A Vector Quantizer Incorporating Both LPC Shape and Energy

L. R. Rabiner
M. M. Sondhi
S. E. Levinson

Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

The theory of vector quantization (VQ) of linear predictive coding (LPC) coefficients has established a wide variety of techniques for quantizing LPC spectral shape to minimize overall spectral distortion. Such vector quantizers have been widely used in the areas of speech coding and speech recognition. The conventional vector quantizer utilizes only spectral shape information and essentially disregards the energy or gain term associated with the optimal LPC fit to the signal being modelled. In this paper we present a method of incorporating LPC spectral shape and energy into the codebook entries of the vector quantizer. To do this we postulate a distortion measure for comparing two LPC vectors which uses a weighted sum of an LPC shape distortion and a log energy distortion. Based on this combined distortion measure we have designed and studied vector quantizers of several sizes for use in isolated word speech recognition experiments. We have found that a fairly significant correlation exists between LPC shape and signal energy; hence a combined LPC shape plus energy vector quantizer with a given distortion requires far fewer codebook entries than one in which LPC shape and energy are quantized separately. Based on isolated word recognition tests on both a 10-digit and a 129 word airlines vocabulary, we have found improvements in recognition accuracy by using the VQ with both LPC shape and energy over that obtained using a VQ with LPC shape alone.

I. Introduction

The idea of quantizing LPC coefficient sets using a vector quantizer, rather than a scalar quantizer, has been studied for several years [1-4]. The "standard" VQ algorithm essentially quantizes the spectral shape of the LPC vector to one of M^* codebook entries, where M^* represents the number of LPC prototype vectors needed to span the space of LPC vectors with a given distortion criterion. This type of vector quantizer disregards the gain or energy associated with the LPC vector; instead it codes only the spectral shape. For LPC vocoder applications the gain of the signal is generally coded independently of the LPC spectral shape; this effectively assumes independence of spectral shape and signal gain. For recognition applications there has been virtually no use of the signal gain information in the standard implementations of isolated word recognition systems [3-5].

Recently Brown and Rabiner [6] were able to show improved performance for an LPC dynamic time warping (DTW) word recognition system by incorporating gain information into the conventional distortion measure. Their results indicated a substantial reduction in word error rates on a moderate size (129) vocabulary of words as used in an airlines reservation and information system.

In this paper we extend the work of Brown and Rabiner by showing how the gain information can be incorporated into the vector quantizer design algorithm yielding a set of codebook entries with both spectral shape and gain information.

II. Review of the LPC Shape VQ Design Algorithm

Assume we are given a section of a speech signal, $s(n)$ $n = 0,1,\ldots,N-1$, with z-transform $S(z)$. From this we derive the pth order LPC model, $\hat{S}(z)$, of the form:

$$\hat{S}(z) = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{1}$$

where $\mathbf{a}' = \{1, a_1, a_2, \ldots, a_p\}$ is the optimal pth order LPC model and G is the model gain. It is readily shown that G can be written in the form

$$G = \sqrt{\sum_n e^2(n)} = \sqrt{a' V a} \tag{2}$$

where $e(n)$ is the error between the true speech samples $s(n)$, and the predicted speech samples (i.e. those obtained from the model) $\hat{s}(n)$, and V is the Toeplitz autocorrelation matrix of the actual speech signal, with the first row given by

$$V(m) = \sum_n s(n)\, s(n+m), \quad m = 0,1,\ldots,p \tag{3}$$

The zeroth autocorrelation coefficient, $V(0)$, is conventionally called the signal energy.

If one wants to compare two LPC models, e.g. $A_T(z)$ and $A_R(z)$, of the forms

$$A_T(z) = \frac{G_T}{1 - \sum_{k=1}^{p} a_k^T z^{-k}} \tag{4a}$$

$$A_R(z) = \frac{G_R}{1 - \sum_{k=1}^{p} a_k^R z^{-k}} \tag{4b}$$

several related LPC distortion measures (distance metrics) have been proposed including:

1. The Itakura-Saito measure of the form

$$d_{IS}(A_T, A_R) = \left(\frac{\mathbf{a}_R' \mathbf{V}_T \mathbf{a}_R}{\mathbf{a}_T' \mathbf{V}_T \mathbf{a}_T} - 1 \right) + \ln \left(\frac{G_T^2}{G_R^2} \right) \tag{5}$$

2. The log likelihood measure of the form

$$d_{LLR}(A_T, A_R) = \ln \left(\frac{\mathbf{a}_R' \mathbf{V}_T \mathbf{a}_R}{\mathbf{a}_T' \mathbf{V}_T \mathbf{a}_T} \right) \tag{6}$$

3. The gain normalized measure of the form

$$d_{GN}(A_T, A_R) = \left(\frac{\mathbf{a}_R' \mathbf{V}_T \mathbf{a}_R}{\mathbf{a}_T' \mathbf{V}_T \mathbf{a}_T} - 1 \right) \tag{7}$$

It is readily seen that d_{LLR} and d_{GN} are essentially identical for values of d close to 0, and differ primarily for large values of d. It is also readily seen that both d_{LLR} and d_{GN} are independent of signal energy since the only term in the expressions of Eqs. (6) and (7) which depends on signal energy is V_T which appears in both the numerator and the denominator; hence it is cancelled out. The d_{IS} measure of Eq. (5) has a signal energy dependent term (G_T^2/G_R^2) and hence contains some energy information; however experimentation by several researchers [2,3] indicates that d_{LLR} and d_{GN} are much better for designing a VQ than d_{IS}. As such in our own work [4] we have used d_{GN} exclusively.

CH1945-5/84/0000-0179 $1.00

Using the distortion measure of Eq. (7) one can define a distortion (distance) between a training LPC vector ($\mathbf{a}_T$) and a VQ codebook vector ($\tilde{\mathbf{a}}_R$). Hence one can devise an algorithm for choosing a set of M^* codebook vectors, $\tilde{\mathbf{a}}_R$, which minimize the distortion of a set of training vectors from the codebook entries, i.e.

$$\bar{D}_I(M^*) = \min_{\{\tilde{\mathbf{a}}_R\}} \left[\frac{1}{I} \sum_{L=1}^{I} \min_{1 \leq m \leq M^*} \left[d_{\mathrm{GN}}(\mathbf{a}_T, \mathbf{a}_R) \right] \right] \tag{8}$$

where we have simplified the distance notation of Eq. (7) to represent distance between a test and reference LPC vector (rather than a test and reference LPC model). Various iterative algorithms for implementing the minimization of Eq. (8) have been proposed and shown to work quite well over a wide range of conditions [1,2]. The optimum codebooks (which we will call spectral shape codebooks) are generated by a method similar to the K-means algorithm. Starting with an initial guess of M^* entries, each LPC vector of the training set is assigned to the closest entry. The centroids of the M^* subsets (clusters) obtained in this manner are used as new trial entries in the codebook, and the iteration continued until some stopping criterion is satisfied. Generally initial solutions for any desired value of M^* are obtained by first obtaining solutions for smaller values of M^* and then splitting some or all codebook entries. In this manner one can start from a value of $M^* = 1$, where the solution is simply the centroid of the training set vectors, and generate codebooks for higher values of M^* by either splitting every cluster [1], or one cluster at a time.

III. Modifications of the Distortion Measure to Include Signal Energy

If we denote any of the LPC shape distortions of Eqs. (6)-(7) as $d_{\mathrm{LPC}}(T,R)$, then a fairly simple and straightforward way of including signal energy in the overall distortion is to form the sum:

$$d(T,R) = d_{\mathrm{LPC}}(T,R) + \alpha f(d_E(T,R)) \tag{9}$$

where $d_E(T,R)$ is an energy distortion, $f(x)$ is a nonlinearity applied to the energy distortion, and α is a multiplicative factor on the energy distortions. If we denote the (unnormalized) test energy as E_T, and the (unnormalized) reference energy as E_R, then

$$E_T = 10 \log_{10}(V_T(0)) \tag{10a}$$

$$E_R = 10 \log_{10}(V_R(0)) \tag{10b}$$

A normalized energy ($\hat{E}_T, \hat{E}_R$) can be defined by making all energy values relative to a local peak energy (e.g., for isolated word recognition we make it relative to the peak energy within a word) Thus

$$\hat{E}_T = E_T - (E_T)_{\mathrm{MAX}} \tag{11a}$$

$$\hat{E}_R = E_R - (E_R)_{\mathrm{MAX}} \tag{11b}$$

and an energy distortion, d_E, can then be defined as

$$d_E(T,R) = |\hat{E}_T - \hat{E}_R| \tag{12}$$

The nonlinearity, $f(x)$, is used to give smaller weight to small energy distortions. The form we have used is

$$f(x) = \begin{cases} 0 & |x| \leq \mathrm{CLIP} \\ x & |x| > \mathrm{CLIP} \end{cases} \tag{13}$$

where CLIP is a threshold chosen by appropriate experimentation.

The combined LPC shape plus energy distortion of Eq. (9) has the property that as α is made small, the properties approach those of the LPC shape distance, and as α is made large, the properties approach those of the energy distortion alone.

3.1 Application of the Combined Distortion Measure to VQ

It is straightforward to use the combined distortion measure of Eq. (9) in the VQ design algorithm of Section II. The resulting VQ codebook vectors are then characterized by an LPC vector, along with a mean normalized log energy value. To understand some of the properties of the codebook vectors, a simple set of experiments was carried out on a set of 10,000 frames of speech derived from spoken isolated words of a 129 word vocabulary of airlines terms. The single words were spoken by 100 different talkers (50 male and 50 female) over a standard dialed-up telephone line.

Figure 1 shows an energy histogram of the 10,000 frames of speech. The peak level of the normalized energy of any frame is, by definition, 0 dB and a dynamic range of about 60 dB for energy can be seen in this figure. The first experiment used the training set to design a conventional LPC shape VQ (i.e. α was set to 0 in Eq. (9)). A VQ with $M^* = 16$ was designed and each of the 10,000 training vectors was assigned to one of the $M^* = 16$ codebook entries. After convergence to the best set of codebook vectors, energy histograms of

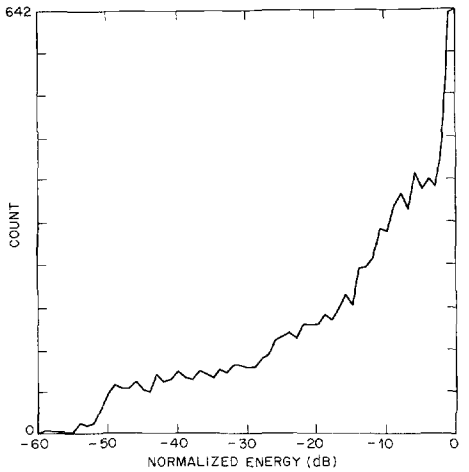

Fig. 1 Energy histogram of 10,000 frames of speech from isolated words.

each of the 16 subsets of training vectors were made, and the results are shown in Figure 2. If signal energy and LPC shape were totally independent, we would expect each of the 16 energy histograms of Fig. 2 to be essentially identical. This is clearly not the case as some of the energy histograms are peaked at near 0 dB (i.e. strong vowels), other histograms are peaked near −50 dB (i.e. silence, weak fricatives), and still other histograms peak somewhere between these upper and lower limits.

The energy histograms of Figure 2 indicate a fairly high degree of correlation between LPC spectral shape and normalized signal energy; hence one would *a priori* expect that a VQ designed from the combined distortion measure would be more efficient than using a separate VQ for LPC shape and a separate quantizer for energy.

A second set of experiments was run on the 10,000 vector training set in which the average LPC distortion, $\bar{d}_{\mathrm{LPC}}$, was determined as a function of M^* (the VQ size) for the case of $\alpha = 0$ (no energy in the distortion measure), and similarly the average energy distortion $\bar{d}_E$, was determined as a function of M^* for the case of $\alpha = \infty$ (no LPC in the distortion measure) and CLIP = 0. The results obtained are given in Table I. The last column of Table I gives the value of α^* such that

$$\alpha^* \bar{d}_E = \bar{d}_{\mathrm{LPC}} \tag{14}$$

i.e. the value of α (as a function of M^*) such that the average distortions due to energy and LPC would be equal. The results in Table I show that the average LPC distortion decreases slowly as M^*

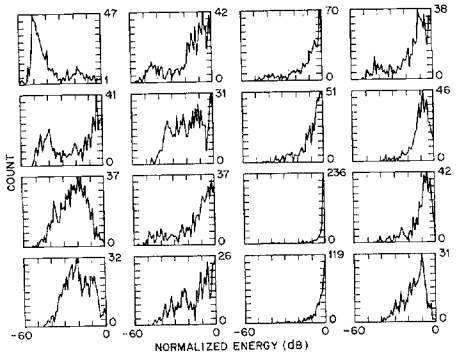

Fig. 2 Energy histograms of the 16 codewords of an $M^* = 16$ shape LPC vector quantizer.

M^*	$\bar{d}_{LPC}$	$\bar{d}_E$	α^*
2	.784	5.88	.13
4	.579	2.97	.19
8	.428	1.47	.29
16	.317	.75	.42
32	.218	.37	.59
64	.196	.19	1.0
128	.149	.09	1.65

TABLE I
Average distortions as a function of M^*

increases, whereas the average energy distortion essentially halves with each doubling of M^*. The halving of the average distortion with each doubling of the size of the quantizer for a scalar variable is a well known effect for scalar quantizers. The last column in Table I shows that α^* increases dramatically as M^* increases. Hence for small values of M^*, α values have to be very small or the VQ essentially becomes an energy quantizer; for larger values of M^* the value of α is not overly important since the LPC distortion dominates.

Based on the above discussion, combined LPC shape plus energy VQ's were designed for 3 sets of conditions, namely:

1. $\alpha = 0.1$, CLIP = 0
2. $\alpha = 0.3$, CLIP = 0
3. $\alpha = 0.3$, CLIP = 6 (dB)

and the results $(\bar{d}_{LPC}, \bar{d}_E)$ as a function of M^* are shown in Table II. For the first set of conditions the resulting VQ achieves compromise values of the shape and energy distortions. For example, when $M^* = 64$ the LPC shape average distortion is comparable to that of an $M^* = 16$ VQ based on LPC shape alone (see Table I), and the energy average distortion is comparable to that of an $M^* \cong 10$ VQ based on energy alone. When α is raised to 0.3 (the second set of conditions), the energy distortion is lowered at the expense of increased LPC shape distortion for a given M^*. Thus for $M^* = 64$ the LPC shape average distortion is now comparable to that of an $M^* \cong 7$ LPC shape VQ, and the energy average distortion to that of an $M^* \cong 18$ energy VQ. When a reasonable clipping threshold is used (CLIP=6), the influence of the energy distortion term is significantly reduced since all vectors within CLIP dB of each other (with similar LPC shapes) contribute zero energy distance. Hence for $M^* = 64$ the LPC average distortion is comparable to that of an $M^* \cong 29$ LPC shape VQ, and the energy average distortion is comparable to that of an $M^* >> 128$ energy VQ.

The general trend of the results of Table II is that for small values of M^* the combined VQ tries to reduce energy distortion at the expense of LPC distortion, while at larger values of M^* the VQ primarily reduces LPC distortion. Since energy is correlated with LPC shape (and vice versa), a reduction in one distortion will always bring about a reduction in the other distortion. Figure 3 shows a series of

	$\alpha = 0.1$ CLIP = 0		$\alpha = 0.3$ CLIP = 0		$\alpha = 0.3$ CLIP = 6	
M^*	$\bar{d}_{LPC}$	$\bar{d}_E$	$\bar{d}_{LPC}$	$\bar{d}_E$	$\bar{d}_{LPC}$	$\bar{d}_E$
2	.93	8.11	1.28	5.96	1.28	4.40
4	.73	5.22	1.18	3.22	1.20	.99
8	.62	3.29	.79	2.37	.78	.14
16	.49	2.34	.70	1.29	.50	.04
32	.38	1.80	.55	.97	.36	.01
64	.31	1.30	.45	.69	.26	.01
128	.26	.98	.36	.49	.21	.008

TABLE II
Average distortions, as a function of M^*

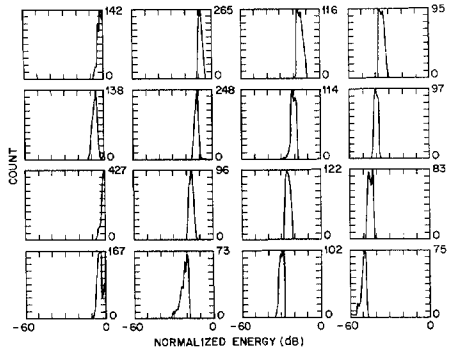

Fig. 3 Energy histograms of the 16 codewords of an $M^* = 16$ combined LPC shape plus energy vector quantizer designed with $\alpha = 0.3$ and CLIP = 0 dB.

energy histograms of the condition 2 VQ (α=0.3, CLIP=0) for the training subsets of the $M^* = 16$ case. The effects of using energy in the combined distortion measure are clearly seen in that each histogram is tight around some average energy for the codeword.

IV. Application of the Combined VQ to Word Recognition

To further evaluate the effectiveness of combining energy plus LPC shape in the VQ, a series of isolated word recognition tests was carried out using the hidden Markov model (HMM) recognition algorithm described in Ref. [4]. In this system an LPC analysis of each speech frame is carried out, and each LPC vector is vector quantized. For each word in the vocabulary an HMM is designed using a training set of VQ outputs for the word. In normal usage each word HMM is scored using a Viterbi algorithm which computes the probability of the sequence of VQ outputs having come from the specified word HMM. The word model with the highest probability score is declared to be the spoken word.

The standard HMM word recognizer can be trivially modified to use the combined LPC plus energy VQ. The only change is in the quantization of the training set and of the unknown test. New HMM models are computed for the combined VQ and the standard scoring algorithm is still used in the recognizer.

4.1 Recognition Results on Digits Vocabulary

For the 10-digits vocabulary a training set of 100 tokens of each digit spoken once by each of 100 different talkers (50 male, 50 female) was used. The test set consisted of a separate set of 100 tokens of each digit spoken once by the same 100 talkers. The test recordings were made about 1 month after the training recordings. All words were recorded over dialed-up telephone lines.

Two sets of VQ parameters were used in the recognition system, namely:

1. $\alpha = 0$ (No energy)

2. $\alpha = 0.1$, CLIP = 6 dB.

For each set of VQ parameters a set of HMM parameters were computed for each word model for the following conditions:

$M^* = 32, 64, 128, 256$

N = Number of states in Markov Model = 5, 8, 10

The results of the recognition tests (in terms of digit error rates) are given in Table III. The baseline LPC shape VQ (Table IIIa) has error rates of from 3 to 6%, depending on N and M^*. Generally the larger the value of M^*, the lower the error rate for the recognizer. No strong dependence on N is seen in these results. The results of Table IIIb show about a 1% reduction in error rate for the recognizers using the combined LPC shape plus energy VQ, for values of M^* of 128 and 256. For the smaller values of M^* (32 and 64) there is no consistent improvement in accuracy with the combined VQ.

4.2 Recognition Results on the 129 Word Airlines Vocabulary

The second recognition test of the combined VQ used a 129 word airlines system vocabulary. The training set again consisted of 100 tokens of each vocabulary word spoken once by each of 100 talkers (50 male, 50 female) over dialed-up telephone lines. The test set was a set of 20 tokens of each word spoken once by each of 20 new talkers (i.e. not included in the training set), again over dialed-up telephone lines.

Four recognition tests were performed under the following conditions:

1. A standard dynamic time warping (DTW) LPC based recognizer without VQ.
2. A DTW recognizer with an LPC shape VQ using $M^* = 128$.
3. An HMM recognizer with an LPC shape VQ using $M^* = 256$, with $N = 10$ states in each Markov model.
4. An HMM recognizer with the combined LPC shape and energy VQ using $M^* = 128$, with $N = 10$ states in each Markov model.

For each test the average word error rate γ of the recognizer was measured as a function of the best β candidates. An error rate of γ for the best β candidates means the correct word was *not* in the β top recognition choices of the system γ% of the time. Results for values of β from 1 (conventional word error rate) to 6 were measured and are shown in Figure 4. The results show that the DTW recognizer without VQ performed the best; however the HMM recognizer with a combined VQ of size $M^* = 128$ performed essentially identically to the DTW recognizer with an LPC shape VQ of size $M^* = 128$, and significantly better than the HMM recognizer with an LPC shape VQ of size $M^* = 256$. Hence the results strongly suggest that energy is a powerful discriminator for polysyllabic vocabularies and in conjunction with an LPC shape VQ of moderate size (equivalently M^* for the shape is about 32) provides better performance than a significantly larger VQ based on LPC shape alone.

V. Summary

Our results with the combined LPC shape and energy VQ indicate that the addition of energy directly into the VQ design algorithm provides an efficient method of incorporating energy constraints into an isolated word recognition system. Tests with isolated word recognizers indicate improved performance using the combined VQ over that for the LPC shape alone, provided that a sufficiently large VQ size, M^*, is used (i.e. $M^* \geq 128$).

Our results on the combined VQ algorithm indicate an improved efficiency of quantization by combining LPC shape and energy into a single VQ, over that obtained for separate LPC shape and energy quantizers. Hence our results could be applied also to LPC voice coding with reductions in bit rate required to achieve desired levels of quantization of the LPC vector and gain terms.

	M^*			
N	32	64	128	256
5	6.0	3.9	3.8	3.8
8	6.1	3.9	4.1	4.0
10	4.8	3.1	3.0	3.0

(a) Digit error rates (%) for $\alpha = 0$ (No energy)

	M^*			
N	32	64	128	256
5	5.7	4.5	2.8	2.9
8	4.1	3.0	2.7	3.2
10	5.4	4.8	2.2	2.2

(b) Digit error rates (%) for $\alpha = 0.1$, CLIP = 6 dB

TABLE III
Digit error rate scores for different values of N and M^*

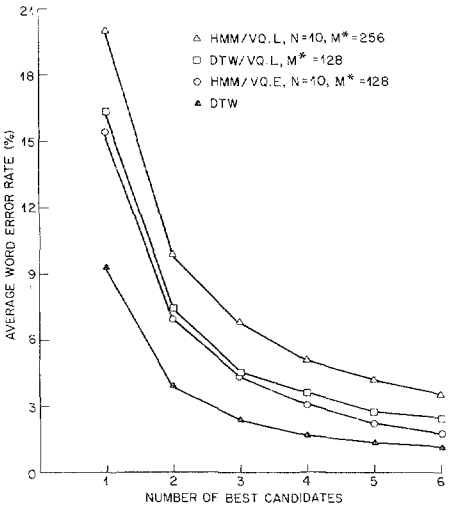

Fig. 4 Average word error rate (%)

References

[1] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantization," *IEEE Trans. Communications*, Vol. COM-28, No. 1, pp. 84-95, January 1980.

[2] B. Juang, D. Wong, and A. H. Gray, Jr., "Distortion Performance of Vector Quantization for LPC Voice Coding," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-30, No. 2, pp. 294-303, April 1982.

[3] J. E. Shore and D. Burton, "Discrete Utterance Speech Recognition Without Time Normalization," *Proceedings ICASSP-82*, pp. 907-910, May 1982.

[4] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker Independent, Isolated word Recognition," *Bell System Tech. Journal*, Vol. 62, No. 4, pp. 1075-1105, April 1983.

[5] L. R. Rabiner and S. E. Levinson, "Isolated and Connected Word Recognition — Theory and Selected Applications," *IEEE Trans. on Communications*, Vol. COM-29, No. 5, pp. 621-659, May 1981.

[6] M. K. Brown and L. R. Rabiner, "On the Use of Energy in LPC-Based Recognition of Isolated Words," *Bell System Tech. J.*, Vol. 61, No. 10, pp. 2971-2987, Dec. 1982.