OPEN LOOP RATE-DISTORTION OPTIMIZED AUDIO CODING

Fredrik Nordén, Mads Græsbøll Christensen, and Søren Holdt Jensen

Department of Communication Technology,
Aalborg University, Denmark

ABSTRACT

This paper addresses complexity reduced rate-distortion optimized audio coding under rate constraint. A technique where distortion minimizing coding templates, chosen from a set of templates, are jointly selected for a set of segments. This optimization requires knowledge of rate-distortion pairs for all segments, and for each coding template, which often are costly to obtain. The proposed framework exchanges true rate-distortion pairs with predicted ones, thereby allowing for complexity reduction. The prediction is based on a property vector extracted for each segment, from which distortion predictions, using Gaussian mixture models, are performed. Here, we evaluate the proposed framework in a sinusoidal coding context. The results show that the proposed framework can increase the distortion performance, compared to a fixed sinusoidal coding scheme.

1. INTRODUCTION

Rate-distortion (R-D) optimization is of interest for audio coding for several reasons. It allows for adaptive coding schemes, where the coder is adapted to user and network constraint as well as source characteristics, thereby increasing the overall distortion performance. For example, parametric coders typically outperform transform coders at low bit-rates, and LPC-based coders perform very well for speech but not for audio. An R-D optimized selection among such a set of coders is thus of interest.

There are a multitude of different applications that can be put into the R-D optimization framework: *1*) Coder selection for specific segments [1], *2*) Distribution of bits over stages in multistage structures [2], *3*) Variable bit-rate (optimal distribution of bits over segments) [3], and *4*) Dynamic time-segmentation [4, 3]. All of these applications require knowledge of the incurred distortion in the current audio segment for all of the coders (coding template, number of sinusoids, etc), in order to perform R-D optimization. For some of the above applications, we end up having to do distortion calculations, which sometimes require both signal analysis and synthesis, for many different coding templates, not necessarily useful in the final coder synthesis.

The complexity of these distortion calculations may be severe, preventing the use of R-D optimized coders in many applications. Thus, we here propose an open loop approach to the R-D optimization problem. We exchange coding distortions with predicted ones, thereby allowing for complexity reduction. For the prediction purpose we employ an open loop framework for distortion prediction proposed in [5]. The framework is based on a property vector extracted from the segment to be coded, from which distortion predictions, using a Gaussian mixture model (GMM) of the joint property-distortion pdf, are performed. We evaluate the proposed framework in a sinusoidal coding context. Based on predicted R-D curves we perform R-D optimized distribution of sinusoids over sets of segments matching a given bit-budget. The results are compared with a sinusoidal coder optimized on original R-D curves, and a sinusoidal coder using a fixed number of sinusoids per segment.

The paper is organized as follows. In Sec. 2 we discuss the basics of R-D optimized coding, and in Sec. 3 we present the prediction framework. This is followed by a presentation of the experimental setup in Sec. 4. In Sec. 5 we evaluate the goodness of the proposed system. Finally, we conclude in Sec. 6.

2. RATE-DISTORTION OPTIMIZATION

The problem of distributing a certain number of bits over a set of segments, $\mathcal{S}$, constituting an optimization *viewport*, can be cast into rate-distortion optimization under rate constraint. This optimization can be stated as the following constrained optimization problem:

$$\begin{aligned} \min \quad & D \\ \text{s.t.} \quad & R \leq R^\star, \end{aligned} \tag{1}$$

where D is the distortion, R is the resulting rate, and $R^\star$ is the target rate. Let $\mathcal{T}_s$ be a finite, discrete set of coding templates (ways of encoding, etc.) for segment s, and $R(\tau)$ and $D(\tau)$ be the rate and distortion associated with coding template $\tau \in \mathcal{T}_s$. The distortion D and the rate R are the sum of distortions and rates over the segments, $\mathcal{S}$, associated with a particular set of coding templates $\boldsymbol{\tau} = [\tau_1 \ \cdots \ \tau_S]$ with $\tau_i \in \mathcal{T}_i$, i.e.

$$D = \sum_{s=1}^{S} D(\tau_s) \quad \text{and} \quad R = \sum_{s=1}^{S} R(\tau_s). \tag{2}$$

The problem (1) can then be written as the following unconstrained problem [4]

$$\min_{\boldsymbol{\tau}} \sum_{s=1}^{S} D(\tau_s) + \lambda R(\tau_s) = \sum_{s=1}^{S} \min_{\tau \in \mathcal{T}_s} D(\tau) + \lambda R(\tau), \tag{3}$$

where λ is the non-negative Lagrange multiplier. The right side follows from assuming that distortions and rates are additive and independent over segments..

This means that the optimization problem can be solved independently for each segment for a particular λ. The Lagrange multiplier λ can be interpreted as the slope of the R-D curve for a certain rate. The problem is then to find the $\lambda^\star$ that leads to the target bit rate $R^\star$. Such a λ cannot be guaranteed to exist for discrete problems such as ours. We can, however, find a solution close to

This research was conducted within the ARDOR project, and was supported by the E.U. under grant IST-2001-34095

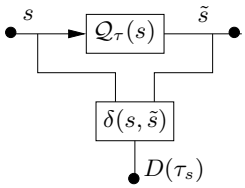

Fig. 1: Illustration of the evaluation of the incurred distortion, $D(\tau_s)$, for one particular coding template, τ, and one particular audio segment, s. $\mathcal{Q}_\tau(\cdot)$ represents the coding or modeling associated with template, τ, and $\delta(\cdot)$ is the distortion criterion.

the optimal one provided that the $\{R(\tau), D(\tau)\}$ points are sufficiently dense. The optimal λ is found by maximizing the concave Lagrange dual function:

$$\lambda^\star = \text{argmax}_\lambda \left[\sum_{s=1}^{S} \left(\min_{\tau \in \mathcal{T}_s} D(\tau) + \lambda R(\tau) \right) - \lambda R^\star \right]. \quad (4)$$

This can be done by sweeping over λ using simple bisection until the rate $R(\lambda)$ is within some range of the target bit rate [4].

Given the optimal $\lambda^\star$, the rate-distortion optimization simply becomes a matter of choosing the optimum coding template for a particular segment s as

$$\tau_s^\star = \text{argmin}_{\tau \in \mathcal{T}_s} \left[D(\tau) + \lambda^\star R(\tau) \right]. \quad (5)$$

For the rate-distortion optimization to result in improvements in perceived quality, the chosen distortion criterion, $\delta(\cdot)$, must reflect human sound perception. In this work we have chosen to work with the distortion criterion proposed in [6], which is further described in Sec. 4.

3. RATE-DISTORTION PREDICTION

To perform R-D optimized coding over a set of segments, $\mathcal{S}$, using a set of coding templates, $\mathcal{T}_s$, we require knowledge of R-D points for each segment and each coding template,

$$\{R(\tau_s), D(\tau_s)\} : \forall \ s \in \mathcal{S}, \ \forall \ \tau_s \in \mathcal{T}_s. \quad (6)$$

Ideally these points are found by coding each segment with each of the coding templates, as visualized in Fig. 1[1]. This approach is highly complex, and in general therefore not feasible. Thus we here suggest an open loop alternative, where distortions, $\{D(\tau_s)\}$, are predicted from the current segment of audio, s, as visualized in Fig. 2. In essence the structure in Fig. 1 is exchanged for the structure in Fig. 2. Below, we discuss the predictor employed to predict the incurred distortion for one particular coding template. In practice we require one predictor, as described below, for each coding template.

3.1. Property Vector Based Prediction

We employ distortion prediction as suggested in [5]. The over all prediction is separated into a property extraction, $f(\cdot)$, and a prediction, $g_\tau(\cdot)$, as visualized in Fig. 2. Each audio segment, s, is processed into a dimension reduced property vector $\mathbf{P}$, from which a prediction, $\hat{D}(\tau_s)$, of the coding distortion, $D(\tau_s)$ is to be found. For simplicity, we below drop segment and coding template indices. The random variable representing the incurred distortion will be denoted $\mathcal{D}$, and the corresponding outcomes will be denoted δ.

[1]The structure in Fig. 1 needs to be processed NxM times, if we perform a joint optimization over N segments, using M coding templates for each segment.

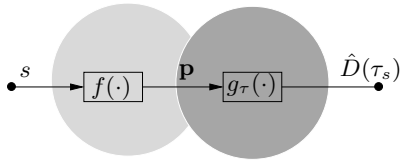

Fig. 2: A framework for prediction of the incurred distortion, $D(\tau_s) = \delta(s, \mathcal{Q}_\tau(s))$, when coding a random vector s, using coding template τ. A dimension reducing property vector extraction, $f(\cdot)$, is followed by a distortion prediction, $g_\tau(\cdot)$.

The selection of a set of properties, $\mathbf{p}$, from the input segment, s, is of great importance for the performance of the proposed framework. The selected set of properties should be a representative for the incurred distortion in the current segment for the given coder. In more theoretical terms, the random input segment, s, is processed into two random variables, the distortion variable, $\mathcal{D}$, with outcomes δ, and the property vector, $\mathbf{P}$. The basic task for the property extractor, $f(\cdot)$, is to extract properties, $\mathbf{P}$, that contain sufficient information about $\mathcal{D}$ for a required predictor accuracy. The amount of information that $\mathbf{P}$ contains about $\mathcal{D}$, or the goodness of a given property vector, can be measured by the mutual information $I(\mathcal{D}; \mathbf{P})$. In this work we have chosen to rely on standard audio properties. Our choice of property vector is further discussed in Sec. 4.

The aim of the predictor, $g(\cdot)$, is to find a prediction, $\hat{\delta}$, of the incurred distortion, δ, based on an observation of the property vector, $\mathbf{P} = \mathbf{p}$. Utilizing a pre-trained GMM for the joint distortion property pdf, $f^{(\mathcal{M})}_{\mathcal{D},\mathbf{P}}(\delta, \mathbf{p})$, we approximate the MMSE at each coding instant as

$$\hat{\delta} = g(\mathbf{p}) = \int \delta f^{(\mathcal{M})}_{\mathcal{D}|\mathbf{P}}(\delta|\mathbf{P} = \mathbf{p}) d\delta, \quad (7)$$

where $f^{(\mathcal{M})}_{\mathcal{D}|\mathbf{P}}(\delta|\mathbf{P} = \mathbf{p})$ is the conditional model pdf, which can be shown to be a mixture of Gaussian densities, and is easily derived from the joint model pdf, $f^{(\mathcal{M})}_{\mathcal{D},\mathbf{P}}(\delta, \mathbf{p})$. In practice, this predictor calculates a weighted sum of conditional means,

$$\hat{\delta} = \sum_{i=1}^{M} \rho_i' \mathbf{m}_{i,\mathcal{D}|\mathbf{P}=\mathbf{p}}, \quad (8)$$

where M is the number of mixture components, and $\{\rho_i'\}$ and $\{\mathbf{m}_{i,\mathcal{D}|\mathbf{P}=\mathbf{p}}\}$ represent the weights and the means of the conditional model pdf, $f^{(\mathcal{M})}_{\mathcal{D}|\mathbf{P}}(\delta|\mathbf{P} = \mathbf{p})$, respectively.

3.2. Performance

The employed prediction scheme is designed to minimize the variance of the prediction error, $Z = \delta - \hat{\delta}$. Assuming an unbiased predictor, the variance of the prediction error can be expressed as

$$\sigma_Z^2 = \text{E}\left[(Z)^2\right] = \text{E}\left[(\delta - \hat{\delta})^2\right]. \quad (9)$$

The *minimum mean square error estimator* (MMSE) for this task, i.e., the one minimizing σ_Z^2, is the conditional mean estimator,

$$\hat{\delta}_{\text{mmse}} = \text{E}[\mathcal{D}|\mathbf{P} = \mathbf{p}] = \int \delta f_{\mathcal{D}|\mathbf{P}}(\delta|\mathbf{P} = \mathbf{p}) d\delta. \quad (10)$$

The employed predictor is an approximation of the MMSE estimator, and the predictor output (8) will approach the true conditional (10), as the model pdf approaches the true pdf.

As discussed above, the performance of the predictor is dependent of the chosen property vector. In [5] the relation between the property goodness, $I(\mathcal{D};\mathbf{P})$, and the overall prediction error, σ_Z^2 was studied. It was shown that for a given property vector, $\mathbf{P}$, the overall prediction error, σ_Z^2, can be bounded as

$$\sigma_{\mathcal{D}}^2 \geq \sigma_Z^2 \geq \frac{1}{2\pi e} 2^{2(h(\mathcal{D})-I(\mathcal{D};\mathbf{P}))}, \qquad (11)$$

where $\sigma_{\mathcal{D}}^2$ is the variance of the distortion variable to be predicted, $h(\mathcal{D})$ is the differential entropy of the distortion random variable $\mathcal{D}$, and $I(\mathcal{D};\mathbf{P})$ is the mutual information between $\mathcal{D}$ and $\mathbf{P}$.

4. EXPERIMENTAL SETUP

Here, we present the experimental framework, separated into the source coding system (sinusoidal coder, R-D optimization , distortion criterion), and the distortion predictor (GMM, property vector, audio database).

4.1. Source Coding System

We employ a *sinusoidal coder* based on a simplified version of psychoacoustic matching pursuit (PAMP) [7]. Using a PAMP based coder, the distortion (12) will decrease in a monotone way as a function of the number of iterations (sinusoids). The analysis/-synthesis is performed for segments of length 35 ms, sampled at 48 kHz. The coder employs a Hanning window and has a 50 % segment overlap. Phases are quantized uniformly using 5 bits per component, whereas amplitudes and frequencies are quantized in the logarithmic domain. Using entropy coding and differential encoding, we obtain perceptually transparent quantization at an average rate of approximately 16 bits/sinusoid.

R-D optimization, c.f. Sec. 2, is here employed to distribute sinusoids (bit-allocation) over optimization viewports, $\mathcal{S}$, of length 1 s, matching a target rate of 25 kbits/s[2]. For each segment the algorithm allocates a number of sinusoids in the range $0-85$. The optimization is performed using the sinusoidal modeling distortion as input, using a cost of 16 bits/sinusoid.

We employ a *distortion criterion*, $\delta(\cdot)$, based on the model in [6]. For a particular segment the distortion can be written as

$$\delta(e(n)) = \int_{-\pi}^{\pi} A(\omega)|\mathcal{F}\{w(n)e(n)\}|^2 d\omega, \qquad (12)$$

where $\mathcal{F}\{\cdot\}$ denotes the Fourier transform, $A(\omega) \in \{x \in \mathbb{R} | x > 0\}$ is a perceptual weighting function and $w(n)$ is the analysis window and $e(n) = \tilde{s}(n) - s(n)$ is the modeling error. The quantization distortion is disregarded in the optimization as the distortion criterion may be overly sensitive to frequency quantization.

4.2. Distortion Predictor

The key component of the *predictor* described in Sec. 3 is a *GMM* for the joint property-distortion pdf, $f_{\mathcal{D},\mathbf{P}}^{(\mathcal{M})}(\delta,\mathbf{p})$, which is to be trained off-line. All GMM's employ 16 mixtures, and the training were conducted using the expectation maximization-algorithm (EM). For GMM training purposed we have extracted a training set, consisting of 180.000 joint property-distortion vectors from the SQAM database (up-sampled to 48 kHz). All test excerpts are disjoint from the training set.

[2] In this context coding templates, referred to in Sec. 2, correspond to a sinusoidal coder using different number of sinusoids.

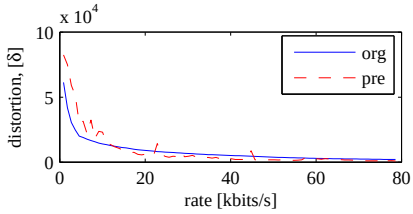

Fig. 3: Original and predicted R-D curves for one segment (35 ms) in the excerpt "glockenspiel".

We have chosen to work with a 4-dimensional *property vector* consisting of: *1) The loudness*, which is calculate as the log of the average energy of the segment, *2) The spectral centroid*, which is calculated as mean of the spectrum with respect to frequency, *3) Spectral bandwidth*, which is calculated as the second moment of the spectrum with respect to frequency, *4) Spectral flatness*, which is calculated as the ratio of the geometric mean and the arithmetic of the power spectrum. We do not claim to have chosen the best property for the task at hand, rather we have chosen to rely on simple (low-complexity) standard audio properties, used in audio classification [8].

5. EXPERIMENTAL RESULTS

We have tested the proposed open loop R-D optimization, for the purpose of R-D optimized bit-allocation (distribution of sinusoids) over optimization viewports, $\mathcal{S}$, c.f. Sec. 2. In the experiments we have exchanged original R-D pairs, c.f. Eq. (6), with predicted R-D pairs. For our particular setup, this means that we have exchanged 86 original R-D pairs, below referred to as a R-D curve, with predicted ones for each segment, as visualized in Fig. 3. Predicted distortion values are only used in the optimization, meaning that presented distortion values are based on original R-D curves. For comparison purposes we have included the performance of a coder with a uniform sinusoidal distribution, i.e. the same number of sinusoids per segment.

In Table 1 we compare the performance of the systems, by averaging the distortion in Eq. (12) over a number of different excerpts. The results show that the proposed system outperforms a uniform sinusoidal distribution for all the excerpts. Naturally, there is a loss compared to the reference system. The gains of optimized systems compared to a system using a fixed number of sinusoids vary. The achievable gain of R-D optimized coding is large for "glockenspiel". The non-stationary character of the signal, results in an R-D optimized bit distribution which is far from uniform, c.f. Fig. 4. The result is a far too high distortion at onsets for the uniform case, c.f. Fig. 5. For the "jazz" excerpt the R-D optimized distribution of sinusoids is not far from uniform, and thus a uniform distribution can compete with the RD optimized, c.f. Table 1. It should be mentioned that the poor performance for the proposed system on the "glockenspiel" excerpt, a 50 % loss, can be traced back to the R-D optimization procedure. Due to the non-convexity of predicted RD-curves, c.f. Fig. 3, the optimization fails in selecting the correct operating point. By simple post processing of predicted R-D curves, smoothing and forcing convexity, the loss can easily be reduced to around 20 %.

An alternative application is up-front coder selection for each optimization viewport, $\mathcal{S}$, i.e. selection of the coder that minimizes the distortion for the current set of segments, $\mathcal{S}$. For this purpose

excerpt	$\mathrm{E}[\delta_{\mathrm{org}}]$	$\Delta\mathrm{E}[\delta_{\mathrm{pre}}]$	$\Delta\mathrm{E}[\delta_{\mathrm{uni}}]$
glockenspiel	$6.55 \cdot 10^2$	50 %	103 %
german speech	$2.42 \cdot 10^4$	3.8 %	9.9 %
castanets	$2.68 \cdot 10^4$	2.7 %	7.6 %
harpsichord	$9.63 \cdot 10^3$	7.6 %	18 %
jazz	$2.79 \cdot 10^4$	3.2 %	4.2 %

Table 1: Average segment distortion, $\mathrm{E}[\delta_{\mathrm{org}}]$, for various excerpts. $\Delta\mathrm{E}[\delta_{\mathrm{sys}}] = \frac{\mathrm{E}[\delta_{\mathrm{sys}}]-\mathrm{E}[\delta_{\mathrm{org}}]}{\mathrm{E}[\delta_{\mathrm{org}}]}$ represents the increase in average distortion compared to an R-D optimized system based on original R-D curves. Here shown for a system using predicted R-D curves(pre), and for a system using uniform bit allocation over the segments (uni).

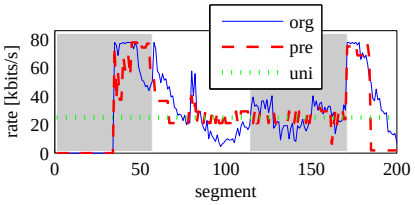

Fig. 4: Bit allocation for the first 200 segments of the excerpt "Glockenspiel". The solid line represents an R-D optimized system based on original R-D curves (org), the dashed line represents an R-D optimized system based on predicted R-D curves (pre), and the dotted line represent a system with uniform bit allocation (uni). The periodically changing shaded and white fields represents optimization viewports.

viewport R-D curves are useful. Viewport R-D curves are achieved by sweeping over $\lambda^\star$, c.f. Eq. (4). In Fig. 6 R-D curves for the first viewport in the "glockenspiel" excerpt are shown. The solid line represents the viewport R-D curve based on original distortion values, the dashed line represents the predicted viewport R-D curve and the dotted line represent the R-D curve for a sinusoidal coder employing a fixed number of sinusoids per segment. Comparing the solid and the dotted curve indicate that we should select the R-D optimized system instead of the fixed system for all rates on this viewport. We can also note that the choice would be the same if we based our decision on the predicted curve, the dashed line, instead of the original. This is obviously a dummy selection, as an optimized system always outperforms a fixed system, but if the dotted curve would have represented for example a waveform coder, such a selection can be of interest. Note that the dashed line represents a prediction of the performance of the R-D optimized system (solid line), and it can therefor indicate a performance better than the performance of the actual system[3].

6. DISCUSSION

In this paper we have studied complexity reduced R-D optimized coding, where R-D curves are exchanged for predicted ones. The proposed framework was applied in a sinusoidal coding context, for the purpose of distributing sinusoids over sets of audio segments. The results show that the proposed framework works, in the sense that the performance is improved compared to a system

[3]Here all figures are based on predicted distortion values, as opposed to above, where predicted distortion values only are used for the optimization, and the results are based on original distortion values.

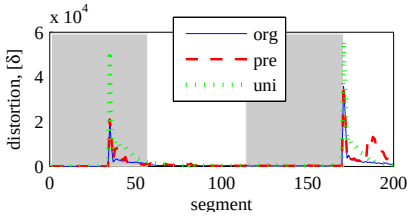

Fig. 5: Distortion distribution for 200 segments of the excerpt "Glockenspiel". The solid line represents an R-D optimized system based on original R-D curves (org), the dashed line represents an R-D optimized system based on predicted R-D curves (pre), and the dotted line represent a system with uniform bit allocation (uni). The periodically changing shaded and white fields represents optimization viewports.

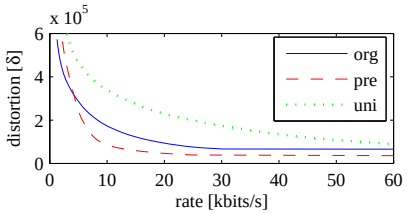

Fig. 6: Viewport R-D curves for the first optimization viewport (1 s) in the excerpt "glockenspiel". The solid line represents an R-D optimized system based on original R-D curves (org), the dashed line represents an R-D optimized system based on predicted R-D curves (pre), and the dotted line represent a system with uniform bit allocation (uni).

with a uniform sinusoidal distribution. It should be noted that we lose compared to an R-D optimized system based on the true R-D curves. This loss can be decreased if our rather raw system is further optimized, meaning a better choice of property vector, and a set of training data better matching the expected audio input.

7. REFERENCES

[1] M. G. Christensen and S. van de Par, "Rate-distortion efficient amplitude modulated sinusoidal audio coding," *to appear in Proc. of the 38th Asilomar Conf. on Sig., Sys., and Comp.*, 2004.

[2] R. Vafin and W. B. Kleijn, "Towards optimal quantization in multi-stage audio coding," in *Proc. ICASSP*, 2004, vol. 4, pp. 205–208.

[3] R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio and speech using psychoacoustical matching pursuits," in *Proc. ICASSP*, 2002, vol. 2, pp. 1809–1812.

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

[5] F. Nordén, S. V. Andersen, S. H. Jensen, and W. B. Kleijn, "Property vector based distortion estimation," *to appear in Proc. of the 38th Asilomar Conf. on Sig., Sys., and Comp.*, 2004.

[6] S. van de Par, S. Kohlrausch, A. Charestan, and R. Heusdens, "A new psychoacoustical masking model for audio coding applications," in *Proc. ICASSP*, 2002, vol. 2, pp. 1805–1808.

[7] R. Heusdens, R. Vafin, and W. B. Kleijn, "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits," *IEEE Signal Processing Letters*, vol. 9, no. 8, pp. 262–265, 2002.

[8] E. Wold *et al.*, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, 1996.