WAVEFORM QUANTIZATION OF SPEECH USING GAUSSIAN MIXTURE MODELS

Jonas Samuelsson

Dept. Signals, Sensors and Systems, Wireless@KTH
KTH (Royal Institute of Technology), Stockholm, Sweden

ABSTRACT

Waveform quantization of speech using Gaussian mixture models (GMMs) is proposed. GMMs are trained directly on the speech waveform, and high dimensional vector quantizers (VQs) that efficiently exploit the redundancy are constructed based on the GMM parameters. Two types of GMMs are studied. The complexity of the scheme is independent of the rate, and the rate can be changed without retraining the VQ. A shape-gain structure improves performance and robustness. Pre- and post-processing using spectral amplitude warping further improves perceptual quality. A 32-dimensional VQ operating at 2 bits/sample reproduces speech sampled at 8 kHz with a PESQ score of 4.2.

1. INTRODUCTION

Multi-rate audio codecs have received considerable attention lately. This is motivated by the advent of more flexible networks that allow operators to allocate variable bandwidths depending on user demands, and network conditions. A technique for data compression based on Gaussian mixture models (GMMs), that fits well in a multi-rate environment, is studied in this work. The compression capacity and design flexibility have been demonstrated previously for the speech spectrum source. Here we propose to use this technique directly on the audio waveform.

Direct waveform quantization has been hindered by the exponential growth in complexity in vector quantizers (VQs) as the dimension and the rate increases. The work in [1] spurred the interest for GMMs as a tool for analysis and design of practical VQs. In [2] the competitive performance, and ease of design when constructing source optimized VQs based on GMMs was demonstrated. The designs in [2] were based on random VQs which are highly competitive, although computationally complex. A structured quantizer based on GMMs was suggested in [3] which makes the complexity independent of the rate but at the price of an increase in distortion. In [4] a method that provides a trade-off between these two extremes was presented. For complexity reasons, and for conceptual simplicity, we will in this work employ the technique of companded GMM-VQ (CGMM-VQ) presented in [3].

We design and evaluate CGMM-VQs that operate on vectors of dimensions 32 and 128. The quantizers are evaluated at rates from 1 to 4 bits per sample. GMMs with unconstrained covariance matrices are used in the baseline system. With the initial motivation to increase the dimension of the VQs, we also study GMMs with auto-regressive (AR) covariance matrices. The virtues of AR GMMs is however not improved performance as will be seen. The SNR of practical designs is contrasted with theoretical bounds obtained via high rate theory. Spectral amplitude warping (SAW) and a shape-gain structure are incorporated in the final design which is evaluated in terms of PESQ scores.

2. GMM WAVEFORM QUANTIZATION

In the following, we model speech as a stochastic sequence of samples $\{X(t)\}$ which are blocked into a sequence of K-dimensional vectors $\{\mathbf{X}(n)\}$. In the following the time index is sometimes dropped for ease of notation. The sequence of vectors is assumed to be stationary, and the pdf of a vector is denoted $p(\mathbf{x})$. Next consider a GMM pdf

$$p_{\mathrm{M}}(\mathbf{x}) = \sum_{m=1}^{M} \rho_m p_m(\mathbf{x}), \tag{1}$$

where p_m is a K-dimensional Gaussian pdf with mean $\mathbf{m}_m$ and covariance matrix $\mathbf{C}_m$. Further, $\mathbf{C}_m = \mathbf{V}_m \Lambda_m \mathbf{V}_m{}^{\mathrm{T}}$ is the eigenvalue decomposition, where $\Lambda_m = \mathrm{diag}(\lambda_{m,1}, \ldots \lambda_{m,K})$. The total number of model parameters is $M + MK + MK(K+1)/2$. The model is often optimized towards minimizing the Kullback-Leibler distance $D(p||p_{\mathrm{M}})$.

Given the model, a CGMM-VQ can be designed instantly, and with any size. By designing a VQ for each mixture component pdf, and creating the total VQ by merging the codevectors from each component, it is ensured that the total VQ has a distribution of codevectors which fits the source pdf. The encoding procedure is done in two steps. First the speech vector is quantized by each of the component quantizers. Transform scalar quantizers are employed as in [5], but here with companded scalar quantizers which makes the computational complexity independent of the rate. The M transform quantizers are designed to be optimal for Gaussian sources with means $\{\mathbf{m}_m\}$ and covariance matrices $\{\mathbf{C}_m\}$. In the second step, the best vector

among the output vectors from the transform quantizers is searched for and an index is created. Next we give the details of the above procedure.

2.1. Companded mixture model VQ

We consider a B-bit, resolution constrained CGMM-VQ that encodes each speech vector into an index $i = \epsilon(\mathbf{x})$, $i \in \{1, \ldots, N\}$, $N = 2^B$. The decoder outputs a reconstructed vector $\tilde{\mathbf{x}} = \delta(i)$. In the encoding, each source vector $\mathbf{x}$ is processed in parallel by the component quantizers in three steps,

$$\mathbf{y}_m = \mathbf{\Lambda}_m^{-1/2} {\mathbf{V}_m}^{\mathrm{T}} (\mathbf{x} - \mathbf{m}_m) \quad (2)$$

$$u_{m,k} = \phi(y_{m,k}/c_c) \quad (3)$$

$$i_{m,k} = \epsilon_{m,k}(u_{m,k}), \quad (4)$$

for $m = 1, \ldots, M$, $k = 1, \ldots, K$. Here, $\epsilon_{m,k}$ is a $N_{m,k} = 2^{B_{m,k}}$ level uniform scalar encoder with granular region in the interval (0,1). The range of $\epsilon_{m,k}$ is the index set $\{1, \ldots, N_{m,k}\}$. Furthermore, ϕ is the Gaussian cumulative distribution function for a scalar random variable with unit variance. The operation in (2) transforms the incoming vector (which is assumed to be Gaussian) to a vector of i.i.d. Gaussians with unit variance, (3) implements the optimal scalar compander for a Gaussian random variable (with $c_c = \sqrt{3}$). Next in the encoding, a candidate vector from each component quantizer is reconstructed like

$$\tilde{u}_{m,k} = \delta_{m,k}(i_{m,k}) \quad (5)$$

$$\tilde{y}_{m,k} = c_c \phi^{-1}(\tilde{u}_{m,k}) \quad (6)$$

$$\tilde{\mathbf{x}}_m = \mathbf{V}_m \mathbf{\Lambda}_m^{1/2} \tilde{\mathbf{y}}_m + \mathbf{m}_m, \quad (7)$$

where $\delta_{m,k}$ is the decoder corresponding to $\epsilon_{m,k}$. The winning candidate is $\tilde{\mathbf{x}}_{m_*}$ where $m_* = \mathrm{argmin}_m \; d(\mathbf{x}, \tilde{\mathbf{x}}_m)$ and $d(\mathbf{x}, \tilde{\mathbf{x}})$ is some suitable distortion measure. Here $d(\mathbf{x}, \tilde{\mathbf{x}}) = (\mathbf{x} - \tilde{\mathbf{x}})^{\mathrm{T}} (\mathbf{x} - \tilde{\mathbf{x}})$. An index i_* is output from the encoder,

$$i_* = 1 + \sum_{m=1}^{m_*-1} 2^{B_m} + \sum_{k=1}^{K} (i_{m_*,k} - 1) \prod_{\alpha=1}^{k-1} N_{m_*,\alpha}, \quad (8)$$

and transmitted over the channel. At the receiver, i_* can be uniquely decoded to obtain m_* and $\{i_{m_*,k}\}$.

2.2. Quantizer design

The design of a CGMM-VQ is a two-step procedure. First a parametric model $p_{\mathrm{M}}(\mathbf{x})$, in our case a GMM, is optimized to fit the source pdf $p(\mathbf{x})$. An important design parameter is the number of mixture components M. There is an obvious trade-off between the accuracy of the model on one hand, and computational complexity, and the risk of over-fitting to training data on the other hand. Unfortunately, there are no theoretical guidelines for the choice of M. Here, models with different number of components are trained, and compared in terms of the performance of the resulting VQ, cf Figure 1. The well-known EM-algorithm is used to fit a GMM to the source pdf.

Given the model, we can instantly design a quantizer to operate at any rate. The design amounts a bit allocation, first among the mixture component quantizers, then among the vector dimensions in each quantizer. Following [3], we allocate bits among the component quantizers according to

$$B_m = B + \log_2 \frac{(\rho_m v_m)^{k/k+2}}{\sum_j (\rho_j v_j)^{k/k+2}}, m = 1, \ldots, M, \quad (9)$$

where $v_m = \prod \lambda_{m,k}^{1/K}$. This formula is based on high rate theory and the assumption that the components are well separated. When using the EM-algorithm however, there is no explicit control of the individual components, and the assumption of separated components is not valid. An alternative allocation which gives similar or better performance is to simply assign $B_m = B - \log_2 M, \forall m$. Given $\{B_m\}$ we next allocate bits among the vector dimensions following [5],

$$B_{m,k} = B_m/K + 0.5 \log_2(\lambda_{m,k}/v_m), k = 1, \ldots, K. \quad (10)$$

2.3. Theoretical analysis

With the aid of high rate theory, the performance of the proposed quantizer can be partially assessed. We will lower bound the distortion of a CGMM-VQ by calculating the expected distortion $D = \mathrm{E}[(\mathbf{x} - \tilde{\mathbf{x}})^{\mathrm{T}} (\mathbf{x} - \tilde{\mathbf{x}})]$ for a quantizer with spherical encoding regions and with a VQ point density function $\gamma = p_{\mathrm{M}}^{K/K+2}$

$$D = N^{-2/K} G \int p(\mathbf{x}) p_{\mathrm{M}}(\mathbf{x})^{-\frac{2}{K+2}} \mathrm{d}\mathbf{x}, \quad (11)$$

where G is the normalized second order moment of the spherical encoding region. To calculate an upper bound on D we can assume a random VQ with the same γ which modifies the factor G in (11), cf. [6]. The distortion can be evaluated using stochastic integration

$$D \approx N^{-2/K} G \frac{1}{T} \sum_{n=1}^{T} p_{\mathrm{M}}(\mathbf{x}(n))^{-\frac{2}{K+2}}. \quad (12)$$

Three things are worth pointing out. Firstly, the two bounds bound the distortion of an optimal design (based on the model which may not be perfect), and the bounds assume an optimal encoding; we fulfill neither of these criteria. Secondly, caution must be used when comparing the bounds to practical performance since they are only valid at high rate. Thirdly, and in passing, we mention that the theoretical bounds can be used as a measure of model quality; the distortion is minimized when $D(p||p_{\mathrm{M}}) = 0$. In Figure 1 the two bounds are compared to practice in terms of SNR.

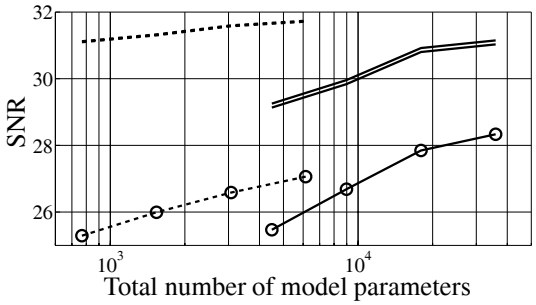

Fig. 1. Solid lines correspond to 32-dimensional full covariance CGMM-VQ, and dashed lines to 128-dimensional AR CGMM-VQ. The lines with circles correspond to practical coders, and the lines without circles are the theoretical bounds from section 2.3. The coders operate on "raw" speech vectors, at a rate of 4 bits/sample.

2.4. Towards a practical coder

The scheme presented so far is a rough and naive attempt at speech compression. In order to improve the performance, and increase the robustness we propose to use a shape-gain VQ structure. We perform the shape-gain encoding open-loop, i.e., the gain $g = \sqrt{\mathbf{x}^{\mathrm{T}}\mathbf{x}/K}$ is quantized to $\tilde{g}$, and the shape $\mathbf{x}/\tilde{g}$ is quantized using the CGMM-VQ. The gain quantizer is a 5 bit uniform scalar quantizer which encodes the gain in a logarithmic domain. Incorporating shape-gain improves PESQ from 3.1 to 3.6 for a 32-dimensional, 64 component CGMM-VQ operating at 2 bits/sample.

To further improve perceptual performance, SAW [7] is incorporated. The analysis window is a 256 sample trapezoidal window with 96 sample slopes on each side, and the overlap and add window is a 64 sample Hann window. The spectrum is compressed using the square root function. When the SAW transform and inverse alone is applied to speech it results in PESQ scores at 4.4 (4.5 is the maximum score). The performance of systems incorporating SAW and shape-gain structure is reported in Table 1. The GMMs are trained on pre-processed and gain-normalized speech.

3. AN ALTERNATIVE MODEL STRUCTURE

The GMMs used above have *full* covariance matrices, i.e., each C_m contains $K(K+1)/2$ unique parameters. Diagonal covariance matrices is one alternative. They do however tend to be less efficient than full covariance matrices when GMMs with the same total number of parameters are compared (especially when the number of parameters increases [8]). Here we propose to use covariance matrices with an AR structure. Each component now represents a *zero-mean*, pth order AR Gaussian process $X(t) = -\sum a_i X(t-i) + U(t)$ where the variance of $U(t)$ is σ^2. If the vector dimension is high, the pdf of a K-dimensional vector is a Gaussian pdf with a $K \times K$ covariance matrix $\mathbf{C}_m = \sigma^2(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$, where $\mathbf{A}$ is a $K \times K$ lower triangular Toeplitz matrix where the first column is $(1, a_1, \ldots, a_p, 0, \ldots, 0)^{\mathrm{T}}$. The AR GMM has the same form as in (1), but with all component mean vectors equal to zero, and with an AR covariance structure. The AR GMM is completely specified by the $M + M(p+1)$ parameters $\{\rho_1, \ldots, \rho_M, \theta_1, \ldots, \theta_M\}$ where $\theta_m = \{\sigma_m, a_{m,1}, \ldots, a_{m,p}\}$.

3.1. Auto-regressive companded mixture model VQ

The encoding procedure for an auto-regressive CGMM-VQ is essentially the same as in section 2.1. Below, we discuss the encoding in mixture component m (the procedure implements the optimal transform coder for a Gaussian vector source with covariance matrix $\mathbf{C}_m$ when K is large). For large K and $p \ll K$, the AR covariance matrices are diagonalized by the Fourier transform, $\mathbf{C}_m = \mathbf{U}^{\#}\mathbf{\Sigma}_m\mathbf{U}$. The elements of the unitary Fourier matrix $\mathbf{U}$ are $U_{k,n} = \frac{1}{\sqrt{K}}\exp(-i2\pi kn/K)$, and $\mathbf{\Sigma}_m$ is a diagonal matrix with the eigenvalues which equal the PSD of the AR process sampled uniformly on the unit circle,

$$\mathbf{\Sigma}_m = \mathrm{diag}(S_m(0), \ldots, S_m(2\pi(K-1)/K)), \qquad (13)$$

$S_m(\omega) = \sigma_m^2 \;/\; \left|\sum_{n=0}^{p} a_{m,n}\exp(-i\omega n)\right|^2$ $(a_{m,0} = 1)$ [9, Ch. 15.9]. Thus, the decorrelation in step (2) in the encoding is replaced by $\mathbf{y} = \mathbf{U}\mathbf{x} = \mathrm{DFT}(\mathbf{x})/\sqrt{K}$. This produces a vector with K complex elements $\{a_k + ib_k\}$, $k = 0, \ldots, K-1$ (dropping the mixture component index for ease of notation). Assuming a Gaussian input with covariance matrix $\mathbf{C}_m$, the variances of the real and imaginary parts are $\mathrm{var}(a_0) = S_m(0)$, $\mathrm{var}(b_0) = 0$, $\mathrm{var}(a_{K/2}) = S_m(\pi)$, $\mathrm{var}(b_{K/2}) = 0$, and $\mathrm{var}(a_k) = \mathrm{var}(b_k) = S_m(2\pi k/K)$, $k \neq 0, K/2$. Further, by the symmetry of the Fourier transform, $a_k = a_{K-k}$, $b_k = -b_{K-k}$, $k \neq 0, K/2$. Thus there are only K unique, real values to quantize, $\mathbf{y}' = (a_0, a_1, \ldots, a_{K/2}, b_1, b_2, \ldots, b_{K/2-1})$. The elements of $\mathbf{y}'$ are normalized to have unit variance yielding $\mathbf{y}'_m$. The remaining steps of the encoding follow those in section 2.1: the elements of $\mathbf{y}'_m$ are companded yielding $\mathbf{u}'_m$ which is quantized by a set of scalar quantizers. The decoder is modified accordingly. Note that the decorrelation operation is common to all mixture components which can result in lower computational complexity. The number of multiplications in the encoding procedure for an AR CGMM-VQ is on the order of $K\log_2 K + MK$, and for full covariance CGMM-VQ, $MK^2 + MK$.

3.2. Design of an AR GMM VQ

For the more general case of a hidden Markov model (HMM) using AR GMMs as state pdfs, the segmental K-means algorithm has proven useful in optimizing the parameters to the speech source. Here, since we have a single state HMM, the segmental K-means essentially reduces to the K-means algorithm using the Itakura-Saito distortion [10].

Given the model, the design of an AR CGMM-VQ follows the procedure in section 2.2. Assuming the mixture components are well separated it can be shown that the optimal allocation between the mixture components follows (9) but with $\lambda_{m,k} = S_m(2\pi(k-1)/K)$, $k = 1, \ldots, K$. Furthermore, for mixture component m, the scalar quantizers that encode (the normalized and companded versions of) $\{a_k\}_{k=0}^{K/2}$ and $\{b_k\}_{k=1}^{K/2-1}$ are allocated $B_m/K + 0.5\log_2(S_m(2\pi k/K)/v_m)$ bits each.

4. EXPERIMENTS

For the experiments, a training set and an evaluation set (with no speaker appearing in both sets) were compiled from the TIMIT database. The speech was resampled at 8 kHz. The training set contains 78 minutes of speech, and the evaluation set consists of 10 sentences (34 s in total), spoken by five male, and five female speakers. The GMMs were trained on 70 000 vectors of speech from the training set. The factor c_c in the encoding and decoding was experimentally tuned to maximize either SNR or PESQ for each model at rate 2 (the same factor was used at the other rates). The performance is measured in SNR $= 10\log_{10}\sigma_X^2/D$ (σ_X^2 and D estimated in simulations), and the PESQ measure.

In all experiments, the dimension of the full covariance GMMs is 32. Attempts to increase the dimension lead to numerical instabilities in training and when using the models. For the AR GMMs different dimensions were tested, and a vector dimension of 128 gave the best results and is used throughout. In Figure 1 the SNR is plotted as a function of the total number of model parameters for systems operating on "raw" speech vectors. It quantifies the suboptimality of the design, and encoding, cf. section 2.3. AR models have the highest potential as indicated by the theoretical curves, but full covariance models perform better in practice. Also note that the performance of a random VQ and a VQ using optimal encoding regions is almost indistinguishable according to theory. In Table 1 PESQ scores for CGMM-VQs operating on SAW pre-processed and gain-normalized vectors are reported for systems using different number of mixture components. At a rate of 2 bits/sample (which includes the overhead for gain quantization), full covariance CGMM-VQ obtains a score of 4.2, and AR CGMM-VQ obtains 4.1. The correlation structures that are possible to model using the AR structure are limited, which explains the slightly worse performance in that case. The computational complexity, and the number of model parameters are lower however. A low number of parameters is important in adaptive schemes were the model may be transmitted together with the compressed data. Finally, a comparison was made with the Enhanced Full Rate (EFR) codec operating at 12.2 kbits/s which resulted in PESQ scores of 4.1 for both systems. Future work includes refinement and tuning of the present system, generalization to HMMs to exploit inter-frame dependency, tests on audio signals (here a model can be trained, e.g., on a piece of music and included with the compressed data; AR GMMs could be especially useful).

Rate	# full components				# AR components			
	8	16	32	64	64	128	256	512
1	3.2	3.4	3.4	3.5	2.9	3.0	3.0	3.2
2	3.9	4.0	4.1	4.2	3.9	4.0	4.0	4.1
3	4.3	4.3	4.4	4.4	4.3	4.3	4.3	4.3

Table 1. PESQ scores for SAW-shape-gain CGMM-VQ using full covariance matrices (left), and AR covariance matrices (right). Speech is sampled at 8 kHz. Rate in bits/sample.

5. CONCLUSIONS

We have presented novel techniques for waveform compression of speech based on Gaussian mixture models. The high dimensionality of the VQs makes it possible to efficiently exploit the redundancy in speech, yielding objective results comparable to those of more mature contemporary speech coders. High rate theory suggests that the design and encoding procedure can be further improved.

6. REFERENCES

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

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

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

[4] T.Z. Shabestary, P. Hedelin, "Spectral quantization by companding," in *Proc. ICASSP*, 2002, pp. 641–644.

[5] J.J.Y. Huang, P.M. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Comm. Syst.*, vol. CS-11, pp. 289–296, 1963.

[6] J.H. Conway, N.J.A. Sloane, "Voronoi regions of lattices, second moments of polytopes, and quantization," *IEEE Trans. Inform. Theory*, vol. IT-28, 1982.

[7] R. Lefebvre, C. Laflamme, "Spectral amplitude warping (SAW) for noise spectrum shaping in audio coding," in *Proc. ICASSP*, 1997, pp. 335–338.

[8] J. Lindblom, J. Samuelsson, "Bounded support Gaussian mixture modeling of speech spectra," *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 88–99, 2003.

[9] S.M. Kay, *Fundamentals of Statistical Signal Processing, Estimation Theory*, Prentice Hall, 1993.

[10] A. Buzo, A. H. Gray, R. M. Gray, J. D. Markel, "Speech coding based upon vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, no. 5, pp. 562–574, 1980.