Linear AM Decomposition for Sinusoidal Audio Coding

Christensen, Mads Græsbøll; Jakobsson, Andreas; Andersen, S. V.; Jensen, Søren Holdt

Published in:
IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05)

Publication date:
2005

Document Version
Early version, also known as pre-print

LINEAR AMPLITUDE DECOMPOSITION FOR SINUSOIDAL AUDIO CODING

Mads Græsbøll Christensen[*†], *Andreas Jakobsson*[‡], *Søren Vang Andersen*[†], *and Søren Holdt Jensen*[†]

[†] Dept. of Communication Technology
Aalborg University, Denmark
{mgc,sva,shj}@kom.aau.dk

[‡] Dept. of Electrical Engineering
Karlstad University, Sweden
andreas.jakobsson@ieee.org

ABSTRACT

In this paper, we present a novel decomposition for sinusoidal audio coding using an amplitude modulation of sinusoids via a linear combinations of arbitrary basis vectors. The proposed method, which incorporate a perceptual distortion measure, is based on the relaxation of a non-linear least squares minimization, and offers benefits in the modeling of transients in audio signals. Further, rate-distortion curves indicate that using a sinusoidal audio coder with the proposed decomposition will be preferable to one without it. This result is also confirmed by listening tests that indicate that, for a given bitrate, significant improvements can be gained over a typical sinusoidal coder.

1. INTRODUCTION

The problem of decomposing a signal into amplitude modulated (AM) sinusoids is encountered in many different applications, for example in parametric audio coding (see, e.g., [1]) where AM sinusoidal models are of interest for handling transients. Even when dynamic time segmentation is employed, there is a need for efficient coding of transients because of limitations on the minimum segment length [2]. In [3], it was shown that perceptually significant improvements can be achieved by applying AM in a frequency dependent way as opposed to singleband AM (see, e.g., [4]). Furthermore, it was shown in [5] that it is indeed efficient in terms of rate-distortion to apply AM on a per component basis. Sinusoidal modeling using both amplitude and frequency modulation, in the form of a linear combination of basis vectors such as low-order polynomials, has been explored for a variety of applications (see, e.g., [6, 7]). Although such models perform well for slowly evolving signals such as voiced speech, they can not satisfactory handle the fast transients often encountered in audio signals. In this paper, we extend the work in [3, 5] by investigating the estimation of parameters for a preselected, linearly independent, set of real-valued basis functions that describe the amplitude modulating function. Furthermore, we examine how to incorporate such a decomposition in parametric audio coding, especially noting that it is not always efficient in terms of rate and distortion to use the AM technique. The rest of the paper is organized as follows: In Section 2, both the signal decomposition and the proposed estimator are presented, followed in Section 3 with the incorporation of a perception-based weighting. Section 4 discuss sinusoidal audio coding using the proposed AM decomposition. Experimental results are presented in Section 5, and Section 6 contains our conclusions.

*The work of M. G. Christensen was conducted within the ARDOR (Adaptive Rate-Distortion Optimized sound codeR) project, EU grant no. IST–2001–34095.

2. PROPOSED DECOMPOSITION

In this paper, we model the signal of interest as a sum of amplitude modulated sinusoids, i.e.,

$$x(n) = \sum_{l=1}^{L} \gamma_l(n) \cos(\omega_l n + \phi_l), \tag{1}$$

where ω_l and ϕ_l denote the lth carrier frequency and phase, respectively, and $\gamma_l(n)$ is the lth amplitude modulating function formed as

$$\gamma_l(n) = \sum_{i=1}^{I} b(n,i) c_{i,l} \tag{2}$$

where $b(n,i)$ and $c_{i,l}$ denote ith basis function evaluated at time instance n and the (i,l)th AM coefficient, respectively; it is assumed that the L carrier frequencies are distinct, so that $\omega_k \neq \omega_l$ for $k \neq l$. Further, to ensure a compact representation, we assume that $IL < N$, i.e., $I < N$ and $L < N$, where N denotes the data length. The additional flexibility in (1), as compared to traditional constant-amplitude models (CA), with $\gamma_l(n) = A_l$, enables improved modeling of transient segments. Thus, the CA model is but a special case of the proposed AM model. Let $x_a(n)$ denote the discrete-time "analytical" signal constructed from $x(n)$ by removing the negative frequency components, such that the resulting signal may be downsampled by a factor two without loss of information [8], assuming that there is little or no signal of interest near 0 and π. Then, $x_a(n)$ can be written as

$$x_a(n) = \sum_{l=1}^{L} \sum_{i=1}^{I} b(n,i) c_{i,l} e^{j\omega_l n + j\phi_l} \tag{3}$$

Without loss of generality, we assume N to be even, and introducing

$$\mathbf{x}_a = \begin{bmatrix} x_a(1) & x_a(3) & \cdots & x_a(N-1) \end{bmatrix}^T, \tag{4}$$

where $(\cdot)^T$ is the transpose operator, the *downsampled* discrete-time "analytical" signal may be expressed as

$$\mathbf{x}_a = [(\mathbf{BC}) \odot \mathbf{Z}]\, \mathbf{a}, \tag{5}$$

where $\odot$ denotes the Hadamard (elementwise) product, i.e., $[\mathbf{E} \odot \mathbf{F}]_{kl} = [\mathbf{E}]_{kl}[\mathbf{F}]_{kl}$, with $[\mathbf{E}]_{kl}$ being the (k,l)th element of $\mathbf{E}$. Further, $\mathbf{Z} \in \mathbb{C}^{N/2 \times L}$ is constructed from the L complex carriers, i.e., $[\mathbf{Z}]_{kl} = e^{j\omega_l(2k-1)}$, $\mathbf{a} = \begin{bmatrix} e^{j\phi_1} & \cdots & e^{j\phi_L} \end{bmatrix}^T$, and

the amplitude modulating function is written using the *known* AM basis vectors, $[\mathbf{B}]_{kl} = b(2k-1,l)$, and the corresponding coefficients, $[\mathbf{C}]_{kl} = c_{k,l}$. Here, $\mathbf{B} \in \mathbb{R}^{N/2\times I}$ and $\mathbf{C} \in \mathbb{R}^{I\times L}$. The problem of interest is given a measured signal, $y(n)$, find an estimate $x(n)$ such that

$$\min_{\mathbf{C},\{\phi_k\},\{\omega_k\}} \sum_{n=1}^{N} |y(n)-x(n)|^2 \tag{6}$$

or, equivalently,

$$\min_{\mathbf{C},\{\phi_k\},\{\omega_k\}} \|\mathbf{y}_a - \mathbf{x}_a\|_2^2 \tag{7}$$

where $\mathbf{y}_a$ is formed similar to $\mathbf{x}_a$, and $\|\cdot\|_2$ denotes the 2-norm. Typically, the nonlinear least squares (NLS) minimization in (7) requires a multidimensional, often multimodal, minimization which is typically computationally infeasible in most practical situations. It is worth noting that for the sinusoidal estimation problem, several suboptimal approaches based on relaxation of the original problem have been suggested to reduce the computational complexity of the minimization, such as the greedy matching pursuit [9] or recursive methods as RELAX [10]. Herein, we propose an iterative method for the minimization of (7) reminiscent to both the above mentioned methods. The suggested method exploits the fact that for given frequencies $\{\omega_k\}_{k=1}^L$, the minimization with respect to $\mathbf{C}$ for fixed $\{\phi_k\}_{k=1}^L$ is quadratic, and conversely the minimization of $\{\phi_k\}_{k=1}^L$ for fixed $\mathbf{C}$. We propose to iteratively estimate $\mathbf{C}$ and $\{\phi_k\}_{k=1}^L$, minimizing the residual for each frequency in a given finite set of frequencies, Ω. Let

$$\mathbf{c}_k = \begin{bmatrix} c_{1,k} & \cdots & c_{I,k} \end{bmatrix}^T. \tag{8}$$

At iteration k, assuming the $k-1$ carriers and corresponding coefficients known (i.e., estimated in prior iterations), we find for each frequency, $\omega \in \Omega$, the model parameters ϕ_k and $\mathbf{c}_k$ minimizing the residual for that particular frequency. The kth carrier is then found as the parameter set minimizing the residual over Ω, i.e.,

$$\hat{\omega}_k = \arg\min_{\omega\in\Omega} \|\mathbf{r}_k - \mathbf{D}_k e^{j\phi_k}\mathbf{B}\mathbf{c}_k\|_2^2, \tag{9}$$

where $\mathbf{D}_k$ is the diagonal matrix constructed from the kth carrier,

$$\mathbf{D}_k = \mathrm{diag}\left(\begin{bmatrix} z_k^1 & z_k^3 & \cdots & z_k^{N-1} \end{bmatrix}\right) \tag{10}$$

with $z_k = e^{j\omega_k}$. Further,

$$\mathbf{r}_k = \begin{bmatrix} r_k(1) & r_k(3) & \cdots & r_k(N-1) \end{bmatrix}^T \tag{11}$$

contains the kth residual, obtained as

$$r_k(n) = y_a(n) - \sum_{l=1}^{k}\sum_{i=1}^{I} b(n,i)\hat{c}_{i,l}e^{j\hat{\omega}_l n + j\hat{\phi}_l}. \tag{12}$$

For each frequency ω, we iteratively solve for ϕ_k and $\mathbf{c}_k$ (with superscript (p) denoting the pth iteration of the alternating minimization); for given $\hat{\mathbf{c}}_k^{(p-1)}$,

$$\hat{\phi}_k^{(p)} = \arg\left\{ \sum_{\substack{n=1,\\ n \text{ odd}}}^{N} \sum_{i=1}^{I} b(n,i)\hat{c}_{i,l}^{(p-1)} e^{-j\omega n} r_k(n) \right\}. \tag{13}$$

Given $\hat{\phi}_k^{(p)}$, the estimate of the AM coefficients reduces to

$$\hat{\mathbf{c}}_k^{(p)} = \mathbf{B}^+ \mathbf{u}_k^{(p)}, \tag{14}$$

with

$$\mathbf{B}^+ = \left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T, \tag{15}$$

which can be pre-computed. Here, the vector $\mathbf{u}_k^{(p)}$ is defined as

$$\mathbf{u}_k^{(p)} = \begin{bmatrix} u_k^{(p)}(1) & u_k^{(p)}(3) & \cdots & u_k^{(p)}(N-1) \end{bmatrix}^T \tag{16}$$

which is the real part (recall that $c_{i,l} \in \mathbb{R}$) of the residual shifted towards DC by the carrier, i.e.,

$$u_k^{(p)}(n) = \mathrm{Re}\left\{ r_k(n) e^{-j\omega n - j\hat{\phi}_k^{(p)}} \right\}. \tag{17}$$

The estimates in (13) and (14) are then found alternately, given the other, until some stopping criterion is reached. Experiments indicate that the algorithm converges to a global maximum. We note that for the special case of constant amplitude (DC) basis, the estimates (9), (13) and (14) reduce to those of a matching pursuit with complex sinusoids [9].

3. INCORPORATING PERCEPTUAL DISTORTION

As is well-known, the 2-norm error measure does not correlate well with human sound perception; the problem of finding a suitable distortion measure is one of computational complexity and mathematical convenience and tractability. On one hand, we desire a measure that considers as much of the human auditory processing as possible (such measure exist), while on the other, it is preferable to have a measure which defines a mathematical norm leading to efficient, simple, estimators and quantizers. Here, we apply the perceptual distortion measure defined in [11]; for a particular segment, the distortion D can then be written as

$$D = \int_{-\pi}^{\pi} A(\omega) \left|\mathcal{F}\left[w(n)e(n)\right]\right|^2 d\omega, \tag{18}$$

where $\mathcal{F}[\cdot]$ denotes the Fourier transform, $A(\omega) \in \{x \in \mathbb{R}, x > 0\}$ is a perceptual weighting function, $w(n)$ is the analysis window, and $e(n) = y(n) - x(n)$ is the modeling error. When the weighting function is chosen as the reciprocal of the masking threshold, the resulting error spectrum will be shaped like the masking threshold. This measure is inherently based on waveform matching as it operates on the Fourier transform of the time domain error. As a result, for example pre-echos will be punished by the measure; the actual distortion values for nonstationary segments, with respect to audibility, should be interpreted with care. In practice, the integral in (18) is calculated as a summation of point-wise multiplications in the frequency domain, corresponding to a circular filtering in the time domain [12], i.e.,

$$D \approx \|\mathbf{H}\mathbf{W}(\mathbf{y}-\mathbf{x})\|_2^2, \tag{19}$$

where $\mathbf{H}$ is an circular Toeplitz matrix constructed from the impulse response of the filter corresponding to $A(\omega)$ and $\mathbf{W}$ is a diagonal weighting matrix containing the elements of the analysis/synthesis window, e.g., a von Hann window; see [12] for further details. Using this perceptual distortion allows us to minimize (9) in a perceptually preferable way. However, doing so makes

the pseudo-inverse $\mathbf{B}^+$, defined in (15), both frequency and segment dependent, forcing it to be re-calculated for each frequency and segment. Via extensive simulations and experiments, we have found that using the perceptual distortion measure is much more important in the frequency estimation in (9) than in the estimation of the AM coefficients in (14) or in the phase estimation in (13). Thus, in an effort to reduce complexity, we suggest that the perceptual distortion weighting is only applied in (9).

4. AUDIO CODING USING THE AM DECOMPOSITION

Many audio segments are well-modeled using a CA sinusoidal model, and applying the proposed AM decomposition is not always preferable from a rate and distortion perspective. Rather, to enable efficient coding of both stationary and transient segments, we propose the use of combined coder, containing both a CA sinusoidal coder and a coder based on the AM decomposition. Herein, the AM decomposition has been incorporated into the experimental coder described in [5]; based on rate-distortion optimization, it is determined in each segment whether a AM or CA sinusoidal model should be used. We refer to such a combined coder as the AM/CA coder, using the term CA coder for the pure CA-based coder. The problem of rate-distortion optimiziation under rate constraint (i.e., finding the optimum distribution of $R^\star$ bits over I segments) can be written as (see [13] for further details)

$$\min \sum_{i=1}^{I} \left[D_i + \lambda R_i \right] = \sum_{i=1}^{I} \min \left[D_i + \lambda R_i \right], \qquad (20)$$

with $\lambda \geq 0$, where the right side follows from the assumption that the (nonnegative) distortions D_i and rates R_i are independent and additive over the ith segment, meaning that the cost function can be minimized independently for each segment. For a finite, discrete set of coding templates, $\mathcal{T}_i = \{c_1, \ldots, c_{L_c} a_1, \ldots, a_{L_a}\}$ with c_k being

(Can the below be right?)

k constant amplitude sinusoids and a_k being k amplitude modulated sinusoids for the ith segment with associated rates $R(\tau)$ and (perceptual) distortions $D(\tau)$,

(or am I just not getting it? :-)

the rate-distortion optimization reduce to selecting the optimum coding template as

$$\tau_i^\star = \arg \min_{\tau \in \mathcal{T}_i} \left[D(\tau) + \lambda^\star R(\tau) \right], \qquad (21)$$

when the λ leading to the target bitrate $R^\star$, denoted $\lambda^\star$, has been found. This is found by maximimizing the concave Lagrange dual function:

$$\lambda^\star = \arg \max_{\lambda} \left(\sum_{i=1}^{I} \min_{\tau \in \mathcal{T}_i} \left[D(\tau) + \lambda R(\tau) \right] \right) - \lambda R^\star. \qquad (22)$$

which is typically done by sweeping over a λ (using some fast method exploiting the convexity of $R(D)$) until the rate $R(\lambda)$ is sufficiently close to the target bitrate. **(Add reference!)** The selection between using an AM or CA sinusoidal model is decided using the following criterion

$$\min_k \left[D(a_k) + \lambda^\star R(a_k) \right] < \min_k \left[D(c_k) + \lambda^\star R(c_k) \right], \qquad (23)$$

implying that the AM coding templates a_k are chosen when these are the optimal choice among $\mathcal{T}_i$ for a particular segment. **(I don't understand what you mean here)**

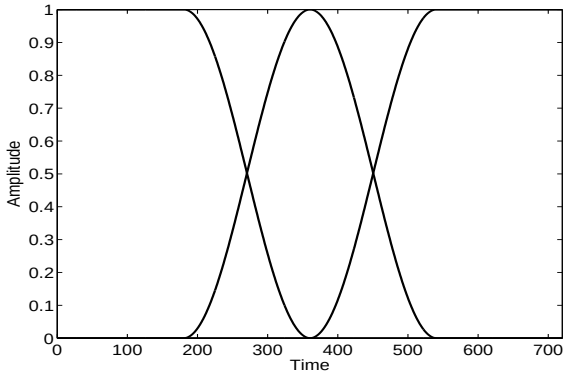

Fig. 1. AM bases.

5. EXPERIMENTAL RESULTS

5.1. Configuration

We have configured the AM/CA coder as follows; 50% overlapping von Hann windows of length 30 ms were used for both analysis and overlap-add synthesis. Further, the sinusoidal phases were quantized uniformly using 5 bits/component, whereas both the amplitudes and the frequencies were quantized using a logarithmic scale, with the amplitude quanitzer also used to quantize the AM coefficients, using the AM bases shown in Figure 1. As entropy (lossless) coding of the quantization indices are commonly used in audio coding, the rates are estimated as the entropies of the quantization indices, yielding approximately 9 bits/component for the frequencies and 6 bits/component for the amplitudes[1]. The quantizers were found to produce perceptually transparent results when compared to un-quantized signals.

(Don't like the below sentence!)

The rates of the coding templates of the rate-distortion optimization are estimated mean-rates of CA and AM sinusoids, being approximately 20 and 30 bits/component, respectively, with the distortions found using (18).

5.2. Informal Evaluation

Informal listening tests indicate that the AM/CA results in high perceived quality of coded excerpts for both stationary and transient parts. Generally, the type of signals that benefit from AM codings are signals exhibiting sharp onsets and stops, percussive sounds and changing signal types, such as transitions from unvoiced to voiced in speech signals. Often the improvements are perceived as an increase in bandwidth of the decoded signal. An illustrative way of showing the improved coding is by observing the rate-distortion tradeoffs; Figure 2 show the rate-distortion curves (or more correctly the distortion-rate) of a typical CA coder with and without the proposed AM subcoder. These curves were found by sweeping over λ in (20), finding the associated optimal rate and distortion pairs. As can be seen from the figure, there is a significant improvement in the rate-distortion tradeoff resulting from the proposed decomposition. It can also be seen that both curves saturates at higher rates. **(Add explanation?)**

[1] We note that these rates can be reduced significantly by differential encoding [14].

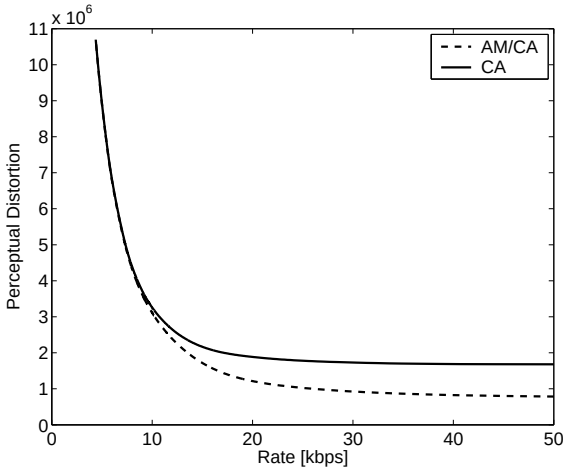

Fig. 2. The rate-distortion curves of a standard CA sinusoidal coder (solid) and that of the AM/CA coder (dashed) for glockenspiel.

Results of Listening Tests			
	Preference [%]		
Excerpt	AM/CA	CA	Significant
Glockenspiel			Yes
Castanets			Yes
Bass Guitar			Yes
English Female			Yes
Total			Yes

Table 1. Results of AB-preference tests.

5.3. Listening Test

An AB preference test was carried out using 7 different excerpts and 8 experienced listeners. The listeners were asked to choose between the CA coder and the AM/CA coder, both operating at a bitrate of approximately 30 kpbs. Each experiment was repeated 4 times in a balanced way such that each subject performed a total of 56 comparisons. The results are shown in Table 1. Significance was determined using a one-sided test with a level of significance of 0.05. The rate-distortion optimization was done such that the total entropy of the quantized parameters (including the envelopes) of the AM/CA coder was lower than of the CA coder. The tests clearly show that performance can indeed be improved using the proposed AM/CA coder for low bitrate audio coding.

6. CONCLUSION

In this paper, we have proposed a linear decomposition technique for amplitude modulated sinusoidal signals, showing that such a method might be used for high quality audio coding. Experiments indicate that a significantly higher rate of convergence, in terms of rate-distortion, can be achieved for transient segments when incorporating the proposed the method in a combined coder. This is also confirmed by listening tests, showing that, for a given bitrate, significant improvements are gained for the coder using the proposed decomposition. These results are promising for applications of amplitude modulation in low bitrate audio coding.

7. REFERENCES

[1] T. Painter and A. S. Spanias, "Perceptual Coding of Digital Audio," in *Proc. IEEE*, Apr. 2000, vol. 88(4).

[2] P. Prandoni, *Optimal Segmentation Techniques for Piecewise Stationary Signals*, Ph.D. thesis, Ecole Polytechnique Federale de Lausanne, 1999.

[3] M. G. Christensen, S. van de Par, S. H. Jensen, and S. V. Andersen, "Multiband amplitude modulated sinusoidal audio modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2004.

[4] E. G. P. Schuijers, A. W. J. Oomen, A. C. den Brinker, and A. J. Gerrits, "Advances in parametric coding for high-quality audio," in *Proc 1st. IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA-2002)*, 2002.

[5] M. G. Christensen and S. van de Par, "Rate-distortion efficient amplitude modulated sinusoidal audio coding," in *Proc. Asilomar Conf. Signals, Systems, and Computers*, 2004.

[6] G. Li, L. Qiu, and L. K. Ng, "Signal representation based on instantaneous amplitude models with application to speech synthesis," in *IEEE Trans. Speech, Audio Processing*, 2000, vol. 8(3).

[7] F. Myburg, A. C. den Brinker, and S. van Eijndhoven, "Sinusoidal analysis of audio with polynomial phase and amplitude," in *Proc. ProRISC*, 2001.

[8] S. L. Marple, "Computing the discrete-time "analytic" signal via FFT," in *IEEE Trans. Signal Processing*, Sept. 1999, vol. 47.

[9] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," in *IEEE Trans. Signal Processing*, Dec. 1993, vol. 40.

[10] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with application to target feature extraction," in *IEEE Trans. Signal Processing*, Feb. 1996, vol. 44(2).

[11] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psychoacoustical masking model for audio coding applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001.

[12] J. Jensen, R. Heusdens, and S. H. Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," in *IEEE Trans. Acoust., Speech, Signal Processing*, Mar. 2004, vol. 12(2).

[13] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," in *IEEE Trans. Acoust., Speech, Signal Processing*, Sept. 1988.

[14] J. Jensen and R. Heusdens, "A comparison of differential schemes for low-rate sinusoidal audio coding," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 2003.