LINEAR AM 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 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 non-linear least squares minimization. It offers benefits in the modeling of transients in audio signals. We compare the decomposition to constant-amplitude sinusoidal coding using rate-distortion curves and listening tests. Both indicate that, at the same bit-rate, perceptually significant improvements can be achieved using the proposed decomposition.

1. INTRODUCTION

The problem of decomposing a signal into amplitude modulated sinusoids is encountered in many different applications, for example in parametric audio coding (see, e.g., [1]) where modulated sinusoidal models are of interest for handling transients. Even when dynamic time segmentation [2, 3] is employed, there is a need for efficient modeling of transients. In [4], it was shown that perceptually significant improvements can be achieved by applying amplitude modulation (AM) in a frequency dependent way as opposed to single-banded AM (see, e.g., [5]). Furthermore, it was shown in [6] that frequency dependent AM achieves lower distortions compared to constant-amplitude (CA) sinusoidal coding at the same rate. 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., [7, 8]). Although such models perform well for slowly evolving signals like voiced speech, they do not handle the transients often encountered in audio signals well.

In this paper, we extend the work in [4, 6] by introducing a signal decomposition based on a set of preselected, linearly independent, real-valued basis vectors that describe the amplitude modulating signal. 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 solution to the associated minimization problem are presented, followed in Section 3 with the incorporation of a perceptual distoriton measure. Section 4 describes sinusoidal audio coding using the proposed AM decomposition. Experimental results are presented in Section 5, and Section 6 concludes on our work.

*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 the proposed decomposition, the signal of interest is modeled 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 amplitude modulating signal formed as the linear combination

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

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 models with $\gamma_l(n) = A_l$, gives improved modeling of transient segments. We note that the constant-amplitude model is a special case of the modulated model, with the amplitude modulating signal being DC. 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 [9] provided that there is little or no signal of interest near 0 and π. 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 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 down-sampled discrete-time "analytical" signal may be put into matrix-vector notation

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

where $\odot$ denotes the Schur-Hadamard 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, $\mathbf{Z} \in \mathbb{C}^{N/2 \times L}$ with $L < N/2$ is constructed from the L complex carriers, i.e., $[\mathbf{Z}]_{kl} = e^{j\omega_l(2k-1)}$, $\mathbf{a} = \begin{bmatrix} e^{j\phi_1} & \cdots & e^{j\phi_L} \end{bmatrix}^T$. 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

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

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. This problem is nonlinear in the frequencies $\{\omega_k\}_{k=1}^L$, and is thus called a nonlinear least squares (NLS) minimization. Typically, this type of problem requires a multidimensional minimization which is computationally infeasible in most situations. 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 [10] or recursive methods such as RELAX [11]. 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 $\{\omega_k\}_{k=1}^L$, 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.,

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

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

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

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 $\hat{\mathbf{c}}_k^{(p-1)}$,

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

Given $\hat{\phi}_k^{(p)}$, the minimization wrt. the AM coefficients reduces to

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

with

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

which can be pre-computed. 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 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, convex function of the number of iterations. 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 [10] with complex sinusoids.

3. INCORPORATING PERCEPTUAL DISTORTION

It is well-known that 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 would like to have a measure that takes as much as possible of the processing in the human auditory system into account, while on the other hand, we would like to have a measure that defines a mathematical norm and leads to efficient, simple estimators and quantizers. Here we apply the perceptual distortion measure presented in [12]. For a particular segment, the distortion D can be written as

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

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. While this measure is a spectral one, it is still inherently based on waveform matching since it operates on the Fourier transform of the time domain error, meaning that preechos, for example, will not go unpunished by the measure. With respect to audibility, the actual distortion values for non-stationary segments should be interpreted with care. In practice the spectral weighting function $A(\omega)$ is a discrete function, as is the error spectrum, and the distortion (18) is calculated as a summation of point-wise multiplications in the frequency domain. This corresponds to a circular filtering in the time domain. Putting this into matrix-vector notation, we get [13]

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

where $\mathbf{H}$ is an circular matrix constructed from the impulse response of the filter corresponding to $\sqrt{A(\omega)}$ and $\mathbf{W}$ is a diagonal weighting matrix containing the elements of the analysis window $w(n)$. Depending on the filter length, it may still be advantageous

to implement the filtering operation in the frequency domain. For further details on this procedure, we refer to, e.g., [13]. Using the perceptual distortion allows us to minimize a perceptually more meaningful measure than the 2-norm. However, doing so makes the pseudo-inverse $\mathbf{B}^\dagger$, defined in (15), frequency and segment dependent, forcing it to be re-calculated for each frequency and segment. Experimentally, we have found that the use of the perceptual distortion measure is much more important when minimizing wrt. the frequency in (9) than when solving for the AM coefficients in (14) and the phase in (13). Minimizing the perceptual distortion measure in (9) leads to the selection of the perceptually most important sinusoids. Thus, in order to minimize the complexity, we only apply the perceptual distortion measure in (9).

4. AUDIO CODING USING THE 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-distortion point of view. 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 [6]. Based on rate-distortion optimization, 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. Let $\mathcal{T}_s$ be a finite, discrete set of coding templates for segment s and $R(\tau)$ and $D(\tau)$ be the rate and distortion associated with coding template τ. Then, the problem of rate-distortion optimization under rate constraint (i.e., finding the optimum distribution of $R^\star$ bits over S segments) can be written as the following unconstrained problem (see [14, 2] for further details)

$$\sum_{s=1}^{S} \min_{\tau \in \mathcal{T}_s} \left[D(\tau) + \lambda R(\tau) \right], \tag{20}$$

with $\lambda \geq 0$. This follows from the assumption that the (non-negative) distortions and rates are independent and additive over the segments s. This means that the cost function can be minimized independently for each segment, for a given λ. Here we use the coding templates $\mathcal{T}_s = \{\psi_1, \ldots, \psi_{L_\psi}, \chi_1, \ldots, \chi_{L_\chi}\}$ with ψ_k being k constant-amplitude sinusoids and χ_k being k amplitude modulated sinusoids for segment s. When the optimal λ that leads to the target bit-rate $R^\star$, denoted $\lambda^\star$, has been found, the rate-distortion optimization simply becomes a matter of choosing the optimum coding template as

$$\tau_s^\star = \arg \min_{\tau \in \mathcal{T}_s} \left[D(\tau) + \lambda^\star R(\tau) \right]. \tag{21}$$

The optimal λ is found by maximizing the concave Lagrange dual function:

$$\lambda^\star = \arg \max_{\lambda} \left(\sum_{s=1}^{S} \left[\min_{\tau \in \mathcal{T}_s} D(\tau) + \lambda R(\tau) \right] - \lambda R^\star \right). \tag{22}$$

Typically, this is done by sweeping over λ (using some fast method exploiting the convexity of $R(D)$) until the rate $R(\lambda)$ is within some range of the target bit-rate [2]. We then chose between AM and CA using the following criterion

$$\min_k \left[D(\chi_k) + \lambda^\star R(\chi_k) \right] < \min_k \left[D(\psi_k) + \lambda^\star R(\psi_k) \right]. \tag{23}$$

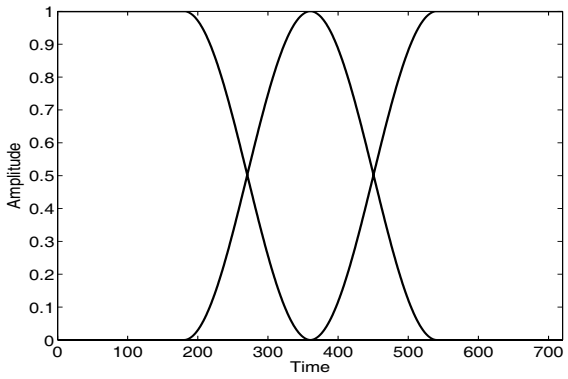

Fig. 1. AM bases used in the experiments.

Thus, AM coding template χ_k is chosen when it is the rate-distortion optimal choice among $\mathcal{T}_s$ for a particular segment.

5. EXPERIMENTAL RESULTS

5.1. Configuration

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. Since entropy coding of the quantization indices is commonly used in audio coding, 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 constant-amplitude. The quantizers were found to produce perceptually transparent results compared to unquantized parameters. In the rate-distortion optimization, distortions are calculated using unquantized values as the measure (18) may be overly sensitive to frequency quantization. Note that the rates can be reduced significantly by differential encoding [15].

5.2. Informal Evaluation

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, such as transitions from unvoiced to voiced in speech signals. Often, the improvements are perceived as an increase in bandwidth. In Figure 2, the rate-distortion curves (or more correctly the distortion-rate curves) of the CA coder and the AM/CA coder are shown. These were found by sweeping over λ in (20) and finding the associated optimal rate and distortion point. It can be seen that there is a significant improvement in the rate-distortion tradeoff resulting from the proposed decomposition. It can also be seen that the curve saturates at higher rates, meaning that lower distortions can be achieved.

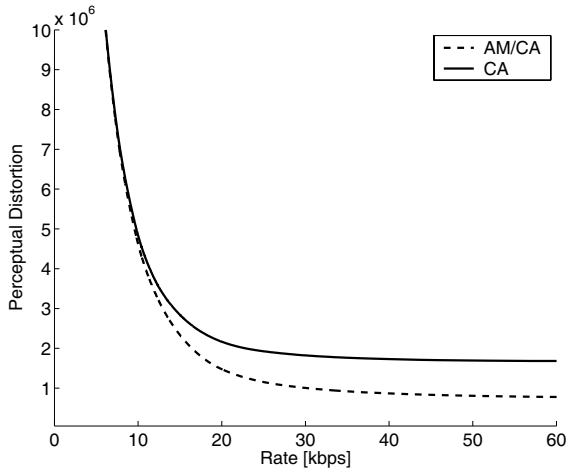

Fig. 2. The rate-distortion curves of the CA coder (solid) and that of AM/CA coder (dashed) for for the excerpt Glockenspiel.

Results of Listening Tests			
	Preference [%]		
Excerpt	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

Table 1. Results of AB-preference test.

5.3. Listening Test

A blind AB preference test with reference was carried out on headphones using 6 different transient excerpts from SQAM[1] with 7 inexperienced listeners participating. The listeners 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 1. 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.

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 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. These results are promising for applications of amplitude modulation in low bit-rate audio coding.

[1]The coded excerpts are currently available on the Internet at `http://kom.aau.dk/~mgc/projects/ldam/`.

7. REFERENCES

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

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

[3] R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio using psychoacoustical matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. 1809–1812.

[4] 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, vol. 4, pp. 169–172.

[5] E. G. P. Schuijers, A. W. J. Oomen, A. C. den Brinker, and J. Breebart, "Advances in parametric coding for high-quality audio," in *Audio Engineering Society, 114th Convention*, 2003, Preprint 5852.

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

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

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

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

[10] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41(12), pp. 3397–3415, Dec. 1993.

[11] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with application to target feature extraction," *IEEE Trans. Signal Processing*, vol. 44(2), pp. 281–295, Feb. 1996.

[12] 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, vol. 2, pp. 1805 – 1808.

[13] J. Jensen, R. Heusdens, and S. H. Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," *IEEE Trans. Speech, Audio Processing*, vol. 12(2), pp. 121–132, Mar. 2004.

[14] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 1445–1453, Sept. 1988.

[15] 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, pp. 205–208.