OPTIMAL PARAMETER ESTIMATION FOR MODEL-BASED QUANTIZATION

Alexey Ozerov[*]

Institut TELECOM,
TELECOM ParisTech, CNRS LTCI
37-39 rue Dareau, 75014 Paris, France
alexey.ozerov@telecom-paristech.fr

W. Bastiaan Kleijn[†]

ACCESS Linnaeus Center, Electrical Engineering
KTH - Royal Institute of Technology
Stockholm, Sweden
bastiaan.kleijn@ee.kth.se

ABSTRACT

We address optimal model estimation for model-based vector quantization for both the constrained resolution (CR) and constrained entropy (CE) cases. To this purpose we derive under high-rate (HR) theory assumptions the rate-distortion (RD) relations for these two quantization scenarios assuming a Gaussian model. Based on the RD relations we show that the maximum likelihood (ML) criterion leads to optimal performance for CE quantization, but not for CR quantization. We introduce a new model estimation criterion for CR quantization that is optimal (under HR theory assumptions) in terms of the RD relation. Our experiments confirm that the proposed criterion for model identification outperforms the ML criterion for a range of conditions.

Index Terms— Constrained resolution, model-based quantization, model estimation, rate-distortion relation, high-rate theory.

1. INTRODUCTION

Transmission networks are becoming increasingly heterogeneous. In the area of source coding this has led to the need for scalable quantizers that can adapt to any transmission rate. Recently several practical methods were proposed that facilitate the creation of such scalable quantizers given a source described by a probabilistic model. Solutions for both the *constrained resolution* (CR) and the *constrained entropy* (CE) [1] constraints have been given. For the case of a Gaussian mixture model (GMM) of the source, Subramaniam and Rao [2] introduced a practical CR quantization scheme, while [3] and [4] proposed practical solutions for the CE case. All these methods are based on high-rate (HR) theory approximations [5], but, as experiments show, they give satisfactory results for low rates as well [6]. In audio and speech coding, the techniques were successfully applied to the quantization of the line spectral frequencies (LSF) based on a GMM [2, 4], to direct quantization of the signal based on a GMM [7], and to signal quantization using an adaptive autoregressive (AR) model [8, 6].

To our best knowledge, the optimality of the model estimation criterion in terms of the quantization performance has not been addressed in a general sense. The maximum likelihood (ML) criterion is generally used for this purpose [2, 4, 6, 7, 8].

In this paper we derive the optimal criteria for model parameter estimation, under HR approximations. We first derive the rate-distortion (RD) relations (given a data sequence to quantize) for CE and CR quantizers (as in [2] and [4]) based on a single-Gaussian model that varies over time. Analyzing these RD relations we show that the ML criterion results in optimal performance for the CE case but not for the CR case. For the CE case, the result is consistent with the minimum description length (MDL) principle [9, 10] but the result for the CR case is new. We call the new model estimation criterion for CR quantization *CR-MDL*. Using two coding schemes, our experiments confirm that the CR-MDL criterion outperforms the ML criterion. We analyze the case of a single Gaussian model varying in time, covering both the GMM-based quantization [2, 3] and the AR model-based quantization [6]. However, the framework is general, and can be extended to include other distributions, such as the generalized Gaussian distribution (GGD) , for which the ML criterion is also not optimal in the CR case.

Related to our proposal is the recent work of Duni and Rao [11] that aims to design optimal CR GMM-based quantizers. In contrast, we optimize the probabilistic data model using as criteria the performance of CR and CE quantizers based on the estimated model. Our theory is not restricted to GMM-based systems.

This paper is organized as follows. In section 2 we briefly describe the Gaussian model-based CR and CE quantization schemes considered here. In section 3 we derive the RD relations, we introduce CR-MDL estimation criterion, and we give a practical optimization method for this criterion. The results are given in section 4, and the conclusions are drawn in section 5.

2. MODEL-BASED QUANTIZATION

In this section we describe the principles of Gaussian-model based quantization under HR assumptions in the CR and CE cases (see e.g. [1]). These methods are based on scalar quantization in the mean-removed Karhunen-Loeve transform (KLT) domain [12].

We consider a k-dimensional random Gaussian vector $S = [S_1, \ldots, S_k]^T$ with mean vector μ and covariance matrix Σ, i.e., $S \sim \mathcal{N}(\mu, \Sigma)$. Let the source vector s be a particular realization of the random vector S. Let $\Sigma = U\Lambda U^T$ be the eigenvalue decomposition of the covariance matrix, where U is an orthogonal matrix ($U^T U = I$) and $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_k\}$ is a diagonal matrix of eigenvalues. The linear transform U^T decorrelating the random vector S is the KLT.

Given a fixed budget of R bits per vector, CR scalar quantization minimizing the mean squared error (MSE) distortion of source vector s consists of the following steps (see e.g., [2]):

1. Remove the mean, apply the KLT, normalize for standard deviation $\sqrt{\lambda_i}$, and apply the optimal Gaussian scalar compressor for each dimension i (Eq. (1)), where $\phi(\cdot)$ is the cumula-

[*] A major part of this work was done while A. Ozerov was with the School of Electrical Engineering, KTH, Stockholm, Sweden.

[†] This work was supported in part by the European Union under Grant FP6-2002-IST-C 020023-2 FlexCode.

tive distribution function for a Gaussian random variable with zero mean and unit variance.

2. Quantize u_i with a scalar quantizer $Q^u_{L_i} : u_i \to \hat{u}_i$ uniform on the interval $(0, 1)$ with L_i levels computed using Eq. (2).
3. Reconstruct the quantized source vector $\hat{s}$ (Eq. (3)).

$$x = \Lambda^{-1/2} U^T (s - \mu), \qquad u_i = \phi(x_i/\sqrt{3}) \tag{1}$$

$$\log_2 L_i = R/k + 0.5 \log_2 \left(\lambda_i / \prod_{l=1}^{k} \lambda_l^{1/k} \right) \tag{2}$$

$$\hat{x}_i = \sqrt{3} \cdot \phi^{-1}(\hat{u}_i), \qquad \hat{s} = U \Lambda^{1/2} \hat{x} + \mu \tag{3}$$

For the CE case with MSE distortion, uniform quantization is asymptotically optimal [5]. As in [4], we consider here scalar uniform quantization with a fixed step size Δ in the mean-removed KLT domain, which can be summarized as follows:

1. Remove the mean and apply the KLT (Eq. (4)).
2. Quantize each dimension y_i with a uniform scalar quantizer $Q_\Delta : y_i \to \hat{y}_i$ having a constant step size Δ. Using an arithmetic coder as an entropy coder [4], the effective codeword length l (in bits) is given by Eq. (5), where $N(\cdot; \mu, \lambda)$ denotes the probability density function (pdf) of Gaussian random variable with a mean μ and variance λ.
3. Reconstruct the quantized source vector $\hat{s}$ (Eq. (6)).

$$y = U^T (s - \mu) \tag{4}$$

$$l = -\sum_{i=1}^{k} \log_2 \int_{\hat{y}_i - \Delta/2}^{\hat{y}_i + \Delta/2} N(y_i; 0, \lambda_i) dy_i \tag{5}$$

$$\hat{s} = U \hat{y} + \mu \tag{6}$$

3. RD RELATIONS AND NEW ESTIMATION CRITERION

In this section we first provide the RD relations (Sec. 3.1) that we have already introduced in the partial case of AR model in [6]. Then, we introduce the new model estimation criterion (Sec. 3.2), for which we give a practical optimization method in its general form (Sec. 3.3) and in the partial case of gain estimation (Sec. 3.4).

We consider a sequence of source vectors $\mathbf{s} = \{s^n\}_{n=1}^N$ and we assume that each vector s^n is quantized as described in section 2 using a Gaussian model $\theta_n = \{\mu_n, \Sigma_n\}$. Thus, we have to deal with a sequence of Gaussian models $\theta = \{\theta_n\}_{n=1}^N$, called hereafter *model*, that must be transmitted as well to the decoder as a side information [1]. Note that such a formulation covers both the GMM-based quantization [2, 3], where a mixture index is transmitted as side information, and the AR model-based quantization [6], where quantized model parameters are transmitted.

3.1. Rate-Distortion Relations

To study the optimal model-parameter estimation we first need a practical expression for the RD relation. It can be shown that in both the CR and CE cases and under HR theory assumptions the (average) rate R (in bits per vector) is related to the (average) distortion D (per dimension) [2] as:

$$R = -\frac{k}{2} \log_2 D + \psi(\mathbf{s}, \theta), \tag{7}$$

where the term $\psi(\mathbf{s}, \theta)$ is independent of the rate R and distortion D, and depends only on data $\mathbf{s}$ and model θ.

In the CR case the term $\psi(\mathbf{s}, \theta)$ is [6]:

$$\psi_{\mathrm{CR}}(\mathbf{s}, \theta) = \frac{k}{2} \log_2 \left(\frac{3(2\pi)^{2/3} C}{k} \right) + \\ + \frac{k}{2} \log_2 \frac{1}{N} \sum_{n=1}^{N} \left[\prod_{l=1}^{k} \lambda_{n,l}^{\frac{1}{k}} \sum_{i=1}^{k} \lambda_{n,i}^{-\frac{1}{3}} N(y_i^n; 0, \lambda_{n,i})^{-\frac{2}{3}} \right], \tag{8}$$

where $\Sigma_n = U_n \Lambda_n U_n^T$ is the eigenvalue decomposition of the covariance matrix Σ_n, $C = 1/12$ is the coefficient of quantization of a scalar quantizer, and $y^n = U_n^T (s^n - \mu_n)$ is the mean-removed decorrelated source vector s^n.

In the CE case the term $\psi(\mathbf{s}, \theta)$ can be represented as [6]:

$$\psi_{\mathrm{CE}}(\mathbf{s}, \theta) = \frac{k}{2} \log_2 C - \frac{1}{N} \log_2 \prod_{n=1}^{N} p_{S_n}(s^n | \theta_n), \tag{9}$$

where $p_{S_n}(\cdot | \theta_n)$ is the pdf of random vector S_n modeled by θ_n.

3.2. Proposed Model Estimation Criterion

We see from equation (7) that under HR theory assumptions the RD relation (for both CR and CE cases) relating the rate and the logarithm of distortion is a linear function with slope $-k/2$ and intercept $\psi(\mathbf{s}, \theta)$. Thus, to minimize the distortion D for any (high) rate R, one must look for a model θ minimizing the term $\psi(\mathbf{s}, \theta)$.

The ML criterion, which is usually used for model estimation [2, 4], can be written as:

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} \prod_{n=1}^{N} p_{S_n}(s^n | \theta_n), \tag{10}$$

We see that the ML criterion is equivalent to minimizing the term $\psi_{\mathrm{CE}}(\mathbf{s}, \theta)$ defined by equation (9), that is the MDL principle [10]. However, the ML criterion is, in general, not equivalent to the minimization of the term $\psi_{\mathrm{CR}}(\mathbf{s}, \theta)$ defined by (8). Thus, in the CR case we introduce the following new model estimation criterion, called *CR-MDL*:

$$\theta_{\mathrm{CR\_MDL}} = \arg\min_{\theta} \phi(\mathbf{s}, \theta), \tag{11}$$

where $\phi(\mathbf{s}, \theta)$ is defined as:

$$\phi(\mathbf{s}, \theta) = \log \sum_{n=1}^{N} \left[\prod_{l=1}^{k} \lambda_{n,l}^{\frac{1}{k}} \sum_{i=1}^{k} \exp \left\{ \frac{1}{3} \frac{(y_i^n)^2}{\lambda_{n,i}} \right\} \right], \tag{12}$$

and obtained by simplifying the term $\psi_{\mathrm{CR}}(\mathbf{s}, \theta)$ (removing an additive constant and multiplying by a positive constant) such that minimizing $\phi(\mathbf{s}, \theta)$ is equivalent to minimizing $\psi_{\mathrm{CR}}(\mathbf{s}, \theta)$.

3.3. Practical Implementation by Newton's Method

Unfortunately, even for the single Gaussian model considered and in contrast to the ML criterion (10), the CR-MDL criterion (11) cannot be solved in a closed form. Thus, for practical optimization of criterion (11) we use Newton's method [14], which consists of the application of the iteration:

$$\theta^{m+1} = \theta^m - \gamma \left[H_\theta \phi(\mathbf{s}, \theta^m) \right]^{-1} \nabla_\theta \phi(\mathbf{s}, \theta^m), \tag{13}$$

where θ^m are the model parameters obtained on the m'th iteration, γ is a small positive constant, and $\nabla_\theta \phi(\mathbf{s}, \theta^m)$ and $H_\theta \phi(\mathbf{s}, \theta^m)$ are respectively the gradient (vector of first-order partial derivatives) and the Hessian matrix (matrix of second-order partial derivatives) computed for the model parameters θ^m.

[1] In this work we do not analyze the rate spend for model parameters transmission, since we know from [13] that under HR theory assumptions the optimal rate for model transmission is fixed, i.e., independent of the total rate.

[2] In the CR case the rate R is constant and the distortion D is computed on average, while in the CE case the rate R is computed on average and the mean distortion $D \approx \Delta^2/12$ is the same for each quantization cell [3].

3.4. Partial Case of Gain Estimation

In this section we derive the Newton's method (13) for gain estimation. We consider a Gaussian model $\theta_n = \{\mu_n, \Sigma_n\} \triangleq \{\mu_n, \sigma_n^2 \Sigma_{a_n}\}$, where σ_n is a scalar gain. Assuming that μ_n and Σ_{a_n} are already estimated, our goal is to estimate the gains σ_n $(n = 1, \ldots, N)$ using the CR-MDL criterion (11). Let $\Sigma_{a_n} = U_{a_n} \Lambda_{a_n} U_{a_n}^T$ be the eigenvalue decomposition of Σ_{a_n}. Then the covariance matrix Σ_n is $\Sigma_n = U_{a_n} \sigma_n^2 \Lambda_{a_n} U_{a_n}^T$. It can be shown that for gain estimation the maximization of the global term $\phi(\mathbf{s}, \theta)$ defined by (12) is equivalent to independent maximization over σ_n of the following terms:

$$\phi_{\mu_n, \Sigma_{a_n}}(s^n, \sigma_n^2) = \log \sigma_n^2 + \log \sum_{i=1}^{k} \exp \left\{ \frac{1}{3} \frac{(y_i^n)^2}{\sigma_n^2 \lambda_{a_n,i}} \right\}, \quad (14)$$

where $y^n = U_{a_n}^T (s^n - \mu_n)$ is independent on σ_n^2. The partial (first and second) derivatives of the term $\phi_{\mu_n, \Sigma_{a_n}}(s^n, \sigma_n^2)$ with respect to σ_n, needed for implementation of Newton's method (Sec. 3.3), are:

$$\begin{aligned} \frac{\partial}{\partial \sigma_n} \phi_{\mu_n, \Sigma_{a_n}}(s^n, \sigma_n^2) &= -\frac{2 g_2}{3 \sigma_n^3 g_0} + \frac{2}{\sigma_n}, \\ \frac{\partial^2}{\partial^2 \sigma_n} \phi_{\mu_n, \Sigma_{a_n}}(s^n, \sigma_n^2) &= \frac{2[9 \sigma_n^2 g_0 g_2 + 2 g_0 g_4 - 2 g_2^2]}{9 \sigma_n^6 g_0^2} - \frac{2}{\sigma_n^2}, \end{aligned}$$

with $g_l = \sum_{i=1}^{k} \left(\frac{y_i^n}{\sqrt{\lambda_{a_n,i}}} \right)^l \exp \left\{ \frac{1}{3} \frac{(y_i^n)^2}{\sigma_n^2 \lambda_{a_n,i}} \right\}$ $(l = 0, 2, 4)$.

4. RESULTS

To evaluate the proposed CR-MDL criterion (11) for the case of gain estimation (Sec. 3.4), and to compare it with conventional ML criterion, we consider two speech-coding schemes.

4.1. Coding Schemes

4.1.1. AR model based scheme with KLT

The first scheme is a flexible coding scheme we recently reported in [6]. This scheme is based on the AR model and the KLT (the corresponding optimal adaptive transform). The scheme, referred to as *AR-KLT*, is schematized in figure 1 (A). It is assumed that every k-dimensional signal time block (frame) s^n is described by a order-p AR model consisting of an excitation variance σ_n^2 and of a set of AR model coefficients $a_n = \{a_{n,j}\}_{j=1}^p$. Thus, s^n is modeled as a realization of a multivariate Gaussian distribution $\mathcal{N}(\bar{0}, \Sigma_n)$ with zero mean [3] and the covariance matrix $\Sigma_n = \sigma_n^2 \Sigma_{a_n}$, where Σ_{a_n} is a Toeplitz matrix, having as first column the autocovariance function of a signal generated with the AR model $A_n(z) = 1 + a_{n,1} z^{-1} + \ldots + a_{n,p} z^{-p}$ (see [13] for details).

In [6] the AR model coefficients a_n are estimated from several fixed-length signal blocks (*frames*) in the ML sense, and then interpolated, while the variance σ_n^2 (or gain σ_n) is estimated from only one frame in the ML sense. In this paper we study the only application of the CR-MDL criterion for variance estimation, while the AR model coefficients are always estimated in the ML sense. The estimation of σ_n^2 with CR-MDL criterion (11) is achieved by minimizing the term $\phi_{\bar{0}, \Sigma_{a_n}}(s^n, \sigma_n^2)$ (see (14)). The gain estimation (either in the ML sense or with the CR-MDL criterion) is represented in figure 1 (A) by the "Gain estim." block.

[3] In contrast to [6], we do not consider here the "ringing" (or zero impulse response) subtraction. Rather, we consider the model used in [13].

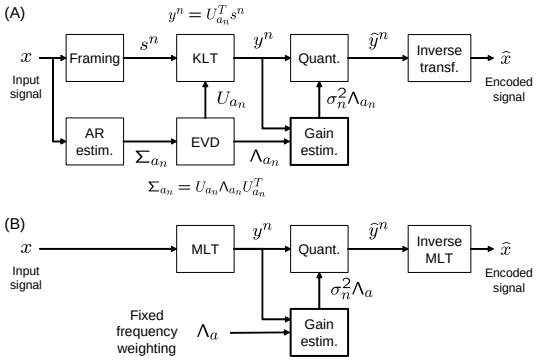

Fig. 1. AR-KLT (A) and MLT-FFW (B) coding schemes.

The transformed vector y^n is ssumed to be distributed as $\mathcal{N}(\bar{0}, \sigma_n^2 \Lambda_{a_n})$ and is quantized (using CR or CE model-based quantizer) as described in section 2. The encoded signal $\hat{x}$ is obtained from the quantized sequence of vectors $\{\hat{y}^n\}_n$ by applying the corresponding inverse transforms (i.e., inverse KLT and frames concatenation).

4.1.2. MLT based scheme with a fixed frequency weighting

The second coding scheme is based on a modulated lapped transform (MLT), which is orthogonal, and a fixed (the same for every frame) weighting in the transformed domain. We call this weighting *Fixed Frequency Weighting* (FFW) and the coding scheme is referred hereafter as *MLT-FFW*. This scheme, schematized in figure 1 (B), is close in spirit to the AR-KLT coding scheme, except that: (i) the orthogonal transform (MLT) is fixed, while in the AR-KLT scheme $U_{a_n}^T$ is adapted to the local signal statistics described by the predictor coefficients a_n, (ii) the transformed coefficients weighting Λ_a (see Fig. 1 (B)) is fixed, while in the AR-KLT the weighting Λ_{a_n} is adapted to the local signal statistics. Thus, we expect that the MLT-FFW scheme is less efficient than AR-KLT, since the statistical model employed is less adapted to the local signal behavior. The detailed description of the MLT-FFW scheme is omitted here, since it is very similar to the description of AR-KLT (see Fig. 1).

4.2. Simulations

For the experimental evaluation we used 10 narrow-band speech signals randomly selected from the TIMIT database evaluation set. For AR-KLT we have chosen the frame length $k = 40$ (5 ms) and the AR model order $p = 10$. For MLT-FFW the MLT was computed with offset $k = 80$ (10 ms).

For both coding schemes we performed simulations in different scenarios, and the results are shown in figure 2 and 3, respectively. For each scenario we plotted the experimental results for a set of rates (circles, triangles or squares) together with the HR theory predicted RD curves given by equation (7) (lines). The following three scenarios were considered:

(i) CR quantization using gain estimated with the ML criterion (10) (circles and dashed line on Fig. 2 and 3),

(ii) CR quantization using gain estimated with the CR-MDL criterion (11) (triangles and solid line on Fig. 2 and 3),

(iii) CE quantization using gain estimated with the ML criterion, which is optimal for the CE case (squares and dotted line).

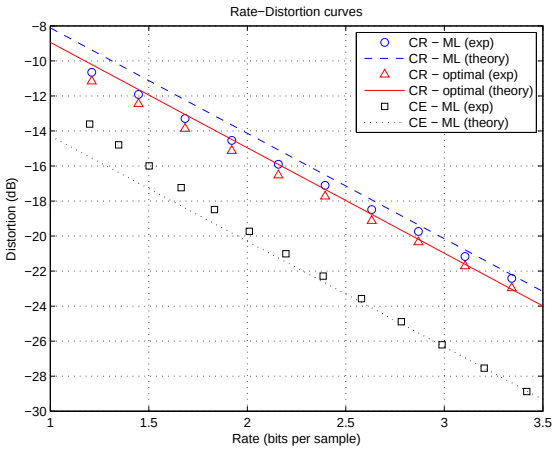

Fig. 2. AR-KLT scheme simulation results.

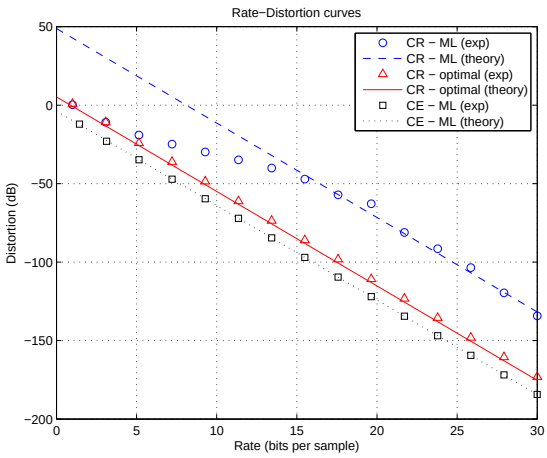

Fig. 3. MLT-FFW scheme simulation results.

We see that for both coding schemes and for all tested scenarios the experimental results approach the theoretical RD relation (7) asymptotically with increasing rate, which confirms the validity of our theoretical results. In the CR case the theoretical distortion improvement obtained with the proposed CR-MDL criterion, as compared to the ML criterion, (diff. between dashed and solid lines) is about 1 dB for the AR-KLT scheme and about 43 dB for the MLT-FFW scheme. In practice this improvement (diff. between circles and triangles) is at least of 0.5 dB for all tested rates for AR-KLT and is about 43 dB starting from the rate of 15 bits per sample for MLT-FFW.

A significant difference exists between improvements obtained using proposed CR-MDL criterion for these two schemes. We observed that while for AR-KLT the distribution of the KLT-transformed and normalized vectors x (see Eq. (1)) is close to Gaussian (due to the Gaussian assumption used for AR model estimation), for MLT-FFW this distribution is close to Laplacian. We conclude that the larger the mismatch between data distribution and model distribution, the greater the improvement due to the proposed optimal criterion, as compared to the ML criterion.

5. CONCLUSION AND FURTHER WORK

Our experimental study performed for two different coding schemes shows that the CR-MDL criterion improves CR quantization performance for both schemes. Moreover, we conclude from experimental evidence that the larger the mismatch between the actual data distribution and model distribution, the greater the performance improvement. Thus, it is particularly worthwhile to use the proposed optimal criterion instead of ML, when mismatch between data and model exists, which is usually the case in the real-world applications.

We have investigated the CR-MDL criterion for variance estimation for the Gaussian model. As for further research, this criterion can be applied for estimation of the remaining parameters in the Gaussian case (i.e., mean vectors and full covariance matrices) and it can be extended to other distributions (e.g., GGD).

6. REFERENCES

[1] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2325–2383, 1998.

[2] A. Subramaniam and B. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 2, pp. 130–142, March 2003.

[3] W. B. Kleijn, "A basis for source coding," Nov. 2006, lecture notes KTH, Stockholm.

[4] D. Zhao, J. Samuelsson, and M. Nilsson, "GMM-based entropy-constrained vector quantization," in *IEEE ICASSP'07*, vol. 4, 15-20 April 2007, pp. 1097–1100.

[5] R. M. Gray, *Source coding theory*. Kluwer Academic Press, 1990.

[6] A. Ozerov and W. B. Kleijn, "Flexible quantization of audio and speech based on the autoregressive model," in *Proc. Asilomar CSSC'07*, Nov. 2007, pp. 535–539.

[7] J. Samuelsson, "Waveform quantization of speech using Gaussian mixture models," in *IEEE ICASSP '04*, vol. 1, May 2004.

[8] M. Li and W. B. Kleijn, "A low-delay audio coder with constrained-entropy quantization," in *Proc. IEEE WASPAA'07*, Nov. 2007, pp. 191–194.

[9] J. Rissanen, "Modeling by the shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[10] A. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. 37, no. 4, pp. 1034–1054, 1991.

[11] E. R. Duni and B. D. Rao, "A high-rate optimal transform coder with Gaussian mixture companders," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 15, no. 3, pp. 770–783, Mar 2007.

[12] J. Y. Huang and P. M. Schultheiss, "Block quantization of correlated gaussian random variables," *IEEE Trans. Commun. Systems*, vol. 11, pp. 289–296, 1963.

[13] W. B. Kleijn and A. Ozerov, "Rate distribution between model and signal," in *Proc. IEEE WASPAA'07*, 2007, pp. 243–246.

[14] M. Avriel, *Nonlinear Programming: Analysis and Methods*. Dover Publications, 2003.