INCORPORATING PERCEPTION INTO LSF QUANTIZATION - SOME EXPERIMENTS

Ronald P. Cohn and John S. Collura

U.S. Department of Defense, 9800 Savage Road STE 6516, Ft Meade MD 20755-6516
E-mail: rpcohn@alpha.ncsc.mil, jscollu@alpha.ncsc.mil

ABSTRACT

In the context of vector quantization (VQ) of the line spectrum frequency (LSF) parameters, we determine experimentally a spectral distribution of quantization error perceived to be "balanced", i.e., error at all frequencies contributing equally, on average, to the perceived distortion. Quantizers which have a balanced distribution should outperform those which don't, given the same number of bits. We examine the spectral error distributions produced by various weighted Euclidean distance measures in the LSF domain and develop one which produces a quantizer having an approximately balanced distribution. This quantizer's performance is compared with that of others having different error distributions.

1. INTRODUCTION

The short term speech spectrum contains important information and is therefore quantized and transmitted by many speech coding algorithms. This information is often determined by linear prediction and can be represented by many spectral parameter sets such as predictor coefficients, log area ratios, reflection coefficients, LSF's, etc. Accurate representation of the spectrum using as few bits as possible is desired. Prior research has shown that VQ of the LSF parameters is a relatively efficient method for minimizing the number of bits needed to represent the speech spectrum. An example is the new 2400 bps MELP coder which uses a 25 bit, multi-stage vector quantizer (MSVQ) to encode 10 LSF coefficients [1].

While LSF VQ's perform well enough to be used in real-time coders, numerous aspects of their design and evaluation could benefit from additional research. One of these is how to make better use of human perceptual characteristics which cause decreased sensitivity to spectral error as the error frequency increases. Some VQ work has used this property in a quantizer performance measure [1][2]. However, we are unaware of any research to determine the amount of spectral error needed for audibility as a function of frequency.

This paper describes experiments which investigate relationships among spectral error, perception, and weighted Euclidean distance measures. We determine experimentally a spectral error distribution that is approximately "balanced", i.e., error at all frequencies contributing equally, on average, to the perceived distortion. Balanced spectral error is desirable since it should produce the least audible distortion for a given number of bits. We then design and evaluate a series of MSVQ's to verify that these quantizers can be designed to have the desired error characteristics and to determine how many bits are needed for spectral quantization to be inaudible.

1.1 LSF VQ Background

The speech derived from a quantized spectrum should sound as similar as possible to speech derived from an unquantized spectrum, with the goal being no audible difference. A common measure of spectral distortion for a single spectrum is the RMS error, D, in dB [3]:

$$D^2 = \frac{1}{\pi}\int_0^{\pi} [10\log_{10}P(\omega) - 10\log_{10}\hat{P}(\omega)]^2 d\omega , \qquad (1)$$

where ω is the radian frequency, and $P(\omega)$ and $\hat{P}(\omega)$ are the linear prediction power spectra:

$$P(\omega) = \left|1 + \sum_{k=1}^{10} a_k e^{-j\omega k}\right|^{-2}, \ \hat{P}(\omega) = \left|1 + \sum_{k=1}^{10} \hat{a}_k e^{-j\omega k}\right|^{-2}, \qquad (2)$$

where a_k and $\hat{a}_k$ are the linear prediction coefficients corresponding to the unquantized and quantized LSF's, respectively. This measure does not fully account for perceptual characteristics, since it gives equal weight to error regardless of where it occurs in frequency.

To utilize the frequency-dependent perceptual property, some research has examined a weighted distortion measure in which the spectral error is given a frequency-dependent weighting [1][2]. This distortion is the Bark-weighted RMS error, D_B, in dB:

$$D_B^2 = \frac{1}{\pi W_0}\int_0^{\pi} W_B(\omega)^2 [10\log_{10}P(\omega) - 10\log_{10}\hat{P}(\omega)]^2 d\omega , \qquad (3)$$

where W_0 normalizes $W_B(\omega)/W_0$ to unity RMS. $W_B(\omega)$ is a weighting function based on the derivative of the Bark scale:

$$W_B(\omega) = \frac{1}{25 + 75\left(1 + 1.4\left(\frac{F_s\,\omega}{2000\pi}\right)^2\right)^{0.69}}, \qquad (4)$$

where F_s is the sampling frequency in Hz. $W_B(\omega)$ is shown in Figure 3. D_B has greater correlation with subjective evaluation than does the unweighted measure of

Eq. (1), presumably because D_B down-weights higher frequencies [1][2]. A weighted distortion measure such as Eq. (3) should be used in the VQ evaluation *and* design processes. A quantizer for which a weighted distortion such as Eq. (3) is minimized will have a spectral error distribution roughly proportional to the inverse of that weighting; a balanced distribution is desired.

Unfortunately, the evaluation of distortion given by Eqs. (1) or (3) is too computationally complex for VQ use in a real-time speech coder. Instead, a weighted Euclidean distance (WED) is used to approximate the distortion between a given LSF vector and each of the quantizer's code book vectors. A WED, d, has the form:

$$d^2(\boldsymbol{f}, \hat{\boldsymbol{f}}) = (\boldsymbol{f} - \hat{\boldsymbol{f}})^T \boldsymbol{W} (\boldsymbol{f} - \hat{\boldsymbol{f}}), \quad (5)$$

where $\boldsymbol{f}$ and $\hat{\boldsymbol{f}}$ are column vectors of the original and quantized LSF vectors and $\boldsymbol{W}$ is a diagonal weighting matrix which may depend on $\boldsymbol{f}$. Ideally, the WED should be proportional to perceived distortion.

While use of a WED (rather than spectral distortion) for VQ design and search generally means that spectral distortion is not minimized, recent research has found weights for which the distortion, D, given by Eq. (1), is equal to the WED for small distances, i.e.,

$$D^2 = (\boldsymbol{f} - \hat{\boldsymbol{f}})^T \boldsymbol{W} (\boldsymbol{f} - \hat{\boldsymbol{f}}), \quad (6)$$

when cubic and higher terms can be neglected [4]. These weights depend on $\boldsymbol{f}$, but their calculation is simple enough for real-time use. We call these the Gardner weights, after the principal investigator. Use of the Gardner weights and choosing $\hat{\boldsymbol{f}}$ to minimize Eq. (5) therefore minimizes the spectral distortion given by Eq. (1), if the quantizer's average distortion is sufficiently small. Quantizers for speech coders usually meet this requirement. A relationship such as Eq. (6) has not been found for weighted distortion measures such as Eq. (3).

Paliwal and Atal reported good VQ performance using a WED with weights determined experimentally [3]:

$$w_i = \{c_i [P(f_i)]^{0.15}\}^2, \quad 1 \le i \le 10, \quad (7)$$

where $P(f_i)$ is the unquantized linear prediction power spectrum at the frequency of the i^{th} LSF component. The c_i help account for the decreased sensitivity to spectral error at higher frequencies and are fixed:

$$c_i = \begin{cases} 1.0, & 1 \le i \le 8 \\ 0.8, & i = 9 \\ 0.4, & i = 10 \end{cases}. \quad (8)$$

2. PERCEPTUAL EXPERIMENTS

Several experiments were conducted to determine the spectral error distribution produced by the above distance metrics and to estimate the spectral error distribution perceived to be balanced.

2.1 The Experimental Procedure

The following procedure was used to incorporate LSF quantization into a speech coding context. An input speech signal was high pass filtered with a 4^{th} order Chebychev Type II filter, having a 60 Hz cutoff frequency, and 30 dB stopband attenuation. The resulting signal was analyzed every 20 msec using 10^{th} order linear prediction, and the prediction residual was calculated. The LSF's were perturbed by adding a simulated error vector; if the perturbed LSF vector was invalid, the error vector was scaled by 0.9 until the LSF vector was valid. The LSF's were then adjusted for a minimum spacing of 50 Hz, if necessary. An output signal was reconstructed using the prediction residual and a synthesis filter corresponding to the perturbed LSF's. The linear prediction used the autocorrelation method with a 25 msec Hamming window, and a 0.994 bandwidth expansion factor; an LSF component spacing of 50 Hz was enforced. For a given distance d, quantizer error was simulated by an error vector, $\boldsymbol{e} = \boldsymbol{f} - \hat{\boldsymbol{f}}$, having a uniform distribution inside a hyper-ellipsoid volume defined by $d^2 \ge \boldsymbol{e}^T \boldsymbol{W} \boldsymbol{e}$. The test speech signal came from eight males and eight females speaking two sentences each, yielding 32 sentences for evaluation. The output speech was evaluated subjectively to determine the effects of the LSF errors.

The RMS distortion was calculated by evaluating the original and distorted spectra given by Eq. (2) for 1024 discrete frequencies, converting the distortion to dB at each frequency, and then taking the RMS value of the distortion at each frequency over all measured spectra. Spectra corresponding to silence were excluded.

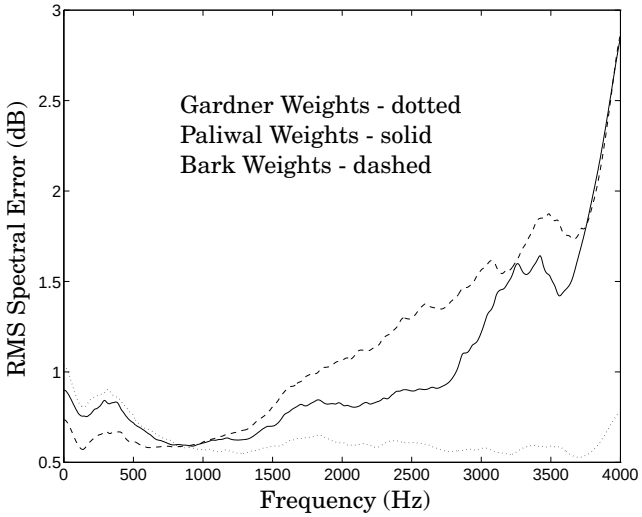

Figure 1. Just Audible Spectral Error Distributions

2.2 Distribution Produced by the Gardner Weights

The experimental procedure described above was followed using the Gardner weights. An initial run was done with a small distortion level, about 0.5 dB, and the output speech was evaluated for audible distortion. The distortion level was increased and the process was repeated until the distortion was audible on about half of the 32 test sentences. The RMS spectral distortion at this just

audible level is plotted in Figure 1. This curve is nearly flat, except for the rise below 500 Hz. A flat curve is expected, since the Gardner weights minimize a uniformly weighted distortion, Eq. (1). The distortion produced by the Gardner weights was audible only in the low frequencies, hence this is not a balanced distribution.

2.3 Distribution Produced by the Paliwal Weights

This experiment was identical to the prior one except for the use of the Paliwal weights. The RMS spectral distortion at the just audible level is plotted in Figure 1. The distortion produced by the Paliwal weights was audible only in the low frequencies.

2.4 Distribution Produced by the Bark Weights

In similar fashion, another experiment was conducted using weights calculated to minimize the Bark-weighted spectral distortion given by Eq. (3). There is no analytic expression for these weights; they were determined numerically. For a given LSF vector, the weight for each component was determined by adding a 4 Hz error, e, to that component, measuring the resulting distortion, D_B, by evaluating Eq. (3) using a 2048 point FFT, and by presuming $w_i = D_B^2/e^2$. This was repeated for error in the opposite direction, and the two weights were averaged. The RMS spectral distortion at the just audible level is plotted in Figure 1. As expected, this curve resembles the inverse of the Bark weighting curve given by Eq. (3). Distortion was audible in the middle and upper frequencies, but not in the lower frequencies.

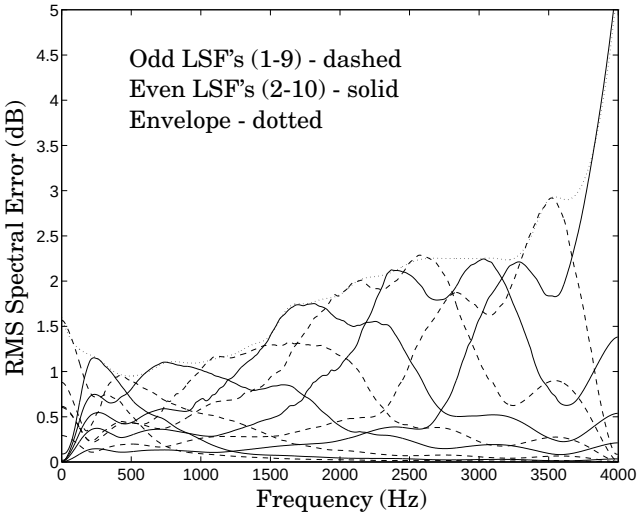

Figure 2. Just Audible Spectral Error for Each LSF

2.5 Perturbation of Individual LSF's

Since none of the above weighting methods produced a balanced spectral error distribution, another experiment was conducted to help determine this distribution. This experiment was similar to the others, except that error was added to the LSF vector one component at a time. The error magnitude for the i^{th} LSF component was determined by Eq. (6) to be $\left|f_i - \hat{f}_i\right| = D/w_i^{0.5}$, where D is the desired distortion, and w_i is the Gardner weight. The error sign was chosen randomly. Error was added to the first LSF component to produce a desired distortion level, initially set to be just audible to a single listener. It was then increased or decreased by 0.05 dB until the spectral error was audible to a small panel of listeners for about half of the 32 test sentences. This process was repeated for all LSF components. The RMS distortion at these just audible levels is shown in Figure 2. The corresponding distortion settings, D, in dB, for LSF components 1 to 10 were 0.35, 0.35, 0.4, 0.6, 0.75, 0.95, 1.2, 1.2, 1.2, and 1.4, respectively.

Figure 2 also shows the envelope of these curves. This envelope is roughly the just audible distortion as a function of frequency. The envelope also represents an approximately balanced spectral error distribution, if interaction among distortion components can be ignored. The inverse of this envelope, which we denote $W_E(\omega)$, is a reasonable alternative to the Bark weighting given by Eq. (4). These functions, normalized to an average value of 1, are shown in Figure 3.

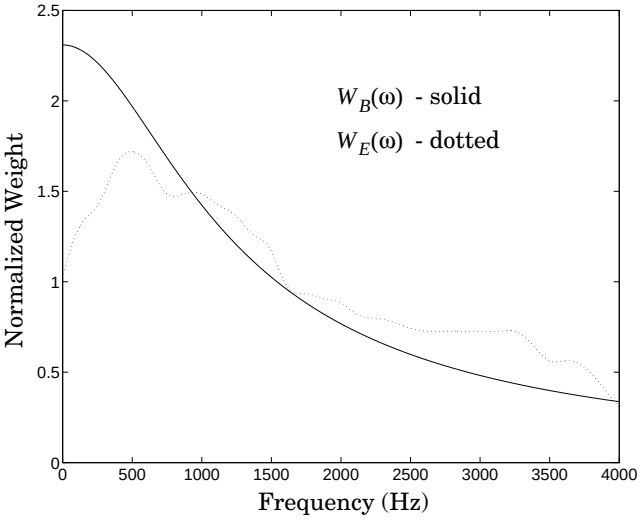

Figure 3. Bark and Experimental Weight Functions

3. PERCEPTUALLY RELATED WED'S

We now need to determine a weighted Euclidean distance which is related to a given weighted distortion function. The analytical approach is to follow that in [4], but to start from a weighted distortion, rather than the unweighted distortion in Eq. (1). An alternative is to determine the weights by direct measurement of the sensitivity as described in Sec. 2.4, but this method involves too much computation for real-time use. We pursued a third option in which the weight for the i^{th} LSF component is the product of the Gardner weight, G_i , and a correction factor determined by the weighting function, $W(\omega)$, selected for use in the distortion measure, e.g., in Eq. (3), $W(\omega) = W_B(\omega)$. These weights are given by:

$$w_i = W(f_i)^2 G_i \tag{9}$$

where f_i is the value of the i^{th} LSF component. For the

$W_B(\omega)$ and $W_E(\omega)$ functions, we observed good correlation between measured weighted distortion and WED's using weights determined by Eq. (9) in experiments using simulated quantization error.

4. QUANTIZER PERFORMANCE EVALUATION

Next, we conducted experiments which compared the performance of actual MSVQ's trained and searched using WED's whose weights were specified by: 1) Paliwal, Eq. (7); 2) Eq. (9), using $W_B(\omega)$; and Eq. (9), using $W_E(\omega)$. The procedure described in Sec. 2.1 was followed, using the actual quantizer rather than simulated error.

4.1 MSVQ Training

The training vectors were generated using the linear prediction technique described in Sec. 2.1, with speech selected for diverse spectral content. An example of how this diversification was achieved is through the use of equal gender content, multiple languages and acoustic background environments, etc. In order to ensure a random distribution of the training data, the inclusion of adjacent analysis vectors was prohibited, and once a vector was selected, there was only a 40 percent probability of selecting each successive vector. In addition to these constraints on data selection, vectors with a low energy were also discarded. There were approximately 4 million vectors in the training database prior to these constraints, yielding approximately 1 million vectors for training. An agglomeration-based algorithm was used for the selection of seed points followed by a generalized Lloyd iteration for the generation of an initial code book. The unweighted Euclidean distance was used. Given these initial code books, additional code books were then trained using the joint optimization procedure with a search depth of M=8 as described in [5].

4.2 Evaluation Procedure and Results

For each of the three WED's, the 32 test sentences were quantized and compared with the input sentences through a double-blind A-B test. The ordering was random, and pairs were presented in both orders. Listeners were told that one of the pairs was input and the other was processed; they were instructed to identify the input.

Unfortunately, our MSVQ training program was not working in time to meet the publication deadline. We sincerely apologize for this and regret any inconvenience to the audience. Complete results and a revised copy of this paper will be available at our poster session.

To obtain preliminary results, the 32 test sentences were quantized using the initial 28 bit codebook, searched to minimize the three WED's described above. The resulting spectral error distributions are shown in Figure 4. Each curve has the expected shape. The quantized sentences were evaluated by five listeners using the A-B test described above. The input sentence was misidentified 10.6%, 10.6%, and 15.0% of the time for the Paliwal, Bark, and experimental WED's, respectively. A higher score means that the quantization is less audible, so the experimental WED performed the best in this test. We expect that the experimental WED will also perform better than the Paliwal WED for quantizer training.

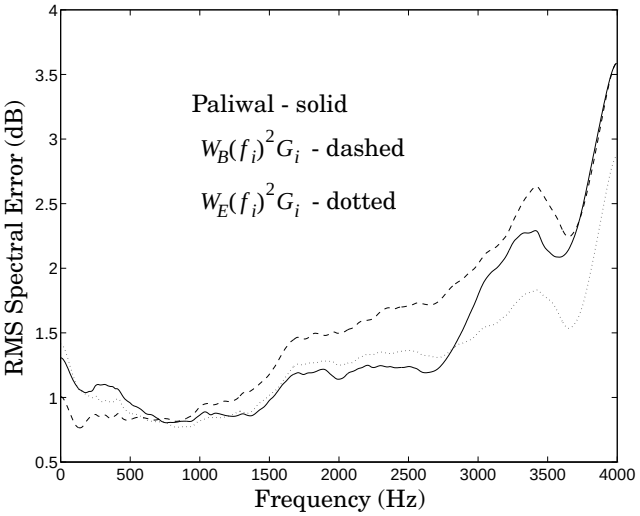

Figure 4. Quantizer Spectral Error Distributions

5. CONCLUSIONS

We have shown that several weighting schemes do not produce a balanced error distribution. We have experimentally determined a spectral error distribution which is approximately balanced. We will show that a VQ can have a balanced error distribution. We will determine whether or not the experimental WED outperforms the Paliwal WED and how many bits are needed for inaudible quantization.

6. REFERENCES

[1] A. McCree, K. Truong, E. B. George, T. P. Barnwell, V. Viswanathan, "A 2.4 kbit/s MELP Coder Candidate for the New U.S. Federal Standard," *Proceedings of IEEE ICASSP 1996*, pp. 200-203.

[2] J. S. Collura, A. V. McCree, T. E. Tremain, "Perceptually Based Distortion Measurements For Spectrum Quantization," *Proceedings of the 1995 IEEE Workshop on Speech Coding for Telecommunications*, Annapolis, MD.

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

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

[5] W. P. 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 Transactions on Speech and Audio Processing*, Vol. 1, No. 4, October 1993, pp. 373-385.