GMM-BASED ENTROPY-CONSTRAINED VECTOR QUANTIZATION

David Y. Zhao[†]*, Jonas Samuelsson*[‡] *and Mattias Nilsson*[†]

[†]KTH (Royal Institute of Technology), School of Electrical Engineering
10044 Stockholm, Sweden.
Email: { david.zhao, mattias.nilsson}@ee.kth.se

ABSTRACT

In this paper, we present a scalable entropy-constrained vector quantizer based on Gaussian mixture models (GMMs), lattice quantization, and arithmetic coding. We assume that the source has a probability density function of a GMM. The scheme is based on a mixture component classifier, the Karhunen Lòeve transform of the component, followed by a lattice quantization. The scalar elements of the quantized vector are entropy coded using a specially designed arithmetic coder. The proposed scheme has a computational complexity that is independent of rate, and quadratic with respect to vector dimension. The design is flexible and allows for adjusting the desired target rate on-the-fly. We evaluated the performance of the proposed scheme on speech-derived source vectors. It was demonstrated that the proposed scheme outperforms a fixed-rate GMM based vector quantizer, and performs closely to the theoretical optimum.

Index Terms— Entropy constrained vector quantizer (ECVQ), lattice, Gaussian mixture model (GMM), arithmetic coding

1. INTRODUCTION

In this work, we consider a practical design of entropy-constrained vector quantizer (ECVQ) using Gaussian mixture models (GMMs), lattice quantization, and arithmetic coding. Existing design and application of ECVQ have been limited to low-rates and low-dimensional vector quantizers (VQ), e.g. [1], due the exponentially increasing computational complexity and memory requirement with rate and dimensionality.

For a high-rate, the optimal quantization point density for an ECVQ is uniform [2]. This well-known result motivates a coding structure consisting of lattice quantization [3, 4] followed by entropy coding using, e.g., an arithmetic coder. Designing a practical entropy code for a high-dimensional lattice quantizer is however a challenging problem. The traditional approach based on ordering the code vectors and keeping track of their corresponding cumulative mass function (CMF) becomes unmanageable for high vector dimensions.

Here we propose an ECVQ scheme based on a parametric description of the source probability density function (PDF) provided by a Gaussian mixture model (GMM). The CMF of the code vectors necessary for the entropy code is computed on-the-fly. Our ECVQ design has a computational complexity that is constant with respect to rate and quadratic with respect to vector dimensionality. Hence, the proposed scheme allows for variable rate vector quantization in rates and dimensions that were not possible in the past.

GMM has been successfully applied to quantization, e.g., [5–8], for fixed-rate vector quantization. Potential approaches for GMM based variable-rate VQ have been discussed in the unpublished works of [9, 10]. In this paper, we propose a practical GMM based ECVQ scheme that allows for a straightforward implementation. We show that the proposed scheme achieves near optimal performance with a low computational complexity.

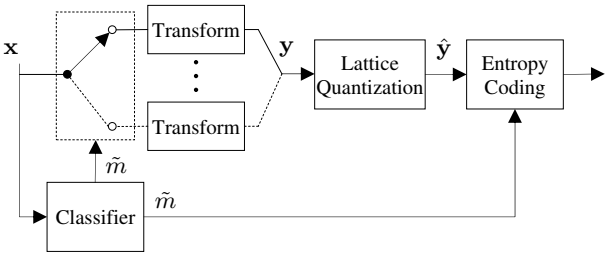

Fig. 1. Schematic diagram of the proposed encoder.

2. PRELIMINARIES

Let $\mathbf{x} = [x_1, \ldots, x_K]^T$ denote a K-dimensional source vector drawn from a sequence of I.I.D. random vectors, and $\hat{\mathbf{x}} = [\hat{x}_1, \ldots, \hat{x}_K]^T$ denote the corresponding vector after quantization. Let $f(\mathbf{x})$ denote the probability density function (PDF) of $\mathbf{x}$, and $p(\hat{\mathbf{x}})$ denote the probability mass function (PMF) of $\hat{\mathbf{x}}$. The quantized vectors are entropy coded and the average rate of the generated bit stream is denoted by $\hat{H}$,

$$\hat{H} = \sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}})\ell(\hat{\mathbf{x}}), \tag{1}$$

where $\ell(\hat{\mathbf{x}})$ denotes the codeword length of $\hat{\mathbf{x}}$. The optimal entropy code has an average rate approaching the entropy of $\hat{\mathbf{x}}$, denoted H_{lower},

$$H_{\text{lower}} = -\sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}}) \log_2 p(\hat{\mathbf{x}}). \tag{2}$$

The distortion measure we considered in this paper is the mean square error (MSE) distortion measure, denoted D,

$$D = \frac{1}{K} \int f(\mathbf{x}) ||\mathbf{x} - \hat{\mathbf{x}}||^2 d\mathbf{x}. \tag{3}$$

The optimal entropy-constrained vector quantizer has a distribution of quantization points that minimizes D, under the constraint that H_{lower} equals a desired target rate R. The optimal quantization point density then fulfills the extended criterion

$$\eta \quad = \quad D + \lambda \left(H_{\text{lower}} - R \right), \tag{4}$$

[‡]J. Samuelsson is with Coding Technologies, Stockholm, Sweden. Email: js@codingtechnologies.com

where λ is the Lagrange multiplier for the entropy constraint.

The optimization problem can be solved under a high-rate assumption, which implies that the data PDF is considered uniform within the boundary of a quantization cell. The optimal quantization point density is shown to be a constant [2]. Hence, uniform quantization points are optimal if the point indices are subsequently coded using an entropy code. This well-known result forms the theoretical basis for using a lattice structured codebook in an ECVQ.

Lattice quantization is an attractive approach for high-dimensional vector quantization due to the low complexity encoding and the minimum memory requirement for the codebook storage. However, designing a practical entropy code for a lattice quantizer has been a challenging problem. The goal of this work is to design a practical low-complexity vector quantizer and an entropy code that perform closely to the theoretically optimal system.

Our ECVQ design is based on the assumption that $f(\mathbf{x})$ is modeled by a GMM with M mixture components,

$$f(\mathbf{x}) = \sum_{m=1}^{M} \rho_m f_m(\mathbf{x}), \tag{5}$$

where m denotes the component index, ρ denote the mixture weights, and $f_m(\mathbf{x})$ are the component Gaussian PDFs with means $\boldsymbol{\mu}_m$ and covariance matrices $\mathbf{C}_m$.

Let $\mathbf{C}_m = \mathbf{V}_m \boldsymbol{\Sigma}_m^2 \mathbf{V}_m^T$ be the eigenvalue decomposition, where $\boldsymbol{\Sigma}_m^2 = \text{diag}[\sigma_1^2, \ldots, \sigma_K^2]$ are the eigenvalues of $\mathbf{C}_m$. $\mathbf{V}_m^T$ is a decorrelating transform, also known as the Karhunen Lòeve transform (KLT), of the mth mixture component.

3. QUANTIZER DESIGN

The proposed scheme consists of a mixture component classifier and per-component quantizers. The quantization is a two step procedure: the input vector $\mathbf{x}$ is first classified into one mixture component with index $\tilde{m}$. Next, $\mathbf{x}$ is quantized to $\hat{\mathbf{x}}$ using the $\tilde{m}$th per-component quantizer. Both the component index $\tilde{m}$ and the quantized vector $\hat{\mathbf{x}}$ are to be entropy coded using an arithmetic coder. A schematic diagram of the proposed scheme is shown in Fig. 1.

The index of the mixture component, $\tilde{m}$, to which the data vector $\mathbf{x}$ belongs to, can be determined using the maximum a posteriori (MAP) classifier,

$$\tilde{m} = \arg\max_m p(m|\mathbf{x}) = \arg\max_m \rho_m f_m(\mathbf{x}). \tag{6}$$

The MAP classifier determines which per-component quantizer to be used for the given $\mathbf{x}$. We will show in Section 4 that using the MAP classifier gives a theoretical performance that is close to the optimum.

The optimal uniform point density motivates usage of lattice structured codebooks in the per-component quantizers. To simplify the design of the arithmetic coder, we transform $\mathbf{x}$ by subtracting the mean, and applying the KLT of the $\tilde{m}$th mixture component,

$$\mathbf{y} = \mathbf{V}_{\tilde{m}}^T (\mathbf{x} - \boldsymbol{\mu}_{\tilde{m}}). \tag{7}$$

The resulting vector $\mathbf{y}$ is a vector of zero-mean and independent Gaussian scalar variables.

In the transformed domain, a lattice structured component codebook, Λ, is generated through a scaled generating matrix $\mathbf{G}$, $\Lambda = \{c\,\mathbf{G}^T\mathbf{u} : \mathbf{u} \in \mathbb{Z}^k\}$, where c is a scaling factor, and $\mathbf{G}$ is the generating matrix, e.g., [4]. The theoretical result from high-rate suggests that c is a constant for a given R, and the same c applies for all component codebooks. Consequently, the same lattice quantizer applies for all mixture components after transformation.

The transformed variable $\mathbf{y}$ is first quantized to the closest code vector in the lattice codebook, $\hat{\mathbf{y}} = c\,\mathbf{G}^T\hat{\mathbf{u}}$, for a particular $\hat{\mathbf{u}}$. Finally, $\hat{\mathbf{y}}$ is entropy coded using an arithmetic code together with the component index $\tilde{m}$. Details regarding the encoding and decoding of the arithmetic code are given in section 5.1.

4. THEORETICAL ANALYSIS

4.1. Distortion analysis

Using a lattice structured codebook, all quantization cells (except those located on the classification boundary) have the same lattice cell shape and volume. Under a high-rate assumption, e.g., [11], D is approximated by

$$D \approx \frac{1}{K}\text{vol}(\hat{\mathbf{x}})^{-\frac{K+2}{K}} \int_{S(\hat{\mathbf{x}})} ||\mathbf{x} - \hat{\mathbf{x}}||^2 d\mathbf{x} \cdot g(c)^{-\frac{2}{K}} = \mathcal{C}_K \cdot g(c)^{-\frac{2}{K}}, \tag{8}$$

where $S(\hat{\mathbf{x}})$ denotes the quantization cell for a code vector $\hat{\mathbf{x}}$, $\text{vol}(\hat{\mathbf{x}})$ denotes the volume of $S(\hat{\mathbf{x}})$, $\mathcal{C}_K$ is the normalized moment of inertia of $S(\hat{\mathbf{x}})$, and the quantization point density, denoted $g(c)$, is a function of the lattice scaling factor c,

$$g(c) = \text{vol}(\hat{\mathbf{x}})^{-1} = |c\mathbf{G}|^{-1}. \tag{9}$$

We note that contributions from quantization cells located on the classification boundaries have been neglected in the analysis above.

4.2. Rate analysis

For a given $\hat{\mathbf{x}}$, the component index $\tilde{m}$ is first entropy coded with average codeword length $-\log_2 \rho_{\tilde{m}}$. The quantized vector $\hat{\mathbf{x}}$ is then entropy coded using the $\tilde{m}$th mixture component with average codeword length $-\log_2 p_{\tilde{m}}(\hat{\mathbf{x}})$. The resulting average rate is then

$$\hat{H} = -\sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}})(\log_2 \rho_{\tilde{m}} + \log_2 p_{\tilde{m}}(\hat{\mathbf{x}})). \tag{10}$$

Under a high-rate assumption, e.g., [11], $\hat{H}$ can be approximated by

$$\hat{H} \approx -\int f(\mathbf{x}) \log_2 \rho_{\tilde{m}} f_{\tilde{m}}(\mathbf{x}) d\mathbf{x} + \log_2 g(c). \tag{11}$$

We note that the rate is lower bounded by the entropy of $\hat{\mathbf{x}}$,

$$\hat{H} \geq -\int f(\mathbf{x}) \log_2 f(\mathbf{x}) d\mathbf{x} + \log_2 g(c) \approx H_{\text{lower}}. \tag{12}$$

Since $\rho_{\tilde{m}} f_{\tilde{m}}(\mathbf{x}) \geq \rho_m f_m(\mathbf{x})$ for all m, the rate is further upper bounded by

$$\hat{H} \leq -\int \sum_{m=1}^{M} \rho_m f_m(\mathbf{x}) \log_2 \rho_m f_m(\mathbf{x}) d\mathbf{x} + \log_2 g(c). \tag{13}$$

The theoretical rate of the proposed scheme lies between the two bounds. The maximum performance loss in theory, denoted by H_{diff}, can be determined by the difference between the bounds. By rearranging the terms, H_{diff} can be written as

$$\begin{aligned} H_{\text{diff}} &= -\int \sum_{m=1}^{M} \rho_m f_m(\mathbf{x}) \log_2 \frac{\rho_m f_m(\mathbf{x})}{f(\mathbf{x})} d\mathbf{x} \\ &= H_M - \sum_{m=1}^{M} \rho_m h(f_m || f), \end{aligned} \tag{14}$$

where $H_M = -\sum_{m=1}^{M} \rho_m \log_2 \rho_m$ is the entropy of the component index, and $h(f_m||f)$ denotes the Kullback-Leibler distance between the mth component PDF and the mixture PDF.

The maximum H_{diff} occurs when all mixture components have the same PDF, then $h(f_m||f)$ equals zero and $\max H_{\text{diff}} = H_M$. H_{diff} decreases with increased separation of the GMM mixture components. If the mixture components are well-separated, such that the Bayesian classification error of the MAP classifier is neglectable, then $h(f_m||f)$ approaches $-\log_2 \rho_m$ and H_{diff} approaches zero. In this case, the proposed scheme approaches the theoretically optimal performance.

4.3. Quantizer optimization

One advantage of the proposed scheme is the flexibility of changing the desired target rate R on-the-fly. This can be done by modifying the lattice scaling factor, c, to obtain different rate-distortion operating points. This allows for quick and seamless adaptation to, e.g., varying communication channel conditions.

Setting the rate $\hat{H}$ to equal R, we get

$$c = \left(|\mathbf{G}| 2^{R + \int f(\mathbf{x}) \log_2 \rho_{\tilde{m}} f_{\tilde{m}}(\mathbf{x}) d\mathbf{x}} \right)^{-\frac{1}{K}}. \tag{15}$$

The integral term in (15) can be solved through numerical integration. Fortunately, the integration term is independent of quantizer rate, and can be evaluated off-line once the GMM is given.

5. IMPLEMENTATION

In this section, we discuss implementational aspects of the proposed scheme, and in particular the entropy coding. For ease of presentation, we present the coding procedure for the Z lattice only, and the scheme can be generalized to use an arbitrary lattice [12].

5.1. Arithmetic coding

The component index $\tilde{m}$ and the quantized vector $\hat{\mathbf{y}}$ are the elements to be entropy coded. In this work, we design a arithmetic coder for this purpose. Our classified scheme simplifies the design of a practical arithmetic coder, even for a high dimensional vector quantizer. After classification and the KLT, $\mathbf{y} = [y_1, \ldots, y_K]$ is a vector of independent Gaussian scalar variables. Instead of encoding the index of the quantized vector, we can, with almost no loss in performance, encode a sequence of scalar indices consisting of the component index and the scalar indices of quantized vector, $[\tilde{m}, \hat{y}_1, \ldots, \hat{y}_K]$. The component index is coded according to the mixture weights, because the signal model (GMM) assumes that the component generation is independent of the subsequent generation of a vector from that component. In the following sections, only arithmetic coding of quantized vectors is discussed.

To simplify the following analysis, we further normalize $\mathbf{y}$ to create unit-variance Gaussian components,

$$\mathbf{y}' = \mathbf{\Sigma}_m^{-1} \mathbf{y}, \tag{16}$$

and the quantized vector is scaled similarly,

$$\hat{\mathbf{y}}' = c\, \mathbf{\Sigma}_m^{-1} \mathbf{G}^T \hat{\mathbf{u}}. \tag{17}$$

5.1.1. Encoding

The Z-lattice has the generating matrix $\mathbf{G} = \mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix. While a Z-lattice based quantizer loses the space-filling advantage of the optimal vector quantizer due to the sub-optimal lattice cell shape, it allows for simple implementation of the arithmetic coder.

The scalar elements of $\hat{\mathbf{y}}'$ are encoded sequentially, and the encoding of the kth element is described here. The kth scalar element, $\hat{y}'_k$, is located on a grid with points at integer multiples of Δ_k, where

$$\Delta_k = \frac{c}{\sigma_k}. \tag{18}$$

The decision interval for the arithmetic code is then determined by the CDF of y'_k evaluated at $\hat{y}'_k - \frac{1}{2}\Delta_k$ and $\hat{y}'_k + \frac{1}{2}\Delta_k$.

Due to the normalization (16), y'_k is Gaussianly distributed with zero-mean and unit-variance for all k. Therefore, the same CDF, $\Phi(\cdot)$, applies for all dimensions,

$$\Phi(y'_k) = \int_{-\infty}^{y'_k} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}z^2} dz = \frac{1}{2}\mathrm{erf}(2^{-\frac{1}{2}} y'_k) + \frac{1}{2}, \tag{19}$$

where $\mathrm{erf}(\cdot)$ is the error function. Thus, the interval $\Phi(\hat{y}'_k - \frac{1}{2}\Delta_k)$ to $\Phi(\hat{y}'_k + \frac{1}{2}\Delta_k)$ is used in the arithmetic encoder.

5.1.2. Decoding

For decoding of dimension k, the $\Phi(\cdot)$ function needs to be inverted to map a point within the decision boundary back to $\hat{y}'_k$. $\Phi(\cdot)$ is a strictly increasing function and an inverse exists. The output of $\Phi^{-1}(\cdot)$ is rounded to the nearest grid point located on the grid given by (18) to obtain $\hat{y}'_k$.

5.2. Lattice truncation

The PMF of each code-vector must be greater than a minimum allowed probability to allow for practical implementation on a finite-precision computer. If a 31 bits integer is used, this threshold is $\delta = 2^{-29}$ [13]. For the kth dimension, we require

$$\frac{1}{2}(\Phi(\hat{y}'_k + \frac{1}{2}\Delta_k) - \Phi(\hat{y}'_k - \frac{1}{2}\Delta_k)) > \delta, \tag{20}$$

for all $\hat{y}'_k$. The additional $\frac{1}{2}$ is due to the rounding of the inverse Φ function in the decoder.

Assuming no numerical error in evaluations of Φ and its inverse, the boundaries for truncation of $\hat{y}'_k$ can be solved to

$$-\sqrt{\log \frac{\Delta_k^2}{8\delta^2\pi}} + \frac{1}{2}\Delta_k \le \hat{y}'_k \le \sqrt{\log \frac{\Delta_k^2}{8\delta^2\pi}} - \frac{1}{2}\Delta_k. \tag{21}$$

If Φ and its inverse are implemented numerically, e.g., through a table lookup, the truncation bound should be adjusted such that (20) is fulfilled for all $\hat{y}'_k$.

6. EXPERIMENTS

In this section, we describe the experimental setup for evaluation of the proposed coding scheme. The experiments were performed in the Matlab environment. The built-in function for $\mathrm{erf}(\cdot)$ and its inverse were used in our implementation.

The source vectors consist of ten-dimensional line spectral frequency (LSF) vectors, extracted from the TIMIT speech database

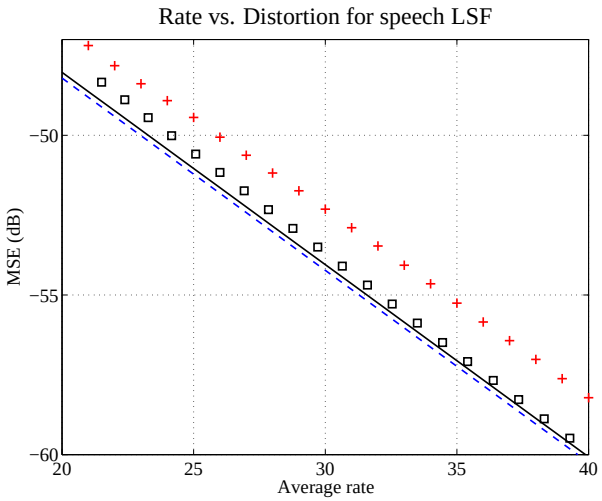

Fig. 2. MSE distortion as a function of rate for the LSF source. The solid line represents the high-rate predicted performance (11, 8) of the proposed scheme, and the dashed line the theoretical optimal performance (12, 8). The squares represent the actual performance of the proposed scheme, and pluses represent the reference system.

using the Adaptive Multi-Rate (AMR) speech coder [14]. The speech signals were downsampled to 8kHz, and unquantized LSF vectors were extracted from the AMR encoder. The complete TIMIT database were used in the experiments. The training set consists of 711867 LSF vectors and the evaluation set consists of 260129 LSF vectors. A GMM with $M = 16$ mixture components with full covariance matrices was optimized over the training set using the expectation-maximization (EM) algorithm [15].

We considered the fixed-rate GMM-VQ of [8], adapted to used the MSE measure, as the reference system. For a fair comparison, the same GMM was used for both the proposed system and the reference system. We evaluated the MSE distortion of quantizers with rates ranging from 20 to 40 bits per vector over the evaluation set.

7. RESULTS AND DISCUSSIONS

Fig. 2 shows the evaluated MSE distortions as functions of rate. Both the experimental results and theoretical predictions are plotted in the figure. The high-rate predicted rate (11) is about 0.3 bit per vector higher than the theoretically optimal rate (12). The actual performance of the proposed scheme is near the theoretical prediction, particularly for rates above 35 bits. We conclude that the proposed quantization scheme achieves near optimal performance at rates useful for practical coders.

Compared to the fixed-rate GMM-VQ system, the proposed scheme performs consistently better. The reduction in distortion for a similar rate is significant, particularly for high rate. For a rate around 30 bits, the proposed scheme has an MSE that is one dB lower below the reference scheme.

In our experiments, we observed that the actual rates (1) are often higher than the theoretically predicted rates (11), and exceed the entropy constraint R. The difference is larger for low rates and decreases with increased rate. Interestingly, this mismatch is smaller for GMMs with less number of components trained over the same training data. This behavior is due to that a LSF-GMM is more likely to contain vanishing dimensions (with near zero eigenvalues) when the number of mixture components increases. When an eigenvalue is small compared to the quantization step size, the assumption of high-rate is no longer valid, and a mismatch between the theoretical and practical rates appears. Therefore, in a practical application, Eq. (15) needs to be modified by locally varying c to compensate this mismatch.

The computational complexity of the proposed scheme is now discussed. Compared to the codebook based scheme [1], our method has a significantly lower complexity. The arithmetic coding based on a sequence of scalar elements has a complexity of order $O(K)$, linear with respect to dimensionality. Using the Z lattice, complexity of the quantization step is also of order $O(K)$. The classification and transformation steps are more complex, and are of order $O(K^2)$ for a GMM with full covariance matrices. If a diagonal-covariance-matrix based GMM is used, this complexity is reduced to $O(K)$. The overall computational complexity of the proposed scheme is therefore no more than order $O(K^2)$, quadratic with respect to dimensionality. Note that the overall complexity is independent of rate.

8. REFERENCES

[1] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 1, pp. 31–42, Jan. 1989.

[2] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inform. Theory*, vol. 25, no. 4, pp. 373–380, Oct. 1979.

[3] J. H. Conway and N. J. A. Sloane, "Fast quantizing and decoding algorithms for lattice quantizers and codes," *IEEE Trans. Inform. Theory*, vol. 28, no. 2, pp. 227–232, Mar. 1982.

[4] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Trans. Inform. Theory*, vol. 48, no. 8, pp. 2201–2214, Aug. 2002.

[5] M. Tasto and P. Wintz, "Image coding by adaptive block quantization," *IEEE Transactions on Communication Technology*, vol. COM-19, no. 6, pp. 957–972, 1971.

[6] P. Hedelin and J. Skoglund, "Vector quantization based on gaussian mixture models," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 4, pp. 385–401, Jul. 2000.

[7] J. Samuelsson and P. Hedelin, "Recursive coding of spectrum parameters," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, pp. 492–503, Jul. 2001.

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

[9] B. D. Rao, "Speech coding using Gaussian mixture models," Technical report, MICRO Project 02-062, 2002-2003.

[10] W. B. Kleijn, "A basis for source coding," Lecture notes, KTH, Stockholm, Jan. 2003.

[11] R. M. Gray, *Source coding theory*, Kluwer Academic Publisher, 1990.

[12] D. Y. Zhao, J. Samuelsson, and M. Nilsson, "Entropy-constrained vector quantization using Gaussian mixture models," Manuscript in preparation, 2006.

[13] E. Bodden, M. Clasen, and J. Kneis, "Arithmetic Coding in revealed - a guided tour from theory to praxis," in *Proseminar Datenkompression 2001*. RWTH Aachen University.

[14] "AMR speech codec; transcoding functions," 3GPP TS 26.090, 1999.

[15] A. P. Dempster, N. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.