COMPANDED LATTICE VQ FOR EFFICIENT PARAMETRIC LPC QUANTIZATION

Marie Oger[†], *Stéphane Ragot*[*], *and Roch Lefebvre*[†]

[†] University of Sherbrooke, Dept. of Electrical Engineering, Sherbrooke, QC, J1K 2R1 Canada
[*] France Télécom R&D/DIH/IPS, Av. Pierre Marzin, 22307 Lannion Cedex, France
E-mail: {marie.oger,roch.lefebvre}@usherbrooke.ca, stephane.ragot@francetelecom.com

ABSTRACT

Source coding based on Gaussian Mixture Models (GMM) has been recently proposed for LPC quantization. We address in this paper the related problem of designing efficient codebooks for Gaussian vector sources. A new technique of ellipsoidal lattice vector quantization (VQ) is described, based on 1) scalar companding optimized for Gaussian random variables and 2) rectangular lattice codebooks with fast trellis-based nearest neighbor search. The Barnes-Wall lattice Λ_{16} in dimension 16 is applied to quantize the line spectrum frequencies (LSF) of wideband speech signals. The LSF are computed in a manner similar to the AMR-WB speech coding algorithm. The performance of memoryless and predictive LSF quantization for different GMM orders (4, 8 and 16) is evaluated at 36 and 46 bits per frame. The companded lattice VQ is shown to perform better than its scalar counterpart, with similar complexity.

1. INTRODUCTION

A parametric approach based on Gaussian mixture models (GMM) has been recently developed for the vector quantization (VQ) of linear-predictive coding (LPC) parameters [1, 2]. Although the coding performance is limited *a priori* by the accuracy of the underlying p.d.f. source model, this approach has some interesting features, such as asymptotic bit-rate savings [1], bit-rate scalability and complexity independent of bit rate [2].

In this paper, we address a problem related to GMM-based VQ: the design of efficient codebooks to represent GMM components, i.e. encode Gaussian vector sources. The main contribution of the paper is the development of a new technique of companded lattice VQ to improve the performance of parametric LPC quantization. We address the specific case of 16^{th} order LSF quantization for a 16 KHz sampled signal. This choice has several motivations. First, multistage LPC quantization, for example as in AMR-WB [3], may receive different numbers of bits. In AMR-WB, 36 or 46 bits are allocated to predictive two-stage LSF quantization depending on the coder bit rate. Even if the first stage codebooks are shared, several sets of tables have to be stored for the second stage and for different bit rates. The bit-rate scalability of GMM-based VQ may be exploited to reduce storage requirements. Furthermore, most results on parametric LPC quantization [1, 2] deal with narrowband speech coding. Yet, the performance/complexity advantage of GMM-based VQ over split/multistage VQ should be more apparent in the wideband case, where high LPC orders and bit allocations are used.

This paper is organized as follows. The LPC coding method of [2] is reviewed in Section 2. A new technique of companded lattice VQ is presented in Section 3 for mean-removed KLT coding of Gaussian components. A specific greedy bit allocation algorithm is also described in this section. In Section 4, the performance/complexity of memoryless and predictive LSF quantization at 36 and 46 bits per frame is evaluated for different GMM orders (4, 8 and 16). The potential advantage of GMM-based LPC quantization is also discussed over other quantizer structures. The conclusions are drawn in Section 5.

2. REVIEW: GMM-BASED LSF QUANTIZATION

The line spectrum frequencies (LSF) provide an efficient LPC representation for quantization purposes [4]. The p.d.f. of LSF vectors $\mathbf{x}$ in dimension n can be modeled [2] by a Gaussian mixture model of order M given by

$$f(\mathbf{x}|\Theta) = \sum_{i=1}^{M} \alpha_i f_i(\mathbf{x}|\theta_i),$$

where

$$f_i(\mathbf{x}|\theta_i) = \frac{1}{\sqrt{(2\pi)^n det(\Sigma_i)}}\; e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i)},$$

with the following constraints: $\alpha_i > 0$ and $\sum_{i=1}^{M} \alpha_i = 1$. The dimension n is usually set to 10 for narrowband speech co, 16 in the wideband case. The set of GMM parameters is given by

$$\Theta = \{\alpha_1, \cdots, \alpha_M, \mu_1, \cdots, \mu_M, \Sigma_1, \cdots, \Sigma_M\},$$

where α_i, μ_i and Σ_i are respectively the weight (a priori probability), the mean vector and the covariance matrix of the i-th GMM component. For a given source database, Θ is usually estimated using the E-M algorithm [5].

2.1 GMM-based VQ by mean-removed KLT coding

The memoryless GMM-based VQ of [2] is illustrated in Figure 1. For an input LSF vector $\mathbf{x}$, the quantized LSF vector $\hat{\mathbf{x}}$ is selected among M candidates $\widehat{\mathbf{x}}^{(i)}$, with $i = 1, \cdots, M$, by minimizing a distortion criterion:

$$\hat{\mathbf{x}} = \hat{\mathbf{x}}^{(j)} \text{ where } j = \arg\min_{i=1,\cdots,M} d(\mathbf{x}, \widehat{\mathbf{x}}^{(i)}).$$

The candidate $\hat{\mathbf{x}}^{(i)}$ is the representative of $\mathbf{x}$ in the i-th GMM component (or class). With this point of view, the computation of $\hat{\mathbf{x}}$ can be interpreted as closed-loop classified VQ. The selection criterion d is the log-spectral distortion (LSD) in [2] – a simple weighted Euclidean distance may also be used [6]. The candidates $\hat{\mathbf{x}}^{(i)}$ are computed in [2] by mean-removed Karhunen-Loeve transform (KLT) coding, which is known to be optimal for the quantization of correlated Gaussian sources [7]. This encoding procedure can be easily extended to the case of predictive LSF quantization [2].

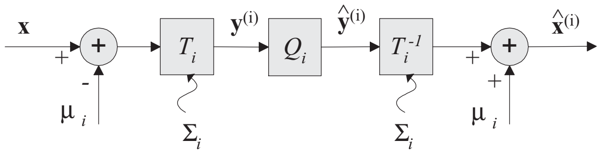


Figure 1: Mean-removed KLT coding of the i-th GMM component.

This work was supported by the NSERC and VoiceAge Corporation. Stéphane Ragot was with Univ. Sherbrooke when this work was done.

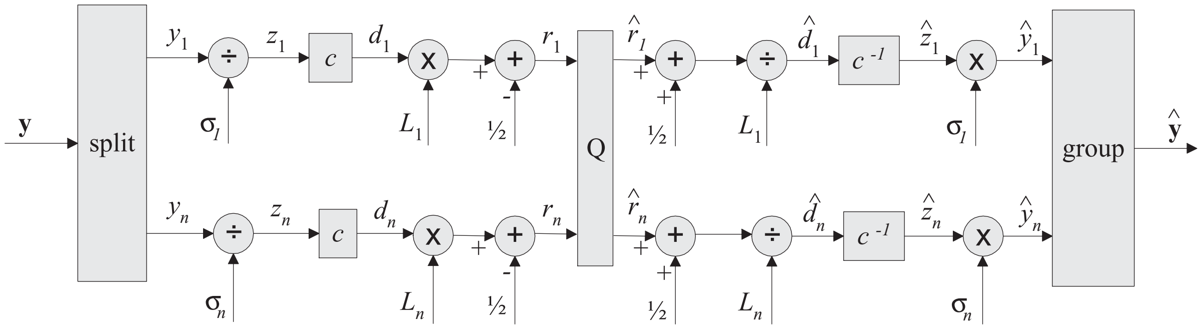


Figure 2: Coding of uncorrelated Gaussian components by companded VQ.

Based on the parameters μ_i and Σ_i of the i-th GMM component, the candidate $\widehat{\mathbf{x}}^{(i)}$ is given by [2]

$$\widehat{\mathbf{x}}^{(i)} = T_i^{-1}\, Q_i\,(T_i(\mathbf{x}-\mu_i)) + \mu_i$$

where T_i is the KLT matrix which decorrelates the i-th GMM component and $Q_i(.)$ is a quantization method optimized for the uncorrelated i-th GMM component. The computation of T_i and the optimization of $Q_i(.)$ require to compute the eigenvalue decomposition of the covariance matrix Σ_i as

$$\Sigma_i = Q_i\ diag(\sigma_{i1}^2, \cdots, \sigma_{in}^2)\ Q_i^T$$

where $\sigma_{i1}^2 \geq \cdots \geq \sigma_{in}^2$ are the eigenvalues of Σ_i (they are assumed ordered without loss of generality) and the matrix Q_i comprises the eigenvectors of Σ_i.

The eigenvalues $\sigma_{i1}^2, \cdots, \sigma_{in}^2$ $(i = 1,..,M)$ and the mixture weights $\alpha_1, \cdots, \alpha_M$ are used in [2] to distribute bits between classes. For a bit budget R_{tot} per vector, the number of bits R_i allocated to the i-th GMM component is constrained so that $2^{R_1} + \cdots + 2^{R_M} \leq 2^{R_{tot}}$. An analytical solution for R_i can be derived assuming high bit rates and good "separation" between classes [2]. An equal bit distribution $(R_1 = \cdots = R_M)$ may also be used [6].

2.2 Coding of uncorrelated Gaussian components by bit-rate-scalable companded scalar quantization

With the method of [2], the design of GMM-based VQ is somehow reduced to the problem of encoding zero-mean uncorrelated Gaussian components. The non-uniform scalar quantization of [2] is illustrated in Figure 2, in the case of a zero-mean Gaussian source $\mathbf{y} = (y_1, \cdots, y_n)$ of covariance matrix $diag(\sigma_1^2, \cdots, \sigma_n^2)$. This technique has two interesting features: bit rate scalability and a complexity independent of bit rate.

In Figure 2, the elements y_i are normalized by σ_i, so as to obtain an i.i.d. zero-mean unit-variance Gaussian source $\mathbf{z} = (z_1, \cdots, z_n)$. An "optimal" scalar compressor $c(.)$ is then applied to z_i. Under the high-rate assumption the optimal compressor for a unit-variance Gaussian random variable u is given by [8]

$$c(u) = \frac{1}{2}(1 + erf(u/\sqrt{6})),$$

where erf is the error function

$$erf(u) = \frac{2}{\sqrt{\pi}} \int_0^u e^{-t^2}\ dt.$$

The inverse operation is given by [8]

$$c^{-1}(u) = \sqrt{6}\ erf^{-1}(2u-1).$$

The source $\mathbf{z} \in [-\infty, +\infty]^n$ is thus mapped into a source $\mathbf{d} \in [0,1]^n$ with $\mathbf{d} = (c(z_1), \cdots, c(z_n))$. Scalar quantization in $[0,1]^n$ is then applied to $\mathbf{d}$. If d_k is quantized with $L_k \geq 1$ scalar levels, the reconstruction $\hat{d}_k$ is given by [9]:

$$\hat{d}_k = ([d_k L_k - \frac{1}{2}] + \frac{1}{2})/L_k$$

where [.] denotes the rounding to the nearest integer.

The scalar quantization described above can be interpreted as follows. We define a codebook $\mathscr{C}(\mathbb{Z}^n)$ as :

$$\mathscr{C}(\mathbb{Z}^n) = \mathbb{Z}^n \cap \mathscr{R}$$

where the region $\mathscr{R}$ of $\mathbb{R}^n$ is given by

$$\mathscr{R} = [0, L_1 - 1] \times \cdots [0, L_n - 1].$$

Then, the reconstruction $\hat{\mathbf{d}} = (\hat{d}_1, \cdots, \hat{d}_n)$ can be written as: $\hat{\mathbf{d}} = ((\hat{r}_1 + \frac{1}{2})/L_1, \cdots, (\hat{r}_n + \frac{1}{2})/L_n)$ where $\hat{\mathbf{r}} = (\hat{r}_1, \cdots, \hat{r}_n)$ is a codevector in $\mathscr{C}(\mathbb{Z}^n)$. This interpretation opens the door to performance improvements by using a "good" lattice instead of $\mathbb{Z}^n$.

3. A TECHNIQUE OF COMPANDED LATTICE VQ

3.1 Preliminaries: binary lattices and error-correcting codes

A *lattice* Λ in $\mathbb{R}^n$ $(n \geq 1)$ is a set of discrete points defined by:

$$\Lambda = \left\{ \sum_{i=1}^{n} \zeta_i \mathbf{v}_i | (\zeta_1, \cdots, \zeta_n) \in \mathbb{Z}^n \right\},$$

where $\mathbf{v}_1, \cdots, \mathbf{v}_n$ are linear independent basis vectors. Two simple examples of lattices, $\mathbf{Z}^n$ and D_n, are illustrated for $n = 2$ in Figure 3, where $\mathbf{v}_1 = (1,0)$, $\mathbf{v}_2 = (0,1)$ for $\mathbb{Z}^2$ and $\mathbf{v}_1 = (2,0)$, $\mathbf{v}_2 = (1,1)$ for D_2. The family of lattices D_n is defined by:

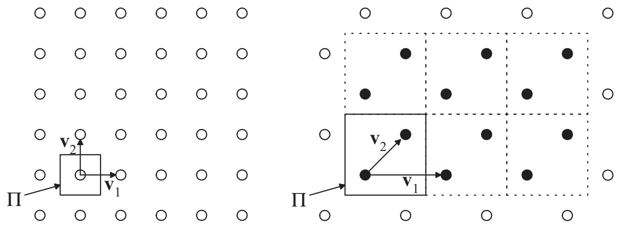
$$D_n = \{(u_1, \cdots, u_n) \in \mathbb{Z}^n | u_1 + \cdots + u_n \text{ even}\}.$$

In this work we will restrict ourselves to *binary lattices*, which are extensively studied in [10]. Binary lattices are connected to block error-correcting codes. For instance, D_n may also be defined as:

$$\begin{aligned} D_n &= 2\mathbb{Z}^n + [n, n-1, 2] \\ &= \{2\mathbf{u} + \mathbf{c} | \mathbf{u} \in \mathbb{Z}^n, \mathbf{c} \in [n, n-1, 2]\}. \end{aligned}$$

where $[n, n-1, 2]$ is the binary parity-check code of length n and Hamming distance 2, having 2^{n-1} codewords. In general, a binary lattice Λ may be decomposed as:

$$\Lambda = p\mathbb{Z}^n + \Pi = \{p\,\mathbf{u} + \mathbf{c} | \mathbf{u} \in \mathbb{Z}^n, \mathbf{c} \in \Pi\}.$$



(a) $\mathbb{Z}^2$ (b) D_2

Figure 3: Examples of 2-D lattices: $\mathbb{Z}^2$ and D_2 – with basic parrallepiped Π.

where $p = 2^m$ (with m integer ≥ 0) and Π are respectively the periodicity and basic parallepiped of Λ. This point of view is illustrated in Figure 3 for $\mathbb{Z}^2$ and D_2. It can be checked that $p = 1$, $\Pi = \{(0,0)\}$ for $\mathbb{Z}^2$ and $p = 2$, $\Pi = [2,1,2] = \{(0,0),(1,1)\}$ for D_2.

In practice, we will use here the so-called Barnes-Wall lattice Λ_{16} in dimension 16 to encode wideband LSF vectors. This binary lattice is defined as [10]:

$$\Lambda_{16} = 4\mathbb{Z}^{16} + 2[16,11,4] + [16,1,16] \quad (1)$$

where $[16,11,4]$ is a 2nd-order Reed-Muller code and $[16,1,16]$ is the binary repetition code of length 16. From Eq. 1, we find that Λ_{16} has a periodicity of 4 and a basic parralepiped $\Pi = 2[16,11,4] + [16,1,16]$ comprising $2^{11+1} = 4096$ points.

3.2 Companded lattice VQ

We define here a fixed-rate codebook from a binary lattice Λ in $\mathbb{R}^n$ as:

$$\mathscr{C}(\Lambda) = \Lambda \cap \mathscr{R}$$

where

$$\mathscr{R} = [0, pm_1 - 1] \times \cdots [0, pm_n - 1], \quad (2)$$

p is the periodicity of Λ and m_k is an integer ≥ 1 ($k = 1, \cdots, n$). Since the region $\mathscr{R}$ is chosen as a parallelipedic region of space, we obtain a rectangular lattice codebook. The sides of $\mathscr{R}$ are restricted to have a length multiple of the lattice periodicity p to simplify the indexing and search algorithms in $\mathscr{C}(\Lambda)$. Following [11], we will refer to m_k as a *multiplicity factor*. An example of codebook $\mathscr{C}(\Lambda)$ is illustrated in Figure 3 (b) for $\Lambda = D_2$, $m_1 = 3$ and $m_2 = 2$. In this figure, the 12 codevectors in $\mathscr{C}(D_2)$ appear as '•' instead of '∘'.

The companded scalar quantization method described in Section 2.2 and Figure 2 can then be extended as follows. Rectangular lattice VQ is applied to the vector $\mathbf{d}$ instead of scalar quantization. The number of quantization levels is given by $L_k = pm_k$ for $k = 1, \cdots, n$. The reconstruction $\hat{d}_k$ is then

$$\hat{d}_k = (\hat{r}_k + \frac{1}{2})/(pm_k),$$

where $\hat{\mathbf{r}} = (\hat{r}_1, \cdots, \hat{r}_n)$ is the nearest neighbor of $\mathbf{r} = (pd_1m_1 - \frac{1}{2}, \cdots, pd_nm_n - \frac{1}{2})$ in $\mathscr{C}(\Lambda)$.

3.3 Indexing $\mathscr{C}(\Lambda)$ and nearest-neighbor search in $\mathscr{C}(\Lambda)$

The problems of indexing $\mathscr{C}(\Lambda)$ and finding the nearest neighbor in $\mathscr{C}(\Lambda)$ are solved in [10], in the case $m_k = 1$, $k = 1, \cdots, n$. The algorithms of [10] can be readily extended to the general case $m_k \geq 1$, $k = 1, \cdots, n$. The indexing of $\hat{\mathbf{r}} \in \mathscr{C}(\Lambda)$ can be split into the computation of two sub-indices, as follows [11]:

- Find $\mathbf{u} = (u_1, \cdots, u_n) \in \mathbf{Z}^n$ with $0 \leq u_k < m_k$ ($k = 1, \cdots, n$) such that $\hat{\mathbf{r}} \in \Pi + p\mathbf{u}$.
- Compute the sub-index of $\mathbf{u}$ using $\lceil \log_2 \prod_{k=1}^{n} m_k \rceil$ bits.
- Compute the sub-index of $(\hat{\mathbf{r}} - p\mathbf{u}) \in \Pi$ using the error-correcting codes defining Π. (In this work, we employ $\Lambda = \Lambda_{16}$. The sub-index of $\hat{\mathbf{r}} - p\mathbf{u}$ is therefore represented with 12 bits.)

An *optimal search* procedure in $\mathscr{C}(\Lambda)$ is described in [10] based on a trellis description and *coset code* construction of the binary lattice Λ – it boils down to computing the branch metrics and parsing the trellis of Λ with the Viterbi algorithm. The metrics are computed here using the mean-square error criterion. An important property of this search procedure is that the *overload* in $\mathscr{C}(\Lambda)$ is *implicit* when computing the metrics and parsing the trellis. Note that the trellis of Λ_{16} has 4 sections and 16 states [12, p. 1769]. The trellis-based search in $\mathscr{C}(\Lambda_{16})$ has a higher complexity than rounding in $\mathbb{Z}^n$, yet the increase in complexity is very limited.

3.4 Optimization of the multiplicity factors

A modified version is provided here to allocate the multiplicity factors m_k specifying the region $\mathscr{R}$ of Eq. 2 and to encode the source $\mathbf{y}$ described in Section 2.2. Given a bit budget R and the covariance matrix $diag(\sigma_1^2, \cdots, \sigma_n^2)$ of $\mathbf{y}$, the allocation procedure consists of the following steps:

1. Reserve bits to index Π : $R' := R - R_\Pi$ (for instance, $R_\Pi = 0$ for $\mathbb{Z}^n$, $n-1$ for D_n, 12 for Λ_{16}). Initialize: $m_k = 1$, $k = 1, \cdots, n$.
2. While $\prod_{k=1}^{n} m_k \leq 2^{R'}$, $m_j := m_j + 1$ where $j = \arg\max_{k=1,\cdots,n} (\sigma_k/m_k)^2$
3. While $\prod_{k=1}^{n} m_k \leq 2^{R'}$, $m_j := m_j + 1$ where j tests all positions from 1 to n sorted according to $(\sigma_k/m_k)^2$, $k = 1, \cdots, n$.

Note that in this work this algorithm is also used to allocate the number of quantization levels L_k to companded scalar quantization since $m_k = L_k$ for $\Lambda = \mathbb{Z}^n$.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental setup

The database used for this work is the NTT-AT wideband speech material (sampled at 16 kHz) which is multilingual, multi-speaker and lasts 5 hours – this material is stored on four CDs. The down-sampling to 12.8 kHz and linear-predictive analysis of AMR-WB [3] is used to extract LSF vectors of dimension 16. Note that silence frames are discarded. Three CDs are selected to build a training database comprising 607,386 LSF vectors, while the other CD is used to generate a test database of 202,112 LSF vectors. The E-M algorithm [5] is applied to the training database to estimate the GMM parameters for an order M =4, 8 and 16 (with full covariance matrices, and means initialized by the generalized Lloyd-Max algorithm). The LSF vectors in the test database are quantized with the GMM-based method of [2]. Memoryless and AR(1) predictive GMM-based VQ are tested. In the AR(1) predictive case, the GMM parameters are trained on the (open-loop) prediction residual. The AR(1) prediction matrix, which is constrained to be diagonal, is estimated in open-loop assuming a perfect reconstruction. The GMM components are quantized by mean-removed KLT coding, using $\mathscr{C}(\mathbb{Z}^{16})$ or $\mathscr{C}(\Lambda_{16})$. The bit allocation to GMM components (or classes) is done according to [2]. The number of quantization levels L_k and multiplicity factors m_k are optimized with the modified greedy algorithm described in this paper.

4.2 Spectral distortion statistics

The performance of LSF quantization is evaluated with the well-known spectral distortion (SD) [4]. The SD statistics obtained for memoryless and AR(1) predictive cases can be found in Tables 1 and 2, respectively. Two bit rates (R_{tot} =36 and 46 bits) and different GMM orders (M =4, 8 and 16) are tested. The histograms of SD are also provided in Figures 4 and 5 for R_{tot} =46 bits and $M = 16$ only.

The results show that the companded lattice VQ based on Λ_{16} improves the performance compared to companded scalar quantization. The gain in average SD is small (around 0.05–0.08 dB) in all cases, and the amount of outliers is typically reduced by 30–50 %. The rectangular lattice VQ developed here provides no shaping gain over scalar quantization, only a granular gain – the granular gain of Λ_{16} over $\mathbb{Z}^{16}$ is around 0.86 dB. This small granular gain in the LSF domain has a limited impact in terms of SD. The bit allocation R_{tot} has no influence on the performance gap between $\Lambda = \mathbb{Z}_{16}$ and Λ_{16} – the average allocation to LSF is indeed 2.25 and 2.88 bits per sample at 36 and 46 bits, respectively.

The performance improves with the GMM order M. The complexity (computation load, storage) increases linearly with M. A trade-off has to be found to be competitive with existing LPC quantization techniques.

The histogram of SD in the memoryless case is bimodal. In fact, the SD is different depending on which GMM component is quantized: the i-th conditional histogram of SD computed with the input LSF vectors coded in the i-th GMM component, $i = 1, \cdots, M$ has specific mean and shape. The results imply that the bit allocation to GMM components developed in [2] may be improved – the underlying assumptions (e.g. clear separation of GMM components) are not always valid, at least in the memoryless case.

Table 1: Results for memoryless GMM-based LSF quantization.

(a) Results at $R_{tot} = 36$ bits per frame.

Λ	M	avg. SD (dB)	$SD \geq 2$ dB (%)	$SD > 4$ dB (%)
$\mathbb{Z}^{16}$	4	1.38	14.00	0.0262
	8	1.28	9.38	0.0067
	16	1.23	6.72	0.0035
Λ_{16}	4	1.31	10.71	0.0341
	8	1.24	7.53	0.0044
	16	1.19	5.57	0.0045

(b) Results at $R_{tot} = 46$ bits per frame.

Λ	M	avg. SD (dB)	$SD \geq 2$ dB (%)	$SD > 4$ dB (%)
$\mathbb{Z}^{16}$	4	0.93	0.95	0.0034
	8	0.86	0.53	0.0005
	16	0.83	0.29	0.0005
Λ_{16}	4	0.85	0.54	0.0019
	8	0.81	0.27	0.0005
	16	0.77	0.13	0

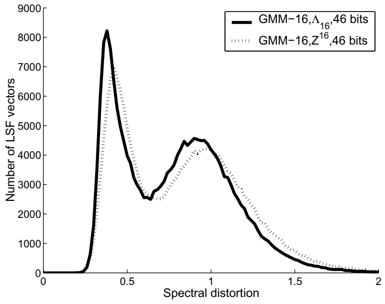


Figure 4: Histograms of SD (memoryless case, 46 bits).

5. CONCLUSIONS

We presented a new technique of companded lattice VQ which extends the scalar quantization of [2]. This generalization keeps the advantages of bit rate scalability and complexity independent of bit rate. The results show that the spectral distortion of GMM-based VQ can be reduced by using rectangular lattice codebooks instead of scalar codebooks. The performance gain is however small since we exploited only the granular gain of Λ_{16} over $\mathbb{Z}^{16}$. Current developments focus on designing MA(1) prediction for the GMM-based LSF quantization presented in this paper, in order to compare the LPC quantization of AMR-WB and GMM-based LSF quantization.

Table 2: Results for AR(1) predictive LSF quantization.

(a) Results at $R_{tot} = 36$ bits per frame.

Λ	M	avg. SD (dB)	$SD \geq 2$ dB (%)	$SD > 4$ dB (%)
$\mathbb{Z}^{16}$	4	1.12	5.12	0.0282
	8	1.08	3.91	0.0074
	16	1.04	2.89	0.0029
Λ_{16}	4	1.06	3.53	0.0118
	8	1.03	2.97	0.0064
	16	1.00	2.30	0.0009

(b) Results at $R_{tot} = 46$ bits per frame.

Λ	M	avg. SD (dB)	$SD \geq 2$ dB (%)	$SD > 4$ dB (%)
$\mathbb{Z}^{16}$	4	0.74	0.51	0.0035
	8	0.72	0.30	0.0005
	16	0.69	0.17	0
Λ_{16}	4	0.69	0.06	0.0020
	8	0.67	0.17	0.0005
	16	0.64	0.09	0

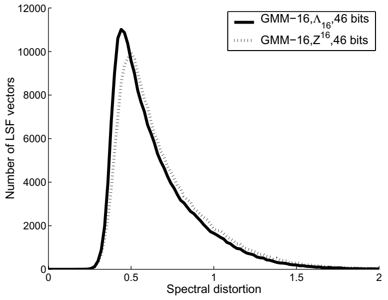


Figure 5: Histograms of SD (AR(1) predictive case, 46 bits).

REFERENCES

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

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

[3] B. Bessette, R. Salami, C. Laflamme, and R. Lefebvre, "The Adaptive Multirate Wideband Speech Codec (AMR-WB)," *IEEE Trans. on Speech and Audio Proc.*, vol. 10, no. 8, pp. 620–637, Nov. 2002.

[4] K.K. Paliwal and W.B. Kleijn, *Quantization of LPC Parameters*, pp. 433–466, in *Speech Coding and Synthesis*, W.B. Kleijn and K.K. Paliwal eds., Elsevier Science, 1995.

[5] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *JRSSB*, vol. 39, no. 1, pp. 1–38, 1977.

[6] S. Ragot, H. Lahdili, and R. Lefebvre, "Wideband LSF quantization by generalized Voronoi codes," in *Proc. Eurospeech*, Sep. 2001, pp. 2319–2322.

[7] Y. Huang and P.M. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Commun. Syst.*, vol. C-10, pp. 289–296, Sep. 1963.

[8] J.A. Bucklew and N.C. Gallagher, "A note on the computation of optimal minimum mean-square error quantizers," *IEEE Trans. Com.*, vol. 30, no. 1, pp. 298–301, 1982.

[9] A.D. Sumbramaniam and B.D. Rao, "Source coding with minimal and rate-independent search and memory complexity," in *Proc. DCC*, 2001.

[10] G.D. Forney, "Coset codes. II. Binary lattices and related codes," *IEEE Trans. Inf. Th.*, vol. 34, no. 5, pp. 1152–1187, Sept. 1988.

[11] J.-P. Adoul, *Lattice and Trellis-Coded Quantization for Efficient Coding of Speech*, chapter 57, pp. 405–422, in *Speech Recognition and Coding, New Advances and Trends*, A. Rubio and J.M. Lopez eds., NATO series, Springer-Verlag, 1995.

[12] G.D. Forney, "Density/Length Profiles and Trellis Complexity of Lattices," *IEEE Trans. Inf. Th.*, vol. 40, no. 6, pp. 1753–1772, Nov. 1994.