VARIABLE-DIMENSION QUANTIZATION OF SINUSOIDAL AMPLITUDES USING GAUSSIAN MIXTURE MODELS

Jonas Lindblom and Per Hedelin

Chalmers University of Technology, School of Electrical Engineering,
SE-412 96 Göteborg, Sweden.
{jonas.lindblom, per.hedelin}@elmagn.chalmers.se

ABSTRACT

In this paper, Gaussian mixture (GM) models are used to design variable-dimension quantizers according to a weighted distortion criterion. A general method for combining a variable-to-fixed dimension transform, with GM modeling and quantization, is proposed. The method provides a convenient and efficient way to encode the amplitudes in a sinusoidal speech coder. Quantizers designed according to the proposed scheme are evaluated both according to weighted distortion criteria, and with respect to a high-rate bound approximation of the distortion. Informal listening tests suggest that the amplitudes can be encoded without subjective loss in a wideband, harmonic coder, at a rate around 40 bits per frame (for the amplitudes only).

1. INTRODUCTION

The dimensionality of the amplitudes and phases in a harmonic sinusoidal speech coder is time-varying. This is because the fundamental frequency f_0 is time varying, and the number of harmonics is inversely proportional to f_0. The dimensionality can be quite large, and varies over a wide range. In a wideband (16 kHz sampling frequency) coder, for example, the dimensionality can easily range from 20 up to above 130. It is therefore non-trivial to vector quantize the parameters in an efficient and convenient manner. An optimal VQ approach would involve having one fixed dimension codebook for each possible dimension [1], which would cost much with respect to storage requirements, but also impose a complex training problem. A very large amount of training data would be required.

A common method to approach the problem with varying dimensionality is to transform the variable-dimension vector with dimension d into a fixed dimension m, prior to coding. It may be the case that m is larger or smaller than d, and the decoder converts the resulting m dimensional vector to the original dimension d. Many different variable-to-fixed dimension transforms have been proposed. In [2], a variable-dimension vector quantization (VDVQ) approach is suggested, where the frequency axis is divided into "frequency bins" and each parameter is mapped to its closest bin to form a fixed-dimension vector. A discrete all-pole model is used in [3] to model the spectral envelope in between the harmonics using a fixed number of parameters. In [4], sample conversion techniques are used to obtain fixed dimensionality. It is shown in [5] that all linear dimension conversion techniques, such as the ones in [2] and [4], but also simple zero-padding and truncation techniques, can be treated as special cases of a general approach for linear dimension conversion, denoted variable sized non-square transform (NST).

This work is supported by the Swedish strategic research program *Personal Computing and Communication* (PCC).

In this work, NST is adopted, but instead of using trained codebooks, a variable dimension quantization scheme, based on Gaussian mixture models (VDGMMQ), is proposed. Quantization based on Gaussian mixture models has received a lot of attention lately, and GM models have for example found use for theoretical evaluation [6] and efficient implementation [7] of vector quantizers.

The paper is organized such that first, a general approach to variable dimension coding based on GM models is presented, then a specific, computationally efficient scheme is proposed for coding of sinusoidal amplitudes in a harmonic speech coder.

2. VARIABLE DIMENSION GM MODELING

In NST, a variable-dimension vector $\mathbf{x}$ with dimension d, is converted to a vector $\mathbf{y}$ with fixed dimension m according to

$$\mathbf{y} = \mathbf{T}^{\mathrm{T}}\mathbf{x}, \tag{1}$$

where $\mathbf{T}$ is a transformation matrix with dimension $d \times m$. For a linear transform with orthonormal basis functions $\mathbf{T}\mathbf{T}^{\mathrm{T}} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix, and $m \geq d$, it holds that

$$\mathbf{x} = \mathbf{T}\mathbf{y}. \tag{2}$$

There are many transformations with orthonormal basis functions available. Examples are the discrete cosine transform, the Karhunen-Loeve transform, the VDVQ approach from [2], and also simple zero-padding (ZP). The VDVQ and ZP transformation matrices (which are employed in this study) are for example given by [5]:

$$\left[\mathbf{T}^{\mathrm{T}}_{\mathrm{VDVQ}}\right]_{i,j} = \begin{cases} 1, & i \leq j\frac{m}{d} < (i+j) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$\left[\mathbf{T}^{\mathrm{T}}_{\mathrm{ZP}}\right]_{i,j} = \begin{cases} 1, & i = j \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The GM-approach proposed in this paper is based on modeling fixed-dimension vectors, $\mathbf{y}$, with Gaussian mixtures

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{i=1}^{M} \rho_{\mathbf{Y}}^{(i)} f_{\mathbf{Y}}^{(i)}(\mathbf{y}), \tag{5}$$

where $\rho_{\mathbf{Y}}^{(i)}, i = 1...M$ are the component weights, summing up to unity, $\sum \rho_{\mathbf{Y}}^{(i)} = 1$. Each density, $f_{\mathbf{Y}}^{(i)}(\mathbf{y})$, is multivariate Gaussian with mean vector $\boldsymbol{\mu}_{\mathbf{Y}}^{(i)}$ and covariance matrix $\mathbf{C}_{\mathbf{Y}}^{(i)}$. Using (2), the parameters of a GM model for $\mathbf{X}$, $f_{\mathbf{X}}(\mathbf{x})$, can now be expressed in terms of the parameters of $f_{\mathbf{Y}}(\mathbf{y})$, since $\boldsymbol{\mu}_{\mathbf{X}}^{(i)} = \mathbf{T}\boldsymbol{\mu}_{\mathbf{Y}}^{(i)}$ and $\mathbf{C}_{\mathbf{X}}^{(i)} = \mathbf{T}\mathbf{C}_{\mathbf{Y}}^{(i)}\mathbf{T}^{\mathrm{T}}$.

3. QUANTIZATION BASED ON THE GM MODEL

In speech coding, a weighted distortion criterion is highly desirable in order to exploit perceptual effects. The weighted distortion between an original variable-dimension vector $\mathbf{x}$, and its quantized counterpart $\tilde{\mathbf{x}}$ is expressed

$$D = (\mathbf{x} - \tilde{\mathbf{x}})^{\mathrm{T}} \mathbf{W} (\mathbf{x} - \tilde{\mathbf{x}}). \quad (6)$$

Assuming that it is possible to decompose $\mathbf{W} = \mathbf{W}^{\frac{1}{2}\,\mathrm{T}} \mathbf{W}^{\frac{1}{2}}$, then

$$D = (\mathbf{x}_w - \tilde{\mathbf{x}}_w)^{\mathrm{T}} (\mathbf{x}_w - \tilde{\mathbf{x}}_w), \quad (7)$$

where $\mathbf{x}_w = \mathbf{W}^{\frac{1}{2}} \mathbf{x}$. Let us for a moment focus on one particular mixture component, i. If $\mathbf{x}$ is Gaussian with mean vector $\boldsymbol{\mu}_{\mathbf{X}}^{(i)}$ and covariance matrix $\mathbf{C}_{\mathbf{X}}^{(i)}$ (denoted $\mathbf{x} \sim N(\boldsymbol{\mu}_{\mathbf{X}}^{(i)}, \mathbf{C}_{\mathbf{X}}^{(i)})$), then

$$\mathbf{x}_w \sim N(\boldsymbol{\mu}_{\mathbf{X}_w}^{(i)}, \mathbf{C}_{\mathbf{X}_w}^{(i)}) = N(\mathbf{W}^{\frac{1}{2}} \boldsymbol{\mu}_{\mathbf{X}}^{(i)}, \mathbf{W}^{\frac{1}{2}} \mathbf{C}_{\mathbf{X}}^{(i)} \mathbf{W}^{\frac{1}{2}\,\mathrm{T}}). \quad (8)$$

The covariance matrix of $\mathbf{x}_w$, $\mathbf{C}_{\mathbf{X}_w}^{(i)}$, is decomposed such that

$$\mathbf{C}_{\mathbf{X}_w}^{(i)} = \mathbf{W}^{\frac{1}{2}} \mathbf{T} \mathbf{C}_{\mathbf{y}}^{(i)} \mathbf{T}^{\mathrm{T}} \mathbf{W}^{\frac{1}{2}\,\mathrm{T}} = \mathbf{V}^{(i)} \boldsymbol{\Lambda}^{(i)} \mathbf{V}^{(i)\mathrm{T}}, \quad (9)$$

where $\mathbf{V}^{(i)}$ is an orthonormal eigenvector matrix, and $\boldsymbol{\Lambda}^{(i)}$ is diagonal with the eigenvalues of $\mathbf{C}_{\mathbf{X}_w}^{(i)}$ on the diagonal. By letting

$$\mathbf{z}^{(i)} = \mathbf{V}^{(i)\mathrm{T}} (\mathbf{x}_w - \boldsymbol{\mu}_{\mathbf{X}_w}^{(i)}) = \mathbf{V}^{(i)\mathrm{T}} \mathbf{W}^{\frac{1}{2}} (\mathbf{x} - \mathbf{T}^{(i)} \boldsymbol{\mu}_{\mathbf{Y}}^{(i)}), \quad (10)$$

a variable $\mathbf{z}^{(i)} \sim N(0, \boldsymbol{\Lambda}^{(i)})$ is created. Furthermore,

$$(\mathbf{z}^{(i)} - \tilde{\mathbf{z}}^{(i)})^{\mathrm{T}} (\mathbf{z}^{(i)} - \tilde{\mathbf{z}}^{(i)}) = \quad (11)$$
$$(\mathbf{V}^{(i)\mathrm{T}} \mathbf{x}_w - \mathbf{V}^{(i)\mathrm{T}} \tilde{\mathbf{x}}_w)^{\mathrm{T}} (\mathbf{V}^{(i)\mathrm{T}} \mathbf{x}_w - \mathbf{V}^{(i)\mathrm{T}} \tilde{\mathbf{x}}_w) = D$$

due to (7) and that $\mathbf{V}^{(i)} \mathbf{V}^{(i)\mathrm{T}} = \mathbf{I}$. Minimizing the l_2-norm in the coordinate system of $\mathbf{Z}^{(i)}$, is hence equivalent to minimizing (6).

3.1. Outline of the quantization scheme

The quantization scheme suggested here operates in the coordinate systems of $\mathbf{Z}^{(i)}$, $i = 1, \ldots, M$, where scalar quantizers and level allocation over both mixture components and dimensions are employed. An incoming variable-dimension vector $\mathbf{x}$ is transformed into all the M $\mathbf{Z}^{(i)}$-coordinate systems, and is quantized with the corresponding quantizers. The general encoding and decoding steps are outlined below. In the next sub-section, a complexity reduced version is presented.

3.1.1. The encoder

Given an incoming variable-dimension vector $\mathbf{x}$, a weighting function $\mathbf{W}$ (the weighting function may be constant, or data dependent), and a transformation matrix $\mathbf{T}$, the encoding steps are:

- Eigenvalue decomposition, cf. (9).
- Form $\mathbf{z}^{(i)} = \mathbf{V}^{(i)\mathrm{T}} \mathbf{W}^{\frac{1}{2}} (\mathbf{x} - \mathbf{T} \boldsymbol{\mu}_{\mathbf{Y}}^{(i)}), \quad i = 1, \ldots, M.$
- Allocate quantization levels. Each mixture i is assigned $l^{(i)} = [\rho^{(i)} L]$ levels, where L is the total number of levels assigned to the current vector. Each dimension is then coarsely assigned $l_k^{(i)}$ levels according to

$$\hat{l}_k^{(i)} = \max \left(\left\lfloor \left(l^{(i)} \right)^{\frac{1}{d}} \sqrt{\boldsymbol{\Lambda}_{k,k}^{(i)} / \overline{\sigma^2}} \right\rfloor, 1 \right), \quad (12)$$

where $\overline{\sigma^2}$ is the geometric mean of the diagonal of $\boldsymbol{\Lambda}^{(i)}$. Based on the coarse assignment, on $\boldsymbol{\Lambda}^{(i)}$, and on tabulated distortions for the $N(0,1)$ variable, a greedy approach in line with [1, p.234] is taken to make adjustments such that $\prod_k l_k^{(i)}$ is close to, but below, $l^{(i)}$.

- Quantize the components of $\mathbf{z}^{(i)}$ using the level allocation, and a tabulated, pdf-optimized quantizer for the $N(0,1)$ variable.
- Select the best candidate component, i^*, and form a joint codeword for i^* and the indices of the scalar quantizers.

3.1.2. The decoder

Given an incoming codeword index for $\mathbf{z}$, a weighting function $\mathbf{W}$, and a transformation matrix $\mathbf{T}$, the decoding steps are:

- Eigenvalue decomposition, cf. (9).
- Perform level allocation as described above.
- Decode $\tilde{\mathbf{z}}$ using the level allocation and the received codeword index.
- Form $\tilde{\mathbf{x}} = \mathbf{W}^{-\frac{1}{2}} \mathbf{V} \tilde{\mathbf{z}} + \mathbf{T} \boldsymbol{\mu}_{\mathbf{Y}}^{(i^*)}$

The mixture index, i^*, does not require separate transmission, since it is implicitly given by the partitioning of the $\mathbf{z}$-codebook.

3.2. Complexity reduction

In the general case described, both the encoder and the decoder need to perform the eigenvalue decomposition (9). Under certain conditions this is not necessary. First of all, $\mathbf{Y}$ can be modeled with diagonal covariances, $\mathbf{C}_{\mathbf{Y}}^{(i)}$. This is common procedure in Gaussian mixture modeling, and it has been noted that correlation within data can still be captured in the model, provided a sufficiently large number of mixtures is used [8]. Moreover, $\mathbf{W}$ may be diagonal. This is the case in the application at hand, but also in many other speech processing applications such as speech spectrum coding [6]. Finally, if $\mathbf{T}$ is the VDVQ (3) or the ZP (4) transformation matrix, $\mathbf{C}_{\mathbf{X}_w}^{(i)}$ is diagonal, such that $\mathbf{V}^{(i)} = \mathbf{I}$ and $\boldsymbol{\Lambda}^{(i)} = \mathbf{W}^{\frac{1}{2}} \mathbf{T} \mathbf{C}_{\mathbf{y}}^{(i)} \mathbf{T}^{\mathrm{T}} \mathbf{W}^{\frac{1}{2}\,\mathrm{T}}$. In these cases, $\mathbf{T}$ can be seen as a "selector" matrix, choosing the appropriate dimensions of $\mathbf{C}_{\mathbf{y}}^{(i)}$ to use, so that when forming $\mathbf{C}_{\mathbf{X}_w}^{(i)}$ from $\mathbf{W}$ and $\mathbf{C}_{\mathbf{Y}}^{(i)}$, matrix multiplication is avoided.

Another way to reduce the computational complexity of the encoder, is to search only the $M' < M$ best candidate components according to $\max_i \left(f_{\mathbf{X}}^{(i)}(\mathbf{x}) \right)$ in the encoding process.

4. PRACTICAL EVALUATION

The variable-dimension coding scheme proposed in this paper was developed for use in a wideband (sampling frequency f_s = 16 kHz) harmonic speech coder called the *Sinusoidal Voice Over Packet Coder* (SVOPC) [9]. In SVOPC, blocks of the linear prediction (LP) residual $r(n)$ are modeled using a harmonic sinusoidal model

$$r(n) = \sum_{k=0}^{d} a_k \sin(2\pi k \frac{f_0}{f_s} n + \phi_k), \qquad n = 0 \ldots N-1, \quad (13)$$

where the parameters are: the fundamental frequency f_0, an amplitude vector $\mathbf{a} = [a_1, \ldots, a_d]^{\mathrm{T}}$ ($a_k > 0$), and a phase vector $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_d]^{\mathrm{T}}$. The number of harmonics, or the dimensionality of $\mathbf{a}$ and $\boldsymbol{\phi}$, is determined by f_0 such that $d = \lfloor f_s / (2 f_0) \rfloor$.

In this study, focus is on coding amplitudes $\mathbf{a}$, from voiced frames, or rather amplitude *shapes* $\mathbf{x} = \mathbf{a}/\sigma$, where $\sigma = \sqrt{\mathbf{a}^{\mathrm{T}} \mathbf{a} / d}$, with the proposed GM based scheme, in a gain-shape VQ approach [1]. The encoding of the other parameters is treated in [9].

The weight matrix, $\mathbf{W}$ (cf. (6)), is similar to the one used in [5]. The diagonal elements are

$$\mathbf{W}_{k,k} = \left| \frac{A(z/\gamma_1)}{A(z)A(z/\gamma_2)} \right|^2_{z=e^{j2\pi k \frac{f_0}{f_s}}}, \qquad (14)$$

where $A(z)$ is the LP prediction filter, $\gamma_1 = 0.9$, and $\gamma_2 = 0.7$.

The TIMIT database [10] was analyzed with the SVOPC encoder, and a training database containing 97 587 variable-dimension amplitude residual vectors with corresponding weight matrices, was created. Similarly, a test database with 6 160 vectors, disjoint from the training database, was extracted.

Coders designed according to the techniques proposed in this paper, were trained, and evaluated with respect to the weighted distortion criteria (6), but also with respect to a *weighted parameter-to-noise ratio* (WPNR) measure

$$\mathrm{WPNR} = 10 \log_{10} \left(\frac{\mathbf{x}^{\mathrm{T}} \mathbf{W} \mathbf{x}}{(\mathbf{x} - \tilde{\mathbf{x}})^{\mathrm{T}} \mathbf{W} (\mathbf{x} - \tilde{\mathbf{x}})} \right), \qquad (15)$$

both averaged over the test database. The ZP transformation was compared to the VDVQ transformation using quantizers based on GM models with 1,8, and 32 mixture components. The GM models were estimated with a version of the Expectation Maximization (EM) algorithm, modified such that only the model parameters corresponding to the non-zero components of a training vector $\mathbf{y}$ are updated (cf. the modified GLA-training in [2] and [5]). According to figure 1, the VDVQ transformation is superior, but the advantage is smaller (less than 1 bit) for higher order GMM quantizers. Therefore, the computationally more efficient zero-padding transform was employed for the remainder of the experiments.

According to the weighted distortion measure (at the rate 50 bits/frame), it is possible to save in the order of 20 bits/frame by employing a 32-mixture quantizer compared to the one-mixture case (which is equivalent to having one scalar quantizer designed for each dimension). According to the weighted PNR measure, the gain is in the order of 10 bits. Informal listening tests indicate that subjectively, the gain is somewhere in between, and that the amplitudes can be encoded transparently (in the sense that the quantization does not introduce audible distortion) in SVOPC using around 40 bits/frame (for the amplitudes only).

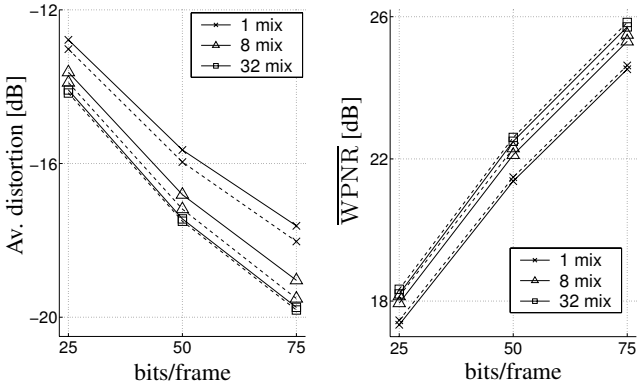

Fig. 1. Average weighted distortion and parameter-to-noise ratio versus the number of bits used to encode each frame with VDGMMQ based on the ZP (solid lines) and the VDVQ (dashed lines) transforms. The weights are normalized such that $\mathrm{trace}(\mathbf{W}) = 1$.

An evaluation of the reduced complexity scheme suggested in section 3.2, can be found in figure 2. The results suggest that there is little to be gained by searching more than the 6-8 most likely quantizers.

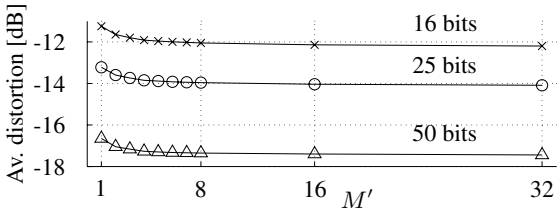

Fig. 2. Average distortion versus the number of searched components M' for the rates 16, 25, and 50 bits/frame.

5. THEORETICAL EVALUATION

Following the work presented in [6], an approximate bound of the high-rate distortion is now derived. Compared to [6], the framework here is quite different. In [6], the weighting matrix $\mathbf{W}$ is dependent on the source vector to be quantized, $\mathbf{x}$. Moreover, the dimensionality d is fixed. Here, the weighting matrix is determined by the LP coefficients, cf. (14), and d varies from 20 up to 132, see figure 3. In order to be able to apply the theory from [6], an

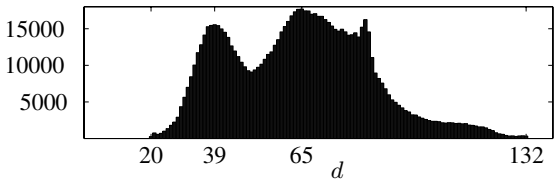

Fig. 3. Histogram illustrating the spread of dimensionality when analyzing the TIMIT database with the SVOPC encoder.

assumption is made that $\mathbf{W}$ depends solely on a auxiliary random variable, $\mathbf{U}$, independent of $\mathbf{X}$. The performance at two common dimensions, $d = 65$ (the most common when processing the whole TIMIT database), and $d = 39$ is studied. Given a dimension d, and an outcome $\mathbf{u}$ of $\mathbf{U}$, the high-rate distortion can be approximately bounded

$$D_{\mathrm{HR}|\mathbf{u}} \gtrapprox \frac{C(d)}{N^{2/d}} \int_{\mathcal{R}^d} f_{\mathbf{X}}(\mathbf{x}) \, |\mathbf{W}(\mathbf{u})|^{\frac{1}{d}} \, \lambda(\mathbf{x})^{-\frac{2}{d}} \mathrm{d}\mathbf{x}, \qquad (16)$$

where $\lambda(\mathbf{x})$ is the point density of the quantizer, N is the number of quantization levels,

$$C(d) = \frac{d}{d+2} V_d^{-\frac{2}{d}}, \text{ where } V_d = \frac{2\pi^{\frac{d}{2}}}{d\Gamma(d/2)}$$

is the volume of a d-dimensional sphere with radius 1. Given a GM model $f_{\mathcal{M}}(\mathbf{x})$ of $\mathbf{X}$, it is possible to express D_{HR} as

$$D_{\mathrm{HR}} \gtrapprox \frac{C(d)}{N^{2/d}} \mathrm{E}_{\mathbf{U}} \left[|\mathbf{W}(\mathbf{u})|^{\frac{1}{d}} \right] \int_{\mathcal{R}^d} f_{\mathbf{X}}(\mathbf{x}) \lambda(\mathbf{x})^{-\frac{2}{d}} \mathrm{d}\mathbf{x}, \qquad (17)$$

where the optimal $\lambda(\mathbf{x}) = f_{\mathcal{M}}^{d/(d+2)}(\mathbf{x}) / \int_{\mathcal{R}^d} f_{\mathcal{M}}^{d/(d+2)}(\mathbf{x}) \mathrm{d}\mathbf{x}$. The integral of (17) can be approximated numerically by means of stochastic integration as described in [6], and the expectation over $\mathbf{U}$ is replaced by an average over a test database. The whole TIMIT database contains 24 265 voiced 65-dimensional vectors, and 17 684 of those were used for training, the rest were saved for evaluation. The corresponding numbers for the 39-dimensional case were 20 163 and 14 697, respectively. For the experiments, $f_{\mathcal{M}}(\mathbf{x})$ is represented by 128-component GM models. 32-mixture VDGMMQs were compared to the bound, but also to quantizers

trained *solely* on the 39 or 65-dimensional data. Results are presented in figures 4 and 5. Compared to quantizers trained on fixed dimensional data, not much is lost (in the order of 1 bit) when comparing to the proposed scheme. Compared to the high-rate bound on the other hand, the difference is in the order of 20 bits.

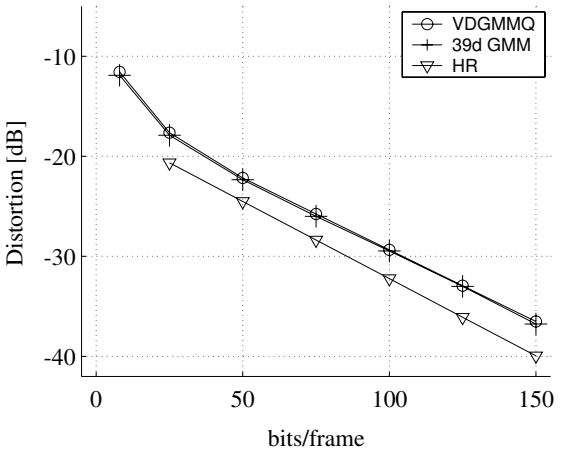

Fig. 4. Evaluation using 39-dimensional vectors from the TIMIT database. 32-mixture quantizers designed with the proposed method are compared to the high-rate bound and to quantizers trained *solely* on 39-dimensional data.

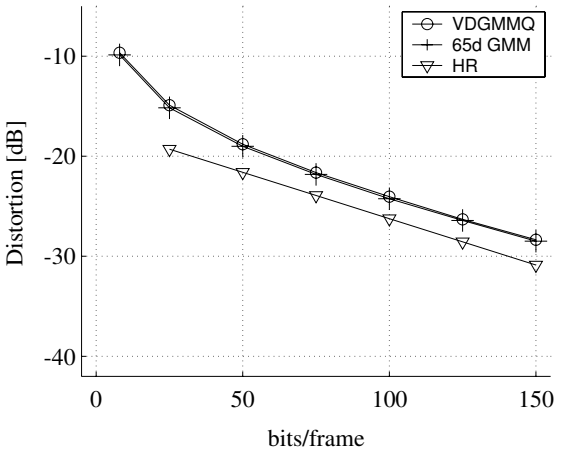

Fig. 5. Evaluation using 65-dimensional vectors from the TIMIT database. 32-mixture quantizers designed with the proposed method are compared to the high-rate bound and to quantizers trained *solely* on 65-dimensional data.

6. DISCUSSION

A general scheme for combining the variable size non-square transform as formulated in [5] with Gaussian mixture modeling and quantization, is proposed. For the ZP or VDVQ transforms, the scheme provides a convenient and computationally efficient way to encode sinusoidal amplitudes in a wide-band harmonic speech coder.

In a fixed-rate scenario, and according to weighted distortion criteria, some 10-20 bits can be saved per frame by employing a 32-component model instead of a one-component model. Informal listening tests indicate that in a sinusoidal coder (SVOPC), the residual amplitudes can be encoded without introducing additional distortion at a rate around 40 bits/frame (for the amplitudes only).

The VDGMMQ scheme is also evaluated in a more theoretical sense. Focusing on two common dimensionalities, it is observed that the performance of VDGMMQ is within one bit from a quantizer designed especially for that particular dimension. This indicates that not much is lost by using only one model for all dimensionalities. An attempt is also made to compare the performance of the VDGMMQ scheme to a high-rate bound approximation of the distortion. The results indicate that, at a high rate, the performance of a 32-mixture VDGMMQ is some 20 bits away from the bound. The high-rate bound approximation is based on the assumption that the source vectors and their corresponding weights are independent, which is of course not completely true. Furthermore, the present implementation of VDGMMQ is based on scalar quantizers. By introducing a union of quantizers (a lattice), it is probably possible to approach the bound approximation (at the expense of increased computational complexity). It is however interesting to note that the distortion-rate curves for the bound and the practical quantizers have the same slope.

It might seem tempting to consider using GM models with full covariance matrices as in e.g. [7]. But with a weighted distortion criterion, it is then necessary to perform an eigenvalue decomposition per coded vector, cf. (9), increasing the computational complexity of the scheme significantly. Furthermore, estimating the full covariance matrices based on the variable dimension data is a non-trivial task.

In the scope of this work, the VDGMMQ scheme is evaluated in a fixed-rate scenario. The scheme is however well suited for variable-rate applications. The level allocation is performed "on-the-fly" on a per-vector basis, based on a bit quota assigned to each vector. Moreover, the complexity of the coder is essentially independent of the rate.

7. REFERENCES

[1] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Boston, 1992.

[2] A. Das *et al.* , "Variable-dimension vector quantization of speech spectra for low-rate vocoders," in *Proc. IEEE Data Compression Conference*, 1994, pp. 421–429.

[3] R. McAulay and T. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science Publishers, 1995, pp. 121–173.

[4] M. Nishiguchi *et al.* , "Vector quantized MBE with simplified V/UV division at 3.0 kb/s," in *Proc. IEEE ICASSP*, 1993, pp. 151–154.

[5] C. Li *et al.* , "Coding of variable dimension speech spectral vectors using weighted nonsquare transform vector quantization," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 6, pp. 622–631, 2001.

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

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

[8] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[9] J. Lindblom, "Coding speech for packet networks," Ph.D. dissertation, Chalmers University of Technology, November 2003, ISBN: 91-7291-374-6.

[10] *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*. NIST, 1990.