RATE-DISTORTION EFFICIENT AMPLITUDE MODULATED SINUSOIDAL AUDIO CODING

Mads Græsbøll Christensen

Dept. of Communication Technology
Aalborg University, Denmark
mgc@kom.aau.dk

Steven van de Par

Digital Signal Processing Group
Philips Research Labs, The Netherlands
steven.van.de.par@philips.com

ABSTRACT

In this paper, an improved parametric audio coder is presented. This coder addresses an important issue in audio coding, namely handling of transients. We propose a dedicated coder for transients based on amplitude modulated sinusoids. This coder is then combined with a constant-amplitude sinusoidal coder, and by rate-distortion optimization we choose which of the two is used for each segment. We show by rate-distortion curves and listening tests that the proposed coder offers significant improvements as compared to the constant-amplitude coder.

1. INTRODUCTION

One of the major challenges in audio coding is efficient handling of non-stationarities. The underlying signal models or transform bases are typically chosen such that a high coding efficiency is achieved for stationary signal parts, and, as a consequence, coding of non-stationary parts becomes highly inefficient. Typical solutions to these problems are e.g. adaptive segmentation using window-switching [1] or rate-distortion (R-D) optimal segmentation [2, 3]. In parametric audio modeling and coding, amplitude modulated (AM) sinusoidal models are of interest for capturing the features of transient sounds in an efficient way, e.g. for recordings of castanets. Damped sinusoids have received some attention for this purpose in the context of audio modeling [4, 5, 6]. Examples of AM in audio coding are [7] and [8], which are singlebanded in their definition, detection and encoding of transients, meaning that the envelope is the same for all components. In [9] it was demonstrated that significant improvements are gained by allowing different sinusoidal components to have different amplitude modulating signals. Since this study focused only on the modeling of audio signals, the question remains whether frequency dependent modulation methods are also efficient in terms of bit-rate. This issue is addressed in this paper where we present an amplitude modulated sinusoidal audio coder, which is efficient in terms of rate-distortion, meaning that at a given rate, it achieves a lower distortion compared to a conventional sinusoidal coder. The rest of the paper is organized as follows: In Section 2 the proposed signal model is presented followed by, in Section 3, the rate-distortion optimization used for coder switching. Sections 4 and 5 deal with the estimation and quantization of sinusoidal parameters, respectively. Experimental results are presented in Section 6 and in Section 7 we discuss the relation to existing work. Finally, Section 8 concludes on our work.

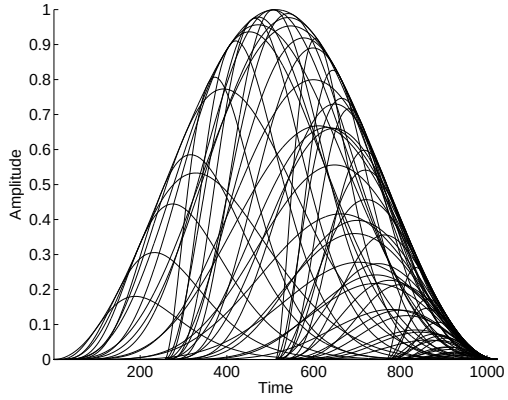

Fig. 1. Examples of windowed gamma envelopes.

2. SIGNAL MODEL

In this paper, we use the following amplitude modulated sinusoidal signal model for $n = 0, \ldots, N-1$:

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

where A_l, ω_l, and ϕ_l are the amplitude, frequency and phase of the l'th sinusoids, respectively. $\gamma_l(n)$ is the amplitude

M. G. Christensen was with the DSP group, Philips Research Labs, Eindhoven, The Netherlands, as a visiting researcher during this work. This research was supported by the ARDOR (Adaptive Rate-Distortion Optimized sound codeR) project, EU grant no. IST–2001–34095.

modulating signal or envelope for $\gamma_l(n) \geq 0$. Here we use a particular model for the envelopes called gamma envelopes:

$$\gamma_l(n) = u(n - n_l)\,(n - n_l)^{\alpha_l}\, e^{-\beta_l(n-n_l)}. \qquad (2)$$

Each envelope is characterized by an onset $n_l \in \mathbb{Z}$, an attack parameter $\alpha_l \in \mathbb{N}$, and a decay parameter $\beta_l \in \mathbb{R}^+$. Moreover, $u(n)$ is the unit step-function. We note that for $\alpha_l = 0$, $\beta_l = 0$ and $n_l = 0$, the model reduces to the conventional constant-amplitude (CA) model, i.e. $\gamma_l(n) = A_l$. In practice an analysis/synthesis window is also used on top of the gamma envelopes. Examples of windowed gamma envelopes are shown in Figure 1. Compared to the damped sinusoids of [4, 5, 6], this model has the additional flexibility of the attack parameter, and for the special case of $\alpha_l = 0$ and $\beta_l \neq 0$, the model reduces to damped sinusoids. In order to efficiently code both stationary segments as well as transients, we propose a coder that consists of two separate subcoders: the AM coder and the CA coder. These are based on the AM and the CA models, respectively. The combination of the two is termed AM/CA coder. We then switch between the two subcoders on a segment-to-segment basis using rate-distortion optimization.

3. RATE-DISTORTION OPTIMIZATION

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 with $\lambda \geq 0$ being the Lagrange multiplier

$$\sum_{s=1}^{S} \min_{\tau \in \mathcal{T}_s} \left[D(\tau) + \lambda R(\tau) \right]. \qquad (3)$$

$\mathcal{T}_s$ is a finite, discrete set of coding templates, i.e. ways of encoding, for segment s. $R(\tau)$ and $D(\tau)$ are the rate and distortion associated with coding template τ. For further details and proofs we refer to [10, 2]. Equation (3) follows from the assumption that the (non-negative) distortions and rates are independent and additive over the segments s. As a result, 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. Note that the AM coding templates may also include CA components, but the CA coding templates contain only CA components. 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

$$\tau_s^\star = \operatorname*{argmin}_{\tau \in \mathcal{T}_s} \left[D(\tau) + \lambda^\star R(\tau) \right]. \qquad (4)$$

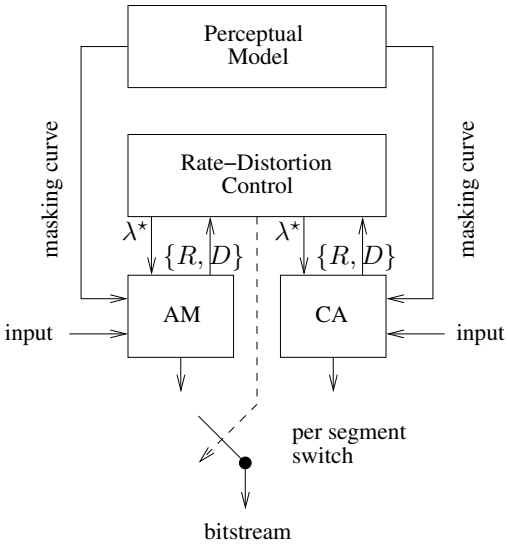

Fig. 2. AM/CA encoder.

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

$$\lambda^\star = \operatorname*{argmax}_{\lambda} \sum_{s=1}^{S} \left[\min_{\tau \in \mathcal{T}_s} D(\tau) + \lambda R(\tau) \right] - \lambda R^\star. \qquad (5)$$

This can be done by sweeping over λ using simple bi-section until the rate $R(\lambda)$ is within some range of the target bit-rate. It should be noted that for a discrete problem such as ours, we cannot guarantee that a solution exists that leads exactly to $R^\star$, and, as a consequence, the found solution may be suboptimal, but for a dense set of coding templates the gap will be small. The AM coding template χ_k is chosen when it is the rate-distortion optimal choice among $\mathcal{T}_s$ for a particular segment, i.e.,

$$\min_k D(\chi_k) + \lambda^\star R(\chi_k) < \min_k D(\psi_k) + \lambda^\star R(\psi_k). \qquad (6)$$

This principle is illustrated in Figure 2. Each segment is analyzed using both subcoders, and the resulting rate and distortion pairs are reported back to the rate-distortion control mechanism. For this procedure to work in the context of audio coding, the distortion measure must reflect the sensitivity of human auditory system. For this purpose we use the perceptual distortion measure proposed in [11].

4. SINUSOIDAL ESTIMATION

The parameters for each sinusoid can be found from finite, discrete sets using psychoacoustic matching pursuit (PMP) [12], which can be implemented using FFTs also for the AM case. This would guarantee convergence in the distortion as

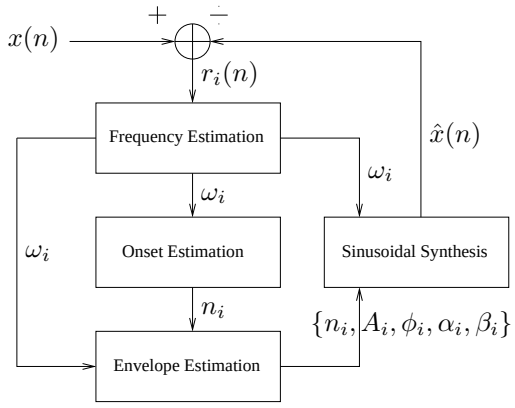

Fig. 3. The iterative AM estimation procedure. Sinusoids are found one at the time and subtracted from the input.

a function of the number of components. It would, however, be very expensive with respect to computational complexity. Instead, we here employ a simpler estimation procedure by noting that the number of different combinations of parameters will be dominated by the number of different frequencies and onset points. Thus, we break the estimation process into three steps: frequency estimation, onset estimation, and, finally, estimation of the envelope parameters and the corresponding phase and amplitude. A block diagram of the estimation procedure is shown in Figure 3. Both the encoder and the decoder use von Hann windows of length 35 ms, and the signal is reconstructed in the decoder using overlap-add with 50% overlap.

The objective of the estimation procedure is to find parameters such that the following perceptual distortion measure is minimized [11]:

$$D = \int_{-\pi}^{\pi} A(\omega) |\mathcal{F}\left[w(n)e(n)\right]|^2 d\omega, \tag{7}$$

where $\mathcal{F}[\cdot]$ denotes the Fourier transform, $A(\omega)$ is a real, positive perceptual weighting function, $w(n)$ is the analysis window, and $e(n) = x(n) - \hat{x}(n)$ is the modeling error with $x(n)$ being the observed signal. In order to shape the error spectrum like the masking treshold, the weighting function $A(\omega)$ is set to the reciprocal of the masking threshold. Here, we derive the masking threshold from [11]. Although this measure is inherently only strictly valid for stationary signals, it does not ignore temporal aspects completely as it is based on waveform matching. As a consequence, temporal errors, such as pre-echos, will not go unpunished by the measure. The measure has been found to comprise a reasonable tradeoff between complexity and correlation with perceived quality for coding purposes.

For the frequency estimation we use a fast method somewhat reminiscent to the weighted matching pursuit [13]. The algorithm operates on the residual, which at iteration $i+1$ is formed as

$$r_{i+1}(n) = r_i(n) - w(n)\gamma_i(n)A_i \cos(\omega_i n + \phi_i) \tag{8}$$

with $r_1(n) = w(n)x(n)$. Let $P_i(\omega) = R_i^*(\omega)R_i(\omega)$ be the power spectrum of the residual at iteration i, i.e. $R_i(\omega) = \mathcal{F}\left[r_i(n)\right]$. We then estimate the frequency as

$$\begin{aligned} \omega_i &= \underset{\omega}{\operatorname{argmax}}\, A(\omega)P_i(\omega) \\ \text{s.t.} \quad & \frac{\partial P_i(\omega)}{\partial \omega} = 0 \quad \text{and} \quad \frac{\partial^2 P_i(\omega)}{\partial \omega^2} < 0 \end{aligned} \tag{9}$$

This estimation criterion can be seen as an asymptotic PMP criterion with $N \to \infty$ for the CA case. The constraints ensure that the frequency will be a peak in the spectrum. This is a reasonable restriction as the amplitude modulating signals all have low-pass characteristics.

A coarse estimate of the integer onset n_i is found in order to limit the search space using the following simple method: Given a model where a sinusoidal component of frequency ω_i is modulated by a unit step-function $u(n-n_i)$, the modeling error can be written as

$$r_i(n) - w(n)u(n-n_0)A_i \cos(\omega_i n + \phi_i). \tag{10}$$

Using analytic signals, this error is minimized in a least-squares sense by maximizing the inner product (with proper normalization) between the modulated sinusoid and the residual. Defining $p(n) = r_i(n)w(n)\exp(-j\omega_i n)$ for $n = 0, \ldots, N-1$, which is calculated only once, the inner product can be written as

$$\Psi(n_0) = \frac{1}{N-n_0}\left|\sum_{n=n_0}^{N-1} p(n)\right|^2, \tag{11}$$

i.e., only the summation limit change over n_0. We then find the onset as the maximizer of this, i.e.

$$n_i = \underset{n_0}{\operatorname{argmax}}\, \Psi(n_0). \tag{12}$$

Given the frequency and the coarse onset, the combination of envelope parameters (including a refined onset estimate) is found as the minimizer of the distortion measure (7). This corresponds to performing a PMP on the subset of the dictionary. Assuming that all the dicitionary elements have been scaled for a particular segment such that

$$\int_{-\pi}^{\pi} A(\omega)\Gamma_k^*(\omega-\omega_i)\Gamma_k(\omega-\omega_i)d\omega = 1 \quad \forall k, \tag{13}$$

with $\Gamma_k(\omega) = \mathcal{F}\left[w(n)\gamma_k(n)\right]$ being the Fourier transform of a windowed envelope k in the dictionary, the envelope is then chosen as

$$\Gamma_i(\omega) = \underset{\Gamma_k(\omega)}{\operatorname{argmax}} \left|\int_{-\pi}^{\pi} A(\omega)\Gamma_k^*(\omega-\omega_i)R_i(\omega)d\omega\right|. \tag{14}$$

From this inner product, the phase and amplitude of the i'th sinusoid can also be found as the modulus and the argument, i.e.

$$A_i \exp(j\phi_i) = 2 \int_{-\pi}^{\pi} A(\omega)\Gamma_i^*(\omega - \omega_i)R_i(\omega)d\omega. \quad (15)$$

It is straightforward to extend (14) and (15) to the real case. In practice the spectra are discrete and the integration is performed as a summation over point-wise multiplications. As most of the spectral energy of $\Gamma_i(\omega - \omega_i)$ is concentrated in a small region around ω_i, the integration range can also be reduced without much loss of accuracy.

5. PARAMETER QUANTIZATION

The phases of the sinusoidal components are quantized uniformly using 5 bits, while amplitudes and frequencies are quantized in the logarithmic domain using the following quantizer (with θ denoting the parameter to be quantized and $\hat{\theta}$ denoting the quantized parameter)

$$\hat{\theta} = \exp\left(\left\lfloor \frac{\log(\theta)}{\log(1+\Delta)} + 0.5 \right\rfloor \log(1+\Delta)\right). \quad (16)$$

With a step-size Δ of 0.161 for the amplitudes and 0.003 for the frequencies, the quantizers were found to produce transparent results compared to the original parameters. For the gamma envelopes we have found 8-10 bits/component to produce good results with most of the bits being spent on the onset grid. In the following tests, an envelope dictionary size of 8 bits were used with $\alpha_l \in \{4, 3\}$, $\beta_l \in \{0.01, 0.005\}$ and an onset n_l step-size of approximately 0.5 ms. Estimated entropies of the quantized parameter sets were used for the rates in the R-D optimization and as a measure of rate in the experiments to follow. For CA this was estimated as approximately 16 bits/component (assuming differential encoding [14]) and 24 bits/sinusoid for AM. Additionally, a 1 bit AM switch is used per component for the AM coding templates. This allows efficient coding of CA components. As the perceptual distortion measure (7) may be overly sensitive to frequency quantization, we use the original parameters in determining the distortions.

6. EXPERIMENTAL RESULTS

In Figure 4 the rate-distortion curves for a representative transient signal, Glockenspiel, are shown for the CA coder (solid) and the AM/CA coder (dashed). It can be seen that there is a clear improvement in terms of reduction of distortion at the same rate. Also, the proposed coder saturates at lower distortions than the CA subcoder. Informal listening tests reveal that pre-echos are clearly reduced and that the transients are better modeled. The types of signals that benefit from the AM coder are signals that exhibit fast onsets,

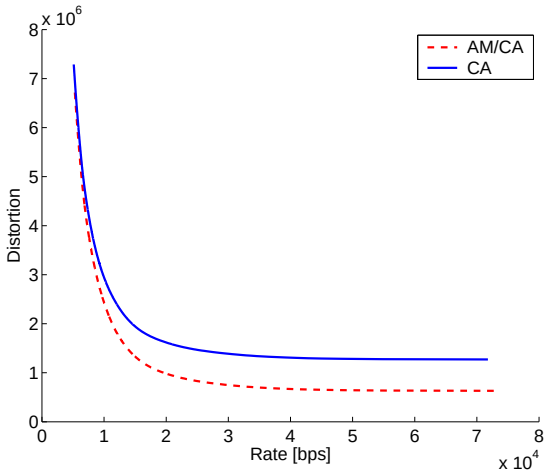

Fig. 4. The rate-distortion curves of the CA coder (solid) and that of AM/CA coder (dashed) for the excerpt Glockenspiel.

impulse-like signals, transitions between different types of signals, and percussive instruments. Any mixture of these types of signals with stationary ones may also benefit from it. A formal listening test has also been performed in the form of a blind AB preference test with reference. The test was performed on speakers in a listening room. Seven different excerpts sampled at 48 kHz were use, and each experiment was repeated 8 times in a randomized, balanced way. Eight experienced listeners were asked to choose between the AM/CA coder and the CA coder at a bit-rate of approximately 30 kbps. In Table 1, the results are shown. A one-sided test based on a binomial distribution with a level of significance of 0.05 was used for determining significance. The listening test shows that performance is improved significantly using the proposed method.

7. DISCUSSION

In audio coding, transient detectors have been used for switching between different window lengths and shapes [1]. The problem with this approach is that these detectors may not reflect the human auditory system well and there may be a mismatch between the classification of transients and the encoding techniques. Based on R-D optimization we gain robustness against such problems, as the transient coder is only used when it is the optimum choice. An alternative, or complement, to AM is R-D optimal time-segmentation [2], which has received attention in the context of parametric audio coding, see e.g. [3]. While the problem of time-segmentation can be solved optimally, it still has some drawbacks. The overlap between adjacent segments is typically fixed to half the minimum segment size. Such small over-

Results of Listening Test			
	Preference [%]		
Excerpt	AM/CA	CA	Significant
Abba	64	36	Yes
Glockenspiel	90	10	Yes
Castanets	98	2	Yes
Harpsichord	76	24	Yes
Bass Guitar	78	22	Yes
Lemon Tree	56	44	No
English Female	78	22	Yes
Total	77	23	Yes

Table 1. Results of AB-preference test at 30 kbps.

laps may increase sensitivity to quantization errors. Also, the complexity and delay of the R-D optimal time-segmentation may be prohibitive for some application. Compared to the singlebanded AM of e.g. [8], the proposed model has the advantage that different envelopes are allowed for different sinusoids, which is a particular advantage for mixtures of sources (see e.g. [9]).

8. CONCLUSION

In this paper, we have presented a new parametric audio coder. This coder consists of two complementary subcoders, namely a constant-amplitude sinusoidal coder and the proposed amplitude modulated coder. The latter uses a model where each sinusoidal component is modulated by a signal known as a gamma envelope, which is characterized by an onset, an attack parameter and a decay parameter. We then switch between the subcoders using rate-distortion optimization and a perceptual distortion measure. From listening tests and rate-distortion curves we conclude that the proposed method improves significantly on a sinusoidal coder at the same bit-rate.

9. REFERENCES

[1] B. Edler, "Codierung von audiosignalen mit überlappender transformation und adaptiven fensterfunktionen," *Frequenz*, pp. 1033–1036, 1989.

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

[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] 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.

[5] M. M. Goodwin, "Matching pursuit and atomic signal models based on recursive filter banks," *IEEE Trans. Signal Processing*, vol. 47(7), pp. 1890–1902, July 1999.

[6] R. Boyer and K. Abed-Meraim, "Audio modeling based on delayed sinusoids," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 12(2), pp. 110 – 120, Mar. 2004.

[7] B. Edler, H. Purnhagen, and C. Ferekidis, "ASAC – Analysis/Synthesis Audio Codec for Very Low Bit Rates," in *100th Conv. Aud. Eng. Soc., preprint 4179*, May 1996.

[8] 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.

[9] 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.

[10] 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.

[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, vol. 2, pp. 1805 – 1808.

[12] R. Heusdens, R. Vafin, and W. B. Kleijn, "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits," *IEEE Signal Processing Lett.*, vol. 9(8), pp. 262–265, Aug. 2002.

[13] T. S. Verma and T. H. Y. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1999, vol. 2, pp. 981–984.

[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, pp. 205–208.