RATE ADJUSTABLE SPEECH CODING BY LATTICE QUANTIZATION

Fredrik Nordén, Turaj Zakizadeh Shabestary, and Per Hedelin

Information Theory, Chalmers University of Technology,
Göteborg, Sweden

ABSTRACT

We address speech coding architectures, suited for channels such as packetized networks. We adopt the analysis-by-synthesis paradigm. To achieve operability at a continuum of rates, all explicit perceptual filtering in the codebook search is eliminated, and quantization is performed with lattice quantizers.

In this particular implementation, rate-distortion performance is improved by entropy coding of the lattice indices. The result is a competitive coder applicable to both speech and music, where subjective as well as objective performance scales well with the rate of the coder.

1. INTRODUCTION

For many years, coding at low rates has been a major goal of speech coding research. Today, we see a paradigm shift where the carrier of speech data is shifting towards packetized networks. This shift changes the constrains that we put on our coders. There is a new demand for high quality coders with the ability to handle both speech and audio at a continuum of rates, without an overwhelmingly complexity increase with rate.

A dominant approach for speech coding is *linear prediction analysis-by-synthesis* (LPAS) [1]. Here, speech is modeled as the response of a time-varying all-pole filter, where both the excitation signal and the filter are coded and transmitted. For long, *code-excited linear predictive* (CELP) coding [2, 3] has been the prevailing approach for coding of the excitation. The good performance of the CELP concept is much due to the utilization of high-dimensional vector quantizers (VQ), being searched in a perceptual domain. The drawback of the conventional CELP approach is a high increase in complexity as rate is increased. To reach a rate adjustable coder, the perceptual filtering and the codebook search need to be addressed. One approach to solve the search complexity problem is the *algebraic* CELP (ACELP) [4], where the codebook is constructed from sparse algebraic codes.

Here, we propose a coding framework having the ability of coding both speech and audio with a high subjective quality at a wide range of rates. The proposed framework is based on lattice quantizers, addressing the search complexity problem. Furthermore, a new analysis-by-synthesis structure is proposed that eliminates all explicit perceptual filtering in the codebook search. Based on these ideas, we here present a coder where quantization is followed by an entropy coding of the VQ indices. The result is a variable rate coder with a performance that scales well with average rate, i.e. good performance is achieved for a wide range of average rates.

In Section 2 we discuss the basics of the proposed coder architecture, including the proposed synthesis structure, and in Section 3 the core coding unit, common to both the excitation coding and the spectrum coding, is presented. In Section 4 and in Section 5 the excitation and the spectrum coding are discussed in more detail, respectively. Finally, experiments and conclusions are presented in Section 6 and 7, respectively.

2. CODING FRAMEWORK PRINCIPLES

We adopt the conventional source-filter concept, where the input signal is modeled as the response of a time-varying all-pole filter. For each input frame, $\mathbf{s}$, filter parameters, $\mathbf{a}$, are determined with a standard LPC method. Subsequently, the excitation, $\mathbf{r}$, and the spectrum parameters, $\mathbf{a}$, are coded, and transmitted, c.f. Fig. 1.

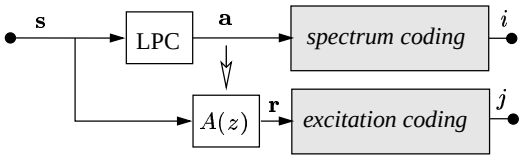

Fig. 1: A basic view of the encoder.

We seek a rate adjustable coder, a coder that has the ability of changing rate without retraining of codebooks. Further, the coder should have an objective and subjective performance that both scale well with rate, potentially lending itself to a layered or embedded implementation.

In order to achieve these demands, we employ *i*) *lattice quantizers*, and *ii*) an architecture *free of all explicit perceptual filtering*. Below, we start with a short discussion on lattice quantization, and continue with a presentation of the proposed analysis-by-synthesis architecture.

2.1. Lattice Quantization

Lattice quantization is tractable due to the availability of fast search algorithms. Distortion optimal quantization using lattices can be achieved in at least two ways; companding of the source into a uniform space [5, 6], or by an entropy coding of the indices subsequent to the lattice quantization [7]. Proposals in a companded environment for the excitation and the spectrum were given in [8] and [9], respectively. In [11] two different usages of lattice quantizers in entropy constrained vector quantization were suggested.

Here, we employ lattice quantizers with a subsequent entropy coding of the indices. The resulting variable rate coder is rate adjustable, i.e. the average rate can easily be adjusted, without any retraining of codebooks or complexity increase.

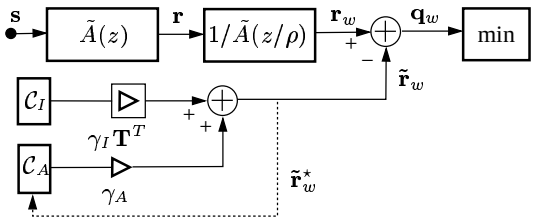

Fig. 2: The search architecture of the proposed coder.

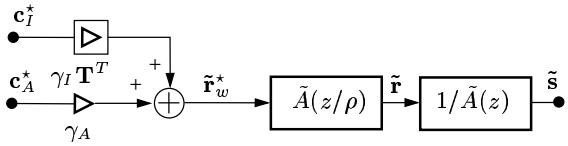

Fig. 3: The synthesis architecture of the proposed coder.

2.2. Analysis-by-Synthesis Architecture

The proposed analysis-by-synthesis architecture, see Fig. 2, is structured to avoid perceptual filtering in the codebook search, still attaining minimization in a perceptual domain [1]. This architecture is facilitated by a decorrelating transform, $\mathbf{T}$, and a long term prediction (LTP) codebook, $\mathcal{C}_A$, operating in a perceptual domain. The perceptual excitation feedback, $\tilde{\mathbf{r}}_w^\star$, is illustrated by the dotted line in Fig. 2.

The excitation is selected in a weighted domain where the error of concern, $\mathbf{q}_w = \mathbf{r}_w - \tilde{\mathbf{r}}_w$, is a filtered version of $\mathbf{s} - \tilde{\mathbf{s}}$. The LTP and the innovation are selected one at time in a multistage fashion, starting with the LTP according to

$$\mathbf{c}_A^\star = \underset{\mathbf{c}_A \in \mathcal{C}_A}{\operatorname{argmin}} \|r_w - \gamma_A \mathbf{c}_A\|, \tag{1}$$

where γ_A is the gain of the selected LTP vector $\mathbf{c}_A^\star \in \mathcal{C}_A$. Subsequently the perceptual innovation codevector, $\mathbf{c}_I$, is selected according to

$$\mathbf{c}_I^\star = \underset{\mathbf{c}_I \in \mathcal{C}_I}{\operatorname{argmin}} \|\mathbf{T}(r_w - \gamma_A \mathbf{c}_A^\star) - \gamma_I \mathbf{c}_I\|, \tag{2}$$

where γ_I is the gain of the selected innovation vector $\mathbf{c}_I^\star \in \mathcal{C}_I$. Here, we have omitted to address the gain selection issue.

Having selected the perceptual LTP and the perceptual innovation as above (the coding of the perceptual innovation is further treated in Section 4). The total perceptual excitation is constructed as

$$\tilde{\mathbf{r}}_w^\star = \gamma_A \mathbf{c}_A^\star + \gamma_I \mathbf{T}^T \mathbf{c}_I^\star, \tag{3}$$

and a synthesized signal, $\tilde{\mathbf{s}}$, is obtained by filtering $\tilde{\mathbf{r}}_w^\star$ with $\tilde{A}(z/\rho)/\tilde{A}(z)$, see Fig. 3.

3. QUANTIZATION AND CODING

Both the coding of the perceptual innovation and the spectrum coding use the same principle in quantization and entropy coding. The difference is in the model of data. A Gaussian mixture model (GMM) governs the modeling of the spectrum vector see Section 5. The perceptual innovation signal is modeled by a single Gaussian, adapted for each subframe based on the decoded spectrum, see Section 4.

Below, we present the quantization for a *d*-dimensional decorrelated Gaussian source, the core unit for both the excitation and the spectrum coding. The decorrelation of the excitation and the spectrum vector are addressed in Section 4 and 5, respectively.

[1] In Fig. 2: note the lack of filters, $A(z/\rho)$, in between the codebooks and the minimization operator, min. Neither of the codebooks provide white vectors.

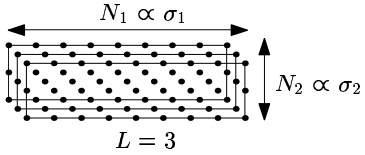

Fig. 4: An illustration of a union of 3 Z-lattices, where $\{\sigma_i\}$ and $\{N_i\}$ represent the standard deviation of the source and the bit allocation, respectively.

3.1. Lattice Codebook Structure

We seek a lattice that is fast searched, having good quantization properties. We have chosen to work with a lattice built by a union of translated Z-lattices (see Fig. 4),

$$\mathcal{C} = \Delta \bigcup_{l=1}^{L} (\mathbf{z} + \mathbf{r}_l), \tag{4}$$

where L is the number of Z-lattices, Δ is the step size of the lattices, $\mathbf{z}$ is a Z-lattice (cubic lattice), and $\mathbf{r}_l$ is a vector offset relative to the origin. The offset, $\mathbf{r}_l$, can be chosen randomly, or in a structured fashion. The latter approach is noticeably superior to the former if the dimension is not high (less than ten). Here, we use the structured approach [8].

By using Z-lattices, the major issues in lattice quantization, i.e. truncation, indexing, and search, break down into scalar problems. Moreover, the interaction among the elements of the union produces Voronoi regions which are more efficient than a hypercube. So, simplicity in design and a reasonable performance is achieved at the same time, making this approach attractive.

3.2. Bit Allocation

To shape the lattices, we do bit allocation among dimensions, see Fig. 4. The allocation is based on the variances, $\{\sigma_i\}$, of the source to be coded according to

$$N_i \propto \frac{\sigma_i}{\left(\prod_{j=1}^{d} \sigma_j\right)^{1/d}}, \tag{5}$$

where N_i is the number of reconstruction points for dimension i. Based on the resulting bit allocation, the step-size, Δ, is calculated c.f. [10].

3.3. Encoding

The encoding in the proposed scheme is performed in two steps. First, we search each Z-lattice for the code vector which minimizes the distortion criterion (scalar-wise for each dimension), we call this code vector a *hot candidate*. The next step is to find the best vector among the hot candidates.

For a given source vector $\mathbf{x}$, the hot candidate vector $\mathbf{c}_i$ which minimizes the Lagrangian cost function [12]

$$J(\mathbf{x}, \mathbf{c}_i, \lambda) = d(\mathbf{x}, \mathbf{c}_i) + \lambda l(\mathbf{c}_i), \tag{6}$$

is chosen. Here, $d(\mathbf{a}, \mathbf{b})$ is a distortion measure, $l(\mathbf{c}_i)$ is the length of the codeword (Huffman codeword) assigned to $\mathbf{c}_i$, and λ is a Lagrange multiplier, providing a mechanism for rate control. By changing λ one has a trade off between performance and bit rate. For the excitation coding we use the local SNR as the distortion measure, while in the spectral coding we use spectral distortion (SD) as the distortion measure.

3.4. Entropy Coding

As mentioned above, the encoding consists of two steps. As a consequence, the indexing also includes two steps. The first is an index, l, pointing to the Z-lattice that the code vector belongs to (the winning hot candidate). This is a fixed length code. The second part is the vector index, $\mathbf{k}$, of the code vector in the Z-lattice, which is subject to entropy coding.

The assumption of a decorrelated Gaussian source allows us to design separate codes for each dimension. Consider dimension i of the lth lattice. The probability of the kth reconstruction point is calculated according to the marginal pdf of the source, $f_i(y)$, and the given step-size, Δ,

$$p_{i,k} = \int_{r_{i,l}+k\Delta}^{r_{i,l}+k\Delta+\Delta} f_i(y)dy, \tag{7}$$

where $r_{i,l}$ is the offset for lattice l, and k is the indexing in the Z-lattice, both for dimension i. Subsequently, the probabilities $\{p_{i,k}\}_{k=1}^{N_i}$ are used to generate a Huffman code for dimension i, where N_i is the number of points in dimension i.

There is one more practical issue to consider for the design of an efficient Huffman code. If the number of points, N', along some dimension is small, then the resulting Huffman code is inefficient. To prevent this from happening, we group dimensions, $i_1, \ldots, i_k$, together resulting in a larger alphabet, $\prod_{j=1}^{k} N_{i_j}$. Thus a more efficient Huffman coding is achieved.

4. PERCEPTUAL INNOVATION CODING

The target for the perceptual innovation coder is the resulting error after LTP coding,

$$\mathbf{e}_w = \mathbf{r}_w - \gamma_A \mathbf{c}_A^\star, \tag{8}$$

where $\mathbf{r}_w$, γ_A, and $\mathbf{c}_A^\star$ are defined in Section 2.2. The coding consists of a "decorrelating" transform $\mathbf{T}$ and a coding unit, as described in Section 3. All parts, visualized in Fig. 5, are adapted on a sub-block basis. Sub-block adaption is based on the information given by the quantized spectrum vector, $\tilde{\mathbf{a}}$.

Below, we discuss properties of the innovation, and the choice of the decorrelating transform, $\mathbf{T}$.

4.1. Decorrelation and Coding

In order to perform an efficient quantization (bit allocation) and entropy coding (probability estimates) of the innovation vector we need a "source" model. A reasonable approximation is to take the LTP residual,

$$\mathbf{e} = \mathbf{r} - \gamma_A \mathbf{W}^{-1} \mathbf{c}_A^\star, \tag{9}$$

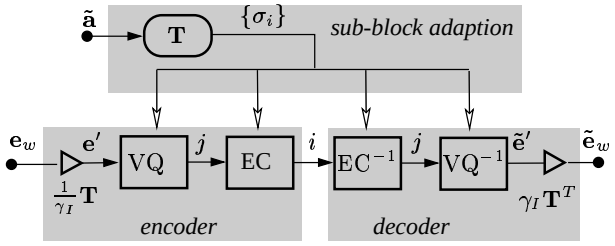

Fig. 5: Innovation encoder/decoder structure.

as white. Here $\mathbf{W}$ is a matrix containing shifted versions of the impulse response of the filter $1/\tilde{A}(z/\rho)$, given by the spectrum vector. The target for the perceptual innovation coder, $\mathbf{e}_w$, is the LTP residual filtered with $1/\tilde{A}(z/\rho)$, in matrix notation $\mathbf{e}_w = \mathbf{W}\mathbf{e}$. Thus, the covariance matrix of the random vector $\mathbf{E}_w$ can be expressed as $\mathbf{C}_{Ew} = \delta^2 \mathbf{W}\mathbf{W}^T$, where δ^2 is the variance of the random vector $\mathbf{E}$, previously assumed white.

Given the covariance matrix $\mathbf{C}_{Ew}$ (also available at the decoder) we seek both a decorrelating transform, $\mathbf{T}$, and the relative scale of the components along each dimension in the new basis, $\{\sigma_i\}$. Mapping $\mathbf{e}_w$ on the basis $\mathbf{T}$, we achieve a new decorrelated vector

$$\mathbf{e}' = \frac{1}{\gamma_I}\mathbf{T}\mathbf{e}_w, \tag{10}$$

where γ_I is a gain normalization. Now, the component variances $\{\sigma_i\}$ for the decorrelated vector, $\mathbf{e}'$, are given by the diagonal elements of the matrix $\mathbf{T}\mathbf{C}_{Ew}\mathbf{T}^T$.

An obvious choice of decorrelating transform, is to perform an eigenvalue decomposition, and assign the eigenvectors to the rows of $\mathbf{T}$ (KLT). The KLT is source dependent and therefore complex. A more attractive choice of orthonormal basis, which we have chosen to work with, is the *discrete cosine transform* (DCT) [7]. The DCT has shown to have almost as good quantization properties as the KLT in terms of the decorrelation ability.

4.2. Rate Adaption

Leaving the fixed-rate paradigm, allows us to distribute bits over time, such that perceptually important frames achieve a higher rate [13]. The innovation coder adapts the number of reconstruction levels of the lattice quantizer, according to properties of the input signal. The number of levels are increased for strong sequences, and for onsets. Further, if the gain of the LTP is not sufficiently high, the LTP is turned off.

5. SPECTRUM CODING

The target for the spectrum coding is the LPC spectrum vector, $\mathbf{a}$ (see Fig. 1), transformed into the line-spectral-frequency (LSF) domain. As described in Section 3 the quantization requires a source model to perform the rate allocation among the dimensions. For the spectrum coding, this is governed by a GMM of the source pdf,

$$f_{\mathcal{M},\mathbf{x}}(\mathbf{x}) = \sum_{i=1}^{M} \rho_i f_i(\mathbf{x}), \tag{11}$$

where $\mathbf{x}$ is the LSF vector, ρ_i are the component weights, and $f_i(\mathbf{x})$ are Gaussian densities. As opposed to the innovation model,

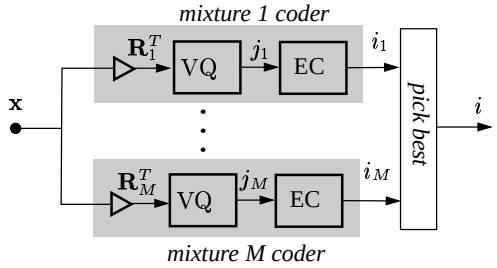

Fig. 6: Spectrum encoder structure. M coders, one for each mixture, work in parallel.

the spectrum model is fixed, and is trained beforehand by e.g. the EM-algorithm [14].

5.1. Decorrelation and Coding

The actual coding is, as illustrated in Fig. 6, performed in parallel using one coder (described in Section 3) for each mixture. The spectrum vector, $\mathbf{x}$, is decorrelated by the pre-processing $\mathbf{R}_i^T\mathbf{x}$, see Fig 6, where $\mathbf{R}_i$ is an orthonormal matrix containing the eigenvectors of the covariance matrix, $\mathbf{C}_i$, for mixture i. The corresponding eigenvalues, $\{\sigma_i\}$, give the strength of each dimension, and are used in the bit allocation described in Section 3.2.

There is one more issue to consider; the rate allocation among mixtures. We allocate rates depending on the weight, ρ_i, and the covariance, $\mathbf{C}_i$, of the mixture, as discussed in [15].

6. EXPERIMENTS

In this section we present results from experiments on the proposed coding system, starting with a short description of a particular instance of the system for 8 kHz sampling.

We employ a frame size of 159 samples, each frame is further subdivided into 3 subframes of 53 samples each. An LPC model of order 12 is extracted once for each frame, using the autocorrelation method with a 25 ms Hamming window. A standard 20 Hz bandwidth expansion is applied to each pole of the spectrum polynomial. The spectrum coding, operating in the LSF domain, is based on a GMM with 32 mixtures, utilizing a union of 109 Z-lattices. Excitation is determined once for each subframe, and utilizes a 9 bit LTP codebook based on upsampling, with a maximum lag of 154 samples. Further, the innovation coder is based on a union of 109 Z-lattices. Both the LTP gain (γ_A) and the innovation gain (γ_I) are coded using scalar quantizers at 6 bits per subframe. Another 6 bits per frame are used for a frame gain, and 1 bit for LTP on/off selection.

In the experiments we evaluate performance as a function of the rate. In Fig. 7a, we can see how the system scales with average rate. Moreover, rate variability is illustrated in Fig. 7b. Only informal listening tests have been performed. Our impression is that the proposed system performs well both for speech and music. We compared the proposed system operating at various rates (see Fig. 7a) with an LD-CELP (ITU-T G.728 16 kbit/s). At rates higher than 9-11 kbit/s the proposed system was preferred.

7. CONCLUSIONS

We propose a coding system that is adjustable in rate and of moderate complexity. The scheme requires no training of codebooks,

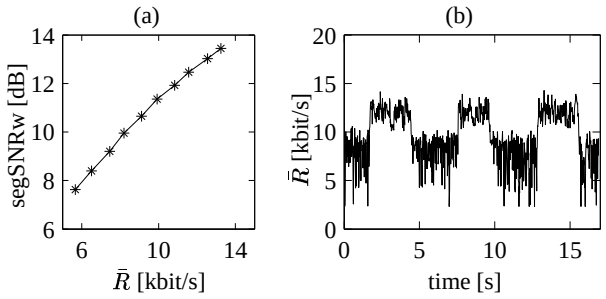

Fig. 7: (a) SNR in the weighted domain as a function of average rate. (b) Rate as a function of time for an average rate of 10.65 kbits/s.

and shows a competitive distortion performance for both speech and audio. Subjective and objective tests point at a coder particularly useful at rates in the range 6-16 kbit/s.

8. REFERENCES

[1] P. Kroon and W. B. Kleijn, "Linear-prediction based analysis-by-synthesis coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., pp. 79–119. Elsevier Science Publishers, Amsterdam, The Netherlands, 1995.

[2] B. S. Atal and M. R. Schroeder, "Stochastic coding of speech signals at very low bit rates," in *Proc. Int. Conf. on Com.*, (Amsterdam, The Netherlands), 1984, pp. 1610–1613.

[3] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Proc. ICASSP*, 1985, vol. 2, pp. 937–940.

[4] C. Laflamme, J-P. Adoul, H.Y. Su, and S. Morisette, "On reducing computational complexity of codebook search in celp coder through the use of algebraic codes," in *Proc. ICASSP*, 1990, vol. 1, pp. 177–180.

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

[6] J.A. Bucklew, "Companding and random quantization in several dimensions," *IEEE Trans. on Info. Theory*, vol. 27, no. 2, pp. 207–211, 1981.

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

[8] F. Nordén and P. Hedelin, "Scalable innovation coding," in *Proc. ICASSP*, 2002, vol. 1, pp. 653–656.

[9] T. Z. Shabestary and P. Hedelin, "Spectral quantization by companding," in *Proc. ICASSP*, 2002, vol. 1, pp. 641–644.

[10] N. Farvardin and F. Y. Lin, "Performance of entropy-constrained block transform quantizers," *IEEE Trans. on Info. Theory*, vol. 37, no. 5, pp. 1433–1439, 1991.

[11] S. F. Simon and W. Niehsen, "On entropy coded and entropy constrained lattice vector quantization," in *Proc. ICIP*, 1996, vol. 3, pp. 419–442.

[12] R. M. Gray, T. Linder, and J. Li, "A lagrangian formulation of Zador's entropy-constrained quantization theorem," *IEEE Trans. on Info. Theory*, vol. 48, no. 3, pp. 695–707, 2002.

[13] T. Painter and A. Spanias, "A review of algorithms for perceptual coding of digital audio signals," in *Proc. Int. Conf. on Dig. Sig. Proc.*, (Santorini, Greece), 1997, vol. 1, pp. 179–208.

[14] A. P. Dempster, N. M.Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. of Royal Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[15] Robert M: Gray, "Gauss mixture vector quantization," in *Proc. ICASSP*, 2001, vol. 3, pp. 1769 –1772.