SPECTRAL DYNAMICS IS MORE IMPORTANT THAN SPECTRAL DISTORTION

H. Petter Knagenhjelm and W. Bastiaan Kleijn

Speech Coding Research Department
AT&T Bell Laboratories, Murray Hill, NJ 07974, USA

ABSTRACT

Linear prediction coefficients are used to describe the power-spectrum envelope in the majority of low-bit-rate coders The performance of quantizers for the linear-prediction coefficients is generally evaluated in terms of spectral distortion. This paper shows that the audible distortion in low-bit-rate coders is often more a function of the dynamics of the power-spectrum envelope than of the spectral distortion as usually evaluated. Smoothing the evolution of the power-spectrum envelope over time increases the reconstructed speech quality. A reasonable objective is to find the smoothest path that keeps the quantized parameters within the Voronoi regions associated with the transmitted quantization index. We demonstrate increased quantizer performance by such smoothing of the line-spectral frequencies.

1. INTRODUCTION

Linear prediction (LP) is used in the majority of speech coders. Its primary function is the description of the power-spectrum envelope (PSE). For many low-bit-rate coders, the LP coefficients require a significant share (e.g. 50 %) of the overall bit rate. Thus, efficient coding of the LP coefficients is of great practical importance to speech coding and much work has been devoted to increasing quantizer performance, e.g. [1, 2, 3]. A static measure is generally used to evaluate the performance of the quantizers. This measure evaluates the root-mean square (rms) distance between the log-power spectrum corresponding to the original LP coefficients for a frame i, $P_i(\omega)$, and the log-power spectrum corresponding to the quantized LP coefficients, $\hat{P}_i(\omega)$. This distance is

$$\mathrm{SD} = \left(\frac{1}{2\pi}\int_{-\pi}^{\pi}[\ln(P_i(\omega)) - \ln(\hat{P}_i(\omega))]^2 d\omega\right)^{\frac{1}{2}}. \qquad (1)$$

It is commonly accepted that a mean value of 1 dB for SD corresponds to transparent speech quality [1]. However, for a small segment of speech, the mean value of SD is generally not very indicative of the perceived distortion. In fact, a segment with an SD of 1 dB may have low quality and a segment with an SD of 3 dB may have high quality. There are two possible reasons for this. First, the *static* nature of the measure is correct, but the form of eqn. 1 does not model the properties of the human auditory system accurately. The second reason is that the very assumption that a static measure can accurately represent perceived distortion is incorrect, because it ignores the dynamics of the PSE. As will be shown in this paper, the dynamics of the PSE does indeed play a significant role in the perceived distortion. This implies that the efficiency of existing quantizers can be increased significantly, albeit at the expense of additional delay.

It is interesting to note that the static measure can be considered an indirect measure of the dynamics of the reconstructed signal (if conventional quantizers are used). In the conventional interpretation the mean of the static measure determines the mean distance between the quantized and the unquantized PSE. However, because of the high effective dimensionality of the space of the LP coefficients, the mean of the static measure is very similar in value to the mean distance between adjacent quantized spectra. Thus, the mean of the static measure also provides an estimate of the step size between successive, quantized PSE (assuming conventional quantization procedures).

Although the dynamics of the PSE is not considered in current quantization procedures, it is commonly considered in another aspect of LP-based coding. Most low-bit-rate coders have an update rate of the LP coefficients which is between 33 and 100 Hz. To bridge the difference between successive updates, the LP coefficients are generally interpolated on a subframe-by-subframe basis, a subframe being between 2.5 and 7.5 ms in length. Various LP-coefficient representations have been compared for interpolation performance [5, 6, 7]. A good interpolation of the LP coefficients results in a perceptually reasonable evolution between transmitted PSEs. For example, linear interpolation of the line spectral frequencies (LSFs) [8] usually leads to a smoothly evolving PSE, which is desirable. Methods which result in excursions of the PSE are not appropriate. More generally, a good method for LP-coefficient interpolation maintains the original *dynamics* of the PSE. The results obtained with the static distortion measure and LP-coefficient interpolation point towards the importance of the dynamics of the PSE for subjective speech quality. In this paper, it will be shown that proper reconstruction of the PSE dynamics leads to a significant improvement in the performance of LP-coefficient quantizers.

2. PERCEPTUAL DISTORTION

In many speech coders, the LP coefficients are quantized using memoryless quantization about once every 20 to 30 ms. The quantization introduces noise in the parameters which manifests itself as an increased rate of change of the PSE. Because the average distance between adjacent

0-7803-2431-5/95 $4.00

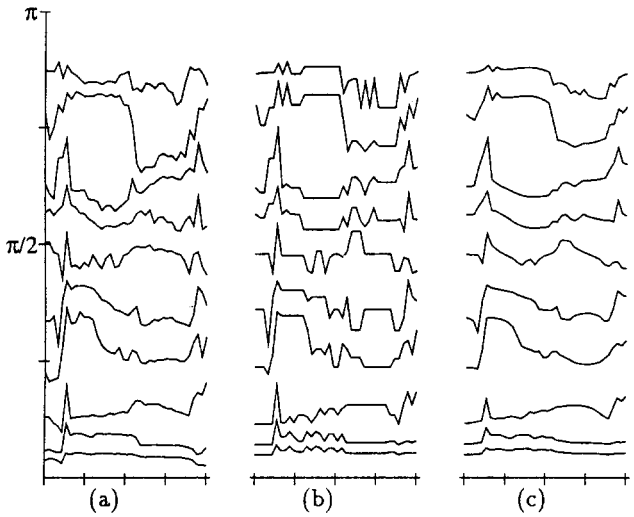


Figure 1: LSF trajectories for the speech segment in fig. 4 a) unquantized (b) quantized and (c) smoothed LSFs.

sets of quantized LP coefficients decreases with increasing quantizer performance, this increase in the rate of change is smaller for better quantizers. Thus, a static performance measure has strong correlation with the rate of change of the PSE.

A plot of the spectral distortion as a function of time, typically shows peaks with a magnitude of many times the mean of SD. Often however, speech segments of high subjective distortion have a low SD. Similarly, speech segments of low subjective distortion often have a high SD. High subjective quality in spite of high SD usually corresponds to regions of speech with rapid changes of the PSE. It can be argued that the quantization noise is masked by the rapid change of the PSE. From a large number of informal tests, we concluded that speech segments with a low SD measure are, in fact, the main source of subjective distortion caused by existing LP-coefficient quantizers. Typically this type of distortion occurs in vowels of long duration, where the PSE is relatively constant. This is consistent with the notion that biological receptory systems are sensitive to small changes in an otherwise steady-state situation.

The LSF are commonly used for quantization, and have desirable interpolation properties. They are a good low-dimensional representation of the PSE. For example, when the PSE is relatively constant, the LSF are relatively constant as well. In the following we will often use the LSF representation for the PSE, but other good representations of the spectrum are expected to lead to similar results.

Estimation errors in the LP analysis will introduce some noise in the estimated PSE. One reason for the estimation errors is nonpitch-synchronous analysis. A typical trajectory (for the word "dune") of the LSF is shown in fig. 1a (the associated waveform is shown in fig. 4a). The analysis was performed every 20 ms. Re-analysis of the signal with a 10 ms offset maintains the general shape of the trajectory, but with different local variations.

Fig. 1b shows the same LSF trajectories after quantization. The quantization results in increased variations of the PSE. When an original parameter (e.g. an LSF) is close to a border of the space partitioning, a Voronoi facet, small parameter variations are likely to cause the quantizer to switch between indices of neighboring centroids. An example of this effect is clearly visible for the 9th LSF in figs. 1a and 1b.

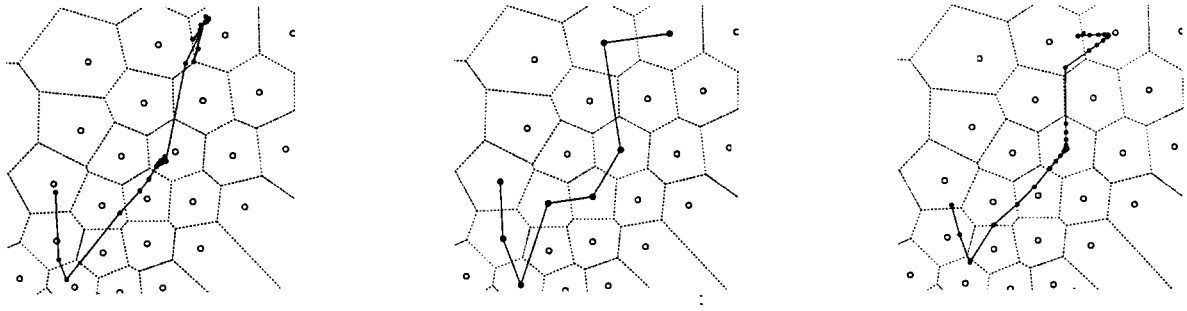
In high resolution quantizers, switching between neighboring centroids will result in small changes in the PSE of the reconstructed speech. However, for coarse quantizers the switching between neighboring centroids often results in relatively large changes in the PSE, and thus in perceived distortion. With a conventional decoding approach, the only solution to this problem is to use high-resolution quantizers. However, the realization that the incorrect rate of change of the PSE rather than the absolute error of the PSE cause much of the subjective distortion suggests that more efficient decoding procedures may exist. To this purpose, the next section explores smoothing of the evolution of the PSE.

3. SPECTRAL EVOLUTION SMOOTHING

Since the PSE dynamics is important to speech quality, it must be considered carefully in the design of a speech coder. To counteract the increase in the rate of change of the PSE caused by the quantization process, a PSE smoothing process must be used. This smoothing process should exploit both perception and the properties of the quantizer. During the quantization process, each original PSE is mapped into a quantized PSE, which corresponds to the centroid of a Voronoi region in the parameter domain [9]. All unquantized parameters within a Voronoi region are mapped to the centroid. Thus, when a certain quantization index is used for reconstruction, it is known that the original parameter was located within the Voronoi region associated with the centroid corresponding to the index. A proper smoothing procedure constrains the reconstructed parameters to fall within the same Voronoi region as the original parameter.

We explored a number of possibilities for smoothing the PSE at the decoder. As a first implementation, we used straightforward low-pass filtering of the differential LSF. The apparent disadvantage of this method is that the formants, particularly formants at higher frequencies, may be displaced from their original locations. However, it is found experimentally that this displacement is not of perceptual significance, while the resulting spectral evolution smoothing results in improved speech quality. In general, we found that low-pass filtering of the differential LSF improved the reconstructed-speech quality in regions where the original PSE changes slowly. This result confirmed that the effect of quantization on the dynamics of the PSE is important.

The filtering procedure does not satisfy the constraint that the reconstructed parameters should fall within the same Voronoi region as that of the original PSE. This was noticeable for rapid onsets, which are smoothed in an undesirable manner by filtering. Whereas filtering improved the subjective speech quality in steady-state regions, it decreased the quality for transitions. To prevent such problems we constructed a method which performs smoothing under the constraint that the original and reconstructed PSE fall within the same Voronoi region.



a) Original b) Quantized c) Smoothed

Figure 2: Space partitioning (Voronoi regions), centroids and LSF trajectories in the $LSF_1 - LSF_2$ plane for a 2-3-5 split VQ using 6 bits in each block. Observe that the quantized trajectory contains multiple coinciding points.

3.1. Smoothing Implementation

Let us again consider the LSF trajectory for the word "dune". Figure 2a shows the trajectory of the original LSF in the $LSF_1 - LSF_2$ plane, for a 2-3-5 split VQ. The figure also shows the centroids and the corresponding Voronoi regions of the quantizer. The corresponding quantized trajectory is shown in fig. 2b. Many of the steps between successive LSF are significantly larger in the quantized case than for the original case. In this section, we will describe a procedure to smooth the quantized trajectory, subject to the constraint that the LSF remain within the same Voronoi region. The result of this method is shown in fig. 2c.

Our smoothing procedure is motivated by a physical equivalent. Each parameter is represented as a vector (two-dimensional in the case of fig. 2). For the reconstructed parameter (RP), we will call these vectors $\mathbf{r}_i$, where i is the time index. In the physical equivalent, each of the RP is considered to represent a point-like object which is subject to forces. These forces are defined such that, in equilibrium, *i*) the distances between adjacent $\mathbf{r}_i$ are small (ensuring a smooth trajectory), and *ii*) the constraint that each point remains within the Voronoi region is reasonably well satisfied.

The attractive force between two RP is set to be proportional to the distance between the RP. Let $\mathbf{F}_{i,i+1}$ be the force on $\mathbf{r}_i$ from $\mathbf{r}_{i+1}$. This force is $\mathbf{F}_{i,i+1} = \gamma \cdot \frac{\mathbf{r}_{i+1}-\mathbf{r}_i}{|\mathbf{r}_{i+1}-\mathbf{r}_i|}$ Each parameter i is also subject to a force pulling towards the centroid. A weak force, α, is present if the RP is inside the Voronoi region. This ensures that the RP moves towards the centroid if no neighboring RP are within another Voronoi region. The centroid force is strong, β, if the RP is outside the Voronoi region. The relative magnitudes are: $\alpha < \gamma \ll \beta$. We approximate the Voronoi region by the largest inscribed hypersphere centered at the centroid, and define its radius as $\mathrm{R_{max}}$. Thus, the centroid force is: $\mathbf{F}_{i,c} = k \cdot \frac{\mathbf{y}_c-\mathbf{r}_i}{|\mathbf{y}_c-\mathbf{r}_i|}$, where $\mathbf{y}_c$ is the centroid vector, and where $k = \begin{cases} \alpha & \text{if } |\mathbf{y}_c - \mathbf{r}_i| < \mathrm{R_{max}} \\ \beta & \text{otherwise} \end{cases}$

The overall force operating on each RP is simply the sum of all forces: $\mathbf{F}_i = \mathbf{F}_{i-1,i} + \mathbf{F}_{i+1,i} + \mathbf{F}_{i,c}$ An example of the forces is shown in fig. 3.

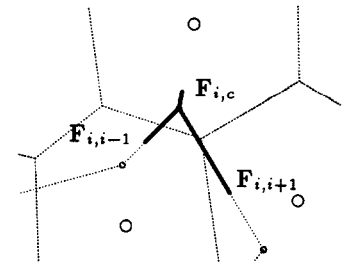


Figure 3: Forces on the i:th reconstruction vector

The equilibrium situation is obtained by means of an iterative procedure. The procedure moves each RP once per iterative loop. For each RP move, the overall force is evaluated and the RP moved in the direction of the force, over a distance proportional to the strength of the force. The procedure was found to converge rapidly.

In our current implementation, no delay constraints were applied. However, it appears that most of the advantages of the method can be maintained with look-ahead of only one RP. In a real-time speech coder this implies that the method may require an added delay of 20 to 25 ms.

4. EXPERIMENT

To demonstrate the effects of the smoothing procedure described above, we quantized the time signal shown in fig. 4a using a coarse quantizer. The LP-residual was computed using the unquantized LP coefficients, and the speech signal was reconstructed using the quantized LP coefficients. The LP update rate was 50 Hz and the LP coefficients were interpolated in the LSF domain using 5 ms subframes. To evaluate the spectral evolution we measured the spectral steps,

$$\Delta\mathrm{PSE} = \left(\frac{1}{2\pi}\int_{-\pi}^{\pi}[\ln(P_{i+1}(\omega) - \ln(P_i(\omega))]^2 d\omega\right)^{\frac{1}{2}}. \quad (2)$$

The spectral steps before and after quantization are shown in figs. 4b and 4c. The spectral steps after quantization can be seen to mimic the steps of the original in transient regions, but are significantly larger in the steady-state regions. The mean spectral step over the utterance is 2.2 dB and 2.9 dB for the unquantized and quantized PSE, respectively. The spectral distortion due to quantization is 2.2 dB. Filtering of the LSF (using a 4-tap FIR filter with cut-off frequency of 12.5 Hz) is shown in fig. 4c. The performance is enhanced in the steady state regions, but at the cost of smearing out regions with large spectral steps. The result of the smoothing procedure described in section 3.1 is shown in fig. 4d. It is clearly shown that the stepsize is essentially preserved in the transition region while the stepsize is very small in the steady-state region. The slightly smaller step size than that observed in the original is the result of removal of all small variations. As was mentioned before, these variations in the original LSF may, in fact, be caused by estimation errors.

The results are further illustrated in fig. 1. Figure 1a shows the dynamics of the original LSF parameters (in radians), $LSF_i, i = 1 \ldots 10$, whereas fig. 1b shows the behavior of the same set of LSF after quantization with a 15-bit split-VQ quantizer [1]. The quantizer has a 3-3-4 split and an equal number of bits for each block. It is evident from the figures that the rate of change of the LSF is increased by the quantization process. It is this rate of change that the smoothing must eliminate. Perceptually most important in fig. 1b is the evolution over time of the three first coefficients LSF_1, LSF_2 and LSF_3, which represent a low-frequency formant. The coefficients are close and noisy, which causes the formant to vary both in frequency and bandwidth. Figure 1c shows the effect of the smoothing with $\alpha = 0.04$, $\gamma = 1$, and $\beta = 4.0$. The resulting LSF trajectories match the original in fig. 1a very well, considering that they are derived form the LSFs of fig. 1b.

The proposed approach was evaluated for various split-VQ configurations using 15 to 24 bits on a number of utterances. The use of spectral evolution smoothing resulted in a significant improvement of the subjective quality in steady state regions. Moreover, no occasions could be found where the smoothing degraded transitions.

In certain cases the improvements was also visible on graphically displayed speech signals. Using an unsmoothed, coarse quantizer can lead to excursions of the filter gain. When this occurs for the dominant formants, the energy contour of the output signal becomes ragged. The spectral evolution smoothing removed these visible quantization artifacts.

5. CONCLUSION

The dynamics of the power-spectrum envelope (PSE) is very important to perception. An LP-coefficient quantizer results in a higher rate of change of the PSE. Such a higher rate of changes is often audible in signal segments where the original PSE changes slowly. Experimental evidence shows that a perceptually motivated smoothing of the quantized PSE leads to a significant increase in subjective performance. This will be increasingly important as the number of bits used for describing the PSE is reduced.

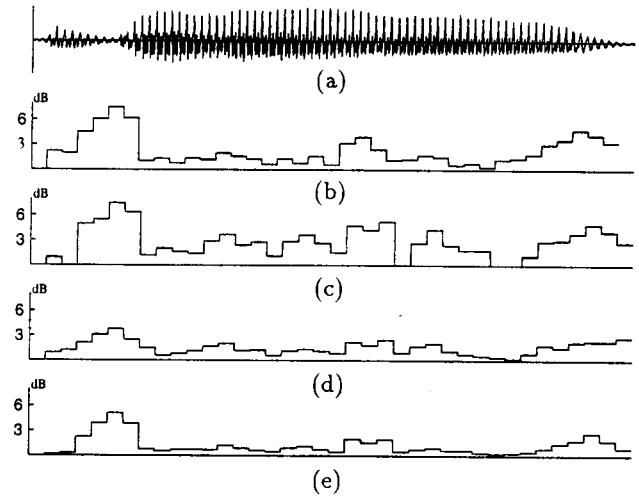


Figure 4: (a) The acoustic waveform (0.76 s) for utterance /du:n/. Spectral steps of adjacent frames for (b) unquantized, (c) quantized, (d) filtered and (e) smoothed LSFs.

6. REFERENCES

[1] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 1, pp. 3–14, 1993.

[2] W. F. LeBlanc, B. Bhattacharya, S. A. Mahmoud, and V. Cuperman, "Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kb/s speech coding," *IEEE Trans. Speech and Audio Process.*, vol. 1, no. 4, pp. 373–385, 1993.

[3] R. Hagen and P. Hedelin, "Robust vector quantization in speech coding," *Proc. Int. Conf. Acoust., Speech, Sign. Process.*, pp. 13–16, 1993.

[4] B. C. J. Moore, *An Introduction of the Psychology of Hearing*, Academic Press, London, 1989.

[5] T. Umezaki and F. Itakura, "Analysis of time fluctuating characteristics of linear predictive coefficients," *Proc. Int. Conf. Acoust., Speech, Sign. Process.*, pp. 1257–1261, 1986.

[6] B. S. Atal, R. V. Cox, and P. Kroon, "Spectral quantization and interpolation for CELP coders," *Proc. Int. Conf. Acoust. Speech, Sign. Process.*, pp. 69–72, 1989.

[7] J. S. Erkelens and P. M. T. Broersen, "Analysis of spectral interpolation with weighting dependent on frame energy," *Proc. Int. Conf. Acoust., Speech, Sign. Process.*, pp. I481–I484, 1994.

[8] F. Itakura, "Line spectrum representation of linear predictive coefficients," *J. Acoust. Soc. Am.*, vol. 57 supp., no. 1, p. S35, 1975.

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