QUANTIZATION OF THE SPECTRAL ENVELOPE FOR SINUSOIDAL CODERS

Thomas Eriksson, Hong-Goo Kang and Yannis Stylianou

AT&T Labs-Research, SIPS
180 Park Avenue, Florham Park, NJ 07932
[eriksson, goo, styliano]@research.att.com

ABSTRACT

In an effort to efficiently code the spectral envelope of speech signals for wideband speech coding based on sinusoidal models, a robust computation of discrete cepstrum coefficients and their quantization is investigated. A parameterization of the spectral envelope has been proposed which is based on discrete cepstral coefficients using regularization techniques. This paper presents an efficient quantization scheme for these coefficients in order to use them in applications like speech coding. We present results which show a 35% reduction in bitrate when compare to simple scalar quantization. To verify the efficiency of the proposed quantization schemes, informal listening tests were performed in the context of a sinusoidal coder.

1. INTRODUCTION

The estimation of a continuous spectral envelope when only discrete values of this envelope are specified is a subject of considerable importance with applications to speech coding and speech synthesis. In speech coding, an efficient parameterization of the spectral envelope is desirable. In speech synthesis, where pitch modifications are required, the amplitudes at the new harmonics can be obtained by resampling the continuous spectral envelope. In the context of sinusoidal coders it is desirable to find a representation method that leads to an envelope which passes through the measured sine-wave amplitudes. While a number of techniques may be used for estimating the spectral envelope including linear prediction or simple cepstral estimation techniques, none of these methods satisfy the above criterion. An attempt to use standard cepstral analysis on an interpolated spectral envelope has been reported in [1]. However, a large number of cepstral coefficients has been used for an accurate envelope fitting in this case. The same problem has been addressed in [2] where a set of nonlinear equations were derived which required the use of a costly iterative procedure. Galas *et.al.* [3] estimated the cepstral coefficients by minimizing a frequency-domain least-squares criterion (discrete cepstrum coefficients). Although this method proves to be very efficient it is plagued with ill-conditioning problems. Cappé, *et.al.* [4] proposed a *regularized* technique to achieve a well-behaved spectral envelope using discrete cepstrum coefficients, avoiding ill-conditioning problems.

The regularized cepstrum coefficients have proven to be a good candidate for an efficient parameterization of the spectral envelope. They have already been used in speech modification [5] and in voice conversion [6]. The quantization of the regularized cepstrum coefficients is an important issue if they are to be used in speech coding.

This paper extends the work of Cappé, *et.al.*, by introducing the quantization of the regularized cepstrum coefficients. To estimate the number of bits that should be allocated for a regularized cepstrum vector while maintaining transparent speech quality, a sequence of quantization methods is used. First, a set of scalar quantizers is optimized based on the probability density function of the regularized cepstrum coefficients. This gives a first estimate of how many bits should be allocated per cepstrum vector. To reduce the number of bits while maintaining the same quality a perceptual weighting criterion is proposed. Based on this criterion the optimum order of the cepstrum analysis can be determined. Further saving of bits can be realized by reducing the intra-correlation of the cepstrum vectors, based on the Karhunen-Loeve Transform, KLT, and by exploiting the inter-correlation of the cepstrum vectors using vector prediction. A safety-net quantizer is used in parallel with the predictive quantizer for low correlated vectors, e.g. transition from voiced/unvoiced to unvoiced/voiced. Finally, two methods of vector quantization, split and multistage, have been studied in order to further reduce the bit rate. Listening tests have been carried out in the context of the Harmonic plus Noise Model, HNM [5] and the results show the efficiency of the proposed quantization scheme.

The paper is organized in three major parts. First, we describe the regularization technique for the estimation of discrete cepstrum coefficients which was proposed by Cappé, *et.al.*. Next, we address the problem of the quantization of the regularized discrete cepstrum coefficients. The last section shows results from an experimental study where the quantization scheme has been tested on a large speech database. Results from listening tests using the Harmonic plus Noise Model, HNM, are also reported.

2. COMPUTATION OF THE REGULARIZED CEPSTRUM COEFFICIENTS (RCC)

In [4], the spectral envelope of a speech signal is represented by discrete cepstrum coefficients. Given a set of L sine-wave amplitudes a_k, measured at the normalized frequencies f_k, the discrete cepstrum is obtained by minimizing the squared error between a measured magnitude a_k and a magnitude $|S(f_k)|$ in the log-spectral domain,

$$\epsilon = \sum_{k=1}^{L} (\log a_k - \log |S(f_k)|)^2. \quad (1)$$

The spectral envelope $|S(f)|$ is related to the discrete cepstrum coefficients by

$$\log|S(f)| = c_0 + 2\sum_{i=1}^{p} c_i \cos(2\pi f i) = \mathbf{M}\mathbf{c}, \tag{2}$$

where c_i are the coefficients of the cepstrum vectors $\mathbf{c}$, p is the order of the cepstrum, and the matrix $\mathbf{M}$ is defined as

$$\mathbf{M} = \begin{bmatrix} 1 & 2\cos(2\pi f_1) & 2\cos(2\pi f_1 2) & \dots & 2\cos(2\pi f_1 p) \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 2\cos(2\pi f_L) & 2\cos(2\pi f_L 2) & \dots & 2\cos(2\pi f_L p) \end{bmatrix} \tag{3}$$

The optimal $\mathbf{c}$ that minimizes ϵ is given by

$$\mathbf{c} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{a} \tag{4}$$

where $\mathbf{a} = [\log(a_1)\dots\log(a_L)]^T$ are the specified log amplitudes. The problem with the solution given by (4) is that the matrix $\mathbf{M}^T\mathbf{M}$ is ill-conditioned when p approaches L (singular when $p \geq L$). Regularization techniques are well-known for obtaining well-behaved solutions to over-parameterized estimation problems [7]. The regularized discrete cepstrum coefficients are obtained by minimizing the error criterion

$$\epsilon_r = \sum_{k=1}^{L} (\log a_k - \log|S(f_k)|)^2 + \lambda\mathcal{R}[S(f)] \tag{5}$$

The first term is the error criterion as it is given in (1). The parameter λ controls the degree of regularization, and should be increased as p approaches L. A classical smoothness constraint $\mathcal{R}[S(f)]$ penalizes rapid variations in the spectral envelope. The functional form of this constraint is:

$$\mathcal{R}[S(f)] = \int_{-1/2}^{1/2} [\frac{d}{df}\log|S(f)|]^2 df \tag{6}$$

In [4], it is shown that the minimum value of ϵ_r is obtained if $\mathbf{c}$ is selected as

$$\mathbf{c} = [\mathbf{M}^T\mathbf{M} + \lambda\mathbf{R}]^{-1}\mathbf{M}^T\mathbf{a} \tag{7}$$

where $\mathbf{R}$is a diagonal matrix with diagonal elements $8\pi^2[0, 1^2, 2^2, \dots, p^2]$.

3. QUANTIZATION OF THE REGULARIZED CEPSTRUM COEFFICIENTS

In this section, we study quantization of the regularized cepstrum coefficients, and various methods to reduce the required number of bits while still maintaining a perceptually transparent quantization. The following methods were applied sequentially: Scalar quantization, perceptual weighting, Karhunen-Loeve transformation, predictive quantization, and vector quantization. The results are summarized in Section 4.

The cepstrum vectors were computed from speech sampled at 16 kHz. The order of the cepstrum coefficients was determined by listening tests, and we found that an order of 32 is enough to represent the spectral envelope for wideband speech signals. For the simulations in this section we used a training database with more than 200,000 cepstrum vectors. An independent database was used for evaluation.

3.1. Scalar quantization

The simplest possible quantization method is scalar quantization. The design of scalar quantizers for the 32-dimensional cepstrum vectors consists of two steps: first an appropriate number of bits is found for each quantizer, and then these quantizers are optimized for the probability density of the cepstrum vector.

The bit allocation procedure should allocate the available bits over the scalar quantizers to maximize the SNR (signal-to-quantization-noise-ratio), defined as

$$\text{SNR} = 10\log_{10}\left(\frac{\mathrm{E}\left[\sum_{k=0}^{p} c_k^2\right]}{\mathrm{E}\left[\sum_{k=0}^{p}(c_k - \tilde{c}_k)^2\right]}\right), \tag{8}$$

where c_k and $\tilde{c}_k$ are the kth components of the cepstrum vector and the quantized cepstrum vector, respectively.

Based on *rate-distortion theory* [8] for Gaussian variables, an approximate bitallocation for the (non-Gaussian) cepstrum vectors can be derived. For scalar quantization, Gaussian rate-distortion theory tell us that the bits should be allocated to make the distortion of each quantizer equal, and that an approximate bit allocation for quantizer k is given by

$$R_k = R\frac{\log_2\sigma_k^2}{\sum_{i=0}^{p}\log_2\sigma_i^2},\ k = 0..p-1 \tag{9}$$

that is, the optimal rate R_k is proportional to the logarithm of the variance σ_i^2. R is the total number of bits that is available. With the above Gaussian bit allocation formula, and with additional fine-tuning to increase the SNR for the (non-Gaussian) cepstrum vectors, we get the bit allocation shown as a solid line in Figure 1.

Using the bit allocation in (9), the scalar quantizers are optimized for the pdf of the cepstrum vectors (using e.g. Max-Lloyd training [9]). Listening tests reveal that an SNR of 35 dB or higher is required for inaudible quantization distortion. With the scalar quantization scheme discussed above, a total of 100 bits are required to reach the desired SNR.

3.2. Cepstrum perceptual weighting

When we tested different bit allocations for the scalar quantizer, and performed listening tests to determine the subjective quality, it was discovered that the first few components of the cepstrum vectors were more important for the subjective quality than the high-indexed components. We therefore decided to introduce a perceptually weighted SNR measure for the cepstrum vectors. The weighted SNR, WSNR, is computed as

$$\text{WSNR} = 10\log_{10}\left(\frac{\mathrm{E}\left[\sum_{k=0}^{p} c_k^2 w_k\right]}{\mathrm{E}\left[\sum_{k=0}^{p}(c_k - \tilde{c}_k)^2 w_k\right]}\right), \tag{10}$$

where w_k is the weighting function. We propose a simple weighting function, given by

$$w_k = C^k. \tag{11}$$

The constant C must be determined by listening tests. Values of C in the interval $0.6 - 0.7$ were found to give perceptually good results. In Figure 1 the resulting bitallocations for the 32 components of the cepstrum vectors are depicted, for the two cases $C = 1$ (unweighted) and $C = 0.65$. Only the first 21 cepstrum components were necessary to quantize in our experiments. With a bit allocation determined by the weighting function proposed above

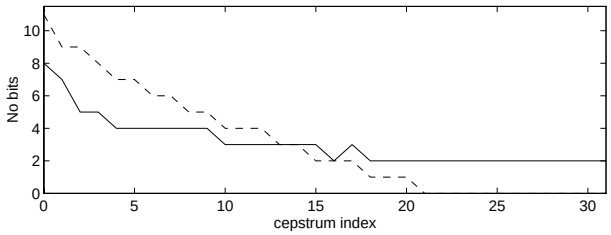

Figure 1: Bit allocation for scalar quantization of unweighted (solid) and weighted (dashed) $C = 0.65$ cepstrum vectors.

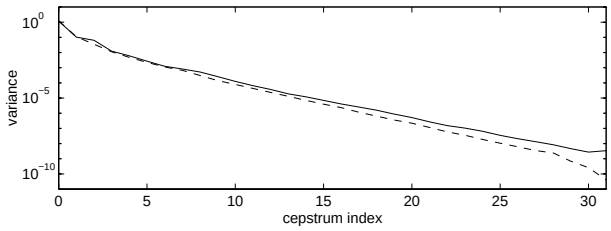

Figure 2: Variances for the cepstrum vectors (solid), and for the KLT-decorrelated vectors (dashed). The variances directly gives the bit allocation.

($C = 0.65$) we were able to reduce the number of bits to 87, and still achieve the same perceptual quality as in the unweighted case using 100 bits. The corresponding WSNR is 39 dB.

3.3. Karhunen-Loeve transform

To further improve the performance of the scalar quantization scheme, the *Karhunen-Loeve transform* (KLT) can be employed. The KLT decorrelates the incoming vectors, thereby leading to bit savings in scalar quantization. The KLT matrix $\mathbf{T}$ is composed by the eigenvectors $\mathbf{l}_k$ of the process,

$$\mathbf{T} = [\mathbf{l}_0, \mathbf{l}_1, ..., \mathbf{l}_p]^T, \tag{12}$$

and these eigenvectors can be found by solving the system of equations

$$\mathbf{R}_{00}\mathbf{l}_k = \lambda_k \mathbf{l}_k \text{ for } k = 0, 1, ..., p, \tag{13}$$

where $\mathbf{l}_k$ are the eigenvalues, and $\mathbf{R}_{00}$ is the autocorrelation matrix, defined as in (16). The vectors $\mathbf{y} = \mathbf{T}\mathbf{x}$ are then uncorrelated. In Figure 2 we see the variances for the cepstrum vectors and for the decorrelated vectors. It can be seen after KLT, the variance is more localized to low indices. However, the difference is small since the cepstrum vector components are fairly uncorrelated to begin with. With KLT, the total number of bits can be reduced by 6, to 81, for the same WSNR as in direct quantization of the weighted cepstrum parameters.

3.4. Vector prediction

Vector prediction has been proposed as a method to exploit inter-vector correlation of vector processes. A linear vector predictor of order K for a vector process $\mathbf{x}_n$ can be written

$$\hat{\mathbf{x}}_n = \sum_{k=1}^{K} \mathbf{A}_k \tilde{\mathbf{x}}_{n-k}, \tag{14}$$

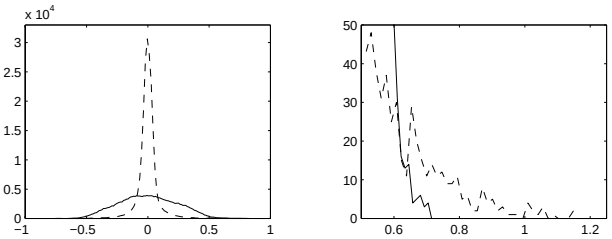

Figure 3: Histogram of a cepstrum coefficient (solid) and of the prediction residual for the same coefficient (dashed). The variance of the prediction residual is much lower, but the number of "outliers" (the tails of the histogram) is high.

where $\hat{\mathbf{x}}_n$ is the one-step-ahead prediction vector, $\tilde{\mathbf{x}}_{n-k}$ are earlier quantized input vectors, and $\mathbf{A}_k$ are the prediction matrices. The optimum (in a MMSE sense) prediction matrices can be found by solving a system of linear matrix equations. The experiments in this document are restricted to first order prediction, and for this case the optimum prediction matrix is given by

$$\mathbf{A}_1 = \mathbf{R}_{01}\mathbf{R}_{11}^{-1} \tag{15}$$

where $\mathbf{R}_{ij}$ are the correlation matrices,

$$\mathbf{R}_{ij} = \mathrm{E}[\mathbf{x}_{n-i}\mathbf{x}_{n-j}^T]. \tag{16}$$

The correlation matrices are estimated by use of the database described previously.
When the prediction matrix $\mathbf{A}_1$ is computed, a set of scalar quantizers is optimized for quantization of the prediction error, $\mathbf{e}_n = \mathbf{x}_n - \hat{\mathbf{x}}_n$.
The prediction gain was much lower than expected (estimated by rate-distortion theory) in our experiments. In Figure 3, we can see the explanation; the variance for the prediction error is much lower than the variance of the cepstrum vector, but the tails of the prediction error histogram are very wide. This is due to a "2-mode" behaviour of speech signals, with voiced and unvoiced segments. A large part of the time the input vectors are highly correlated and the predictor works well, but occasionally the input vector is uncorrelated with the previous, and the predictor is unable to perform well. This results in "outliers", prediction residual vectors with hard-to-quantize high-energy components. In the next subsection, we discuss a solution to this problem.

3.5. Safety-net

The *safety-net quantizer* [10] was proposed as a solution to the problem of low-correlation vectors discussed in the previous subsection. A safety-net quantizer is a memoryless quantizer working in parallel with a predictive quantizer (or any other quantizer exploiting correlation between consecutive vectors). The advantage of the safety-net scheme is that the memoryless quantizer takes care of the low-correlation "outliers", and the predictive quantizer can concentrate on the highly correlated vectors. Another advantage is that error propagation, introduced by bit errors when the transmission channel is noisy, is canceled by the memoryless quantizer, but in this document we assume error free transmission. A safety-net extended predictive quantization scheme is depicted in Figure 4. In this report, the predictive and the safety-net quantizers were trained independently; the performance can be further

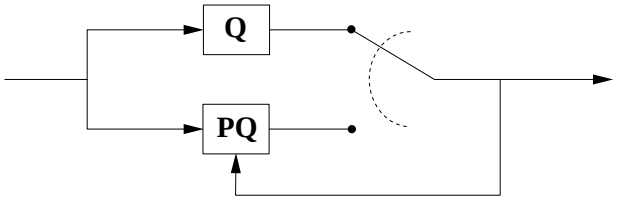

Figure 4: A predictive quantizer (PQ) extended with a safety-net quantizer (Q). One bit is used to select which of the quantizers to use.

improved by simultaneous optimization.
The safety-net scheme leads to additional reductions of 11-12 bits, and we need a total of 70 bits to achieve transparent cepstrum quantization.

3.6. Vector quantization

Vector quantization (VQ) has been proven to be the optimal quantization scheme, in the sense that for a given delay, no other scheme can perform better than VQ [11]. Since single-stage, full-size VQ is very complex, a large number of complexity reduction methods have been proposed. We have studied two methods for "divide-and-conquer"-quantization of cepstrum and prediction residual vectors: split VQ and multistage VQ [11].
With *split VQ*, the input vectors are divided into a set of smaller vectors, which are subsequently quantized independently. The complexity is considerably reduced compared to the complexity of full-size VQ, but some performance loss is inevitable.
The 32-dimensional vector is split into 8 subvectors of various dimension, and each of these are quantized with a VQ with 10 bits or less. This scheme leads to a reduction of 5 bits compared to the scalar quantization schemes.
The basic idea of *multistage VQ* is to divide the quantization into successive stages, where the first stage performs a relatively crude quantization, the second stage quantize the error vector between the original and the quantized first stage output, and so on. Multistage VQ is preferable when there is high correlation between the components in the vector, and a split scheme cannot perform well. In our experiments, the multistage VQ (with 7 stages) performs approximately as good as the split VQ, with a 5 bit reduction compared with scalar schemes. The complexity is considerably higher than the complexity of the split VQ, however.

4. SUMMARY AND DISCUSSION

To verify the efficiency of the proposed quantization schemes listening tests have been carried out with a sinusoidal speech coder, HNM [5]. In our listening tests, a WSNR of about 39 dB or higher was required for transparent quantization of the cepstrum vectors. In table 4, we present the required number of bits using the different quantization schemes. Note that each scheme builds on the results of the previous, i.e. the KLT was applied on perceptually weighted vectors, the predictor was designed for the KLT-decorrelated vectors and so on.
By employing more and more sophisticated methods, we were able to reduce the number of bits for a transparent quantization by a total of 35 bits (from 100 bits per vector to 65 bit per vector). The largest gains were due to the use of a weighted distortion measure, and from applying a safety-net-extended predictive quantization scheme.

Table 1: The required number of bits for the different quantization schemes, for a perceptually transparent quantization

quantizer scheme	*bits*
scalar quantization	100
perceptual weighting	87
KLT	81
prediction+safetynet	70
vector quantization	65

Preliminary tests to fully quantize an HNM *wideband speech* coder were carried on. About 80 % of the encoded bit-stream in the HNM coder comes from spectral envelope information. With the above quantization scheme and a frame size of 10 ms, the rate for the HNM *wideband speech* coder will be below 8 kbit/s. Some preliminary experiments with a frame size of 20 ms were also done, and the results suggest that a rate below 5 kbit/s is within reach.

5. REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), ch. 4, pp. 121–173, Elsevier, 1995.

[2] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modelling," *Proc. IEEE*, vol. 39, pp. 411–423, 1991.

[3] T. Galas and X. Rodet, "An improved cepstral method for deconvolution of source-filter systems with discrete spectra: Application to musical sound signals," in *Proc. of International Computer Music Conference*, (Glasgow), pp. 82–84, 1990.

[4] O. Cappé and E. Moulines, "Regularization techniques for discrete cepstrum estimation," *IEEE Signal Processing Letters*, vol. 3(4), pp. 100–102, April 1996.

[5] Y. Stylianou, J. Laroche, and E. Moulines, "High-Quality Speech Modification based on a Harmonic + Noise Model.," *Proc. EUROSPEECH*, 1995.

[6] Y. Stylianou, O. Cappé, and E. Moulines, "Statistical methods for voice quality transformation," *Proc. EUROSPEECH*, 1995.

[7] A. Tikhonov and V. Arsenin, *Solutions of Ill-Posed Problems*. Washington: Winston, 1977.

[8] R. M. Gray, *Source coding theory*. Kluwer Academic Publishers, 1990.

[9] S. P. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, pp. 129–137, mar 1982.

[10] T. Eriksson, J. Lindén, and J. Skoglund, "Exploiting interframe correlation in spectral quantization - a study of different memory vq schemes," in *Proc.IEEE ICASSP*, vol. 2, pp. 765–768, 1996.

[11] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Kluwer Academic Publishers, 1992.