SINUSOIDAL MODELING USING FRAME-BASED PERCEPTUALLY WEIGHTED MATCHING PURSUITS

Tony S. Verma and Teresa H.Y. Meng

Department of Electrical Engineering, Stanford University
verma@furthur.stanford.edu, http://www.stanford.edu/~darkstar

ABSTRACT

We propose a method for sinusoidal modeling that takes into account the psychoacoustics of human hearing using a frame-based perceptually weighted matching pursuit. Working on blocks of the input signal, a set of sinusoidal components for each block is iteratively extracted taking into consideration perceptual significance by using extensions to the well known matching pursuits algorithm. These extensions allow including information about the time-varying masking threshold of the input signal during the pursuit. The blocks overlap-add together to reconstruct the entire signal. Although the perceptually weighted matching pursuit on each block can iterate until the error between the original and the reconstructed signal is zero, lower order approximations are possible by stopping the pursuit when the error becomes imperceptible to the human ear or by stopping the pursuit after a number of the perceptually most significant sinusoidal elements are found. The proposed sinusoidal model finds use in many applications including signal modifications and compression.

1. INTRODUCTION

Sinusoidal modeling has seen extensive use in a wide range of speech and audio applications including compression, signal modifications, and audio scene analysis. Many formulations for sinusoidal modeling exist, some based on spectral peak picking algorithms [1, 2] while others use frame-based analysis-by-synthesis techniques [3]. Although the implementations of these techniques are quite different, the goal is the same: for the l^{th} frame of the input, find a set of K sinusoidal signals parameterized by *amplitude*, *frequency* and *phase*. Although each formulation has its advantages, analysis-by-synthesis techniques guarantee convergence in the sense that the error between the original and the reconstructed signal can be forced to zero. Current analysis-by-synthesis formulations converge in a way that minimizes the mean-square of the error at each iteration. As such, each iteration of the analysis-by-synthesis algorithm finds and removes the sinusoidal component in a frame of the input signal that contains the greatest energy. Here we present a formulation for analysis-by-synthesis sinusoidal modeling that takes into account the psychoacoustics of the human hearing system by using perceptually weighted matching pursuits. In an iterative fashion, the algorithm finds and removes the perceptually most significant sinusoidal components in a frame of input signal. Therefore, instead of removing the sinusoidal component that contains the greatest energy at each iteration, our algorithm removes the sinusoidal component that contains the greatest perceived energy at each iteration.

The first section of the paper describes both parsing of the input signal into frames and how these frames are put back together to produce a reconstructed signal. On each of the frames, a perceptually weighted matching pursuit is performed. Use of the matching pursuits algorithm [4] as a general framework for analysis-by-synthesis techniques allows previous results in the literature to facilitate the proof of convergence of our algorithm. Section 3 gives a brief review of the matching pursuit algorithm before examining how to make matching pursuits equivalent to an analysis-by-synthesis sinusoidal modeling technique that includes psychoacoustic phenomena. The section ends by showing that the perceptually weighted matching pursuit algorithm has an intuitive interpretation in terms of the Discrete Fourier Transform (DFT) and can be efficiently computed via the Fast Fourier Transform (FFT). Section 4 shows an example of the algorithm and the final section gives conclusions.

2. OVERLAP-ADD FORMULATION

In our formulation of sinusoidal modeling, frames of the input signal x are represented as a combination of sinusoidal signals. The combination of sinusoids for each frame, as will be described, is found via perceptually weighted matching pursuits. These frames are combined in an overlap-add fashion to reconstruct the entire signal. Mathematically, we take $x = \{x[n]; n \in \mathcal{Z}\}$ and make an ensemble of timelimited signals $x_l = \{x_l[n]; n \in \mathcal{Z}\}$ by hopping a rectangular window over signal. Let the l^{th} windowed signal be

$$x_l[n] = \sqcap_N [n - lp]\, x[n] \tag{1}$$

where N is the length of the window, p is stride length of the window constrained so that $p \leq N$, and the rectangular window is

$$\sqcap_N[n] = \begin{cases} 1 & n = 0, 1, \ldots, N-1 \\ 0 & \text{otherwise} \end{cases}$$

Each of these timelimited signals can then be considered a finite duration signal in $\mathcal{R}^N$ to which weighted matching pursuits are applied. Although we consider each timelimited signal a finite duration signal, we keep track of the time location of each frame to ensure proper reconstruction. As such, the matching pursuit reconstruction of each frame, $\hat{x}_l = \{\hat{x}_l[n]; n \in \mathcal{Z}\}$, is once again considered an ensemble of timelimited signals. Finally, the approximation to x, $\hat{x} = \{\hat{x}[n]; n \in \mathcal{Z}\}$, is completed with a windowed overlap-add reconstruction of the form:

$$\hat{x}[n] = \sum_l v\,[n - lp]\, \hat{x}_l[n] \tag{2}$$

where the timelimited reconstruction window v has the constraint

$$\sum_l v\,[n - lp] = 1 \tag{3}$$

If the error of the matching pursuit on each of the x_l signals converges to zero, the formulation allows perfect reconstruction which is immediate from plugging (1), which is the error-free matching pursuit decomposition of each frame, into equation (2) and using the window constraint from equation (3).

3. PERCEPTUALLY WEIGHTED MATCHING PURSUIT FORMULATION

With the overlap-add framework described, we now focus on a perceptually weighted matching pursuit. In this section, we will consider the matching pursuits algorithm applied to one frame, i.e., one of the x_l signals, and consider that signal to be a column vector $\mathbf{x}_l$ that belongs to $\mathcal{R}^N$. Notationally, we will use lowercase bold letters to denote column vectors and uppercase bold letters to denote matrices. The notation $x[n]$ and $w[n,m]$ denotes an element of vector $\mathbf{x}$ and matrix $\mathbf{W}$ respectively. In addition, we will use $X\,[k/M]; k = 0, 1, \ldots, M-1$ to denote the M point DFT of $x[n]; n = 0, 1, \ldots, N-1$ (if $M > N$ zero-padding is assumed).

3.1. The Matching Pursuits Algorithm

Matching pursuits refers to an iterative method for computing signal decompositions in terms of a linear combination of vectors from a highly redundant dictionary [4]. The M elements of the dictionary, $\mathcal{D} = \{\mathbf{g}_m\}\,; m = 0, 1, \ldots, M-1$, span $\mathcal{R}^N$ and are restricted to have unit norm, $\|\mathbf{g}_m\| = 1$ for all m. The algorithm is greedy in that at each stage the vector in the dictionary that best matches the current signal is found and subtracted to form a residual. The algorithm then continues on this residual signal. More specifically, at the k^{th} iteration of the algorithm, the k^{th} index m_k is found corresponding to the dictionary element which has the largest correlation with the k^{th} residual $\mathbf{r}_k$. This index maximizes $|\,\langle \mathbf{g}_m, \mathbf{r}_k \rangle\,|$ over all m. The projection onto this dictionary element is then subtracted from the current residual to form the next stage residual. Thus at the k^{th} iteration, for $k \geq 0$, the next stage residual is $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k \mathbf{g}_{m_k}$, where $\alpha_k = \langle \mathbf{g}_{m_k}, \mathbf{r}_k \rangle$. The algorithm is initialized by setting $\mathbf{r}_0 = \mathbf{x}$. Therefore, the decomposition consists of a set of correlation terms $\{\alpha_0, \alpha_1, \ldots\}$ and indices $\{m_0, m_1, \ldots\}$. The signal reconstruction is the weighted linear combination of the dictionary elements found during the decomposition, which is, if the decomposition runs for K iterations, $\sum_{k=0}^{K-1} \alpha_k \mathbf{g}_{m_k}$.

The energy in the residual converges to zero as the number of iterations approaches infinity [4]. Although exact reconstruction is possible, the matching pursuit is generally stopped by some criterion to allow low order approximations to the input signal. Appropriate stopping criteria for the perceptually based sinusoidal model are given toward the end of section 3.4.

3.2. Sinusoidal Modeling Matching Pursuits Dictionary

It was shown in [5] that by choosing a particular matching pursuits dictionary and by using a generalization of the matching pursuit algorithm, which allows finding optimal sets of dictionary elements (i.e., a dictionary subspace) at each iteration [4, 6], a frame-based matching pursuit resembles a frame-based analysis-by-synthesis sinusoidal model. The dictionary consists of complex exponentials, $\mathbf{g}_m = \left\{ g_m[n] = \frac{1}{N} e^{j2\pi \frac{m}{M} n}; n = 0, 1, \ldots, N-1 \right\}$; $m = 0, 1, \ldots, M-1$, and the dictionary subspace consists of two dictionary elements: a dictionary element and its complex conjugate. Since $\mathbf{x}_l$ is real, the correlation coefficients appear in conjugate pairs [6] and we only need to search (and compute) half of the correlation coefficients for the absolute maximum which gives $\alpha_k = a_k e^{j\theta_k}$, the largest correlation coefficient, at the k^{th} iteration. Thus at the k^{th} iteration the residual signal is:

$$\begin{aligned} r_{k+1}[n] &= r_k[n] - \alpha_k g_{m_k} - \alpha_k^* g_{m_k}^* \\ &= r_k[n] - \frac{2a_k}{N} \cos\left[2\pi \frac{m_k}{M} n + \theta_k \right] \end{aligned} \tag{4}$$

Equation (4) shows that at each iteration, the projection onto the dictionary subspace will be a constant amplitude, constant frequency cosine. The amplitude and phase for each of the cosines are found from the correlation terms, $\{\alpha_0, \alpha_1, \ldots\}$, and the frequency for each is found from the indices $\{m_0, m_1, \ldots\}$ by dividing by the dictionary size, M. Although these are constant amplitude and frequency cosines, the overlap-add formulation from section 2 provides a smooth transition from frame to frame.

3.3. Weighted Matching Pursuits

To make the matching pursuits algorithm include perceptual characteristics of human hearing, we modify the pursuit in two ways. First, we modify each of the dictionary elements by a scaler and allow the dictionary elements to have non-unit norms. Let the dictionary weighting sequence be $\gamma = \{\gamma[m]; m = 0, 1, \ldots, M-1\}$ and restrict $\gamma[m] \neq 0$ for all m. In addition, γ must be real and even $moduloM$. The dictionary now has the form

$$\mathbf{g}_m = \left\{ g_m[n] = \gamma[m] e^{j2\pi \frac{m}{M} n}; n = 0, 1, \ldots, N-1 \right\} \tag{5}$$

with $m = 0, 1, \ldots, M-1$. Secondly, we generalize the inner product to a weighted inner product, $\langle \mathbf{x}, \mathbf{y} \rangle_W \equiv \mathbf{y}^* \mathbf{W} \mathbf{x}$, where $\mathbf{W}$ is a symmetric positive definite matrix. The restrictions on $\mathbf{W}$ are necessary to ensure a valid inner product space [7]. In our formulation, we will choose the dictionary weighting sequence, γ, based on psychoacoustic information and $\mathbf{W}$ equivalent to a window (i.e., $\mathbf{W}$ is a diagonal matrix whose elements on the diagonal are the coefficients of a window).

With these modifications, convergence of the matching pursuit is still guaranteed; however, the rate at which convergence occurs will in general be different. For a matching pursuit to converge in a properly defined inner product space the dictionary vectors must span the space, in this case, $\mathcal{R}^N$. A properly defined inner product space allows projections to make sense and therefore forces energy of the residual to be monotonically decreasing. The energy will decrease until the residual is orthogonal to every dictionary element which will only happen when the residual is zero (since the dictionary is required to span the space). Since we have a valid inner product space and chose non-unit norm dictionary elements that are complete in $\mathcal{R}^N$, convergence is guaranteed. The rate of convergence, however, will be different because using unit norm dictionary elements and a standard inner product, each iteration of the algorithm finds the dictionary element(s) that removes the greatest energy from the current residual [4]. With the modifications, if γ and $\mathbf{W}$ are chosen in a perceptually significant way, each step of the algorithm finds the dictionary element(s) that removes the greatest perceived energy from the current residual.

The matching pursuit algorithm remains essentially the same except the weighted inner product must be included in all definitions and non-unit norm dictionary elements must be accounted for. Again, we need to find two dictionary elements (an element and its complex conjugate) at each iteration. However, as shown in the previous section, and because of the symmetry of γ, we actually need only search half of the correlations and find the absolute maximum because the other is given as its complex conjugate. With this in mind, the k^{th} index is

$$m_k = \max_m{}^{-1} \frac{\left|\langle \mathbf{g}_m, \mathbf{r}_k \rangle_W\right|}{\langle \mathbf{g}_m, \mathbf{g}_m \rangle_W} \tag{6}$$

and $\alpha_k = \langle \mathbf{g}_{m_k}, \mathbf{r}_k \rangle_W / \langle \mathbf{g}_{m_k}, \mathbf{g}_{m_k} \rangle_W$ is the k^{th} correlation coefficient. In the next section, we show how this matching pursuit can be efficiently implemented in terms of the FFT.

3.4. DFT Interpretation

Since each iteration of the weighted matching pursuit requires M weighted correlation calculations, after which the largest absolute weighted correlation must be found, the computational complexity is high for a general unstructured dictionary and weighting matrix $\mathbf{W}$. However, because of the choice of both the weighted dictionary and the weighting matrix, we can use the DFT (or the FFT if the number of dictionary elements is a power of 2) for the correlation computations. At the k^{th} iteration, we must the compute the inner products shown in equation (6). In the following, because $\mathbf{W}$ is diagonal, we address the diagonal elements as $w[n]$ and ignore the rest of the matrix. In addition, since $w[n]$ is a window, $w[n] = 0$ for $n \notin \{0, 1, \ldots, N-1\}$. So we must compute, for $m = 0, 1, \ldots, M-1$,

$$\begin{aligned}
\frac{\left|\langle \mathbf{g}_m, \mathbf{r}_k \rangle_W\right|}{\langle \mathbf{g}_m, \mathbf{g}_m \rangle_W} &= \frac{|\mathbf{g}_m^* \mathbf{W} \mathbf{r}_k|}{\mathbf{g}_m^* \mathbf{W} \mathbf{g}_m} \\
&= \frac{\left|\sum_{n=0}^{N-1} \gamma^*[m] e^{-j2\pi \frac{m}{M} n} w[n] r_k[n]\right|}{\sum_{n=0}^{N-1} \gamma^*[m] e^{-j2\pi \frac{m}{M} n} w[n] \gamma[m] e^{j2\pi \frac{m}{M} n}} \\
&= \frac{\gamma[m] \left|\sum_{n=0}^{N-1} w[n] r_k[n] e^{-j2\pi \frac{m}{M} n}\right|}{\gamma[m]^2 \sum_{n=0}^{N-1} w[n]} \\
&= \frac{\left|\sum_{n=0}^{M-1} w[n] r_k[n] e^{-j2\pi \frac{m}{M} n}\right|}{\gamma[m] \sum_{n=0}^{M-1} w[n]} \\
&= \frac{\left|R_k^w\left[\frac{m}{M}\right]\right|}{\gamma[m] W\left[\frac{0}{M}\right]}
\end{aligned} \tag{7}$$

and find its maximum to find the k^{th} index, m_k. Then the k^{th} correlation is given as

$$\alpha_k = \frac{R_k^w\left[\frac{m_k}{M}\right]}{\gamma[m_k] W\left[\frac{0}{M}\right]} = \frac{a_k e^{j\theta_k}}{\gamma[m_k] W\left[\frac{0}{M}\right]} \tag{8}$$

where a_k and θ_k are the magnitude and phase of $R_k^w[m_k/M]$ respectively. The numerator of equation (7) is the magnitude of the windowed M point DFT of the k^{th} residual: $R_k^w[m/M]$. The denominator modifies each DFT coefficient by $\gamma[m]$, the m^{th} psychoacoustic weighting factor, and a constant, $\sum_n w[n]$, which can be ignored when finding the maximum correlation coefficient. Therefore the dictionary weighting factors cause an inverse amount of importance to be placed on the dictionary elements. If the weighting sequence, γ, is the psychoacoustic masking threshold of $\mathbf{x}_l$, as found using the psychoacoustic model described in [8], equation (7) gives, for $k = 0$, the so-called signal-to-mask ratio for the input signal. Signal-to-mask ratio components less than 1 are perceptually irrelevant, while the maximum of the signal-to-mask ratio is the psychoacoustically most significant spectral element of the signal (although the accuracy of the model in [8] is debatable). Assuming the psychoacoustic model in [8] is least somewhat accurate and choosing γ as the masking threshold of $\mathbf{x}_l$, the weighted matching pursuit will iteratively find the perceptually most significant spectral component in each residual as compared to the masking ability of $\mathbf{x}_l$. Possible choices for γ is unlimited (e.g., one possible choice is absolute threshold of hearing), although most choices will not carry the same perceptual significance as the masking threshold of the input. The use of a weighted inner product with the special choice of $\mathbf{W}$ as a windowing operation is also important not only because windowing is one of the operations required for computing the masking threshold, but because without windowing the algorithm would be effectively searching a rectangularly windowed spectrum for the best weighted correlation which could cause problems because of the poor side-lobe performance of such a window.

Reconstruction for the frame is:

$$\begin{aligned}
\hat{x}_l[n] &= \sum_{k=0}^{K-1} \alpha_k g_{m_k}[n] + \alpha_k^* g_{m_k}^*[n] \\
&= \sum_{k=0}^{K-1} \frac{2a_k}{W\left[\frac{0}{M}\right]} \cos\left[2\pi \frac{m_k}{M} n + \theta_k\right]
\end{aligned} \tag{9}$$

Using equation (9), instead of storing α_k and m_k and using a dictionary that contains the psychoacoustic weighting factors, we can store the parameter triplet of *amplitude*, *frequency* and *phase* as:

$$\{A_k = 2a_k / W\left[0/M\right], f_k = m_k/M, \theta_k = \theta_k\} \tag{10}$$

and use a separate reconstruction dictionary that does not contain the psychoacoustic weighting factors, $\tilde{g}_m[n] = e^{j2\pi \frac{m}{M} n}$; $n = 0, 1, \ldots, N-1$; $m = 0, 1, \ldots, M-1$. Using the parameters in (10) and the reconstruction dictionary $\tilde{\mathbf{g}}_m$ directly gives equation (9). Because the reconstruction dictionary is fixed, there is no overhead associated with storing the psychoacoustic weighting factors, which may vary frame-by-frame or, if desired, during the pursuit. This is important for compression applications.

The matching pursuit computational burden is further reduced by updating the correlations directly at each iteration. Similar to [4], the correlations can be updated as

$$\frac{\langle g_m, r_{k+1} \rangle}{\langle g_m, g_m \rangle} = \frac{\langle g_m, r_k \rangle - \alpha_k \langle g_m, g_{m_k} \rangle - \left(\alpha_k \langle g_m, g_{m_k} \rangle\right)^*}{\langle g_m, g_m \rangle} \tag{11}$$

for all m. Thus for any application of matching pursuits only one set of M correlations need be computed at the start and the following correlations for subsequent residual signals are updated iteratively. Because of the choice of dictionary elements, equation (11) has an interpretation entirely in terms of the DFT. It can be shown that (11) is equivalent to, for all m:

$$\frac{R_k^w\left[\frac{m}{M}\right] - \frac{A_k}{2}\left(e^{j\theta_k} W\left[\frac{m-m_k}{M}\right] + e^{-j\theta_k} W\left[\frac{m+m_k}{M}\right]\right)}{\gamma[m] W\left[\frac{0}{M}\right]}$$

which shows the correlations for the next iteration are found by subtracting two frequency shifted window transforms from the

DFT of the last stage residual, then dividing the spectrum by the perceptual weighting sequence. Therefore the entire l^{th} frame matching pursuit can be performed in the DFT domain as follows, assuming M is a power of 2 and $m = 0, 1, \ldots, M-1$ (although using symmetry of the DFT allows using $m = 0, 1, \ldots, M/2$):

1. Store $W[m/M]$, the M point FFT of the window. Set $\mathbf{r}_0 = \mathbf{x}_l$. Compute the M point FFT of $r_0[n]w[n]$, the current windowed residual transform $R_0^w[m/M]$.
2. Compute $R_k^w[m/M]/\gamma[m]$, the current windowed residual transform divided by dictionary weighting elements and find the absolute maximum of the resulting set. This gives the current sinusoidal parameters, $\{A_k, f_k, \theta_k\}$, as defined in (10).
3. If suitable stopping criterion is met (as discussed below), exit; otherwise compute, $R_{k+1}^w[m/M]$, the next stage windowed residual transform as:

$$R_k^w[\tfrac{m}{M}] - \frac{A_k}{2}\left(e^{j\theta_k} W\left[\tfrac{m-m_k}{M}\right] - e^{-j\theta_k} W\left[\tfrac{m+m_k}{M}\right]\right)$$

   and let this next stage residual window transform be the current stage residual transform and return to step 2.

Although the algorithm could continue until the residual signal converges to zero, there are much better stopping criteria. One possibility is when residual falls below the psychoacoustic masking threshold of $\mathbf{x}_l$ (e.g., when $|\alpha_k| < 1$). With this criterion, although the residual could be very large in a mean-square sense, the reconstruction is perceptually identical to the original. Another possibility is to stop the pursuit after K iterations, which gives the reconstruction with the K perceptually most important sinusoids. Alternatively, the proposed sinusoidal model seamlessly integrates with multi-part models such as sines+noise [2] or sines+transients +noise [5]. When used with multi-part models the stopping criterion is when the residual no longer contains components that correlate well with sinusoids (i.e., transients and/or noise).

4. EXAMPLE

We ran the algorithm on a piece of rock music sampled at $f_s = 32KHz$ shown in figure 1(a). The input was split into blocks of $N = 1024$ samples with an analysis stride of $p = 800$ samples. The dictionary size was $M = 8192$. The weighting matrix $\mathbf{W}$ performed the equivalent of a 1024 point Hamming window and γ for each block was found as the psychoacoustic masking threshold. The perceptually weighted matching pursuit continued until the residual fell below the masking threshold of the input signal for each block. The reconstructed signal shown in figure 1(b) required on average $K = 180$ sinusoidal parameters per frame. Although the error shown in figure 1(c) between the original and the reconstruction is quite large, the original and the reconstructed signals are perceptually identical (as judged by informal listening tests).

5. CONCLUSION

By using extensions to the matching pursuits algorithm, a method for sinusoidal modeling that iteratively extracts the perceptually most significant sinusoids in a signal was presented. This formulation, which can be integrated into multi-part signal models, is beneficial to many sinusoidal modeling applications, but is particularly well suited for scalable compression because the reconstructed signal can be perceptually identical to the original and because the sinusoidal components are ordered in terms of perceptual significance.

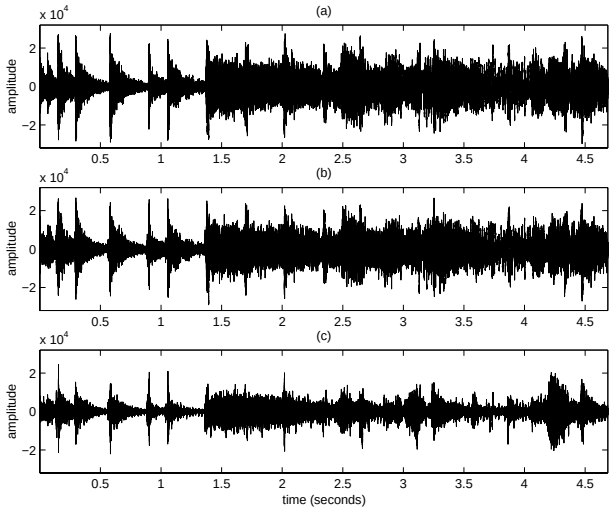

Figure 1: (a) Input signal (b) Reconstructed signal (c) Error signal

6. ACKNOWLEDGMENTS

The authors would like to thank Scott Levine and Julius O. Smith for countless discussions and input on this topic.

7. REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal speech model", *IEEE Trans. ASSP*, pp. 744–754, 1986.

[2] X. Serra and J. O. Smith, "Spectral modeling synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition", *ICMJ*, vol. 14, no. 4, pp. 14–24, Winter 1990.

[3] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones", *JAES*, vol. 40, no. 6, pp. 497–515, June 1992.

[4] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries", *IEEE Trans. SP*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[5] T. Verma and T. Meng, "An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio", in *Proc. ICASSP*, May 1998.

[6] M. Goodwin, "Matching pursuit with damped sinusoids", in *Proc. ICASSP*, April 1997, vol. 3, pp. 2037–2040.

[7] H. Anton and C. Rorres, *Elementary Linear Algebra*, Anton Textbooks, 1991.

[8] ISO/MPEG Committee, "Information technology - coding of moving pictures and associated audio for digital storage media at up to about 5 1.5mbit/s - part 3: Audio", *ISO/IEC 11172-3*.