QUANTIZATION OF THE LPC MODEL WITH THE RECONSTRUCTION ERROR DISTORTION MEASURE

Jan S. Erkelens and Piet M.T. Broersen
Delft University of Technology, Department of Applied Physics
P.O. Box 5046, 2600 GA Delft, The Netherlands
Tel +31 15 2781823 / +31 15 2786419 Fax +31 15 2784263
e-mail: erkelens@tn.tudelft.nl / broersen@tn.tudelft.nl

ABSTRACT

In Linear Predictive Coding algorithms, the coding of the speech signal consists of two separate stages: coding of the LPC model and coding of the excitation. In CELP, the LPC excitation is coded by Analysis-by-Synthesis in the reconstruction domain, not by minimization of the error in the LPC residual domain. Commonly used distortion measures for quantization of the LPC spectral model are the Spectral Distortion and the Likelihood Ratio. For small quantization errors, they belong to a class of similar distortion measures which express an error in the residual domain. A new spectral distortion measure is proposed, the Reconstruction Error Distortion measure, which expresses an error in the reconstruction domain. Preliminary results indicate that about five bits per frame can be gained with this new measure, without a loss in subjective quality.

1 INTRODUCTION

Accurate quantization of LPC models is very important for the quality of low bitrate speech coders. The objective of quantization of the LPC model is transparent quality, i.e., the quantization does not affect the subjective quality of coded speech. It has been reported that transparent quality is achieved if the average of the Spectral Distortion (SD) between original and quantized models is about 1 dB with not too many outliers [1]. To reach this threshold, 24 bits are needed for split Vector Quantization (VQ) [1], and 20 bits for single stage VQ when the spectral tilt is coded separately [2].

This paper is organized as follows. In section 2 the time-domain interpretation of conventional distortion measures is considered. For small quantization errors, SD belongs to a class of similar distortion measures [3,4] to which also Likelihood Ratio (LR) belongs. In the time domain, LR measures an error in the residual domain. In section 3, we propose a new distortion measure, the Reconstruction Error Distortion (RED) measure, which measures an error in the reconstruction domain. It is based on the squared difference between two infinitely long impulse responses, but it can be exactly computed with an expression, in form similar to the Likelihood Ratio. RED doesn't belong to the mentioned class of similar distortion measures. In section 4, it is shown experimentally that RED compares favourably to SD. Some comments on the results are given in section 5 and the paper is concluded in section 6.

2 CONVENTIONAL DISTORTION MEASURES: TIME DOMAIN INTERPRETATION

The Spectral Distortion and the Likelihood Ratio are two distortion measures which are commonly used in speech coding. They are expressed in the frequency domain as follows:

$$SD^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left| \log \left| \frac{\hat{A}(e^{j\omega})}{A(e^{j\omega})} \right|^2 \right|^2 d\omega \quad , \tag{1}$$

$$LR = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left| \frac{\hat{A}(e^{j\omega})}{A(e^{j\omega})} \right|^2 d\omega \quad . \tag{2}$$

A(z) and $\hat{A}(z)$ are polynomials in the backward time-shift operator z^{-1} with as coefficients the LPC parameters and disturbed LPC parameters, respectively. $|1/A(z)|^2$ models the speech spectral envelope in the frequency domain. Recent work [3,4] has shown that, for small quantization errors, Spectral Distortion can be approximated by a weighted squared distortion measure in the coefficients of any LPC representation:

$$SD^2 \approx 2\Delta\lambda^T \mathbf{W}_\lambda \Delta\lambda \quad , \tag{3}$$

where $\Delta\lambda$ is the vector of coefficient errors and $\mathbf{W}_\lambda$ is a sensitivity matrix. The symbol λ denotes an arbitrary representation of the LPC model, e.g. Log Area Ratios, LPC parameters, Line Spectrum Frequencies, etc.

For the LPC parameter representation, $\mathbf{W}_\lambda$ equals $\mathbf{R}$, the autocorrelation matrix corresponding to the LPC model. Equation (3) then expresses the well-known relation between LR and SD for small errors [5]:

$$SD^2 \approx 2(LR - 1) = 2\Delta\boldsymbol{a}^T \boldsymbol{R} \Delta\boldsymbol{a} \quad . \tag{4}$$

For Autoregressive processes, LR has a simple interpretation in the time-domain. It can be shown [6] that the expectation of the squared difference between the residuals e(n) and $\hat{e}(n)$, computed with the true and disturbed (e.g. quantized) parameters, respectively, is described by LR:

$$\mathcal{E}\{ e(n) - \hat{e}(n) \}^2 = \sigma^2 (LR - 1) \quad , \tag{5}$$

where σ^2 is the variance of e(n).

3 THE RECONSTRUCTION ERROR DISTORTION MEASURE

Consider the analogy between coding the LPC excitation and coding the LPC model. The use of LR and similar distortion measures for coding the LPC model would be analogous to a direct comparison of candidate LPC excitations with the LPC residual signal. However, coding the LPC excitation with Analysis-by-Synthesis, as applied in for example CELP coders, gives much better results, both in terms of SNR and in subjective quality. Therefore, we want to investigate whether an LPC distortion measure can be developed, which is analogous to the Analysis-by-Synthesis procedure for coding the LPC excitation. It is important, however, that the quantized model remains close (in terms of a suitable distortion measure) to the original unquantized model. Otherwise, the model may lose its physical interpretation in terms of formants of speech [6]. This is not beneficial for quality, because then models from consecutive frames may differ much from each other, even in stationary parts of the signal. Furthermore, techniques based on this physical interpretation of LPC models, such as error weighting and postfiltering, may lose their effectiveness if this interpretation is no longer valid.

LPC parameters are obtained with autoregressive estimation methods which remove correlation from the signal: the residual signal is whitened. Therefore, the new distortion measure is developed with the statistical assumption that the LPC residuals have a white spectrum. Suppose a zero mean white noise signal e(n) with variance σ^2 is fed through two synthesis filters $H(z)=1/A(z)$ and $\hat{H}(z)=1/\hat{A}(z)$ and that two signals s(n) and $\hat{s}(n)$ are thus obtained. A(z) may be the model obtained from LPC analysis and $\hat{A}(z)$ a candidate model from a codebook. For this white noise input, the expectation of the squared error between s(n) and $\hat{s}(n)$ is equal to σ^2 times the squared error between the impulse responses h(k) and $\hat{h}(k)$ of H(z) and $\hat{H}(z)$, respectively:

$$\mathcal{E}\left\{ s(n) - \hat{s}(n) \right\}^2 = \sigma^2 \sum_{k=0}^{\infty} \left\{ h(k) - \hat{h}(k) \right\}^2 \quad . \tag{6}$$

Now a new distortion measure will be defined, based on (6), which will be called the Reconstruction Error Distortion measure (RED). It is obtained by normalizing (6) with the expectation of $s^2(n)$, which is σ^2 times the energy of h(k):

$$RED = \frac{\sum_{k=0}^{\infty} \left\{ h(k) - \hat{h}(k) \right\}^2}{\sum_{k=0}^{\infty} h^2(k)} \quad . \tag{7}$$

The denominator is simply the power gain R(0) of the filter $H(z)=1/A(z)$. This new measure constitutes an analogy with the Analysis-by-Synthesis excitation selection.

With Parseval's relation, the infinite sum in (7) can be written in the frequency domain as:

$$\sum_{k=0}^{\infty} \left\{ h(k) - \hat{h}(k) \right\}^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{1}{A(e^{j\omega})} - \frac{1}{\hat{A}(e^{j\omega})} \right|^2 d\omega \quad , \tag{8}$$

which can alternatively be written as:

$$\sum_{k=0}^{\infty} \left\{ h(k) - \hat{h}(k) \right\}^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{A(e^{j\omega}) - \hat{A}(e^{j\omega})}{A(e^{j\omega})\hat{A}(e^{j\omega})} \right|^2 d\omega \quad . \tag{9}$$

Consider the following ARMA process x(n):

$$\left\{ A(z)\hat{A}(z) \right\} x(n) = \left\{ A(z) - \hat{A}(z) \right\} \varepsilon(n) \quad . \tag{10}$$

The integral in the right-hand side of (9) can be recognized as the variance of this ARMA process, divided by the innovation variance σ^2. The variance of an ARMA process can be computed easily by separation of the AR and MA parts [7]:

$$\begin{aligned} \{ A(z)\hat{A}(z) \} v(n) &= \varepsilon(n) \quad , \\ x(n) &= \{ A(z) - \hat{A}(z) \} v(n) \quad , \end{aligned} \tag{11}$$

where v(n) is an AR process. It follows that the infinite sum in (7) can be exactly computed in the time domain, with an expression similar to the expression for the Likelihood Ratio:

$$\sum_{k=0}^{\infty} \left\{ h(k) - \hat{h}(k) \right\}^2 = \Delta \boldsymbol{a}^T \boldsymbol{S} \, \Delta \boldsymbol{a} \quad , \tag{12}$$

where **S** is the covariance matrix corresponding to the process described by $1/\{A(z)\hat{A}(z)\}$, divided by σ^2.

4 COMPARISON OF THE DISTORTION MEASURES

In the first experiment a threshold for transparent quantization was determined in terms of RED. From several speech sentences, LPC models have been estimated and the residual signals have been computed. To the Log Area Ratios of the LPC model a vector of white Gaussian numbers is added, scaled in such a way that RED has a fixed value. The Log Area Ratios have been used for this experiment because a stable model is always guaranteed with this transformation of the LPC parameters. Next, a reconstruction is made by feeding the residual signals through the disturbed synthesis filters. It was found that, for a fixed error of 0.08 in terms of RED, there was almost no audible difference between original and reconstruction.

To make a fair comparison, the same experiment was performed for SD, i.e. the error in the vector of Log Area Ratios was scaled to yield a fixed SD. It was found that a slightly higher error than 1 dB could be tolerated in this

experiment. The level that was obtained this way for Spectral Distortion was about 1.3 dB.

The following experiments have been carried out to estimate the number of bits needed to obtain transparent quantization with RED. From speech data a Single Stage 17 bits codebook was obtained. In other speech data LPC models have been estimated in frames and the residual signals have been computed. The LPC models are vector quantized using either SD or RED as the search measure. Fig. 1 shows an example of how the average of RED and SD behave as a function of the codebook size. Results are the average of 722 frames obtained from three male and three female speakers. Fig. 1(a) shows the average of RED for codebook sizes of 12 to 17 bits. The crosses indicate average RED when RED itself is used as a search measure, the circles indicate average RED when SD is used as a search measure and RED is computed for the models selected by SD. A level of about 0.08 is reached with only 16 bits. Fig. 1(b) shows the average of SD for the same codebooks and speech signals. In this figure the crosses indicate average SD when SD itself is used as the search measure and the circles indicate average SD of the models selected by RED. Clearly, the 1.3 dB level for inaudible errors is not attained. A result from high rate quantization theory can be used to estimate the number of bits needed to obtain an average SD of 1.3 dB. The average of a mean squared error measure decreases exponentially with the number of bits. It has been shown [2] that the average SD can be approximated with the following function:

$$\overline{SD} \approx C\,2^{-b/p} \quad , \qquad (13)$$

where C is a constant depending on the vector dimension and on the distribution of the cepstral coefficients in speech, b is the number of bits and p is the model order which is equal to 10 in this case. A least-squares fit of this function to the curve of Fig. 1(b) yields the value 5.4 for C. The continuous line in Fig. 1(b) represents the fitted curve. This formula predicts that about 21 bits are needed to reach an average SD of 1.3 dB, 5 bits more than the number of bits at which the average RED is 0.08.

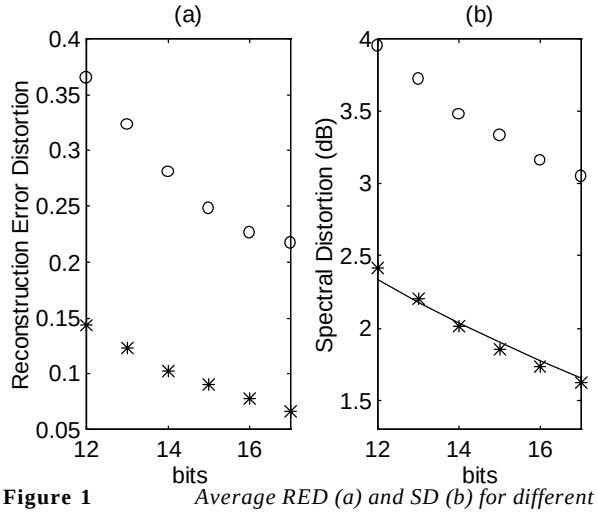

Figure 1 *Average RED (a) and SD (b) for different code book sizes.*

The crosses and circles in these figures are separated by quite a distance. This indicates that RED and SD are really different distortion measures. If LR and SD would have been compared, then the crosses and circles would have been much closer to each other (if LR is transformed properly according to (4)). Fig. 1 indicates that it is possible that models with a good subjective quality have a higher SD than 1 dB. In fact, this can be proven experimentally as follows. Another way to add an error to an LPC model, is by adding white noise to a speech frame. A fixed distortion is obtained between the LPC models from the clean and the noisy speech by appropiately scaling the noise energy in each frame. For a fixed RED of 0.08, it turned out that the average SD was as large as 3.3 dB. In not a single frame SD was smaller than 1 dB. Yet the reconstructions, obtained by feeding the unquantized LPC residuals of the clean speech through the disturbed models from the noisy speech, could not be distinguished from the original clean speech by listening. This means that good models may have a large SD. Therefore, SD sometimes overlooks models in a code book that are subjectively good, but have a large SD.

The reconstruction errors of the signals that are obtained by feeding the LPC residuals through the quantized synthesis filters were much in favour of RED. The reconstruction errors are compared in terms of the Segmental Signal to Noise Ratio (SSNR). When RED was used as a distortion measure, the SSNR was 12.6 dB for the 17 bits codebook. When Spectral Distortion was used, SSNR was only 8.8 dB. Informal listening to the reconstructions revealed a better quality for RED than for SD. However, the quality of the reconstructions was not completely transparent in our experiments; the code book was obtained without training.

The performance of RED in a CELP coder was also better. When the LPC model was not quantized, the average SSNR for 12 sentences was 7.87 dB. When RED was used to select a model from a 12 bits code book, average SSNR was 7.49 dB. When SD was used, average SSNR was 7.28 dB.

5 COMMENTS ON THE RESULTS

Usually, a codebook is obtained by clustering a very large data set into a codebook of the desired size. No clustering at all was performed here and hence the codebook is of rather poor quality. Informal listening to the reconstructions revealed that a slightly better quality was obtained with RED than the quality obtained with SD. The quality of the reconstructions was not transparent, however, because there were too many outliers, i.e., models that could not be quantized with sufficient accuracy with this codebook. These

outliers were clearly noticeable for both distortion measures. It is not known how many outliers can be tolerated with RED or even what defines an outlier for RED. Furthermore, it may be impossible to construct a codebook of 16 bits which ensures both an average RED and a number of outliers small enough for transparent quantization.

The fact that RED achieves a higher SSNR than SD in a CELP coder is very interesting. It shows that, although LPC models are *estimated* by minimizing the prediction gain, best results are *not* obtained by *selecting* a model from a codebook with a high prediction gain (which would have a small Likelihood Ratio). We don't completely understand yet the explanation of this apparent paradox.

RED has been developed by considering the time-domain interpretation of distortion measures. What about the frequency domain interpretation? In SD (1) and LR (2) the *quotient* of the LPC spectra is used. This means that low energy parts of the spectrum are matched with the same *relative* accuracy as high energy parts. It can be seen in (8) that RED is based on the *difference* between the LPC transfer functions, which means that low and high energy parts are matched with the same *absolute* accuracy. This seems a disadvantage of RED, because the human auditory system works roughly on a logarithmic scale. But there is also a disadvantage associated with a relative weighting. If parts of the spectrum are small, these parts are given too much importance by a relative weighting if they are masked; omitting them may even be better. Especially in voiced speech there is more energy at the lower frequencies than at the higher frequencies because voiced spectra typically have a tilt of about 6 dB per octave. Therefore, the difference between Spectral Distortion and RED seems large, particularly in voiced speech. The listening experiments, however, did not reveal a large difference in quality, although the distortions for RED were audible mostly in the high frequencies, and the distortions for SD mostly in the low frequencies.

Two of the weaknesses a distortion measure may have, are the following. Firstly, the distortion measure may select a model from a code book that is good in terms of this distortion measure, but leads to a poor subjective quality. Secondly, the distortion measure may reject models with a high distortion which are subjectively good. In the previous section an experiment was performed, where a fixed error of 0.08 in terms of RED leads to an average SD of 3.3 dB. If these errors would have been audible, this would have meant that RED suffers seriously from the first weakness and it would not be a suitable distortion measure. However, the errors were *not* audible, indicating that SD may suffer from the second weakness. This may be of importance in situations where a coder has to operate under noisy conditions.

In this paper, RED was defined as the normalized squared error between the impulse responses of two LPC synthesis filters. The justification comes from an analogy between coding of the excitation with Analysis-by-Synthesis and coding of the LPC model. However, usually a perceptual weighting filter is applied, which comes down to using a bandwidth expanded synthesis filter for the selection of the excitations. Therefore, RED could alternatively be defined as the normalized squared error between the impulse responses of the bandwidth expanded synthesis filters. Furthermore, the spectral tilt could be coded separately. It is not known at this time if these alternatives will further improve the performance. Using the bandwidth expanded filters has the advantage that the impulse responses are decaying much faster and have essentially died out after a small number of samples, in practice. The sum on the left-hand side of (12) may be truncated to just a small number of terms. Using the truncated sum is computationally less complex than using the right-hand side of (12).

6 CONCLUSIONS

A new distortion measure, the Reconstruction Error Distortion measure, is proposed for the purpose of quantization of the LPC models in low bit rate speech coders. It measures an error in the reconstruction domain, in contrast with conventional distortion measures which express an error in the residual domain. Preliminary experiments indicate that the efficiency of the quantization may improve with five bits per frame with this new distortion measure.

ACKNOWLEDGMENT

This research was supported by the Dutch Technology Foundation under project DTN11.2436.

REFERENCES

[1] K.K. Paliwal and B.S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame", *IEEE Trans. Speech Audio Proc.*, Vol. 1, No. 1, pp. 3-14,1993.

[2] P. Hedelin, "Single Stage Spectral Quantization at 20 Bits", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. I 525-I 528, 1994.

[3] J.S. Erkelens and P.M.T. Broersen, "Equivalent Distortion Measures for Quantisation of LPC Model", *Electron. Lett.*, Vol. 31, No. 17, pp. 1410-1412, 1995.

[4] W.R. Gardner and B.D. Rao, "Theoretical Analysis of the High Rate Vector Quantization of LPC Parameters", *IEEE Trans. Speech Audio Proc.*, Vol. 3, No. 5, pp. 367-381, 1995.

[5] A.H. Gray and J.D. Markel, "Distance Measures for Speech Processing", *IEEE Trans. Acoust., Speech, Signal Process*, Vol. 24, No. 5, pp. 380-391, 1976.

[6] J.S. Erkelens, "Autoregressive Modelling for Speech Coding: Estimation, Interpolation and Quantisation", *Ph.D. Thesis, Delft University of Technology*, 1996.

[7] H.E. Wensink and P.M.T. Broersen, "Practical Aspects of Moving Average Estimation", *Quinzième Colloque sur le Traitement du Signal et des Images - Gretsi*, pp. 201-204, 1995.