Amplitude modulated sinusoidal signal decomposition for audio coding

Christensen, Mads Græsbøll; Jakobsson, Andreas; Andersen, Søren Vang; Jensen, Søren Holdt

Published in:
IEEE Signal Processing Letters

DOI (link to publication from Publisher):
10.1109/LSP.2006.871856

Publication date:
2006

Document Version
Accepted author manuscript, peer reviewed version

Amplitude Modulated Sinusoidal Signal Decomposition for Audio Coding

Mads Græsbøll Christensen*, Andreas Jakobsson, Søren Vang Andersen, and Søren Holdt Jensen

***Abstract*— In this paper, we present a decomposition for sinusoidal coding of audio, based on an amplitude modulation of sinusoids via a linear combination of arbitrary basis vectors. The proposed method, which incorporates a perceptual distortion measure, is based on a relaxation of a nonlinear least-squares minimization. Rate-distortion curves and listening tests show that, compared to a constant-amplitude sinusoidal coder, the proposed decomposition offers perceptually significant improvements in critical transient signals.**

EDICS: COM-CODE

I. Introduction

Perceptual audio coding aims at minimizing the perceived distortion at a given bit-rate. In parametric audio coding this is done by using signal models that capture the signal energy in a few parameters. It is critical that the signal models can represent a wide range of different audio signals, and especially transient signals have proven to be troublesome in that respect over the years. A number of different, and complementary, methods for dealing with transients have been proposed over the years, namely adaptive segmentation, noise-shaping, and variable rate [1]. Among these, adaptive segmentation is by far the most common approach for dealing with transients in audio coding. For an exhaustive discussion of the different methods and how they relate, see [1], [2]. Recently, amplitude modulated (AM) sinusoidal models have demonstrated to lead to improved coding of transients in audio even when rate-distortion optimal time segmentation [3] is employed [2]. Sinusoidal modeling using both amplitude and frequency modulation, in the form of low-order polynomials, has been explored (see, e.g., [4], [5]). Although such models perform well for slowly evolving signals, such as voiced speech, they can not handle the steep transients often encountered in audio signals satisfactorily.

In this paper, we treat a signal decomposition based on a set of preselected, linearly independent, real-valued basis vectors that describe the amplitude modulating signal and its application to audio coding. Specifically, we examine how to incorporate a perceptual distortion measure efficiently in the decomposition.

The rest of the paper is organized as follows: In Section II, both the signal decomposition and the solution to the associated minimization problem are presented. Experimental results are presented in Section III, and Section IV concludes on our work.

II. Proposed Decomposition

In the proposed decomposition, the signal of interest is modeled as a sum of L 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 amplitude modulating signal formed as a linear combination of I basis functions,

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

where $b(n,i)$ and $c_{i,l}$ denote the ith basis function evaluated at time instance n and the (i,l)th AM coefficient, respectively. We will here assume that the L carrier frequencies are distinct, so that $\omega_k \neq \omega_l$ for $k \neq l$. The additional flexibility in (1), as compared to the traditional constant-amplitude (CA) models with $\gamma_l(n) = A_l$, allows for improved modeling of transient segments. We note that the CA model is a special case of the modulated model, with the amplitude modulating signal being constant. 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 down-sampled by a factor two without loss of information provided that there is little or no signal of interest near 0 and π [6]. The signal model $x_a(n)$ can then be written as

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

Choosing the segment length, N, to be even, and introducing

$$\mathbf{x}_a = \left[\begin{array}{cccc} x_a(1) & x_a(3) & \cdots & x_a(N-1) \end{array} \right]^T, \tag{4}$$

where $(\cdot)^T$ denotes the transpose, the down-sampled discrete-time "analytical" signal can be expressed as

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

where $\odot$ denotes the Schur-Hadamard (element-wise) product. Further, $\mathbf{Z} \in \mathbb{C}^{N/2 \times L}$ with $L < N/2$ is constructed from the

Part of this work was presented at the IEEE Int. Conf. on Acoust., Speech and Signal Proc., Philadelphia, March 2005.

This research was supported by the ARDOR (Adaptive Rate-Distortion Optimized sound codeR) project, EU grant no. IST–2001–34095, and the Intelligent Sound project, Danish Technical Research Council grant no. 26–02–0092.

M. G. Christensen, S. V. Andersen, and S. H. Jensen are with the Department of Communication Technology, Aalborg University, Denmark (Contact information: phone: +45 96 35 86 20, email: mgc@kom.aau.dk).

A. Jakobsson is with the Dept. of Electrical Engineering, Karlstad University, Sweden.

L complex carriers with the (k,l)th element being $[\mathbf{Z}]_{kl} = e^{j\omega_l(2k-1)}$ and

$$\mathbf{a} = \begin{bmatrix} e^{j\phi_1} & \cdots & e^{j\phi_L} \end{bmatrix}^T. \tag{6}$$

The amplitude modulating signal 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, $\mathbf{B} \in \mathbb{R}^{N/2\times I}$ with $I < N/2$ and $\mathbf{C} \in \mathbb{R}^{I\times L}$. The problem of interest is, given a measured signal, $y(n)$, find $x(n)$ such that

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

where $\mathbf{W}$ is a perceptual weighting matrix, $\mathbf{y}_a$ is formed similar to $\mathbf{x}_a$, and $\|\cdot\|_2$ denotes the 2-norm. Here, $\mathbf{W}$ is derived from the auditory masking model proposed in [7]. This problem is nonlinear in the frequencies $\{\omega_k\}_{k=1}^L$, and is thus termed a nonlinear least-squares (NLS) minimization. Typically, this type of problem requires a multidimensional minimization which is computationally infeasible in most situations. Herein, we propose an iterative method for the minimization of (7). The method exploits that, for given $\{\omega_k\}_{k=1}^L$, the minimization problem 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 find $\mathbf{C}$ 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., found 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.,

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

where $\mathbf{D}_k$ is the diagonal matrix constructed from the kth carrier, with $z_k = e^{j\omega_k}$, i.e.,

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

Furthermore,

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

contains the kth residual, obtained as

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

In iteration 1, the residual is initialized as $r_1(n) = y_a(n)$. 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 $\hat{\mathbf{c}}_k^{(p-1)}$,

$$\hat{\phi}_k^{(p)} = \angle\left\{\left(\mathbf{c}_k^{(p-1)}\right)^T\mathbf{B}^T\mathbf{D}_k^H\mathbf{G}\mathbf{r}_k\right\}, \tag{13}$$

where $\mathbf{G} = \mathbf{W}^H\mathbf{W}$ and with $\angle(x)$ denoting the argument of x. Given $\hat{\phi}_k^{(p)}$, the minimization with respect to the AM coefficients reduces to

$$\hat{\mathbf{c}}_k^{(p)} = \mathrm{Re}\left\{e^{-j\hat{\phi}_k^{(p)}}\mathbf{\Pi}\mathbf{r}_k^{(p)}\right\}, \tag{14}$$

where

$$\mathbf{\Pi} = \left(\mathbf{B}^T\mathbf{D}_k^H\mathbf{G}\mathbf{D}_k\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{D}_k^H\mathbf{G}, \tag{15}$$

with $\mathrm{Re}(\mathbf{x})$ denoting the element-wise real part of $\mathbf{x}$. The parameters in (13) and (14) are then found alternately, given the other, until some stopping criterion is reached. For a given ω the problem is convex, and the algorithm converges to a global maximum. Hence, the 2-norm of the residual is a non-increasing function of the number of iterations. It also follows from the convexity that the initialization of the parameters in (13) and (14) is not critical since the estimates will converge to the same value regardlessly, but the required number of iterations may vary depending on the initialization.

It is important to understand that there is an inherent tradeoff between computational complexity and perceptual relevance in the choice of the perceptual weighting matrix. For the general case of the weighting matrix $\mathbf{W}$ having no particular structure, as in [8], the complexity of the proposed decomposition may be overwhelming. However, for certain structured matrices, the decomposition can be simplified significantly. For the particular auditory masking model proposed in [7], the weighting matrix $\mathbf{W}$ is a circulant and Hermitian matrix of the form

$$\mathbf{W} = \begin{bmatrix} w_0 & w_{M-1} & \cdots & w_1 \\ w_1 & w_0 & \cdots & w_{M-1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{M-1} & w_{M-2} & \cdots & w_0 \end{bmatrix}, \tag{16}$$

with $M = N/2$ and $w_n = \frac{1}{M}\sum_{m=0}^{M-1}\sqrt{A(m)}e^{-j2\pi mn/M}$ [9]. $A(m)$ is chosen as the reciprocal of the masking curve derived from the model presented in [7]. It follows that the eigenvalue decomposition (EVD) of $\mathbf{G}$ can be written as [10]

$$\mathbf{G} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H, \tag{17}$$

where the eigenvectors are

$$\mathbf{U} = \frac{1}{\sqrt{M}}\begin{bmatrix} \mathbf{u}_0 & \mathbf{u}_1 & \cdots & \mathbf{u}_{M-1} \end{bmatrix}, \tag{18}$$

with $\mathbf{\Lambda} = \mathrm{diag}(\begin{bmatrix} A(0) & \ldots & A(M-1) \end{bmatrix})$ and the columns $\mathbf{u}_f = [\, u_f^0 \ \cdots \ u_f^{M-1}]^T$ being composed from $u_f = e^{j2\pi f/M}$. Using the EVD, the calculation of the phase can be simplified as

$$\hat{\phi}_k^{(p)} = \angle\left\{\left(\mathbf{c}_k^{(p-1)}\right)^T\mathbf{B}^T\mathbf{D}_k^H\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H\mathbf{r}_k\right\}, \tag{19}$$

which can be computed efficiently by calculating $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H\mathbf{r}_k$ once for each k. Similarly, the computational complexity in finding the AM coefficients can be reduced by combining (15) and (17), yielding

$$\mathbf{\Pi} = \left(\mathbf{B}^T\mathbf{D}_k^H\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H\mathbf{D}_k\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{D}_k^H\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H. \tag{20}$$

The significant reduction in complexity stems from $\mathbf{U}$ being known and that all matrix-vector or matrix-matrix products involving $\mathbf{U}$ can be calculated using FFTs. For a thorough treatment on the application of the distortion measure to this problem, we refer to the discussion in [2].

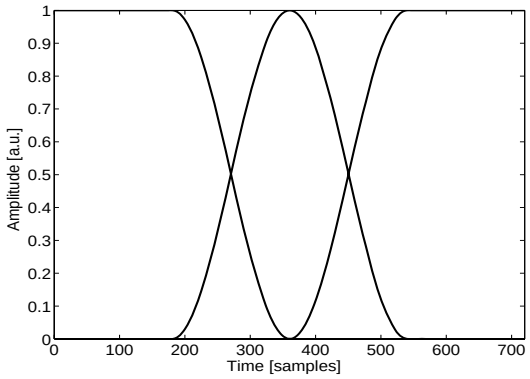

Fig. 1. The AM bases $b(n,i)$ for $n = 1, 3, \ldots N-1$ used in the experiments for a sampling frequency of 44.1 kHz and a segment size of 30 ms.

III. Experimental Details and Results

Many audio segments are well-modeled using a CA sinusoidal model, and applying the proposed AM decomposition is not always preferable from a rate-distortion point of view. Rather, to enable efficient coding of both stationary and transient segments, we propose the use of a 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 [2]. Based on rate-distortion optimization [3], it is determined in each segment whether an 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.

In the experiments to follow, von Hann windows of length 30 ms were used in both analysis and overlap-add synthesis with 50% overlap. Sinusoidal parameters are quantized as follows: phases are quantized uniformly using 5 bits/component, whereas amplitudes and frequencies are quantized in the logarithmic domain. We estimate the resulting rates as the entropies of the quantization indices, which gives approximately 9 bits/component for frequencies and 6 bits/component for amplitudes. The AM coefficients are also quantized using the amplitude quantizer. This leads to an average of 30 bits/component for amplitude modulated sinusoids and 20 bits/component for CA. The quantizers were found to produce perceptually transparent results compared original parameters.

In Figure 1, the particular AM bases used in the experiments are depicted. The AM bases are constructed from tapered cosine functions such that a smooth transitions are achieved. These have been chosen because they posses a number of attractive properties, namely that a linear combination of these vectors will result in a smooth modulating signal, and that they contain CA as a special case. The first property is desirable since non-smooth modulations may results in a perceived roughness in the synthesized signal, while the second property is desirable because not all sinusoidal components may be modulated in a particular segment. Here, we also note that the estimate in (13) is initialized such that the corresponding modulating signal has a constant amplitude and then (14) is found using that estimate.

We have considered two different implementations of the AM/CA coder. One, termed AM/CA (2-norm), where the perceptual distortion measure is only applied in the frequency estimation in (9) and the 2-norm is used in finding the AM coefficients, and one, termed AM/CA (p-norm), where the perceptual distortion measure is also applied in (13) and (14). The former is the least complex of the two, while the latter is expected to result in lower distortions. That this is actually the case is illustrated in Figure 2. The top panel of this figures shows a piece of the castanet signal from SQAM [11]. The middle panel shows the perceptual distortion of the different decompositions as a function of the entropy (number of bits or number of components) for the particular segment indicated by vertical lines in the top panel. This illustrates the convergence of the decompositions. In the bottom panel, the rate-distortion curves (or more correctly the distortion-rate curves) of the CA coder and the AM/CA coder are shown. The curves are calculated as described in [3], using the perceptual distortion measure in [7]. It can be seen that there is a significant improvement in the rate-distortion tradeoff resulting from the proposed decomposition as compared to the CA coder. It can also be seen that the AM/CA (p-norm) implementation is best in terms of perceptual distortion. However, the complexity-reduced implementation AM/CA (2-norm) achieves a performance very close to that of the AM/CA (p-norm). Therefore, this implementation has been used in the following listening tests.

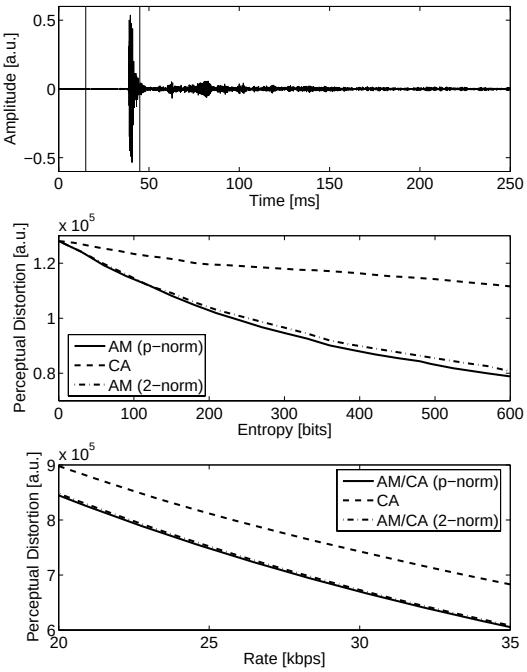

Fig. 2. The top panel shows the castanet signal while the middle panel shows the perceptual distortion as a function of the entropy for a particular segment. The rate-distortion curves for the castanet signal in top panel are depicted in the bottom panel.

Informal listening tests indicate that the combined AM/CA coder results in high perceived quality of coded excerpts for both stationary and transient parts. Generally, the type of signals that benefit from AM are signals that exhibit sharp onsets and stops, percussive sounds and changing signal types,

TABLE I

RESULTS OF AB-PREFERENCE TEST.

Excerpt	Preference [%] AM/CA	CA	Significant
Castanets	100	0	Yes
Claves	80	20	Yes
Glockenspiel	63	37	Yes
Harpsichord	63	37	Yes
Vibraphone	57	43	No
Xylophone	78	22	Yes
Total	74	26	Yes

such as transitions from unvoiced to voiced in speech signals. Often, the improvements are perceived as an increase in bandwidth. Also, a blind AB preference test with reference was carried out on headphones using 6 different critical excerpts from SQAM [11] having a length of 5-10 s. The seven listeners that participated were asked to choose between the CA coder and the AM/CA coder, both operating at a bit-rate of approximately 30 kbps. Each experiment was repeated 8 times in a randomized, balanced way. The results are shown in Table I. Significance was determined using a binomial distribution and a one-sided test with a level of significance of 0.05. The test shows that performance can be improved significantly using the proposed decomposition. We remark that, as shown in [2], both the CA and AM/CA coders may be further improved by the use of optimal segmentation [3]. This comes at the cost of increased delay and complexity, which may be prohibitive for some applications, e.g. [12].

IV. CONCLUSION

We have proposed a signal decomposition based on amplitude modulated sinusoids, and we have demonstrated that this decomposition may 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 method in a combined coder. This is also confirmed by listening tests, showing that for a given bit-rate, significant improvements can be gained for the coder using the proposed decomposition.

REFERENCES

[1] T. Painter and A. S. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88(4), pp. 451–515, Apr. 2000.

[2] M. G. Christensen and S. van de Par, "Efficient parametric coding of transients," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14(4), July 2006.

[3] P. Prandoni and M. Vetterli, "R/D optimal linear prediction," *IEEE Trans. Speech and Audio Processing*, pp. 646–655, 8(6) 2000.

[4] G. Li, L. Qiu, and L. K. Ng, "Signal representation based on instantaneous amplitude models with application to speech synthesis," *IEEE Trans. Speech and Audio Processing*, vol. 8(3), pp. 353–357, 2000.

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

[6] S. L. Marple, "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Processing*, vol. 47, pp. 2600–2603, Sept. 1999.

[7] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. on Applied Signal Processing*, vol. 9, pp. 1292–1304, June 2005.

[8] J. Plasberg, D. Zhao, and W. B. Kleijn, "Sensitivity matrix for a spectro-temporal auditory model," in *Proc. XII European Signal Processing Conf. (EUSIPCO)*, 2004, pp. 1673–1676.

[9] M. G. Christensen and S. H. Jensen, "On perceptual distortion minimization and nonlinear least-squares frequency estimation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14(1), Jan. 2006.

[10] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, 1996.

[11] European Broadcasting Union, *Sound Quality Assessment Material Recordings for Subjective Tests*. EBU, Apr. 1988, Tech. 3253.

[12] E. Allamanche, R. Geiger, J. Herre, and T. Sporer, "MPEG-4 Low Delay Audio Coding based on the AAC Codec," in *106th Conv. Aud. Eng. Soc.*, May 1999, paper preprint 4929.