Efficient Parametric Coding of Transients

Christensen, Mads Græsbøll; van de Par, Steven

Published in:
IEEE transactions on speech and audio processing

Publication date:
2006

Document Version
Accepted author manuscript, peer reviewed version

Citation for published version (APA):
Christensen, M. G., & van de Par, S. (2006). Efficient Parametric Coding of Transients. *IEEE transactions on speech and audio processing*, *14*(4), 1340-1351.

Efficient Parametric Coding of Transients

Mads Græsbøll Christensen*, *Student Member, IEEE*, and Steven van de Par

***Abstract*— In this paper, methods for improved parametric coding of transients are presented. We propose a signal model for coding of transients consisting of a sum of sinusoids each being amplitude-modulated by a different gamma envelope. These envelopes are characterized by an onset time, an attack and a decay parameter. An efficient method for estimating these parameters is presented. Further, methods are proposed that combine this transient model with a constant-amplitude sinusoidal model in order to achieve efficient coding of both stationary and transient signal parts. By rate-distortion optimization using a perceptual distortion measure we combine variable rate bit allocation and segmentation in an optimal way. Formal as well as informal listening tests show that significant improvements can be achieved with the proposed model as compared to a state-of-the-art sinusoidal coder by the combination of optimal segmentation and amplitude modulated sinusoidal audio coding.**

I. Introduction

In the past couple of decades, sinusoidal models for digital processing of speech and audio have received much attention for a wide variety of applications where sinusoidal speech coding and modeling [1]–[4] was among the first and perhaps the most prominent. Also for analysis and synthesis of music [5], [6] the sinusoidal model has been of interest. In recent years, the growth of the Internet and wireless communication has spurred renewed interest in sinusoidal models, this time for coding of audio [7]–[15] at low bit-rates. In perceptual audio coding, compression is achieved by exploiting statistical redundancies as well as perceptual irrelevancies of the source (see e.g. [16]). In parametric audio coding, a compact representation of the source signal is achieved using parametric models and the statistical redundancies and irrelevancies of the model parameters are exploited for efficient coding.

A major challenge in audio coding in general is efficient coding of non-stationary segments (see e.g. [16]). Signal models and 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. Sinusoidal coding using constant-amplitude (CA) sinusoids is an example of this difficulty. The inefficient coding of transients leads to a number of problems. Firstly, errors introduced before onsets are very poorly masked compared to the situation where a simultaneous masker is present [17]. These types of errors are known as pre-echos. Secondly, bad modeling of transients leads to very dull sounding attacks and a perceived lack of bandwidth of the decoded signal. The typical solution to these problems are adaptive segmentation using window switching [18] and window shape adaptation or rate-distortion (R-D) optimal segmentation [14], [19], [20]. Other methods that aim at solving this problem include wavelet-packets [21], temporal noise shaping (TNS) [22], gain modification [23], [24], transient location modification [25], switching from a parametric signal model to a wavelet or transform representation [7], [9], multi-resolution sinusoidal modeling [26] and coding of transients using sinusoidal modeling in the transform domain [27]. In parametric audio modeling and coding, transients can be handled by adapting the signal model to better fit the input signal. A particularly interesting class of such adapted models are the amplitude modulated (AM) sinusoidal models[1] [28]. In these models, the signal is decomposed into a sum of sinusoidal components having a time-varying envelope. The different realizations of damped sinusoids that have been applied to audio modeling in [29]–[33] are examples of this. In audio coding AM has been applied in [8], [13]. Like [5] these use a singlebanded model of the modulating signal meaning that the envelope is the same for all components. In [34] it was demonstrated that significant improvements are achieved by allowing different sinusoidal components to have different amplitude modulating signals. Since this study focused only on modeling of audio signals, the question remains whether frequency-dependent AM methods are also efficient in terms of bit-rate, i.e., whether they achieve a lower distortion, both subjectively and objectively, compared to a conventional sinusoidal coder at the same rate.

In the present paper we seek to answer that question along with some other unanswered questions regarding parametric coding of transients. We present a coder based on a particular model of the amplitude modulating signal known as gamma envelopes. Figure 1 shows the waveform of a sinusoid modulated by a windowed gamma envelope. The gamma envelopes are characterized by an onset time, an attack and a decay parameter. This model differs from existing models used for parametric modeling and coding of audio in that each sinusoid can have a different envelope with an onset at an arbitrary position within a segment, and in that it is characterized by an attack parameter. In addition to the new signal model, the proposed coder incorporates rate-distortion optimal bit allocation and segmentation. Further, we consider different ways of achieving efficient coding of both stationary and transient signal parts. Finally, we quantify, by subjective listening tests, the performance of the different methods for

Part of this work was presented at the 38th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, November 2004.

M. G. Christensen is with the Department of Communication Technology, Aalborg University, Denmark (phone: +45 96 35 86 72, email: mgc@kom.aau.dk, homepage: http://kom.aau.dk/~mgc). When this work was conducted, he was a visiting researcher with the Digital Signal Processing group, Philips Research Laboratories, Eindhoven, The Netherlands.

S. van de Par is with the Digital Signal Processing group, Philips Research Laboratories, Eindhoven, The Netherlands (email: steven.van.de.par@philips.com).

This research was supported by the ARDOR (Adaptive Rate-Distortion Optimized sound codeR) project, EU grant no. IST–2001–34095.

[1] In this text, AM means either amplitude modulation or amplitude modulated depending on the context.

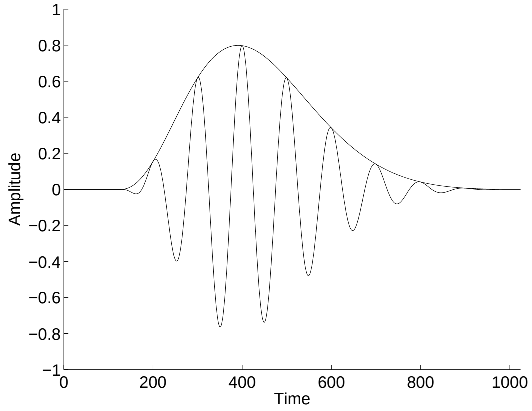

Fig. 1. Illustration of a sinusoid modulated by a windowed gamma envelope. The gamma envelopes are parameterized by an onset, an attack parameter and a decay parameter.

different types of signals.

The main part of this paper is organized as follows: in Section II the proposed signal model and the perceptual distortion measure which is instrumental in this work are presented. The rate-distortion optimization used for allocation and segmentation is presented in Section III, and Sections IV and V deal with the estimation of sinusoidal parameters. Implementation details, the experimental setup for perceptual tests and their results are presented in Sections VI and VII, respectively. In Section VIII we discuss the relation to existing work, and, finally, in Section IX we conclude on our work.

II. Fundamentals

The presented coder can be described as comprising the following steps: in the encoder, the input signal is split into a number of overlapping segments and a window is applied to each segment. The model parameters are then estimated and subsequently quantized, entropy coded and finally put into the bit-stream. In the decoder, the bit-stream is mapped back to the quantized parameters, and the segment is synthesized using overlap-add with an appropriate window.

In this paper, we propose a coder based on the following amplitude modulated sinusoidal signal model for time index $n = 0, \ldots, N-1$:

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

where A_l, ω_l, and ϕ_l are the amplitude, frequency and phase of the l'th sinusoids, respectively. The number of components is denoted L and $\gamma_l(n)$ is the modulating signal or envelope when $\gamma_l(n) \geq 0 \; \forall n$. Here we use a particular model of the envelopes which we shall henceforth refer to as gamma envelopes. This model is derived from the integrand of the gamma function, which is commonly used to characterize the gamma distribution in statistics. The gamma envelopes are given as

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

Each envelope is characterized by an onset time $n_l \in \mathbb{Z}$, an attack parameter $\alpha_l \in \mathbb{N}$, and a decay parameter $\beta_l \in \mathbb{R}^+$. Moreover, $u(n)$ is the unit step sequence. The envelopes composed from all possible combinations of these parameters will henceforth be referred to as the envelope dictionary. Inserting (2) into (1), we get the so-called gamma-tones commonly used as stimuli in psychoacoustical experiments and for modeling of the auditory filters [35]. Here, we rather use it as a signal model that, as we shall see, has been found to perform well for the problem at hand. The distinction between the model parameters α_l and β_l in (2) is only figurative since changing β_l for a fixed α_l will affect the attack and α_l will likewise affect the decay. We note that for $\alpha_l = 0$, $\beta_l = 0$ and $n_l = 0$, the lth sinusoid reduces to a constant-amplitude (CA) sinusoid, i.e. $\gamma_l(n) = 1$. The situation where all components have constant amplitude will be termed the CA model. For $\alpha_l = 0$ and $\beta_l \neq 0$ for all l, the model reduces to the so-called delayed damped sinusoids of [32], and with $\alpha_l = 0$ and $n_l = 0$ it becomes equivalent to the damped sinusoids of [30], [33]. Compared to the different variations of damped sinusoids of [29]–[32], this model has the additional flexibility of the attack parameter. It is well-known that different instruments do have different attacks, and studies show that the attacks are in fact important features in the recognition of musical instruments [36]. This can also be witnessed from the many transient signals on the SQAM disc [37].

In finding the model parameters and in the R-D optimization, it is advantageous to use a perceptual distortion measure since we seek to minimize the perceived distortion. In choosing a distortion measure we face conflicting demands. On one hand we wish to use a distortion measure that takes as much of the human auditory system into account as possible. On the other hand we wish to have a distortion measure that is both of reasonably low computational complexity and defines a norm such that it may be subject to optimization. Consequently, we have chosen the spectral distortion measure of [38], which is defined as

$$D = \int_{-\pi}^{\pi} A(\omega) |E(\omega)|^2 d\omega, \tag{3}$$

where $A(\omega)$ is a real, positive perceptual weighting function, and $E(\omega)$ denotes the discrete-time Fourier transform of the windowed error, i.e.,

$$E(\omega) = \sum_{n=0}^{N-1} w(n) e(n) e^{-j\omega n}, \tag{4}$$

with $w(n)$ being the analysis window, $e(n) = x(n) - \hat{x}(n)$ the modeling error, and $x(n)$ the observed signal. We note in passing that this and all other Fourier transforms will in practice be calculated for discrete values of ω. In order to shape the error spectrum according to the masking threshold, the weighting function $A(\omega)$ is set to the reciprocal of the masking threshold. Here, we derive the masking threshold from [38]. This distortion measure improves on other models in that it takes the spectral integration in the human auditory system into account. Although the measure is strictly only valid for stationary signals, it does not ignore temporal aspects completely as it is based on waveform matching. In order to achieve a low distortion, the phase and temporal envelope of the coded signal must match that of the original. As a consequence, temporal errors, such as pre-echos, will not go unpunished by the measure. The spectral distortion measure has been found to comprise a reasonable tradeoff between complexity and correlation with perceived quality for coding purposes and as we shall see, good results can be achieved using it. Henceforth, when we refer to distortions, we mean the perceptual distortion defined in (3).

The discrete-time Fourier transform of $\gamma_l(n)$ denoted $\Gamma_l(\omega)$ can be shown to be

$$\Gamma_l(\omega) = \sum_{n=0}^{N-1-n_l} n^{\alpha_l} e^{-j\omega n_l} \left(e^{-j\omega-\beta_l}\right)^n \quad (5)$$

$$= j^{\alpha_l} \frac{\partial^{\alpha_l}}{\partial\omega^{\alpha_l}} \frac{e^{-j\omega n_l} - e^{-\beta_l(N-n_l)}e^{-j\omega N}}{1-e^{-\beta_l}e^{-j\omega}}. \quad (6)$$

As indicated by (4), an analysis window is applied to the gamma envelopes. In the decoder, a window is also used in the synthesis, which is performed using overlap-add with a fixed overlap. Both the encoder and the decoder use tapered von Hann windows of the same length. With M denoting the overlap in samples and N being the (even) segment length, the windows are defined for $n = 0, \ldots, N-1$ as

$$w(n) = \begin{cases} v(n), & 0 \quad \leq n < \quad M \\ 1, & M \quad \leq n < \quad N-M \\ v(n-N+2M), & N-M \quad \leq n < \quad N \end{cases} \quad (7)$$

with the even length von Hann window being defined as

$$v(n) = \frac{1}{2} - \frac{1}{2}\cos\left(\frac{\pi(n+0.5)}{M}\right). \quad (8)$$

Let $W(\omega)$ denote the discrete-time Fourier transform of the window $w(n)$. Then the discrete-time Fourier transform of the windowed envelope can be written as the circular convolution

$$Z_l(\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Gamma_l(\omega-\xi)W(\xi)d\xi. \quad (9)$$

Hence, the window, which has low-pass characteristics, smoothes the spectrum. As the windowed gamma envelopes have no discontinuities at segment boundaries the spectrum of the windowed gamma envelopes will generally be more well-behaved than when no window is applied. This is important since the distortion measure will punish spectral distortion due to not only the mainlobe but also the sidelobes. In Appendix I, a closed-form expression of the discrete-time Fourier transform of the windowed gamma envelopes is derived.

III. R-D Optimal Allocation and Segmentation

Since audio signals may exhibit varying degrees of stationarity, it is often advantageous to allow for a flexible segmentation and allow the bit-rate to vary over time. In addition, it is observed that the proposed AM signal model is only efficient in terms of rate-distortion for transient segments, while the CA model is an efficient representation of tonal stationary segments. In order to combine the two models in an optimal way as well as doing optimal segmentation of the input signal, we use rate-distortion optimization. Further, the rate-distortion optimization also results in a rate-scalable coder, which is advantageous in dealing with critical signal parts. For completeness we now briefly review the basic definitions, assumptions and results for solving the problem of optimal segmentation and allocation based on [19], [39]. First, let us start out by introducing some definitions. We define a segment σ_s as having a length of a positive integer multiple $m \in \mathbb{Z}^+$ of a minimum segment length κ, i.e. $\ell(\sigma_s) = \kappa m$, and a segmentation as $\boldsymbol{\sigma} = [\ \sigma_1 \ \cdots \ \sigma_S\ ]$ consisting of S disjoint, contiguous segments that satisfy

$$\sum_{s=1}^{S} \ell(\sigma_s) = \kappa M, \quad (10)$$

where κM is the total length of the signal to be encoded. Each of these segments, say segment σ_s, can then be encoded using a set of coding templates $\mathcal{T}_s$ (different models, model orders, number of bits, etc.). Next, we define $R(\sigma_s, \tau_s)$ and $D(\sigma_s, \tau_s)$ as the non-negative cost in bits and distortion associated with coding template $\tau_s \in \mathcal{T}_s$ for segment σ_s. Assuming that the distortions and cost in bits associated with a particular segmentation $\boldsymbol{\sigma}$ and coding templates $\boldsymbol{\tau} = [\ \tau_1 \ \cdots \ \tau_S\ ]$ are additive over the segments, we can write the total distortion and total number of bits as

$$D(\boldsymbol{\sigma},\boldsymbol{\tau}) = \sum_{s=1}^{S} D(\sigma_s,\tau_s) \quad R(\boldsymbol{\sigma},\boldsymbol{\tau}) = \sum_{s=1}^{S} R(\sigma_s,\tau_s), \quad (11)$$

respectively. The problem of distributing a certain number of bits over a number of quantizers can be cast into the problem of rate-distortion optimization under rate constraint. This can be stated as the following constrained optimization problem:

$$\begin{array}{ll} \min & D(\boldsymbol{\sigma},\boldsymbol{\tau}) \\ \text{s.t.} & R(\boldsymbol{\sigma},\boldsymbol{\tau}) = R^\star, \end{array} \quad (12)$$

with $R^\star$ being the bit budget, i.e. the total number of bits to be distributed. Next, introducing the Lagrange multiplier $\lambda \geq 0$, the constrained optimization problem in (12) can be written as the unconstrained minimization problem [39]

$$J(\lambda) = \min_{\boldsymbol{\sigma}} \min_{\boldsymbol{\tau}} \sum_{s=1}^{S} D(\sigma_s,\tau_s) + \lambda(R(\sigma_s,\tau_s) - R^\star). \quad (13)$$

We now have an outer minimization over the segmentation, and an inner minimization over coding templates given the segmentation. In (11) we assumed that $D(\cdot)$ and $R(\cdot)$ are additive over segments. By also assuming that they are independent over segments, the inner minimization in (13) can be simplified significantly. Specifically, the optimization problem reduces to the following, where the coding templates can be optimized independently for a segmentation and a particular λ [19]:

$$J(\lambda) = \min_{\boldsymbol{\sigma}} \sum_{s=1}^{S} \min_{\tau\in\mathcal{T}_s} \left[D(\sigma_s,\tau) + \lambda R(\sigma_s,\tau)\right] - \lambda R^\star. \quad (14)$$

This leads to the following important result: as the rates and distortions are additive over segments, the outer minimization can be solved using dynamic programming [19]. The optimal

λ that leads to the target rate $R^\star$, denoted $\lambda^\star$, can be found by maximizing the concave Lagrange dual function [40], i.e.,

$$\lambda^\star = \operatorname*{argmax}_{\lambda} J(\lambda) \tag{15}$$

This can be done by sweeping over λ until $R(\boldsymbol{\sigma}, \boldsymbol{\tau})$ is within some range of the bit budget [19]. It should be noted that for a discrete problem such as ours, we cannot guarantee that strong duality holds for the optimization problem, and, as a consequence, the found solution may be suboptimal, but for a dense set of coding templates the gap will be small (see [40]). For a fixed segmentation, i.e. given $\boldsymbol{\sigma}$, the outer minimization disappears, and we only have to minimize over the coding templates. This was the approach used in [41].

IV. Parameter Estimation

The distortion measure (3) defines a norm and is in fact induced by an inner product (see [42]). The parameters for each sinusoid can then be found using a matching pursuit algorithm [43]. This would guarantee convergence in the distortion as a function of the number of components. The psychoacoustic matching pursuit (PMP) [42] is an algorithm that does this, i.e. it performs matching pursuit using the norm (3). The inner products can be found using FFTs also for the AM case. It would, however, be very expensive with respect to computational complexity. Since the R-D optimal segmentation requires that at every segment boundary, all combinations of segment lengths and coding templates are evaluated, it is critical that the estimation procedure is fast. In that spirit, we here employ a simpler procedure than PMP. We start out by noting 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 successive 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 2.

For the frequency estimation we use a fast method somewhat reminiscent of the weighted matching pursuit [44]. The algorithm operates on the residual, which at iteration $i+1$ is formed as

$$y_{i+1}(n) = y_i(n) - w(n)\gamma_i(n) A_i e^{j(\omega_i n + \phi_i)}. \tag{16}$$

The residual is initialized as the discrete-time analytic signal

$$y_1(n) = w(n)x(n) + jw(n)\mathcal{H}\left\{x(n)\right\}, \tag{17}$$

where $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform. This, including windowing, is the preprocessing step in Figure 2. In practice, the Hilbert transform is found using the FFT method. By operating on the analytic signal, we ignore the spectral contents of $x(n)$ for negative frequencies. This is done in order to simplify the estimation procedure. Convergence in the modeling of the analytic signal also ensures convergence in the real signal since

$$\Re\left\{w(n)x(n) + jw(n)\mathcal{H}\left\{x(n)\right\}\right\} = w(n)x(n), \tag{18}$$

however, for a non-zero error, the analytic signal modeling will introduce some error due to the correlation between negative and positive sides of the spectrum.

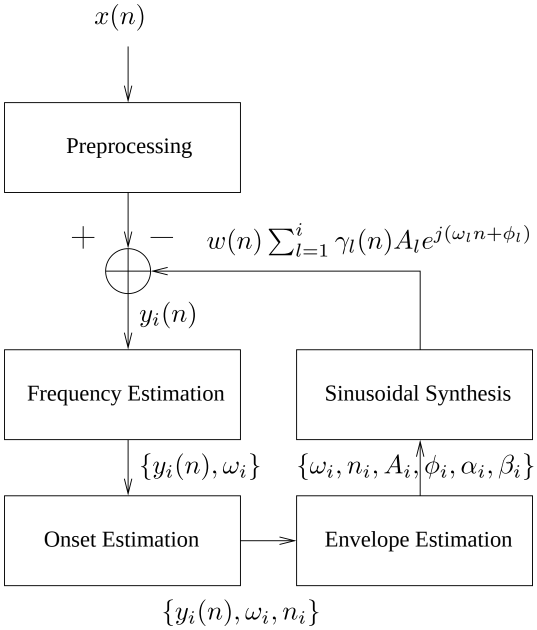
$x(n)$
Preprocessing
$+$
$-$
$w(n)\sum_{l=1}^{i}\gamma_l(n)A_l e^{j(\omega_l n+\phi_l)}$
$y_i(n)$
Frequency Estimation
Sinusoidal Synthesis
$\{y_i(n), \omega_i\}$
$\{\omega_i, n_i, A_i, \phi_i, \alpha_i, \beta_i\}$
Onset Estimation
Envelope Estimation
$\{y_i(n), \omega_i, n_i\}$

Fig. 2. The iterative AM parameter estimation procedure. Sinusoids are found one at the time and subtracted from the input.

Let $P_i(\omega) = Y_i^*(\omega)Y_i(\omega)$ be the squared magnitude of the discrete-time Fourier transform of the residual at iteration i , i.e.,

$$Y_i(\omega) = \sum_{n=0}^{N-1} y_i(n) e^{-j\omega n}, \tag{19}$$

which may be updated efficiently in the frequency domain. Then the frequency is estimated as

$$\begin{aligned} \omega_i &= \operatorname*{argmax}_{\omega} A(\omega)P_i(\omega) \\ \text{s. t.} \quad & \frac{\partial P_i(\omega)}{\partial \omega} = 0 \quad \text{and} \quad \frac{\partial^2 P_i(\omega)}{\partial \omega^2} < 0. \end{aligned} \tag{20}$$

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 also for the AM case as the modulating signals all have low-pass characteristics. We cannot, however, guarantee that the error converges in a convex way.

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 sequence $u(n-\zeta)$, the modeling error can be written as

$$y_i(n) - w(n)u(n-\zeta)A_i e^{j(\omega_i n + \phi_i)}. \tag{21}$$

This error is minimized in a least-squares sense by maximizing the inner product (with proper normalization) between the modulated sinusoid and the residual:

$$\Psi(\zeta) = \frac{1}{\sum_{n=\zeta}^{N-1} w^2(n)} \left| \sum_{n=\zeta}^{N-1} y_i(n) w(n) e^{-j\omega_i n} \right|^2. \tag{22}$$

We note that the product $y_i(n)w(n)e^{-j\omega_i n}$ for $n = 0, \ldots, N-1$ only has to be computed once for each sinusoid. We then find the onset as the maximizer of (22), i.e.,

$$n_i = \underset{\zeta}{\operatorname{argmax}}\, \Psi(\zeta). \tag{23}$$

Given the frequency and the coarse onset, the combination of envelope parameters, including a final onset estimate, is found as the minimizer of the distortion measure (3). This corresponds to performing a PMP on the subset of the dictionary. We assume that all the dictionary elements have been scaled for a particular segment such that they all have unit perceptual norm, i.e.,

$$\int_{-\pi}^{\pi} A(\omega) Z_k^*(\omega - \omega_i) Z_k(\omega - \omega_i) d\omega = 1 \quad \forall k, \tag{24}$$

with Z_k being the discrete-time Fourier transform of the windowed envelope k in the dictionary, i.e. (see Appendix I)

$$Z_k(\omega) = \sum_{n=0}^{N-1} w(n)\gamma_k(n)e^{-j\omega n}. \tag{25}$$

The envelope, i.e. the combination of α_i, β_i and n_i, is then found in an analysis-by-synthesis manner as the minimizer of the perceptual distortion or, equivalently, as the following maximization of the inner product:

$$Z_i(\omega) = \underset{Z_k(\omega)}{\operatorname{argmax}} \left| \int_{-\pi}^{\pi} A(\omega) Z_k^*(\omega - \omega_i) Y_i(\omega) d\omega \right|^2. \tag{26}$$

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.

$$A_i e^{j\phi_i} = \int_{-\pi}^{\pi} A(\omega) Z_i^*(\omega - \omega_i) Y_i(\omega) d\omega. \tag{27}$$

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 $Z_i(\omega - \omega_i)$ is concentrated in a small region around ω_i, the integration range can also be reduced without much loss in accuracy but with considerable reduction of computational complexity.

For the segment lengths used here, the analytic signal model (considering only the positive parts of the spectrum) has been found to perform satisfactorily. We note that it is also possible to account to some extent for the interaction between different components, including the positive and negative sides of the spectrum, in a number of different ways. The different well-known optimizations of matching pursuit (see e.g. [45]) can be applied at the cost of additional complexity since (3) defines a norm.

V. Rate-Regularized Estimation

In section IV, the parameter set of each envelope, denoted $\Omega_i = \{\ \alpha_i\ \beta_i\ n_i\ \}$, was found in iteration i as the minimizer of the distortion

$$\hat{\Omega}_i = \underset{\Omega_i}{\operatorname{argmin}}\, D(\Omega_i), \tag{28}$$

or equivalently as the maximization in (26). Since sinusoids having constant amplitude do not require the envelope parameters to be transmitted, disregarding the rate in the estimation results in a parameter set which is suboptimal in a rate-distortion sense. In [41] every segment was analyzed using a set of constant-amplitude sinusoids and a set of amplitude modulated sinusoids and by rate-distortion optimization the best representation was chosen for each segment. This was done in order to find an efficient representation in terms of rate. Suppose we have an estimate, or a guess, of $\lambda^\star$ denoted ν, the need for multiple analyses can be eliminated by instead minimizing in each iteration of the estimation

$$\hat{\Omega}_i = \underset{\Omega_i}{\operatorname{argmin}} \left[D(\Omega_i) + \nu R(\Omega_i) \right], \tag{29}$$

where $R(\Omega_i)$ denotes the rate associated with the parameters Ω_i. The rate-distortion optimization is still performed outside the estimation such that the rate-constraint is met. The rate-regularized estimation procedure results in coding templates that are optimized for the target bit-rate. As an example, consider the choice in iteration i between an amplitude modulated sinusoid and a constant-amplitude sinusoid. Using the estimation criterion in (28), the amplitude modulated sinusoid may be chosen, while using (29) may result in the constant-amplitude sinusoid being chosen because the amplitude modulated sinusoid is more expensive in terms of rate. The estimation criterion (29), which we from now on shall refer to as the rate-regularized estimation or just regularized estimation, corresponds to optimizing the coding templates for the target bit-rate. The regularization constant ν does not, however, play the role of the Lagrange multiplier in constrained optimization since we do not solve for it. By choosing $\nu = 0$, the estimation criterion will reduce to (28). Using a large ν will result in an estimation that will tend to choose constant-amplitude over amplitude-modulated sinusoids, while for a small ν, the opposite will occur. In the extremes, this will result in a coder containing only constant-amplitude or amplitude modulated sinusoids. It must be stressed that even if $\nu = \lambda^\star$, i.e. if we guessed the optimal ν, the estimation is not optimal as the individual iterations are not independent. It is of course possible to iterate over ν, but this would be costly in terms of complexity. In most practical situations, the actual choice of ν has been found not to be very critical, i.e., it can simply set to a constant value.

VI. Implementation Details

A. Sinusoidal Parameter Quantization and Rate Estimates

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 quantizers.With θ denoting the parameter to be quantized and $\lfloor\cdot\rfloor$ the truncation operation, the quantized parameter $\hat{\theta}$ is calculated as

$$\hat{\theta} = \exp\left(\left\lfloor \frac{\log(\theta + \epsilon)}{\log(1 + \Delta)} + 0.5 \right\rfloor \log(1 + \Delta) \right), \tag{30}$$

with a small positive constant ϵ being added for numerical reasons. With a step-size Δ of 0.161 for the amplitudes and

TABLE I

CODER CONFIGURATION FOR DIFFERENT TEST CASES DENOTED BY CODER ACRONYM.

Coder	Description
CA	The CA coder uses coding templates consisting of constant-amplitude sinusoids only and a fixed segmentation. This is the simplest possible coder.
AM	The AM coder uses amplitude modulated coding templates and a fixed segmentation. This coder uses the rate-regularized estimation procedure using a regularization constant of 100.
AM/CA	A combination of the CA and AM coder operating on a fixed segmentation. It switches between the two on a segment-to-segment basis using R-D optimization. It does not use the rate-regularized estimation procedure, i.e. a regularization constant of 0 is used.
CA+SEG	As the CA coder but with R-D optimal segmentation.
AM+SEG	The same as the AM coder but with R-D optimal segmentation.
AM/CA+SEG	This is the AM/CA coder combined with R-D optimal segmentation.

0.003 for the frequencies, the quantizers were found to produce transparent results compared to the original (non-quantized) parameters, meaning that informal listening tests showed no degradation in the perceived quality due to the quantization. These quantizers are motivated by studies that show that for amplitude and frequency the just noticeable differences are nearly constant on a logarithmic scale [46]. 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. The entropies of the quantized sinusoidal parameters were also found not to be affected much by the AM. For the amplitude, phase and frequency the entropy was estimated as approximately 20 bits/component. Assuming differential encoding [47], this can be reduced to 16 bits/component. Since the perceptual distortion measure (3) may be overly sensitive to frequency quantization, we use the original parameters in determining the distortions. For the same reason the original parameters are used in generating the residual in the estimation (16).

B. Coding Templates and Segment Sizes

In the experiments to follow, a number of different coder configurations were considered. These are listed in order of rising complexity in Table I. The table shows what types of coding templates were used, how they were found and whether R-D optimal segmentation (SEG) was used. The coding templates are defined as $\mathcal{T}_s = \{\chi_0, \ldots, \chi_L\}$, where χ_i means i sinusoids, which may or may not be modulated, depending on the type of coder. For example, the AM/CA coder uses fixed segmentation and contains coding templates found by analyzing a particular segment using a set of AM sinusoids and a set of CA sinusoids. Note that the AM coding templates can contain constant-amplitude components since these are included as a special case of the model (2), while the CA coding templates contain only CA components. In order to efficiently code CA components in the AM coding templates, a one bit AM switch is used per component. This may be more efficiently encoded using run-length coding. The CA+SEG coder is comparable in quality to that of [48], which uses the PMP and R-D optimal segmentation and uses identical quantizers. The segmentation algorithm described in Section III requires that the distortions are additive over segments. For this to be true, the segments have to be disjoint. However, in order to avoid discontinuities at segment boundaries, some amount of overlap must be introduced between adjacent segments. That the errors introduced in the overlapping regions may have non-zero cross-terms is then simply ignored. Since the distortions also have to be independent over segments, the amount of overlap between segments cannot depend on the segment length. Therefore a natural choice for the amount of overlap is half the size of the minimum segment length. It is important that the overlap is not too small since this may cause undesirable artifacts due to quantization and estimation errors. Consequently, a minimum segment length of 10 ms and an overlap of 5 ms is chosen, meaning that all segment sizes are integer multiples of 10 ms and may start on a 5 ms time-grid. Further, for very long segments, the spectral weighting function becomes increasingly inaccurate as the maskers cannot be assumed to be stationary. Therefore a maximum length of 40 ms has been used. For the coders that use a fixed segmentation, a von Hann window of 30 ms with 15 ms overlap was used. In the experiments to follow, we ignore the side information associated with the segmentation, as this can generally be considered small compared to the total rate. Moreover, the critical comparisons are between coders that use the same type of segmentation and thus have the same rate for the side information. The excerpts used in the tests to follow are fairly short, and the rate-distortion optimization has therefor been carried out over the entire length of the signals.

C. Gamma Envelope Dictionary

It has been found that using the perceptual distortion measure (3) in selecting the envelope parameters made the parameter estimation more robust toward introducing artifacts than using a squared error measure. This can be attributed to the fact that the spectral distortion measure takes into account that the wide mainlobe and sidelobes of modulated sinusoids may introduce errors in parts of the spectrum where no masker is present. However, it was also found necessary to limit the steepness of the attack in order to prevent artifacts from being introduced. Namely, we found that for small α_l, the coder was prone to introduce roughness and click artifacts due to the discontinuities introduced by the unit step sequence. We again note that for $\alpha_l = 0$, the model reduces to that of [32]. Hence, the envelope dictionary was designed empirically from the results of informal listening tests. With a more refined distortion measure, the envelope dictionary could be designed using standard vector quantization techniques. In the following tests, an envelope dictionary for a sampling frequency of 48 kHz composed from $\alpha_l \in \{2, 3, 4, 5\}$, $\beta_l \in \{0.003, 0.005, 0.01, 0.02\}$ and an onset n_l step-size of approximately 0.5 ms was used. As a consequence of this the envelope dictionary size varies with the segment lengths. Since

TABLE II

LIST OF EXCERPTS USED IN THE TESTS.

Number	Name	Type	Length
1	Castanets and Guitar	Mixed	6 s
2	Claves	Solo	7 s
3	Glockenspiel	Solo	8 s
4	Grand Piano	Solo	11 s
5	ABBA	Mixed	10 s
7	Bass Guitar	Solo	12 s
8	English Female Speech	Speech	6 s
9	Castanets	Solo	7 s
10	Harpsichord	Solo	9 s
11	Tracy Chapman	Mixed	13 s
12	Triangle	Solo	9 s
13	Xylophone	Solo	8 s

the frequency and envelopes of transients may vary much from signal to signal, no entropy coding of the envelope parameters was assumed in the rate estimates, i.e. the upper bound is used. These are 9, 10, 10 and 11 bits per envelope for 10, 20, 30 and 40 ms segments, respectively. Preliminary experimental results also suggest that differential coding of onset times may lead to a reduction of the average bits per component. The spectra of the windowed gamma envelopes were stored in a lookup table in order to perform fast estimation (equations (26) and (27)) using the spectral distortion measure (3).

VII. EXPERIMENTAL RESULTS

A. Signal Examples

As an example of a coded signal, the xylophone coded at 30 kbps is shown in Figures 3 and 4. It can be seen that the CA coder introduces a pre-echo and that the transient is smeared and has lost its sharpness. In the CA+SEG coder, the pre-echo is much reduced, but the transient is still not as sharp as the original. The AM/CA+SEG coder sharpens the attack further and reduces the pre-echo.

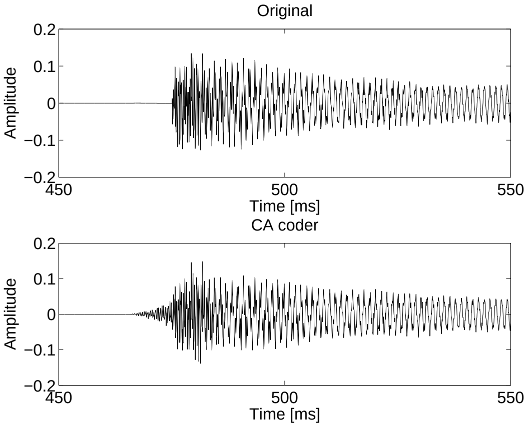

Fig. 3. Signal example, xylophone, original (top) and coded at 30 kbps using the CA coder (bottom).

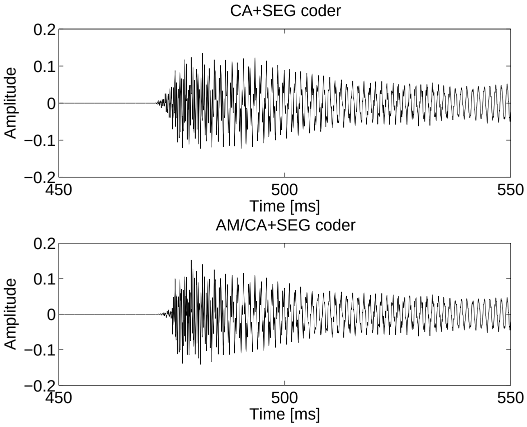

Fig. 4. Signal example, xylophone, coded at 30 kbps using the CA+SEG coder (top) and using the AM/CA+SEG coder (bottom).

In Figure 5 the rate-distortion curves[2] for a representative transient sinusoidal signal, glockenspiel, are shown for the CA coder, the AM/CA coder and the AM coder. Similarly, in Figure 6, the same is shown for the CA+SEG coder, the AM/CA+SEG coder and the AM+SEG coder. The signal has a duration of approximately 10 s and R-D optimization was performed on the entire signal. For the fixed segmentation, it can be seen that there is a clear improvement for the AM and AM/CA coders in terms of a reduction of the distortion compared to the CA coder at the same rate. Also, the proposed coder saturates at lower distortions than the CA coder for glockenspiel. It can also be seen from figure Figure 6, that when R-D optimal segmentation is employed, the rate of convergence is higher for all coders. An interesting observation is also that the rate-regularized coder, the AM coder, performs similarly to the AM/CA coder. This means that the dual analyses of the AM/CA coder can be avoided with very little loss of performance. From these figures, it seems that for this particular excerpts, the glockenspiel, very little is achieved by combining AM and SEG. It looks as if similar performance can be achieved with either AM or SEG, with the AM coder being less complex than the CA+SEG coder. For other signals such as the castanets, though, the R-D curves show that improvements can be gained by the combination of AM and R-D optimal segmentation.

In Figure 7 the R-D optimal segmentation boundaries are shown for the AM coder and the AM/CA coder for 30 kbps for the excerpt Castanets. It can be seen that a higher coding efficiency is achieved as longer segments are chosen around the transients when AM coding templates are included. It was also found that when R-D optimal segmentation was used, there was still an advantage of using the onsets, i.e. improvements were still gained by allowing $n_l \neq 0$ in (2). Constraining $n_l = 0 \, \forall l$, i.e. reducing the model to that of [30], [33], led to shorter segments and a loss in perceived quality. The ability of the model to position onsets of the individual sinusoids at arbitrary positions within each segment has proven to be an important one. The effect of the rate-regularized estimation procedure is illustrated in Figure 8, where the rate-distortion curves of the AM coder for different regularization constants are shown for 2 s of claves. It can be seen that in the region 20-40 kbps, approximately 5 kbps can be saved compared to no regularization. Depending on the signal at hand, this result may vary.

[2] In information theory the relation $D(R)$ is traditionally referred to as the distortion-rate curve. We refer to this relationship using the aesthetically more pleasing term rate-distortion curve.

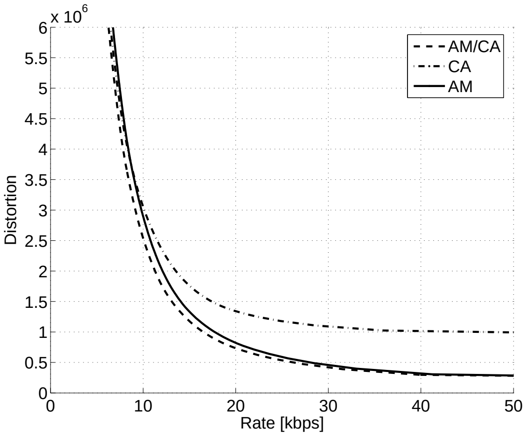

Fig. 5. The rate-distortion curves of the CA coder (dash-dotted), the AM/CA coder (dashed) and the AM coder (solid) using a fixed segmentation for the glockenspiel.

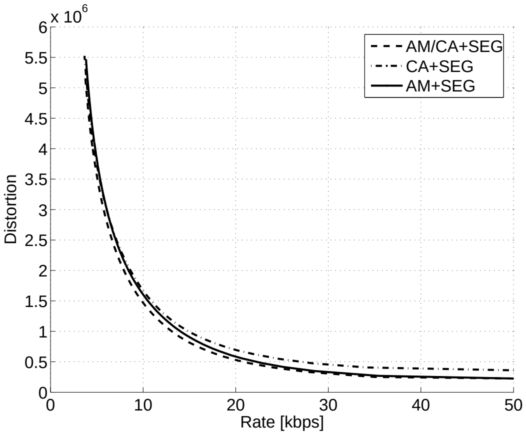

Fig. 6. The rate-distortion curves of the CA+SEG coder (dash-dotted), the AM/CA+SEG coder (dashed) and the AM+SEG coder (solid) using R-D optimal segmentation for the glockenspiel.

B. Test Material

In order to evaluate the proposed method for parametric coding of transients, we conducted a formal listening test. In addition, we report our experience from informal listening tests to give the reader some indications as to the nature of the improvements that were made. In the informal and formal listening tests, the excerpts shown in Table II were used[3]. These represent a wide variety of different types of signals, many of which are known to be critical excerpts in perceptual audio coding [37]. All the signals were monophonic and were 16 bit signals sampled at 48 kHz and they have a length of 6-12 s. Many more signals were used in the development, but these are the ones that have been tested extensively. In ITU-R BS.1534-1 [49] it is recommended to use excerpts that are known to be critical in testing of audio coding algorithms. Problematic transients by no means occur in all excerpts. Consequently, these tests are concerned mainly with excerpts that are known to be critical yet different of type. For example, the glockenspiel excerpt is very tonal and stationary for the most parts but has very steep attacks, while the castanet excerpt has very stochastic and strongly modulated characteristics. The excerpts 5 and 11 are pop music containing mixtures of multiple instruments and vocal.

C. Informal Listening Tests

Informal listening tests revealed that pre-echos are clearly reduced and that the transients are better modeled using the proposed model than with constant-amplitude sinusoids. For many signals, though, the improvements are fairly subtle since they are already handled well using constant-amplitude sinusoids. Often, the improvements are perceived as an increase of bandwidth of the coded signal. For critical excerpts, such as castanets the improvement are clearly audible. The types of signals that benefit from the AM coder are signals that exhibit fast onsets, impulse-like signals, transitions between different stationary parts of signals, and percussive instruments. Any mixture of these types of signals with stationary ones may also benefit from it. It was also found that the AM coder improves the perceived quality of sinusoidally coded speech. Namely, the speech was found to suffer less from the tonal artifact often encountered in sinusoidal speech coding. Experiments showed that the AM coder proved R-D optimal for plosives, in transitions in pitch and in transitions between voiced and unvoiced sounds. For speech, it may also be beneficial to incorporate a model for frequency modulation [50]. Informal listening tests also revealed that the perceptual distortion measure (3) does not fully reflect the perceived improvement caused by the AM. For example, the relative improvement in terms of rate-distortion between the CA coder and the AM coder appears small for the castanets, while the perceived difference is large. This may be explained by the fact that the model [38] was derived for predicting the masking of sinusoidal component, and that the castanets are not very sinusoidal by nature unlike signals like the glockenspiel, claves and xylophone. The perceptual distortion measure (3) does, though, form a robust measure for estimation of model parameters and for the R-D optimization. When the R-D optimal segmentation is employed, the effects of the AM coder are less audible compared to the CA coder for excerpts where the signals exhibit fast onsets. Examples of this are glockenspiel and claves while for castanets, the combination of the AM coder and R-D optimal segmentation results in significant improvements. The use of variable bit-rate and R-D optimization has also been found to improve performance for transients for all the coders, since more bits can be allocated for critical signal parts, such as transients, this way.

D. MUSHRA Test

In order to quantify the improvements gained by the different methods for handling of transients, we use a subjective listening test. We use the MUSHRA test (MUlti-Stimulus test with Hidden Reference and Anchors) [49], which is a double blind test for subjective assessment of intermediate quality level of coding systems. For each excerpt, the listeners were asked to rank 8 differently processed versions relative to a known reference on a score from 0 to 100. These

[3] Some of the processed excerpts are available on first author's homepage at `http://kom.aau.dk/~mgc/projects/gamma`

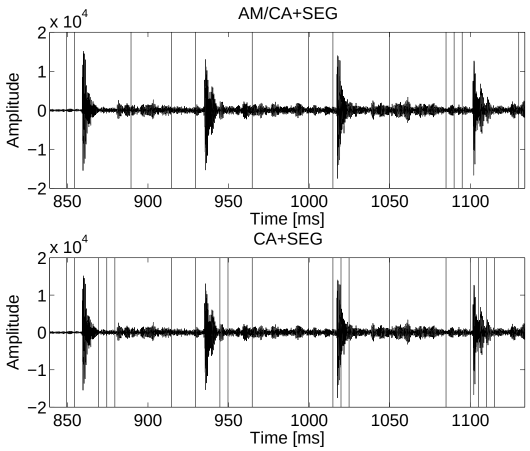

Fig. 7. Example of R-D optimal segmentation boundaries (indicated by vertical lines) for castanets for the AM/CA+SEG coder (top) and the CA+SEG coder (bottom) operating at 30 kbps. Note that both the signals shown are the original.

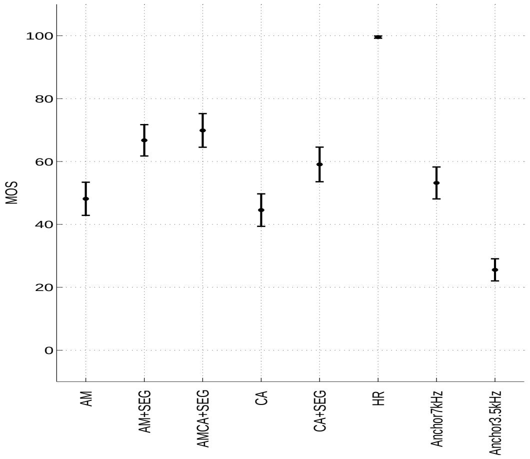

Fig. 9. Results of the MUSHRA listening test. MOS scores for different coders averaged over all excerpts and all listeners. The error bars indicate the 95% confidence intervals.

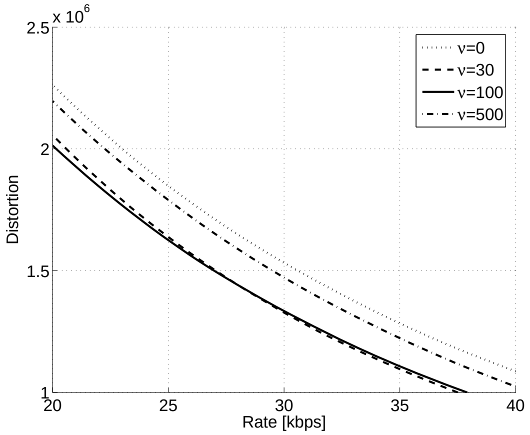

Fig. 8. The rate-distortion curves of the AM coder for different regularization constants ν for claves optimized over 2 s.

included the hidden reference (denoted HR), an anchor low-pass filtered at 7 kHz and an anchor low-pass filtered at 3.5 kHz (denoted Anchor 7 kHz and Anchor 3.5 kHz, respectively). The remaining 5 versions were the AM, CA, CA+SEG, AM+SEG and the AM/CA+SEG coders all operating at 30 kbps. In the MUSHRA test the hidden reference is used to verify the consistency of responses of subjects because a very high score is expected here. The anchors are included to be able to make comparisons between different listening tests and because they constitute a well-defined and simple signal modification. In order to limit the length of the listening test a representative subset of the excerpts listed in Table II was chosen. Nine experts listeners participated in the test (the authors not included). The test was performed on speakers in a listening room. As the proposed coders do no incorporate residual coding and are thus not complete parametric coders, a reference coder has not been included in this test. In MUSHRA tests the hidden reference define known points on the scale. In Figure 9 the resulting MOS (Mean Opinion Score) scores of the different coder configurations averaged over all excerpts and listeners are shown. Since we are dealing with particular critical excerpts, it is of interest to investigate the performance for the individual excerpts. These are shown in Table III with the excerpt being identified by the number in Table II. From Figure 9 we see that the AM/CA+SEG coder scores about 10 points higher at average than the CA+SEG coder, and more than 20 points higher than the CA coder. Although the AM coder does not seem to perform significantly better than the CA coder in this test, the AB preference test in [41] showed a significant preference for the AM/CA coder over the CA coder. In the table, it can be seen that for particular excerpts, such as the castanets (excerpt 9), there is a huge improvement in the combination of AM and the R-D optimal segmentation over the CA coder both with and without optimal segmentation, in fact the R-D optimal segmentation helps very little without the AM model. It can also be seen that there is a fairly small loss on average in the rate-regularized estimation procedure of the AM+SEG coder compared to the AM/CA+SEG, except for the glockenspiel (excerpt 3). Taking the confidence intervals into account, this difference is too small to be of any statistical significance. The reason for the fairly poor performance of the AM+SEG coder compared to the AM/CA+SEG coder for the glockenspiel is that the same regularization constant was used for processing all excerpts, and for the glockenspiel, this constant is not close to the optimal λ. It is interesting to note that the glockenspiel scores the highest among all excerpt. This is not surprising because the glockenspiel signal is very tonal and the AM model is well-suited for handling the non-stationary parts of this signal. This also holds for the very similar signals of SQAM, such as the claves, xylophone, triangle and others.

VIII. Discussion

As can be concluded from the listening test results, the proposed parametric coding of transients in combination with R-D optimal segmentation leads to a significant gain in audio quality as compared to constant-amplitude sinusoidal coding. Switching between different window lengths and shapes or coders (e.g. [9], [18]) has traditionally been achieved by transient detection schemes. However, there may be a mismatch

TABLE III
RESULTS OF THE MUSHRA LISTENING TEST. MOS SCORES FOR DIFFERENT CODER CONFIGURATIONS FOR THE INDIVIDUAL EXCERPTS.

Excerpt	1	3	5	7	9	10	11
AM	42	70	41	45	43	56	39
AM+SEG	67	79	58	71	66	68	58
AM/CA+SEG	65	92	62	68	72	71	59
CA	32	60	41	42	29	65	43
CA+SEG	47	84	64	63	35	65	55
HR	99	99	99	100	100	99	100
Anchor 7 kHz	47	66	56	62	47	42	52
Anchor 3.5 kHz	22	33	24	27	22	24	27

between the classification of transients and the R-D optimal coder. Based on R-D optimization and/or the rate-regularized estimation method robustness against such problems is gained, but this comes at the cost of additional complexity. We also note that the R-D optimal allocation scheme is similar to the so-called bit reservoir method for handling of transients (see [16]). Rate-distortion optimal allocation (variable rate) in itself does not, however, ensure that more bits are spent when transients are present. Rather, it spends the bits where most distortion can be reduced, and hence it depends on the appropriateness of the signal model.

The scores from the MUSHRA test reported here may be further improved by residual coding since noise components are not efficiently coded using sinusoids. Many parametric audio coders employ residual noise coding that only encodes a spectral and a coarse temporal envelope (e.g. [13], [51]). It is also possible to improve performance of parametric audio coders for transient signals by employing waveform approximating residual coding as done in [52], [53]. In such coding schemes, the residual coder may compensate for errors introduced by the sinusoidal coder.

Recently, preliminary results on linearization of the spectro-temporal psychoacoustical model [54] have been reported in [55]. Such a linearization results in a distortion measure that defines a norm and would thus be applicable to the AM estimation problem at the cost of increased complexity. Further, if such a measure is shown to reflect temporal aspects better than (3), this could lead to improved coding of transients as presented here as well as to more refined envelope dictionary design.

Compared to the singlebanded AM of e.g. [15], the model proposed in this paper has the advantage that different envelopes are allowed for different sinusoids, which is a particular advantage for mixtures of sources (see e.g. [34]). Some interesting parallels can be drawn to related work in audio coding. In [25] transient locations are modified in order to achieve more efficient coding of transients. This is, in a sense, what is happening when the onsets are quantized, and seen in the light of [25], onsets should be estimated very precisely and then quantized jointly to a coarse grid. A successful tool in dealing with efficient coding of transients in transform coding is TNS [22]. TNS is based on linear predictive coding of transform coefficients. Since amplitude modulation may just as well be interpreted as a frequency domain filtering, there is an duality in TNS and AM. One conceptual difference between TNS and gain modification [23] as applied in transform coding on the one hand and AM as presented here on the other hand is that TNS and gain modification operate on the input and output signals and hence shape the noise, whereas in AM, the signal model is modified to fit the input signal.

IX. SUMMARY

In this paper, methods for efficient parametric coding of transient audio signals have been presented. We propose a specific model for handling of transients based on amplitude modulated sinusoids. In this model, each sinusoid is modulated by a different envelope known as a gamma envelope each being characterized by and onset, an attack and a decay parameter. These degrees of freedom have proven to be important in efficient coding of transients. Existing methods assume either that the modulating signal is the same for all components, that the onset always occurs at the start of a segment, or that no attack parameter is necessary. Combined with a constant-amplitude sinusoidal model, efficient coding of both stationary and transient signals is achieved using rate-distortion optimization based on a perceptual distortion measure. The rate-distortion optimization leads to optimal allocation and segmentation and therefore eliminates the need for transient detectors. Informal and formal listening tests reveal that for critical excerpts the combination of amplitude modulation and rate-distortion optimal segmentation leads to large improvements over a sinusoidal coder using only the optimal segmentation. This shows that segmentation techniques are not substitutes for good signal models.

X. ACKNOWLEDGMENT

The authors would like to thank A. Kohlrausch and A. Härmä both of Philips Research Laboratories, Eindhoven, The Netherlands, and Søren Holdt Jensen of the Department of Communication Technology, Aalborg University, Denmark, for their constructive comments and suggestions.

APPENDIX I
FOURIER TRANSFORM OF WINDOWED GAMMA ENVELOPE

The estimation of model parameters and calculation of distortions require that the spectra of the windowed gamma envelopes are computed. Doing this by FFTs may be prohibitive for low complexity applications and storing them in memory may also not be feasible. Here, we instead derive a closed-form expression for generating the discrete-time Fourier transform directly in the frequency domain. The discrete-time Fourier transform of the windowed gamma envelope can be found from the following finite sum:

$$Z_l(\omega) = \sum_{n=0}^{N-n_l-1} n^{\alpha_l} e^{-\beta_l n} w(n+n_l) e^{-j\omega(n+n_l)}, \tag{31}$$

with $w(n)$ being the tapered von Hann window (7). In finding the discrete Fourier transform we shall use the following transform pair:

$$n^a x(n) \leftrightarrow j^a \frac{\partial^a}{\partial \omega^a} X(\omega). \tag{32}$$

Assuming that $n_l < M-1$ and splitting the sum (31) up into three different sums having different window parts, we get

$$\begin{aligned} Z_l(\omega) = &\sum_{n=0}^{M-1-n_l} n^{\alpha_l} e^{-\beta_l n} v(n+n_l) e^{-j\omega(n+n_l)} \\ &+ \sum_{n=M-n_l}^{N-M-1-n_l} n^{\alpha_l} e^{-\beta_l n} e^{-j\omega(n+n_l)} \\ &+ \sum_{n=N-M-n_l}^{N-1-n_l} n^{\alpha_l} e^{-\beta_l n} v(n-N+2M+n_l) \\ &\times e^{-j\omega(n+n_l)}. \end{aligned} \tag{33}$$

with v(n) being the modified von Hann window in (8). Tedious calculations now lead to the following closed-form expression of the discrete-time Fourier transform of the windowed gamma envelopes:

$$\begin{aligned} Z_l(\omega) = j^{\alpha_l} \frac{\partial^{\alpha_l}}{\partial \omega^{\alpha_l}} \Bigg(&\frac{1}{2} e^{-j\omega n_l} \frac{1-(e^{-\beta_l - j\omega})^{M-n_l}}{1-e^{-\beta_l - j\omega}} \\ &- \frac{1}{4} e^{-j\omega n_l + j\frac{\pi}{M} n_l + j\frac{\pi}{2M}} \frac{1-(e^{-\beta_l - j\omega + j\frac{\pi}{M}})^{M-n_l}}{1-e^{-\beta_l - j\omega + j\frac{\pi}{M}}} \\ &- \frac{1}{4} e^{-j\omega n_l - j\frac{\pi}{M} n_l - j\frac{\pi}{2M}} \frac{1-(e^{-\beta_l - j\omega - j\frac{\pi}{M}})^{M-n_l}}{1-e^{-\beta_l - j\omega - j\frac{\pi}{M}}} \\ &+ e^{-j\omega n_l} \frac{(e^{-\beta_l - j\omega})^{M-n_l} - (e^{-\beta_l - j\omega})^{N-M-n_l}}{1-e^{-\beta_l - j\omega}} \\ &+ \frac{1}{2} e^{-j\omega n_l} \frac{(e^{-\beta_l - j\omega})^{N-M-n_l} - (e^{-\beta_l - j\omega})^{N-n_l}}{1-e^{-\beta_l - j\omega}} \\ &- \frac{1}{4} e^{-j\omega n_l + j\frac{\pi}{2M} - j\frac{\pi}{M}(N-n_l)} \\ &\times \frac{(e^{-\beta_l - j\omega + j\frac{\pi}{M}})^{N-n_l-M} - (e^{-\beta_l - j\omega + j\frac{\pi}{M}})^{N-n_l}}{1-e^{-\beta_l - j\omega + j\frac{\pi}{M}}} \\ &- \frac{1}{4} e^{-j\omega n_l - j\frac{\pi}{2M} + j\frac{\pi}{M}(N-n_l)} \\ &\times \frac{(e^{-\beta_l - j\omega - j\frac{\pi}{M}})^{N-n_l-M} - (e^{-\beta_l - j\omega - j\frac{\pi}{M}})^{N-n_l}}{1-e^{-\beta_l - j\omega - j\frac{\pi}{M}}} \Bigg). \end{aligned} \tag{34}$$

In evaluating these expressions for particular parameter values and frequencies L'Hospital's rule must be used. For the coder presented in [41], where the window is simply a von Hann window with a fixed length, the corresponding expression is much simpler.

REFERENCES

[1] P. Hedelin, "A tone oriented voice excited vocoder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1981, pp. 205–208.

[2] L. Almeida and J. Tribolet, "Harmonic coding: A low bit-rate, good-quality speech coding technique," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 7, 1982, pp. 1664–1667.

[3] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34(4), pp. 744–754, Aug. 1986.

[4] ——, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 4.

[5] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis-synthesis of musical tones," *J. Audio Eng. Soc.*, vol. 40(6), pp. 497–516, 1992.

[6] J. O. Smith and X. Serra, "Spectral Modelling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition," *Computer Music Journal*, vol. 14(4), pp. 12–24, 1990.

[7] K. N. Hamdy, M. Ali, and A. H. Tewfik, "Low bit rate high quality audio coding with combined harmonic and wavelet representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1996, pp. 1045–1048.

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

[9] S. N. Levine and J. O. Smith III, "A switched parametric & transform audio coder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1999, pp. 985–988.

[10] T. S. Verma and T. H. Y. Meng, "A 6kbps to 85kbps scalable audio coder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2000, pp. 877–880.

[11] H. Purnhagen and N. Meine, "HILN - The MPEG-4 Parametric Audio Coding Tools," in *IEEE International Symposium on Circuits and Systems*, 2000.

[12] ISO/IEC, *Coding of audio-visual objects – Part 3: Audio (MPEG-4 Audio Edition 2001)*. ISO/IEC Int. Std. 14496-3:2001, 2001.

[13] A. C. den Brinker, E. G. P. Schuijers, and A. W. J. Oomen, "Parametric coding for high-quality audio," in *112th Conv. Aud. Eng. Soc.*, 2002, paper Preprint 5554.

[14] 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.

[15] E. G. P. Schuijers, A. W. J. Oomen, A. C. den Brinker, and J. Breebart, "Advances in parametric coding for high-quality audio," in *114th Conv. Aud. Eng. Soc.*, 2003, paper Preprint 5852.

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

[17] L. L. Elliot, "Backward and forward masking of probe-tones of different frequencies," *J. Acoust. Soc. Am.*, vol. 34, pp. 1116–1117.

[18] B. Edler, "Codierung von Audiosignalen mit überlappender Transformation und adaptiven Fensterfunktionen," *Frequenz*, pp. 1033–1036, 1989.

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

[20] P. Prandoni, M. M. Goodwin, and M. Vetterli, "Optimal time segmentation for signal modeling and compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1997, pp. 2029–2032.

[21] M. Erne, G. Moschytz, and C. Faller, "Best wavelet-packet bases for audio coding using perceptual and rate-distortion criteria," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1999, pp. 909–912.

[22] J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)," in *101st Conv. Aud. Eng. Soc.*, 1996, paper preprint 4384.

[23] M. Link, "An attack processing of audio signals for optimizing the temporal characteristics of a low bit-rate audio coding system," in *95th Conv. Aud. Eng. Soc.*, 1993, paper preprint 3696.

[24] T. Vaupel, "Ein Beitrag zur transformationscodierung von Audiosignalen unter Verwendung der Methode der 'Time Domain Aliasing Cancellation (TDAC)' und einer Signalkompandierung im Zeitbereich," Ph.D. dissertation, Universität-Gesamthochschule Duisburg, Germany, 1991.

[25] R. Vafin, R. Heusdens, and W. B. Kleijn, "Modifying transients for efficient coding of audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, pp. 3285–3288.

[26] S. N. Levine, T. S. Verma, and J. O. Smith III, "Alias-free, multiresolution sinusoidal modeling for polyphonic, wideband audio," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 1997, pp. 101–104.

[27] T. S. Verma and T. H. Y. Meng, "An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 6, 1998, pp. 3573–3576.

[28] M. G. Christensen, S. V. Andersen, and S. H. Jensen, "Amplitude modulated sinusoidal models for audio modeling and coding," in *Knowledge-Based Intelligent Information and Engineering Systems*, ser. Lecture Notes in Artificial Intelligence, V. Palade, R. J. Howlett, and L. C. Jain, Eds. Springer-Verlag, 2003, vol. 2773, pp. 1334–1342.

[29] J. Nieuwenhuijse, R. Heusdens, and E. F. Deprettere, "Robust Exponential Modeling of Audio Signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1998, pp. 3581–3584.

[30] 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.

[31] 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.

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

[33] K. Hermus, W. Verhelst, P. Lemmerling, P. Wambacq, and S. van Huffel, “Perceptual audio modeling with exponentially damped sinusoids,” *Signal Processing*, vol. 85, pp. 163–176, 2005.

[34] 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*, vol. 4, 2004, pp. 169–172.

[35] A. Aertsen and P. Johannesma, “Spectro-Temporal Receptive Fields of Audiotory Neurons in the Grass Frog. I. Characterization of tonal and natural stimuli,” *Biol. Cybern.*, vol. 38, pp. 223–234, 1980.

[36] T. D. Rossing, *The Science of Sound*, 2nd ed. Addison-Wesley Publishing Company, 1990.

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

[38] 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*, vol. 2, 2001, pp. 1805 – 1808.

[39] 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.

[40] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[41] M. G. Christensen and S. van de Par, “Rate-distortion efficient amplitude modulated sinusoidal audio coding,” in *Rec. Thirty-Eighth Asilomar Conf. Signals, Systems, and Computers*, 2004, pp. 2280–2284.

[42] 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.

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

[44] 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*, vol. 2, 1999, pp. 981–984.

[45] M. M. Goodwin, “Adaptive Signal Models: Theory, Algorithms, and Audio Applications,” Ph.D. dissertation, University of California, Berkeley, 1997.

[46] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed. Academic Press, 1997.

[47] 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.

[48] R. Heusdens, J. Jensen, W. B. Kleijn, V. kot, O. A. Niamut, S. van de Par, N. H. van Schijndel, and R. Vafin, “Bit-rate scalable intra-frame sinusoidal audio coding based on rate-distortion optimisation,” *J. Audio Eng. Soc.*, 2005, submitted.

[49] *ITU-R BS.1534*, ITU Method for subjective assessment of intermediate quality level of coding system, 2001.

[50] P. Maragos, J. F. Kaiser, and T. F. Quatieri, “Energy Separation in Signal Modulations with Application to Speech Analysis,” *IEEE Trans. Signal Processing*, vol. 41(10), pp. 3024–3051, Oct. 1993.

[51] M. M. Goodwin, “Residual modeling in music analysis-synthesis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1996, pp. 1005–1008.

[52] F. Riera-Palou, A. C. den Brinker, and A. J. Gerrits, “A hybrid parametric-waveform approach to bistream scalable audio coding,” in *Rec. Thirty-Eighth Asilomar Conf. Signals, Systems, and Computers*, 2004, pp. 2250–2254.

[53] R. Vafin and W. B. Kleijn, “Towards optimal quantization in multistage audio coding,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, 2004, pp. 205–208.

[54] T. Dau, D. Püschel, and A. Kohlrausch, “A quantitative model of the effective signal processing in the auditory system. i. model structure,” *J. Acoust. Soc. Am.*, vol. 99(6), pp. 3615–3622, June 1996.

[55] 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.