New Results in Rate-Distortion Optimized Parametric Audio Coding

Christensen, Mads Græsbøll; Jensen, Søren Holdt

Published in:
Proc. of AES 120th Convention

Publication date:
2006

Document Version
Early version, also known as pre-print

Audio Engineering Society

Convention Paper

Presented at the 120th Convention
2006 May 20–23 Paris, France

New Results in Rate-Distortion Optimized Parametric Audio Coding

Mads Græsbøll Christensen[1] and Søren Holdt Jensen[1]

[1]*Dept. of Communication Technology, Aalborg University, DK-9220 Aalborg, Denmark*

Correspondence should be addressed to Mads Græsbøll Christensen (mgc@kom.aau.dk)

ABSTRACT

In this paper, we summarize some recently published methods and results in parametric audio coding. These are all based on rate-distortion optimized coding using a perceptual distortion measure. We summarize how a number of well-known computationally efficient methods for incorporating perception in sinusoidal parameter estimation relate to minimizing this perceptual distortion measure. Then a number of methods for parametric coding of transients are compared and results of listening tests are presented. Finally, we show how the complexity of rate-distortion optimized audio coding can be reduced by rate-distortion estimation.

1. INTRODUCTION

Audio coding is one of the success stories of modern signal processing. It is the art of maximizing the perceived quality of audio signals encoded at a desired bit-rate. Perhaps the most common incarnations of audio coders are transform/sub-band coders such as MPEG-1 Layer III (mp3) or MPEG-2/4 AAC (see, e.g., [1, 2]). Parametric models have been applied successfully to digital processing of audio and speech signals during the past couple of decades and in the past few years, there has been significant interest in so-called parametric coding techniques, e.g. [3–7] and recently, coding standards based on parametric coding have been established [4, 8–10]. These parametric coding techniques have primarily been used as alternatives to transform coders, but some parametric coding techniques have also been successfully combined with the more traditional transform coders. For example, sinusoidal coders have been combined with transform coders in a multi-stage structure in [11, 12]. Also, the recent advances in parametric multi-channel coding such as [13–15] (see also [16]) and the perceptual noise substitution method [17, 18] are examples of this. Parametric coding can be described as coding by means of signal models or perceptual cues. The most common case is perhaps the combination of some sinusoidal model and an auto-regressive (AR) stochastic process, and this is also the focus of the present paper. In one sense, parametric coding can be described as a form of structured vector quantization, where the codebook structure is imposed by the choice of signal model.

This relation is further strengthened by the equivalence between a commonly used method for finding the model parameters in parametric coding, namely matching pursuit (MP) [19] and multi-stage gain-shape vector quantization (see, e.g., [20]).

Audio coders are designed and evaluated in terms of rate, distortion, delay, complexity, robustness, and flexibility and in designed audio coders, all these criteria must be taken into account. Often, however, only rate, distortion and delay constraints are explicit. The design of audio coders have mainly been concerned with achieving the lowest possible distortion at a given bit-rate and also the computational complexity has played an important role. More recently, the delay has also become a factor and dedicated standards for low delay coding have been implemented [21]. In the past few years, though, there has been an increasing interest in scalability, i.e. flexibility in terms of the aforementioned design criteria, and robustness. A valuable tool in achieving flexible and robust solutions is rate-distortion (R-D) optimization. Rate-distortion optimized audio coding is coding that optimizes itself according to time-varying constraints, such as rate or distortion, and the source that is to be encoded.

In this paper, we present some new methods and results in rate-distortion optimized audio coding. These are: 1) computationally efficient sinusoidal parameter estimation based on a perceptual distortion measure, 2) amplitude modulated sinusoidal audio coding, and 3) computationally efficient rate-distortion optimization by rate-distortion estimation. Most parametric coders are based on a sinusoidal model and a residual model. We consider the problem of estimating the parameters of the perceptually most important sinusoids. A number of sinusoidal frequency estimators that incorporate perception have been proposed in the literature. These incorporate perception by various heuristic ways, e.g. by pre-filtering of the input signal [22, 23] or component weighting as done in the weighted matching pursuit [24]. We relate and analyze these estimators a framework based on a perceptual distortion measure and show that they are equivalent to to minimizing the perceptual distortion, as done in [25], under certain conditions.

An important problem in audio coding is efficient coding of transients. A wide range of methods for this have been proposed. Most audio coders deal with this problem using variable segmentation, variable bit-rate, and perceptual noise shaping. Recently, amplitude modulated sinusoidal models have been shown, in listening tests, to offer improved coding of transients, even when combined with the other methods. Three different experimental coders based on amplitude modulated sinusoidal models have been proposed and shown to lead to improved coding compared to a state-of-the-art sinusoidal coder. The proposed coders are based on different models of the amplitude modulating signal, namely a linear combination of basis vectors, a frequency-domain all-pole filter, and so-called gamma envelopes. Here, we compare these methods in terms of perceived quality and computational complexity. The coder based on gamma envelope model has been found to produce the highest perceived quality at high bit-rates and at moderate complexity while the frequency-domain all-pole filter model has very low complexity and performs well at low bit rates but cannot handle very complex mixtures of sources. The linear combination of basis vectors, on the other hand, has been found to produce reasonable quality but suffers from high complexity.

There has been a significant interest in rate-distortion optimized audio coding in recent years, e.g. [26, 27]. Using rate-distortion optimization, an optimal segmentation and allocation of bits can be found for any desired bit-rate, but this requires that distortions are calculated for all allocations and segments. We instead estimate the distortions based only on a number of simple signal features. Specifically, the relationship between these features and distortions are modeled using a Gaussian mixture, and, for a particular segment, the distortions are estimated using a computationally efficient linear Bayesian estimator. MUSHRA listening tests reveal that this principle can be applied to the problem of finding a rate-distortion optimal segmentation in a sinusoidal coder resulting in a complexity reduction by a factor of ten without much loss in perceived quality.

The paper is organized as follows. In Section 2 we review the basic results of rate-distortion optimization and in Section 3 the perceptual distortion measure which is used in this paper is presented. Then, in Section 4, the problem of finding the sinusoidal components that minimize this distortion measure is treated. A number of different methods for efficient coding of transients are presented in Section 5, and in Section 6, the principles of rate-distortion estimation are presented. Finally, in Section 7, we summarize the contributions.

Aside from the information given in this paper, more details can be found in [28] and in the referenced papers.

2. RATE-DISTORTION OPTIMIZATION

In this section, we briefly review the basic results of the rate-distortion optimal allocation and segmentation scheme of [29], which is based on the earlier work on optimal allocation reported in [30]. First, we define a segment σ_s as having a length of $\ell(\sigma_s) = \kappa m$ with $m \in \mathbb{N}$, and a segmentation as $\boldsymbol{\sigma} = [\, \sigma_1 \, \cdots \, \sigma_S \,]$ consisting of disjoint, contiguous segments that satisfy

$$\sum_{s=1}^{S} \ell(\sigma_s) = \kappa G, \tag{1}$$

where κG is the total length of the signal and κ is natural number. Then each of these segments, say segment s, can be encoded using a set of coding templates $\mathcal{T}_s$. Let $D(\tau)$ be the distortion and $R(\tau)$ the number of bits associated with coding template $\tau \in \mathcal{T}_s$. Assuming additivity over the segments in the segmentation $\boldsymbol{\sigma}$ and coding templates $\boldsymbol{\tau} = [\, \tau_1 \, \cdots \, \tau_S \,]$, we can write the total distortion as

$$D(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \sum_{s=1}^{S} D(\sigma_s, \tau_s) \tag{2}$$

and the total number of bits as

$$R(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \sum_{s=1}^{S} R(\sigma_s, \tau_s). \tag{3}$$

The rate-distortion optimization problem can be stated as

$$\text{minimize} \quad D(\boldsymbol{\sigma}, \boldsymbol{\tau}) \quad \text{s.t.} \quad R(\boldsymbol{\sigma}, \boldsymbol{\tau}) = R^\star, \tag{4}$$

where $R^\star$ is the bit budget. This can be written as an unconstrained problem using the Lagrange multiplier method [29, 30], i.e.,

$$\min_{\boldsymbol{\sigma}} \sum_{s=1}^{S} \min_{\tau \in \mathcal{T}_s} \left[D(\sigma_s, \tau) + \lambda R(\sigma_s, \tau) \right] - \lambda R^\star. \tag{5}$$

The optimal λ that leads to the target rate $R^\star$ can be found by sweeping over different λ using simple bisection until the resulting rate is within some desired range of the bit budget. In sinusoidal coding, the coding templates are often chosen to be different number of sinusoids such that the number of sinusoids may vary from segment to segment. Additionally, we here also consider coding templates to include different models, e.g. the amplitude modulated sinusoidal models to be discussed in Section 5.

3. A PERCEPTUAL DISTORTION MEASURE

For the R-D optimization to be effective in audio coding, it is imperative that a distortion measure that reflects the human auditory system is used. The methods and results presented in this paper are all based on the auditory masking model proposed in [31, 32]. A block diagram of this model is shown in Figure 1 with C and B being calibration constants. According to this model, the distortion D for a particular segment can be written as

$$D = \sum_{k=0}^{K} A(k) |E(k)|^2, \tag{6}$$

where $A(k)$ is a real, positive weighting function derived from [31, 32] and $w(n)$ is the analysis window, with

$$E(k) = \sum_{n=0}^{N-1} w(n) \left[x(n) - \hat{x}(n) \right] e^{-j 2\pi \frac{k}{K} n}, \tag{7}$$

and $x(n)$ is the input and $\hat{x}(n)$ is the reconstructed signal. When the perceptual weighting function $A(k)$ is chosen as the reciprocal of the masking threshold, as is the case here, the resulting error spectrum will be shaped according to the masking threshold. This distortion measure can be written using matrix-vector notation as

$$D = \|\mathbf{H}\mathbf{e}\|_2^2 = \|\mathbf{H}\left(\mathbf{x} - \hat{\mathbf{x}}\right)\|_2^2, \tag{8}$$

where the matrix $\mathbf{H}$ is the perceptual weighting matrix having the following circulant structure

$$\mathbf{H} = \begin{bmatrix} h(0) & h(K-1) & \cdots & h(1) \\ h(1) & h(0) & \cdots & h(K-1) \\ \vdots & \vdots & \ddots & \vdots \\ h(K-1) & h(K-2) & \cdots & h(0) \end{bmatrix}, \tag{9}$$

with $h(n) = \frac{1}{K} \sum_{k=0}^{K-1} \sqrt{A(k)} \cos\left(2\pi k n / K\right)$. Additionally, this matrix is also symmetric, i.e. $\mathbf{H}^H = \mathbf{H}$. This distortion measure has been applied to sinusoidal audio modeling and coding in, for example, [25, 26].

4. SINUSOIDAL ESTIMATION

We now present some new insights into the problem of finding the perceptually most important sinusoids that were originally published in [33]. Given a real observed signal $x(n)$ for $n = 0, \ldots, N-1$, find the parameters of the signal of interest $\hat{x}(n)$ in additive noise $e(n)$:

$$x(n) = \hat{x}(n) + e(n). \tag{10}$$

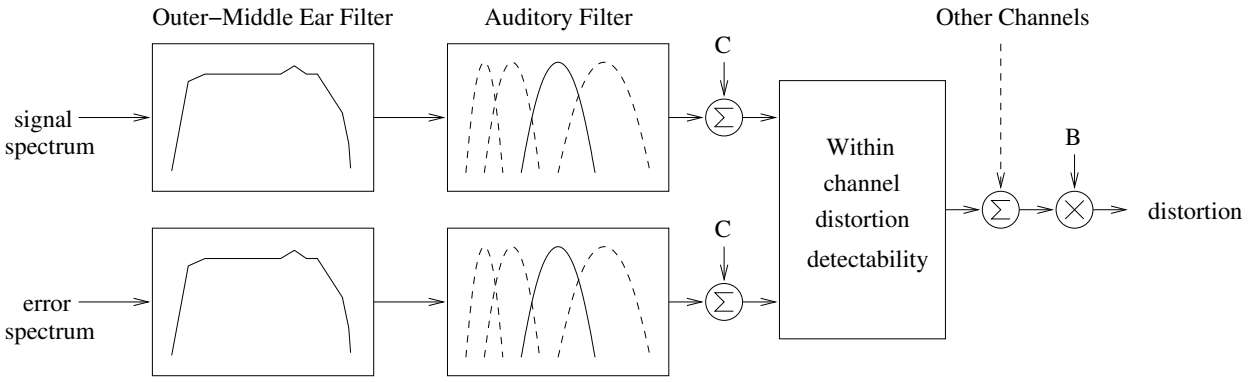

Fig. 1: Block diagram of the masking model used throughout this paper. The model was proposed in [31, 32].

In our case the signal of interest $\hat{x}(n)$ is a sum of sinusoidal components

$$\hat{x}(n) = \sum_{l=1}^{L} A_l \cos\left(\omega_l n + \phi_l\right), \tag{11}$$

with each component having an amplitude A_l, phase ϕ_l, and frequency ω_l. The perceptual nonlinear least-squares (NLS) estimates of the frequencies $\boldsymbol{\omega} = \left[\, \omega_1 \;\cdots\; \omega_L \,\right]^T$ are the set of frequencies that minimize the perceptual distortion, i.e.,

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}}{\operatorname{argmin}} \|\mathbf{H}(\mathbf{x} - \mathbf{Z}\mathbf{a})\|_2^2. \tag{12}$$

Seen in the light of the R-D optimization problem, the estimator in (12) is optimal in the sense that for a given number of bits (with the number of bits being approximately proportional to the number of sinusoids), it minimizes the perceptual distortion. The matrix $\mathbf{Z}$ is a Vandermonde matrix defined as

$$\mathbf{Z} = \left[\; \mathbf{z}_1 \quad \mathbf{z}_1^* \quad \ldots \quad \mathbf{z}_L \quad \mathbf{z}_L^* \;\right], \tag{13}$$

with * denoting complex conjugation and

$$\mathbf{z}_l = \left[\; z_l^0 \quad \ldots \quad z_l^{N-1} \;\right]^T, \tag{14}$$

where $z_l = e^{j\omega_l}$ are the complex poles. Furthermore, we have that $\mathbf{a} = \left[\, a_1 \; a_1^* \;\cdots\; a_L \; a_L^* \,\right]^T$ with $a_l = \frac{A_l}{2} e^{j\phi_l}$. These can be estimated as

$$\hat{\mathbf{a}} = \left(\mathbf{Z}^H \mathbf{H}^2 \mathbf{Z}\right)^{-1} \mathbf{Z}^H \mathbf{H}^2 \mathbf{x}. \tag{15}$$

Now, we can write the optimal estimator that minimizes the perceptual distortion as

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}}{\operatorname{argmin}} \|\mathbf{H}(\mathbf{x} - \mathbf{Z}\mathbf{a})\|_2^2 \tag{16}$$

$$= \underset{\boldsymbol{\omega}}{\operatorname{argmax}}\, \mathbf{x}^H \mathbf{H}^2 \mathbf{Z} \left(\mathbf{Z}^H \mathbf{H}^2 \mathbf{Z}\right)^{-1} \mathbf{Z}^H \mathbf{H}^2 \mathbf{x}. \tag{17}$$

However, solving this is not computationally feasible. Instead several relaxed methods that find sinusoids one at a time have been proposed and we will now relate these to the minimization of the perceptual distortion measure. First, we define the residual vector at iteration i as $\mathbf{r}_i = \left[\, r_i(0) \;\cdots\; r_i(N-1) \,\right]^T$ with

$$r_{i+1}(n) = r_i(n) - \hat{A}_i \cos(\hat{\omega}_i n + \hat{\phi}_i), \tag{18}$$

which is initialized as $r_1(n) = x(n)$. In the perceptual matching pursuit [25], which is a derivative of matching pursuit [19] (see also [34]), the sinusoid that minimizes the perceptual norm is chosen as

$$\hat{\omega}_i = \underset{\omega}{\operatorname{argmin}} \|\mathbf{H}(\mathbf{r}_i - \mathbf{z}a)\|_2^2, \tag{19}$$

with $\mathbf{z} = \left[\, e^{j\omega 0} \;\cdots\; e^{j\omega(N-1)} \,\right]^T$. then we get the greedy frequency estimator

$$\hat{\omega}_i = \underset{\omega}{\operatorname{argmax}} \frac{|\left\langle \mathbf{H}\mathbf{z}, \mathbf{H}\mathbf{r}_i \right\rangle|^2}{\|\mathbf{H}\mathbf{z}\|_2^2}, \tag{20}$$

and the associated optimal scaling (the amplitude and phase) is

$$\hat{a}_i = \frac{\left\langle \mathbf{H}\mathbf{z}, \mathbf{H}\mathbf{r}_i \right\rangle}{\|\mathbf{H}\mathbf{z}\|_2^2}. \tag{21}$$

Note that the perceptual matching pursuit can be solved efficiently using FFTs. Furthermore, it converges in the perceptual distortion meaning that the perceptual distortion decreases as we increase the number of sinusoids and thereby also the number of bits. Consider now the signal model component being an eigenvector of the perceptual weighting matrix

$$\mathbf{H}\mathbf{v} = \lambda\mathbf{v}. \tag{22}$$

This assumption leads to some interesting results and is indeed valid for certain important cases. As is well-known, complex sinusoids are eigenvectors of convolution operators, i.e.,

$$\mathbf{v} = \left[\begin{array}{ccc} e^{j\omega 0} & \cdots & e^{j\omega(N-1)} \end{array} \right]^T. \tag{23}$$

This holds only in general for the asymptotic case $N \rightarrow \infty$. Using the eigenvector assumption the sinusoidal frequency estimation criterion can be significantly simplified:

$$\min \|\mathbf{H}(\mathbf{r}_i - \hat{\mathbf{r}}_i)\|_2^2 = \min \|\mathbf{H}\mathbf{r}_i - \lambda\mathbf{v}a\|_2^2, \tag{24}$$

where a is a complex scale factor. The estimation criterion can now be reduced to the so-called pre-filtering method [22, 23], i.e.,

$$\hat{\omega}_i = \underset{\omega}{\operatorname{argmax}} \frac{|\langle \mathbf{v}, \mathbf{H}\mathbf{r}_i \rangle|^2}{N}. \tag{25}$$

Note that $< \mathbf{v}, \cdot >$ can be found using an FFT. The inner product in the estimation criterion can further be rewritten as

$$\langle \mathbf{v}, \mathbf{H}\mathbf{r}_i \rangle = \mathbf{v}^H \mathbf{H}\mathbf{r}_i = (\lambda\mathbf{v})^H \mathbf{r}_i, \tag{26}$$

whereby frequency estimation criterion then becomes

$$\hat{\omega}_i = \underset{\omega}{\operatorname{argmax}} |\lambda|^2 \frac{|\langle \mathbf{v}, \mathbf{r}_i \rangle|^2}{N}. \tag{27}$$

This is the weighted MP [24]. In weighted MP, the eigenvalue of a complex sinusoids of frequency $\omega = 2\pi\frac{k}{K}$ is approximated as

$$\hat{\lambda} \approx \sqrt{A(k)}. \tag{28}$$

As we have seen in this section, both the weighted MP and the pre-filtering method can be seen as approximations to the perceptual MP and as the segment length N is increased, these approximations become more accurate. The weighted MP is identical to the pre-filtering method and the perceptual MP under certain conditions, namely that the sinusoids are eigenvectors of the perceptual weighting matrix. Additionally, the perceptual MP can be seen as a relaxation of the optimal perceptual nonlinear least-squares method that simultaneously solves for the L sinusoids that minimize the perceptual distortion. It is also interesting to note that asymptotically, all these methods attain the Cramér-Rao bound meaning that the estimated frequencies have the lowest possible variance [33] for sufficiently large N under some mild conditions on the noise and the perceptual weighting function.

5. EFFICIENT CODING OF TRANSIENTS

There exists a number of different and complementary tools for handling transients, namely 1) segmentation 2) variable rate, and 3) perceptual noise-shaping/distortion measure. R-D optimization based on a perceptual distortion measure incorporates all these in an elegant manner. Also, a number of adapted signal models based on amplitude modulation (AM) have been proposed specifically for dealing with transients in audio coding in combination with the methods mentioned above [35–39]. In this section we compare these method, which are all based on the following modified sinusoidal signal model for $n = 0, \ldots, N-1$

$$\hat{x}(n) = \sum_{l=1}^{L} \gamma_l(n) A_l \cos(\omega_l n + \phi_l), \tag{29}$$

where $\gamma_l(n)$ is the amplitude modulating signal or envelope if $\gamma_l(n) \geq 0$ for all n. The papers referenced above then differ in the model they impose on $\gamma_l(n)$ and how the model parameters are found. Aside from the question of how to model $\gamma_l(n)$ and find the associated parameters, there is also the interesting question of how to decide when to use such modified models. Since most audio segments are stationary, these modified models that have additional parameters associated with them will not always be the best choice. However, by R-D optimization, any heuristic switching, as is often seen in audio coders, can be avoided. A block diagram of the encoder and decoder based on rate-distortion optimization and amplitude modulated sinusoidal audio coding is depicted in Figure 2.

In [38], an amplitude modulated sinusoidal audio coder is presented. It is based on a nonlinear model of the mod-

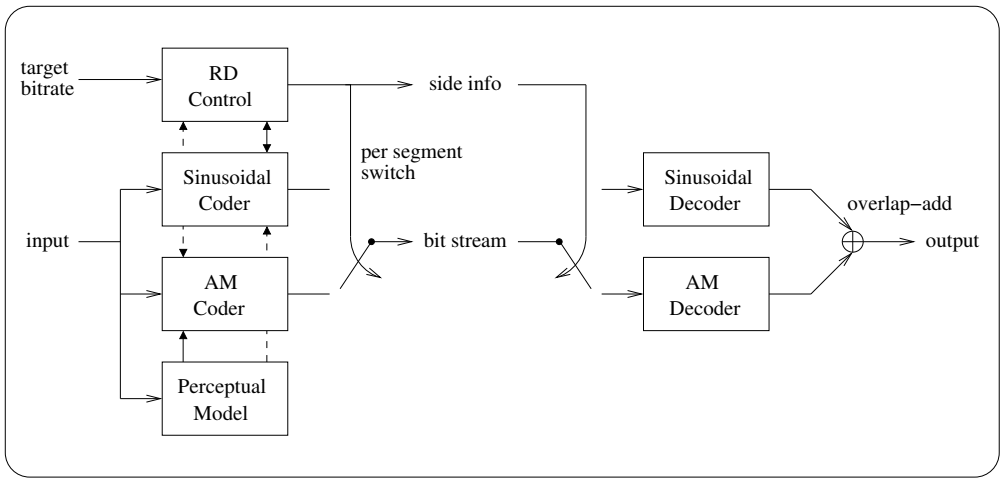

Fig. 2: Block diagram of the encoder and decoder that include an amplitude modulated (AM) coder and a sinusoidal coder based on constant amplitudes.

ulating signal which is characterized by an onset, an attack, and a decay. Each sinusoidal component can have a different envelope. This model is combined with a sinusoidal coder without amplitude modulation in a rate-distortion optimized framework that uses optimal distribution of sinusoids over segments and optimal segmentation [29, 30]. The gamma envelopes are given as

$$\gamma_l(n) = u(n - n_l)\,(n - n_l)^{\alpha_l}\, e^{-\beta_l(n-n_l)}. \qquad (30)$$

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 $\beta_l \in \mathbb{R}^+$. The parameters of this signal model are found by analysis-by-synthesis using the perceptual distortion measure.

In Figure 3 an example of a coded signal is shown. In the top panel, the original, the claves signal from SQAM [40], is shown. In the middle panel, the reconstructed signal is shown for a sinusoidal coder that uses constant amplitudes (CA) and a fixed segmentation while, in the bottom panel, the reconstructed signal for the AM coder combined with the CA coder using optimal allocation and segmentation (AM/CA+SEG) is shown. In both cases, the bit-rate was 30 kbps. The coder uses log-quantization of amplitudes and frequencies, uniform quantization of phases along with segment lengths of 10, 20, 30, 40 ms segments with 5 ms overlap.

A MUSHRA-like test [41] was carried out with 9 listeners participating and using 7 critical transients mono

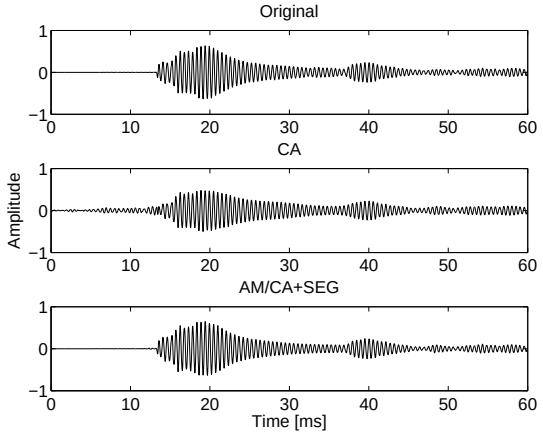

Fig. 3: Example of a signal coded using the AM coder based on the gamma envelopes. In the top panel, the original signal is depicted. In the middle panel, the signal coded by the CA is shown, while a the bottom the synthesized signal of the AM/CA+SEG coder is shown. In both cases, the bit-rate is 30 kbps.

excerpts from the EBU's SQAM [40]. The mean scores and the respective 95 % confidence intervals are listed in Table 1 for a bit-rate of 30 kbps. The results are shown for four different configurations, namely constant amplitude (CA) with a fixed segmentation, constant amplitude

AM Coding Techniques			
	Gamma Envelopes	Linear Combination	FDLP
Model	Nonlinear, attack, decay, onset parameters	Linear combination of vectors	Frequency domain AR process (in subbands)
Estimation	Analysis-by-synthesis	Least-Squares	Linear Prediction
Complexity	Medium	High	Low
Quality	High	Medium	Medium
Flexibility	Medium	High	Low

Table 2: Comparison of various amplitude modulated sinusoidal audio coders.

Results of Listening Test				
Statistic	CA	CA+SEG	AM	AM/CA+SEG
Mean	45	59	48	70
Conf.	5	6	5	5

Table 1: Summary of the results (means and $\pm$ 95 % confidence intervals) of MUSHRA-like listening test in [38] for the amplitude modulated sinusoidal audio coder based on the gamma envelopes for 30 kbps.

with optimal segmentation (CA+SEG), amplitude modulated with a fixed segmentation (AM) and the combination of the AM and CA coders with optimal segmentation (AM/CA+SEG). The results reported in [38] prove that it is indeed efficient in terms of bit-rate to allow different modulating signals for different components and that optimal segmentation and adapted models are complementary coding techniques; furthermore, the optimal segmentation changes with the signal model.

Two other AM coders have also been developed that have different properties. In [37], the amplitude modulating signal is modeled as a linear combination of arbitrary basis vectors. This model is rather different from the other models considered in this section in that the constraints on the amplitude modulating signal being nonnegative is relaxed; sinusoidal frequencies may occur at spectral minima. The model can exploit spectral symmetries for coding purposes and is demonstrated in listening tests to improve upon a sinusoidal coder. Also, this coder has the advantage that since the model parameters are linear they may easily be optimize. In terms of achieving a scalable and flexible coder, this model is desirable. The main downside of the AM coders and their complicated signal models is the complexity associated with finding the parameters. The work presented in [35] aims at finding an alternative that has low complexity. An amplitude modulated sinusoidal audio coder based on the theory of [35] and the results of paper [36] was developed in [39]. It uses frequency-domain linear prediction (FDLP), a principle similar to the temporal noise shaping of [42], as a means for estimation and efficient coding of the modulating signal. The envelopes are found in critically sampled subbands, and given these envelopes, the remaining sinusoidal parameters are found. This coder has very low complexity and requires little memory compared to that of [38], and it is demonstrated in listening tests to improve upon a baseline coder in a delay constrained setup. The strength of the methods used in this paper is that the model of the modulating signal is not very restrictive. On the other hand, the envelope estimator will approximate the squared instantaneous envelope of the subbands, so the method may result in incorrect envelopes for complicated sub-band signals consisting of mixtures of different components. Additionally, it was also shown in [38] that the prediction order should not be chosen too high as the estimator then will model cross-terms that are due to the sinusoidal carriers. In Table 2 an overview of the various AM coders and their properties are shown in terms of the underlying model characteristics, the estimation procedure used for finding the modulating signal, the complexity, and perceived quality. Finally, also the flexibility, by which we mean how restrictive and scalable the model is compared in relative terms. From the table and the discussion in this section, it should be clear that the various methods have their pros and cons and that it depends on the application at hand which is the best solution.

6. RATE-DISTORTION ESTIMATION

In order to find the R-D optimal allocation and segmentation, we need to find $R(\sigma, \tau)$ and $D(\sigma, \tau)$. These are found by encoding and decoding the signal for various

combinations of segments and coding templates and in many cases, this will be done for coding templates and segments that will not be used in the coding of the signal. The optimal segmentation, for example, requires that $V = \frac{G^2+G}{2}$ different segments are evaluated for all different coding templates when segment lengths can be from 1 to G (measured in terms of the minimum segment length, see (1)). If the maximum segment length is limited to H with $G \gg H$ then $V \approx GH$. However, the number of segments used can be no more than G. Likewise, the coder switching structure employed in the AM/CA coder in Section 5 requires that distortions and rates are calculated for both the CA and AM coders. Clearly, this approach will generally be wasteful in terms of computational complexity. In rate-distortion estimation [43], the distortions $D(\sigma, \tau)$ are estimated for different rates $R(\sigma, \tau)$ based on signal features. We now briefly present the basic idea of rate-distortion estimation and its application to a sinusoidal coder. The joint PDF of the distortions $\mathbf{D} = [\, D_1 \, \ldots D_M \,]$ associated with coding templates $\tau_1, ..., \tau_M$ and the feature vector $\mathbf{p}$ is modeled as

$$p(\mathbf{D}, \mathbf{p}) = \sum_{k=1}^{K} w_k p_k(\mathbf{D}, \mathbf{p}) \tag{31}$$

where $p_k(\mathbf{D}, \mathbf{p})$ is a multivariate Gaussian having mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$ while w_k is the mixture weight. These are found using the expectation maximization algorithm Given some observed features in $\mathbf{p}$ we can then estimate the distortion $\hat{D}_i$ by minimizing the Bayesian mean-square error:

$$E = \int \int (D_i - \hat{D}_i)^2 p(\mathbf{D}, \mathbf{p}) d\mathbf{D} d\mathbf{p} \tag{32}$$

$$= \int \left[\int (D_i - \hat{D}_i)^2 p(\mathbf{D}|\mathbf{p}) d\mathbf{D} \right] p(\mathbf{p}) d\mathbf{p}. \tag{33}$$

This is solved by the conditional mean estimator:

$$\hat{\mathbf{D}} = \int \mathbf{D} p(\mathbf{D}|\mathbf{p}) d\mathbf{D}. \tag{34}$$

It turns out that this has a particularly simple form for Gaussian mixtures, i.e.,

$$\hat{\mathbf{D}} = \sum_{k=1}^{K} \tilde{w}_k \tilde{\boldsymbol{\mu}}_k, \tag{35}$$

where $\tilde{\boldsymbol{\mu}}_k$ is the conditional mean. In [44] it was proposed to use only diagonal covariance matrices for a number of reasons. Firstly, the computational complexity associated with the estimator for the diagonal case is much less than that of full covariance matrices, although a higher model order may be needed for the modeling performance of the GMM. Secondly, the training of the GMM and the Bayesian estimator can easily be shown to preserve the desirable properties that the rate-distortion curves are non-increasing and often also convex.

In [45], the principle of rate-distortion estimation was applied to a sinusoidal coder and later, in [44], also to the problem of finding the optimal segmentation. Note that the principle of rate-distortion estimation also can be applied to determining whether an amplitude modulated sinusoidal signal model should be applied. We will now briefly present some results regarding the application of rate-distortion to optimal segmentation in a sinusoidal coder and summarize the results. The features that were used for estimating the distortions in the sinusoidal coder were log-power, spectral flatness, linear prediction flatness, spectral centroid, spectral bandwidth, power stationarity, and spectral stationarity. All these features have in common that they are very simple and have fast implementations. Different number of sinusoids are used as coding templates and segments of 10, 20, 30, and 40 ms were used with 5 ms overlap in the optimal segmentation. We compare the estimated segmentation to the optimal segmentation and a fixed segmentation (30 ms segments) at 30 kbps. A MUSHRA-like test was performed using 6 mono excerpts from SQAM [40] and 8 listeners. In Table 3 the results are shown in the form of means and 95 % confidence intervals. The listening test shows that the estimated rate-distortion pairs can be replaced by estimates with only a moderate loss in quality while simulations indicate a complexity reduction by a factor of 10. It must be stressed that the rate-distortion estimation scheme can take dependencies between target bit-rate and different coding templates on the optimal segmentation into account.

7. SUMMARY

In this paper, some new methods and results in rate-distortion optimized audio coding based on a perceptual distortion measure have been presented. We have shown how various methods for incorporating perception in sinusoidal estimators relate. Various amplitude modulated sinusoidal audio coders having different properties have been proposed. It has been demonstrated that amplitude modulated sinusoidal audio coding can improve parametric coders. Finally, it has been proven that the

Results of Listening Test			
Statistic	Estimated	Fixed	Optimal
Mean	53	33	59
Conf.	8	8	8

Table 3: Summary of the results of MUSHRA-like listening test (means and $\pm$ 95 % confidence intervals) in [44] for the rate-distortion estimation scheme. The excerpts were encoded at 30 kbps.

principle of rate-distortion estimation can alleviate one of the major issues in R-D optimized audio coding, namely complexity.

8. ACKNOWLEDGMENTS

The work of M. G. Christensen was supported by the ARDOR project (Adaptive Rate-Distortion Optimized sound codeR project, EU grant no. IST–2001–34095). The authors thank their various co-authors on the respective parts of the work referenced herein, namely Christoffer A. Rødbro, Fredrik Nordén, Andreas Jakobsson, Søren Vang Andersen, and Steven van de Par for fruitful collaborations and everybody who participated in the listening tests.

9. REFERENCES

[1] K. Brandenburg and G. Stoll, "ISO-MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 31–42.

[2] M. Bosi *et al.*, "ISO/IEC MPEG-2 Advanced Audio Coding," in *101th Conv. Aud. Eng. Soc.*, 1996.

[3] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 4.

[4] 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.

[5] 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.

[6] 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.

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

[9] ISO/IEC, *ISO/IEC 14496-3:2001/AMD2*, Parametric Coding for High-Quality Audio, July 2004.

[10] 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.

[11] R. Vafin and W. B. Kleijn, "Rate-distortion optimized quantization in multistage audio coding," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14(1), pp. 311–320, Jan. 2006.

[12] N. van Schijndel and S. van de Par, "Rate-distortion optimized hybrid sound coding," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 2005, pp. 235–238.

[13] F. Baumgarte and C. Faller, "Binaural Cue Coding–Part I: Psychocaoustic fundamentals and design principles," *IEEE Trans. Speech and Audio Processing*, vol. 11(6), pp. 509–519, 2003.

[14] C. Faller and F. Baumgarte, "Binaural Cue Coding–Part II: Schemes and applications," *IEEE Trans. Speech and Audio Processing*, vol. 11(6), pp. 520–531, 2003.

[15] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *EURASIP J. on Applied Signal Processing*, vol. 9, pp. 1305–1322, June 2005.

[16] J. Herre, "From joint stereo to spatial audio coding–recent progress and standardization," in *Proc. Int. Conf. Digital Audio Effects*, 2004, pp. 157–162.

[17] D. Schulz, "Improving audio codecs by noise substitution," *J. Audio Eng. Soc.*, vol. 7/8, pp. 593–598, Jul/Aug 1996.

[18] J. Herre and D. Schulz, "Extending the MPEG-4 AAC codec by Perceptual Noise Substitution," in *Proc. 104th Conv. Aud. Eng. Soc.*, 1998, paper preprint 4720.

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

[20] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1993.

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

[22] B. Edler and G. Schuller, "Audio coding using a psychoacoustic pre- and post-filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2000, pp. 881–884.

[23] J. Jensen, R. Heusdens, and S. H. Jensen, "A perceptual subspace method for sinusoidal speech and audio modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2003.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] M. G. Christensen, "Estimation and modeling problems in parametric audio coding," Ph.D. dissertation, Aalborg University, 2005.

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

[30] 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.

[31] 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.

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

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

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

[35] 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.

[36] 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.

[37] M. G. Christensen, A. Jakobsson, S. V. Andersen, and S. H. Jensen, "Amplitude modulated sinusoidal signal decomposition for audio," *IEEE Signal Processing Lett.*, vol. 13(7), July 2006.

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

[39] M. G. Christensen and S. H. Jensen, "Computationally efficient amplitude modulated sinusoidal audio coding using frequency-domain linear prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2006, to appear.

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

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

[42] 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.

[43] F. Nordén, S. V. Andersen, S. H. Jensen, and W. B. Kleijn, "Property vector based distortion estimation," in *Rec. Thirty-Eighth Asilomar Conf. Signals, Systems, and Computers*, 2004, pp. 2275–2279.

[44] C. A. Rødbro, M. G. Christensen, F. Nordén, and S. H. Jensen, "Low complexity rate-distortion optimized time-segmentation for audio coding," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 2005, pp. 231–234.

[45] F. Nordén, M. G. Christensen, and S. H. Jensen, "Open loop rate-distortion optimized audio coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, 2005, pp. 161–164.