January 1990.
- [13] James M. Kates, "Speech enhancement based on a sinusoidal model," *Journal of Speech and Hearing Research*, vol. 37, no. 2, pp. 449–464, Apr. 1994.
- [14] M. W. Macon and M. A. Clements, "Speech synthesis based on an overlapped sinusoidal model," *Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3246, May 1995, (A).
- [15] Michael Goodwin and Martin Vetterli, "Time-frequency signal models for music analysis, transformation, and synthesis," in *Time-Frequency Time-Scale Symposium*, 1996.
- [16] J. C. Rutledge and M. A. Clements, "Compensation for recruitment of loudness in sensorineural hearing impairments using a sinusoidal model of speech," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 3641–3644, April 1991.
- [17] T. F. Quatieri and R. J. McAulay, "Peak-to-RMS reduction of speech based on a sinusoidal model," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 273–287, February 1991.
- [18] Robert J. McAulay and Thomas F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," Rep. TR-693, M.I.T., Lincoln Lab., 1985, AD-A157023.
- [19] K. Fitz and L. Haken, "Bandwidth enhanced sinusoidal modeling in Lemur," in *Proceedings of the International Computer Music Conference*, 1995, pp. 154–157.
- [20] Khaled N. Hamdy, Murtaza Ali, and Ahmed H. Tewfik, "Low bit rate high quality audio coding with combined harmonic and wavelet representations," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996.
- [21] Michael Goodwin, *Adaptive Signal Models: Theory, Algorithms, and Audio Applications*, Ph.D. thesis, University of California, Berkeley, 1997.
- [22] Stylianou Ioannis, *Harmonic Plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications, Paris, Jan. 1996.
- [23] Y. Stylianou, J. Laroche, and E. Moulines, "High quality speech modification based on a harmonic + noise model," *Proceedings of EUROSPEECH*, pp. 451–454, September 1995.
- [24] William A. Yost and Donald W. Nielsen, *Fundamentals of Hearing*, Holt, Inc., 1985.
- [25] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, no. 4, pp. 744–754, August 1986.
- [26] David V. Anderson, "Speech analysis and coding using a multi-resolution sinusoidal transform," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 1996, vol. 2, pp. 1037–1040.
- [27] David V. Anderson, *Audio Signal Enhancement Using Multi-resolution Sinusoidal Modeling*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, Mar. 1999.
- [28] Miguel Angel Rogríguez-Hernández and Francisco Javier Casajús-Quirós, "Improving time-scale modification of audio signals using wavelets," in *Proceedings of the 5th International Conference on Signal Processing Applications and Technology*, Dallas, TX, 1994, pp. 1573–7 vol. 2.
- [29] K. Wang and S. A. Shamma, "Modeling the auditory functions in the primary cortex," in *Proceedings of the SPIE. The International Society for Optical Engineering*, 1994, vol. 2242, pp. 692–703.
- [30] Scott N. Levine, Tony S. Verma, and Julius O. Smith III, "Alias-free, multiresolution sinusoidal modeling for polyphonic, wideband audio," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, NY, 1997.
- [31] N. J. Fliege and U. Zölzer, "Multi-complementary filter bank," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1993, pp. III 193–196.
- [32] Stéphane Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, Chestnut Hill, MA, 1998.
- [33] Michael Goodwin and Martin Vetterli, "Atomic decompositions of audio signals," in *IEEE Audio Signal Processing Workshop*, 1997.
- [34] E. B. George and M. J. T. Smith, "A new speech coding model based on a least-squares sinusoidal representation," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1641–1644, April 1987.
- [35] Gilbert Strang and Truong Nguyen, *Wavelets and Filter Banks*, Wellesley–Cambridge Press, Wellesley, MA, 1996.
- [36] Ingrid Daubechies, *Ten Lectures on Wavelets*, SIAM, 1992.
- [37] Richard G. Baraniuk and Douglas L. Jones, "Shear madness: New orthonormal bases and frames using chirp functions," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3543–3549, Dec. 1993.