

## Low Complexity Rate-Distortion Optimized Time-Segmentation for Audio Coding

Rødbro, Christoffer A.; Christensen, Mads Græsbøll; Nordén, Fredrik; Jensen, Søren Holdt

*Published in:*

IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005

*Publication date:*  
2005

*Document Version*  
Accepteret manuscript, peer-review version

[Link to publication from Aalborg University](#)

*Citation for published version (APA):*

Rødbro, C. A., Christensen, M. G., Nordén, F., & Jensen, S. H. (2005). Low Complexity Rate-Distortion Optimized Time-Segmentation for Audio Coding. I *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005* (s. 231 - 234). Electrical Engineering/Electronics, Computer, Communications and Information Technology Association.

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

### Take down policy

If you believe that this document breaches copyright please contact us at [vbn@aub.aau.dk](mailto:vbn@aub.aau.dk) providing details, and we will remove access to the work immediately and investigate your claim.

# LOW COMPLEXITY RATE-DISTORTION OPTIMIZED TIME-SEGMENTATION FOR AUDIO CODING

Christoffer A. Rødbro, Mads G. Christensen, Fredrik Nordén, and Søren Holdt Jensen

Aalborg University  
Dept. of Communication Technology  
9220 Aalborg Ø, Denmark  
{car, mgc, fn, shj}@kom.aau.dk

## ABSTRACT

In this paper, we investigate a reduced complexity approach to rate-distortion optimized time-segmentation in audio coding. Instead of the conventional closed-loop approach for determining the coding distortions, they are estimated from a set of features extracted from the audio signal. Care is taken to ensure that properties such as convex and non-increasing rate-distortion curves carries over from the training data to the estimated rate-distortion pairs. With computational complexity reductions of a factor close to 10, perceptual listening tests reveal a slight signal quality reduction, while maintaining a large improvement over fixed segmentation.

## 1. INTRODUCTION

Adaptive time-segmentation has been shown to be an efficient method for improving the rate-distortion trade-off in speech- and audio coding [1], [2], [3]. These methods usually employ an analysis-by-synthesis procedure in which full encoding-decoding operations are required for each and every candidate segment, including those not actually used in the final signal representation. This is necessary in order to determine the distortion and rate if the segment is used in the signal representation. The distortions are obtained by explicitly comparing the encoded-decoded segments to the corresponding original segments, and the optimal segmentation is then found as the one minimizing the total distortion, usually subject to a rate constraint. If the encoding-decoding processes are computationally extensive, as is the case e.g. in the psychoacoustic matching pursuits (PMP) schemes of [4], [5], this may lead to practically infeasible execution times, even for off-line applications such as audio compression. However, [6] proposed a strategy for *estimating*, at low complexity, the distortion arising from coding a signal segment. In [7], this approach was used to predict the optimal distribution of sinusoidal components in a fixed segmentation PMP coder. In this work we shall use a slightly modified approach to estimate the optimal time-segmentation in the same coder.

The rest of this paper is structured as follows: first, rate-distortion optimized time-segmentation is reviewed in Section 2. Next, Section 3 describes how to incorporate the distortion estimation approach of [6] into such a scheme. Objective as well as subjective results are given in Section 4 before Section 5 concludes on the work.

## 2. RATE-DISTORTION OPTIMIZED TIME-SEGMENTATION

The rate-distortion optimized time-segmentation algorithm of [1] is based on the constrained optimization problem:

$$\begin{aligned} \text{minimize : } & D(\tau, \mathbf{c}(\tau)) \\ \text{s.t. : } & R(\tau, \mathbf{c}(\tau)) \leq R_C. \end{aligned} \quad (1)$$

Here,  $\tau = \{s_1, s_2, \dots, s_{\sigma(\tau)}\}$  denotes the time-segmentation consisting of  $\sigma(\tau)$  variable length segments  $s_i$ , each having a length equal to an integer number of *grids* (e.g. 5 ms). The vector  $\mathbf{c}(\tau) = \{c_1(\tau), c_2(\tau), \dots, c_{\sigma(\tau)}(\tau)\}$  denotes the *coding templates*, (i.e. different ways of encoding each segment in a segmentation  $\tau$ ).  $R_C$  is the target bit budget, whereas  $R$  is the total number of bits used and  $D$  is the total distortion, the latter two found by summation over the segments:

$$R(\tau, \mathbf{c}(\tau)) = \sum_{i=1}^{\sigma(\tau)} r(c_i(\tau)) \quad \text{and} \quad D(\tau, \mathbf{c}(\tau)) = \sum_{i=1}^{\sigma(\tau)} d(c_i(\tau)).$$

Here,  $r(c_i(\tau))$  is the number of bits used for encoding segment  $s_i$  using template  $c_i(\tau)$  and  $d(c_i(\tau))$  is some measure of the distortion between the original segment and the one encoded using template  $c_i(\tau)$ . Usually, the constrained optimization problem (1) is solved by recasting it as an unconstrained problem with cost-function:

$$J(\tau, \mathbf{c}(\tau)) = D(\tau, \mathbf{c}(\tau)) + \lambda R(\tau, \mathbf{c}(\tau)). \quad (2)$$

Now, by setting  $\lambda$  to some value (say  $\lambda_x$ ) and minimizing  $J$  over  $\{\tau, \mathbf{c}(\tau)\}$  we will obtain a pair  $(D_x, R_x)$  optimal for  $\lambda_x$ . Thus,  $\lambda$  can be iterated and  $J$  minimized in each step, until a rate  $R \lesssim R_C$  is obtained. In each iteration, the minimization of  $J$  is a two-step procedure: first, the coding templates  $c_i^*(\tau)$  optimal for  $\lambda_x$  are found for each segment:

$$\forall i, \tau : c_i^*(\tau) = \arg \min_{c_i(\tau)} \{d(c_i(\tau)) + \lambda_x r(c_i(\tau))\}. \quad (3)$$

By denoting  $j_i^*(\tau) = d(c_i^*(\tau)) + \lambda r(c_i^*(\tau))$ , the optimal segmentation  $\tau^*$  is the one minimizing the sum over  $j_i^*(\tau)$ :

$$\tau^* = \arg \min_{\tau} \sum_{i=1}^{\sigma(\tau)} j_i^*(\tau). \quad (4)$$

This minimization is carried out at reasonable complexity using a dynamic programming technique, see [1] for details.

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

The computational problems of the procedure described above appears in (3): in order to find the optimal coding templates, we need the distortion if the coding template is used in the signal representation. Thus, we must encode all segments with all coding templates, even for the segments and coding templates not used in the final representation. If we denote the number of grids in the signal by  $G$  and all segment lengths from 1 to  $G$  grids are allowed, the total number of possible segments in the signal equals  $K = \frac{G^2+G}{2}$ . Alternatively, if the maximum segment length is limited at  $L$  with  $G \gg L$ ,  $K \approx GL$ . This is in contrast to the number of segments actually used in the signal representation,  $\sigma(\tau) \leq G$ .

The number and nature of the coding templates depend directly on the type of coder(s) employed. In the rest of this paper, we shall focus on (psychoacoustic) MP-based coders, e.g. [4], [5], [7]. In such a coder, the signal segments are iteratively decomposed into a weighted sum of basis functions. In each iteration, a basis function is chosen from an over-complete dictionary as the one minimizing a perceptual error norm (a distortion). In that the representation of each basis function requires a certain amount of bits, varying the number of components results in different coding templates with corresponding  $\{r(c_i(\tau)), d(c_i(\tau))\}$  pairs. Due to the MP approach, these pairs will lie on a non-increasing (and sometimes also convex) hull. Unfortunately, the PMP algorithm is computationally extensive, primarily because accurate modeling of certain signal segments such as transients requires quite a large number of iterations, (e.g. [7] applied 0-85 sinusoids in each segment). Even with the efficient implementation of [5] requiring 3 FFTs per iteration, this results in up to 255 high-order FFTs for each of the  $K$  frames; clearly, means for reducing this complexity is called for. One way of doing so is based on the observation that running the MP on all  $K$  possible segments is wasteful, because only  $\sigma(\tau)$  of them are used in the final segmentation. This motivates estimating the distortions  $d(c_i(\tau))$  used in (3) instead of calculating them explicitly.

### 3. DISTORTION ESTIMATION

The principle of [6] is to estimate the coding distortion from a vector of features extracted from each candidate audio segment; the computational complexity required to determine these features should be low, or little complexity reduction is gained. Section 3.1 will account for the features explicitly used, but they should be general signal describing ones, such as spectral information, periodicity, stationarity, power, etc. The  $P$  features are stacked in a vector  $\mathbf{p}_i$ ,  $i$  denoting the candidate segment index. Now, the distortions arising if assigning 1, 2, ...,  $C$  components for representing the segment are added to this vector<sup>1</sup>:

$$\mathbf{o}_i = \begin{bmatrix} d_i^{(1)} & d_i^{(2)} & \dots & d_i^{(C)} & \mathbf{p}_i^T \end{bmatrix}^T = \begin{bmatrix} \mathbf{d}_i \\ \mathbf{p}_i \end{bmatrix} \in \mathbb{R}^{C+P} \quad (5)$$

Now, from a set of training data (we used a subset of the SQAM database [8]), a pdf in the form of a multivariate Gaussian mixture is estimated:

$$\mathbf{o}_i \sim \sum_{m=1}^M w_m \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \quad (6)$$

<sup>1</sup> Actually, the vector is built from normalized distortions,  $d_i^{(c)} / \|\mathbf{s}_i\|_2^2$ , with the estimates being rescaled accordingly. This reduces the dynamic range of the distortions and thus eases the statistical modeling to be described in the following.

where  $M$  is the number of mixture components,  $w_m$  denotes the mixture weights ( $\sum_{m=1}^M w_m = 1$ ), and  $\boldsymbol{\mu}_m$ ,  $\boldsymbol{\Sigma}_m$  are the Gaussian mean vectors and covariance matrices, respectively.  $w_m$ ,  $\boldsymbol{\mu}_m$  and  $\boldsymbol{\Sigma}_m$  are found using the expectation maximization algorithm [9], with a model being build for each possible segment length. In the following, we shall drop the subscript  $i$  leaving the frame index implicit.

At this point, we have obtained a pdf in the form of a Gaussian Mixture Model (GMM) describing the features and the distortion arising from coding jointly. The task at hand is: given the features extracted from a segment and the GMM, estimate the vector of distortions. It can be shown that the conditional MMSE estimator is of the form:

$$\hat{\mathbf{d}} = E[\mathbf{d}|\mathbf{p}] = \sum_{m=1}^M \tilde{w}_m \tilde{\boldsymbol{\mu}}_m, \quad (7)$$

where  $0 \leq \tilde{w}_m \leq 1$  depends on  $\mathbf{p}$  (see [6] for details), and

$$\tilde{\boldsymbol{\mu}}_m = \boldsymbol{\mu}_{m,d} + \boldsymbol{\Sigma}_{m,dp} (\boldsymbol{\Sigma}_{m,pp})^{-1} (\mathbf{p} - \boldsymbol{\mu}_{m,p}), \quad (8)$$

with  $\boldsymbol{\mu}_{m,d} \in \mathbb{R}^C$  and  $\boldsymbol{\mu}_{m,p} \in \mathbb{R}^P$  being sub-vectors of  $\boldsymbol{\mu}_m$ ,

$$\boldsymbol{\mu}_m = \begin{bmatrix} \boldsymbol{\mu}_{m,d} \\ \boldsymbol{\mu}_{m,p} \end{bmatrix}, \quad (9)$$

whereas  $\boldsymbol{\Sigma}_{m,dp} \in \mathbb{R}^{C \times P}$  and  $\boldsymbol{\Sigma}_{m,pp} \in \mathbb{R}^{P \times P}$  are sub-matrices of  $\boldsymbol{\Sigma}_m$ ,

$$\boldsymbol{\Sigma}_m = \begin{bmatrix} \boldsymbol{\Sigma}_{m,dd} & \boldsymbol{\Sigma}_{m,dp} \\ \boldsymbol{\Sigma}_{m,pd} & \boldsymbol{\Sigma}_{m,pp} \end{bmatrix}. \quad (10)$$

In some cases, the approach reviewed above leads to a problem reported in [7], in that there is no guarantee that the estimated distortion vector  $\hat{\mathbf{d}}$  will be a non-increasing sequence. This leads to certain problems, for example, the algorithm does not recognize that adding sinusoidal components never leads to increased distortion. However, this problem is easily circumvented by confining the covariance matrices to a diagonal structure, implying that  $\boldsymbol{\Sigma}_{m,dp} = \mathbf{0}$  so that (8) reduces to:

$$\tilde{\boldsymbol{\mu}}_m = \boldsymbol{\mu}_{m,d}. \quad (11)$$

Now, the estimator in (7) is a positively weighted sum of the GMM mean sub-vectors  $\boldsymbol{\mu}_{m,d}$  and thus non-increasing if the individual  $\boldsymbol{\mu}_{m,d}$  are. This is indeed true because in the EM-algorithm, the  $\boldsymbol{\mu}_m$  updates are positively weighted sums of the training vectors [9]. Thus, because the distortion vectors  $\mathbf{d}_i$  extracted for training are non-increasing, so are  $\boldsymbol{\mu}_{m,d}$ , and in turn  $\hat{\mathbf{d}}$ . Also, note that convexity carries over in the same way, which is a coveted property because it prevents ambiguity in the minimization (3).

Also, it should be noted that constraining the covariance matrices to be diagonal has the beneficial side effect of significantly reducing the computational complexity associated with finding  $\tilde{w}_m$  and  $\tilde{\boldsymbol{\mu}}_m$ . Specifically, the main complexity in calculating  $\tilde{w}_m$  stems from evaluating  $M$  Gaussians in the GMM, which has complexity  $\mathcal{O}(MP^2)$  for full covariance matrices, but only  $\mathcal{O}(MP)$  for diagonals. Also for full covariance matrices, determining  $\tilde{\boldsymbol{\mu}}_m$  for all  $m$  using (8) has complexity  $\mathcal{O}(MPD)$ , whereas the diagonal case of (11) is cost free. On the other hand, a somewhat larger number of mixtures  $M$  will be necessary to obtain an adequately precise modeling.

### 3.1. The feature vector

A problem not addressed in the preceding work is how to select which features to include in the vector  $\mathbf{p}$ . For this, a “deflation” strategy is employed, the idea being to start out with a large number of parameters and then sequentially remove one at a time until the estimation performance begins to degrade on a test set. Such a large number of parameters requires many degrees of freedom in the model, and we therefore used  $M = 320$  mixture components. The initial length  $P = 22$  parameter vector contained the parameters listed in Table 1. Note that some of the parameters are redundant, for example both power and log-power are included. However, it is not obvious which of these to use.

1.	Signal power.
2.	Number of zero-crossings.
3.	Loudness (log-power).
4.	A spectral flatness measure.
5.	A spectral centroid measure.
6.	A spectral bandwidth measure.
7.	A LPC flatness measure.
8.	A periodicity measure.
9.-20.	12 mel-cepstrum coefficients.
21.	A power stationarity measure.
22.	A spectral stationarity measure.

Table 1: The features included in the initial feature vector  $\mathbf{p}$ .

An example illustrating the behavior of the deflation strategy is shown in Figure 1. The left-hand plot seems to indicate that no feature is much more important than any other; the estimated distortion MSEs obtained are quite similar. However, since the case where the signal power is removed gives a slightly better overall performance, this parameter is eliminated. Then, in the next iteration, parameter number 3 (log-power) becomes very important, since the information contained in this parameters is no longer redundant with the rest. Also, this plot indicates that the next parameter to be removed from the model should be number 8, the periodicity measure. Using this approach, the features sequentially removed from the parameter vector were the mel-cepstrum coefficients, the signal power, the periodicity measure, and the number of zero-crossings, resulting in a final parameter vector length of  $P = 8$ . It should be noted that different parameters (and coders) could be applied for different segment lengths; doing so, however, is beyond the scope of this paper.

## 4. EXPERIMENTS

In the following, experimental results will be presented with 4 different segment lengths being allowed in the segmentation: 10 ms, 20 ms, 30 ms and 40 ms (including 5 ms overlap). For fixed segmentation, a window update rate of 15 ms was used, corresponding to the 20 ms window in adaptive segmentation. Through informal listening, these windows were found appropriate for the 30 kbps target bit rate used. For further details, see [7].

An example of the optimal and the estimated segmentations is shown in Figure 2 for a section of the SQAM “claves” signal. We see that the estimation captures the onset, whereas the segmentation deviates in the more stationary signal areas. This is a typical behavior that seems sensible, in that adaptive segmentation has its greatest impact in non-stationary signal areas.

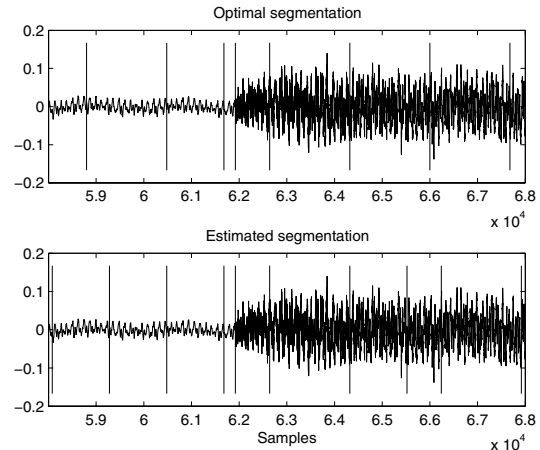


Figure 2: Comparison of the optimal segmentation and that yielded by the proposed method. The vertical lines represent the segmentation.

To access the perceptual degradations (if any) induced by the proposed segmentation approach as compared to optimal segmentation a MUSHRA [10] listening test was carried out. The test set consisted of 6 different audio samples (3 single instrument, 1 solo, 1 orchestra, and 2 pop), none of which were included in the training.

The samples were presented for the proposed method based on estimated distortions (“estimated”), for fixed segmentation (“fixed”), and for rate-distortion optimal segmentation (“optimal”). Moreover, signals low-pass filtered at 3.5 kHz, 7 kHz and 10 kHz were included as anchors 1 to 3, whereas the original was included as a hidden reference (HR). Averaged results for 8 listeners are shown in Figure 3. Typically, the ratio between the total number of segments,  $K$ , and the number of segments actually used,  $\sigma(\tau)$ , lay between 10 and 15.

## 5. CONCLUSION

The scores in Figure 3 indicate that the perceptual quality is slightly degraded for the distortion estimation based approach as compared to optimal segmentation. However, there is still a significant quality gain over fixed segmentation. These results should be compared to computational complexity of the methods. While the optimal segmentation approach requires  $K$  executions of the PMP, the distortion estimation based approach requires only  $\sigma(\tau)$ , with the ratio  $\frac{K}{\sigma(\tau)} > 10$ . On top of this, the distortion estimation approach requires a feature vector extraction as well as the GMM-based estimation procedure described in Section 3 for each of the  $K$  segments. However, the complexity of these steps is low as compared to the +200 FFTs required by the PMP, so realistically the complexity reduction is in the neighborhood of 10. This is supported by the observed Matlab execution times.

## 6. REFERENCES

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

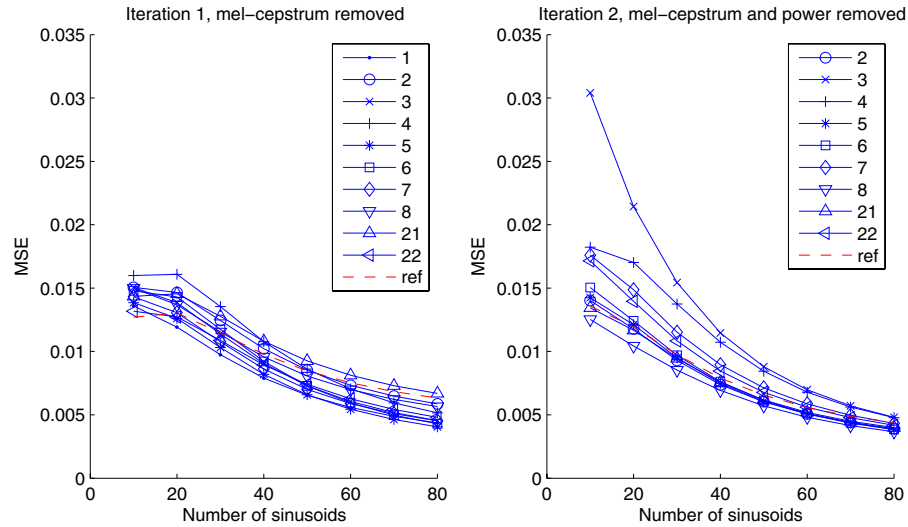


Figure 1: Example illustrating the deflation strategy for a segment length of 30 ms. The legend number refers to which feature has been removed, corresponding to the numbers in Table 1. “ref” represents the MSE when none of the parameters are removed.

2000.

- [2] 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.
- [3] C. A. Rødbro, J. Jensen, and R. Heusdens, “Adaptive time-segmentation for speech coding with limited delay,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 2004, pp. 465–468.
- [4] 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.
- [5] R. Heusdens, R. Vafin, and W. B. Kleijn, “Sinusoidal modeling of audio and speech using psychoacoustic-adaptive matching pursuits,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, pp. 3281–3284.
- [6] F. Nordin, 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.
- [7] F. Nordin, M. G. Christensen, and S. H. Jensen, “Open loop rate-distortion optimized audio coding,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2005, pp. 161–164.
- [8] European Broadcasting Union, *Sound Quality Assessment Material Recordings for Subjective Tests*. EBU, Apr. 1988, Tech. 3253.
- [9] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. Prentice Hall PTR, 2001, ch. 4, pp. 172–175.
- [10] ITU-R BS.1534, ITU Method for subjective assessment of intermediate quality level of coding system, 2001.

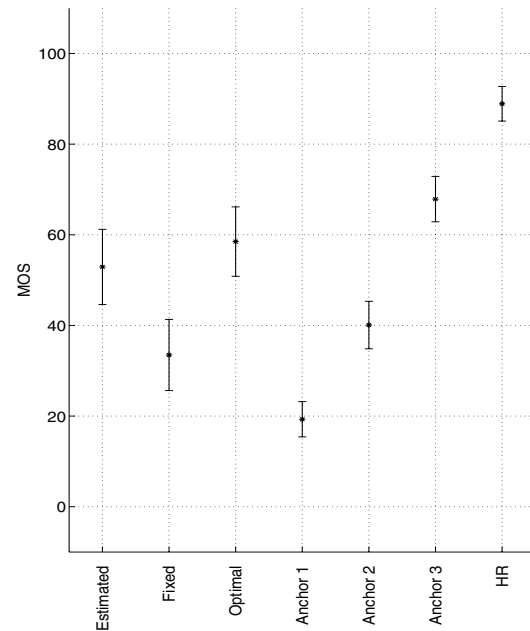


Figure 3: Average scores of MUSHRA listening test. The vertical lines indicate the 95% confidence intervals.