Proc. Biennial Symp. Commun. (Kingston, ON), pp. C.3.8–C.3.11, June 1986

Tree Coding in a Code Excited Linear Predictive Speech Encoder

CHUNG C. CHU AND PETER KABAL

Abstract

Delayed decision coding together with noise masking theory is used in a code excited linear predictive speech coding system. The quantizer design is based on a modified (M, L)-algorithm which implements a multi-path tree encoder. Prediction errors are coded in blocks of N samples. The performance of the coding system as a function of M, L, N, and an additional parameter is given. High objective and subjective performance is obtained at low bit rates.

The use of a delayed decision multi-path tree encoder as a quantizer in a Code Excited Linear Predictive (CELP) speech coder is studied [1]. Input speech is first passed through a time varying linear 12^{th} order formant and 3 tap pitch predictor filter to remove near- and far-sample correlations. The predictor coefficients and other parameters (gain and pitch lag) are sent to the receiver as side information. The quantizer then encodes the prediction residual signals in blocks of N samples. The synthesis filter in the receiver consists of the corresponding inverse prediction filters. The transmission rate for the coding is 2 kb/s. This leaves 2.8 kb/s for coding the side information if a total bit rate of 4.8 kb/s is to be obtained.

1. Quantizer Design

The quantizer is implemented using a modified (M, L)-algorithm [2][3][4]. In the algorithm used in [3][4], F branches radiate from each node available in response to a block of residual samples to form paths L branches long. Each new branch is associated with a codeword. A maximum of M paths are first selected from the set of paths whose number is upper bounded by MF. The index of the first branch of the optimal path is transmitted to the receiver. The output decision is made L blocks after each input block is considered. Only a subset of the chosen paths originating from the optimal branch is kept. In the approach used here, the optimal path is chosen first, then a maximum of M paths are selected among the paths originating from the optimal branch. The number of paths kept in the modified scheme is closer to M than that in the original approach.

Each entry of the codebook is composed of N samples generated by a Gaussian random number generator. The address of the codeword for a branch populated at time i is computed according to the formula

$$A = b_{i-(N_s-1)}F^{N_s-1} + \cdots + b_{i-2}F^2 + b_{i-1}F + b_i \tag{1}$$

where $b_{i-l} \in [0, F-1]$ for $l = 0, 1, ..., N_s - 1$ is the index of the branch forming the path at time $i - l$.

2. Distortion Measures

In addition to the artificial delays, a spectral weighting is used in the quantizer to help find the approximations giving rise to the perceptually best reconstructed signals. The block of candidate output samples on a branch is pitch and formant synthesized using the inverse prediction filters. Then the quantizer shapes the difference between the original and the synthesized signal on each branch using a frequency weighting filter $W(z) = \frac{1-F(z)}{1-F(\gamma^{-1}z)}$ where $0 \leq \gamma \leq 1$ and $F(z) = \sum_{k=1}^{12} a_k z^{-k}$ is the formant predictor. In this study the value of γ is 0.75. Mean-squared values of the spectrally weighted reproduced noise on all branches forming a path are added. The accumulated mean-squared errors associated with all paths are compared to find the minimum. Mathematically, the optimal delayed decision is made for the i^{th} input block according to the

C. C. Chu is with the Department of Electrical Engineering, McGill University, Montreal, Quebec, H3A 2A7

P. Kabal is with the Department of Electrical Engineering, McGill University, Montreal, Quebec, H3A 2A7 and INRS-Télécommunications, Université du Québec, Verdun, Quebec, H3E 1H6.

cumulative error measure

$$E_{\text{opt}} = \min_{j}\Big[\sum_{k=i}^{i+L-1} e_j^2(k)\Big] \qquad 0 \le j \le g-1 \tag{2}$$

with g being the total number of candidate output blocks available for the $(i+L-1)^{st}$ input block. As a result of the minimization, the codeword at time i with the best subjective short time spectrum with respect to the short time spectral envelope of input speech is obtained.

3. Computer Simulation

Two high quality speech sentences spoken by a female and a male speaker are used as inputs to the system. These two sentences are (1) Cats and dogs each hate the other (CATF8) and (2) It's easy to tell the depth of a well (WELLM8), respectively. The formant predictor coefficients are computed using the autocorrelation method. New sets of coefficients are computed for successive frames of 80 samples which correspond to 10 ms at an 8 kHz sampling frequency. The 3 tap pitch predictor coefficients are calculated using a covariance method. For each subframe of 40 samples inside a frame, a set of 3 optimal coefficients is obtained. The update rate is 5 ms. Each subframe of prediction residual samples is further sub-divided into integral number of blocks of N samples.

In the simulation, residual samples in a subframe is normalized by the standard deviation of that subframe to get unit variance to match the statistics of the codes from the random number generator.

Because this study is concerned with the performance of a predictive coding system using a tree quantizer for the residual signals, quantization errors due to side information quantization are not considered. All time varying side information including predictor coefficients, gain factors and pitch frequency is sent to the receiver directly without quantization. A noiseless channel is also assumed.

4. Performance of the System

The objective and subjective performance as a function of N_s, M, L, and N is given in this section. The segmental signal-to-noise ratio (segSNR), which is the averaged value of log scaled signal-to-noise ratios (SNR) of small speech segments—typically 16 ms long, is used to show the objective performance of this system. Informal subjective listening tests were performed in sound proof conditions.

4.1 Performance as a function of N_s

The objective performance of the system as a function of N_s is shown in Fig. 1. The objective performance increases with the value of N_s and then saturates as the value of N_s becomes large. Subjectively, each reconstructed sentence has very high fidelity. Distortion in the forms of clicks on the words "cats" and "hate" are detected. In CATF8, high frequency background noise appears when $N_s = 1$. According to Eq. (1), if the value of N_s is large, most of the addresses of the candidate codewords for a quantizer input block are different. The codes in use are destructured. Better performance is expected with a large value of N_s because a quantizer with less structured codes performs well [2]. A logical explanation for the saturation rests on the branching factor F and the maximum number of paths kept M. The maximum number of branches allowed on a tree at time i is known to be MF and the maximum number of codewords considered according to Eq. (1) is F^{N_s}. It is then obvious that the codewords associated with all branches at time i are different if $F^{N_s} \ge MF$. Maximum randomness is achieved and the performance reaches a constant level if equality holds.

4.2 Performance as a function of M

As the value of M increases by one under the upper bound F^{L-1}, one more path is kept ,and one more node is available on the tree for the next input block. It is clear that F more codewords will be considered in response to the new input block. If the values of N_s is large enough, this additional set of F codewords will be different from the codewords already selected. The overall performance of the system is expected to increase. This expectation is verified by the plot in Fig. 2 which shows the objective performance of the system as a function of M. Subjective quality is very high with no background noise. Also as M increases, some of those codewords with bad short term performance will be kept and considered in the future with new codewords. Figure 2 confirms that such consideration of long term effects does benefit the encoding system.

The plot also shows that the objective performance saturates long before M reaches its maximum value allowed. A full search in a delayed decision encoder with a finite value of L is not worthwhile. Saturation occurs because the value of L is finite. According to Eq. (2), cumulative errors on the paths cannot be averaged out in too long a time span. Bad codewords are retained but rejected very soon.

4.3 Performance as a function of L

It is understood that a short term optimal decision may not be good for the long term performance of the system when these short term decisions are combined together to reconstruct speech signals. The main effect of delayed decision due to L is that quantization error is averaged out and considered over a period of LN samples. Another advantage of delayed decision is the feasibility of multi-path encoding. In short, the delay allows the selection of long term optimal and revocable approximations to quantizer inputs from a large set of choices. The objective performance of the system as shown in Fig. 3 is increasing with L. The relative perceptive quality is reflected by the segSNR also.

4.4 Performance as a Function of Block Size N

According to rate distortion theory, quantization noise decreases as the quantization block size increases for a fixed encoding rate. The objective performance of the system as a function of N is plotted in Fig. 4.(a) and 4.(b). The experiments were performed with the maximum number of codewords available in a codebook upper bounded by 1024 for all block sizes. As expected, the objective performance increases with the quantization block size N. A dramatic increase in performance with an increase in N is found when $(M, L, N_s) = (1, 1, 1)$. When compared to the original sentences, utterances in all reconstructed signals are less loud and sharp. However, except for $(M, L, N_s, N) = (1, 1, 1, 4)$ and $(1, 1, 1, 8)$ all cases are highly intelligible and natural. Subjective quality for $(M, L, N_s, N) = (1, 1, 1, 4)$ is unacceptable. Severe distortion on each word makes the speech very unnatural. When $N = 8$, distortion appears in the reconstructed CATF8 as a background noise instead of as a degradation on the words.

Analyses show that computational complexity and memory increase exponentially with N in both the tree quantizer and its special case the codebook quantizer. In general, the tree quantizer with nonunity M and L is more complex than a codebook quantizer using the same block size. However Fig. 4.(a) and 4.(b) show that with small block size N, the tree quantizer outperforms a single path codebook quantizer. Moreover, the system with the delayed decision multi-path tree quantizer characterized by $(M, L, N_s, N) = (16, 27, 6, 8)$ is much simpler than that with a codebook quantizer characterized by $(1, 1, 1, 40)$[†]. The tree quantizer when characterized by such a combination is superior to the codebook quantizer used in the original CELP design.

5. Conclusion

The CELP system using a delayed decision multi-path tree quantizer implemented by a modified (M, L)-algorithm can code speeches at low bit rates with high fidelity. The tree quantizer is found to be superior to a simple codebook quantizer in terms of computational complexity, memory requirements and output noise.

References

1. M.R. Schroeder and B.S. Atal, "Code Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates", *Proc. Int. Conf. Acoust., Speech, Signal Proc., Vol. 3, Paper 25.1, Mar. 1985*
2. J.B. Anderson and J.B. Bodie, "Tree Encoding of Speech", *IEEE Trans. Information Theory, Vol. IT-21, No. 4, pp. 379-387, July 1975*
3. N.S. Jayant and S.A. Christensen, "Tree-Encoding of Speech Using the (M, L)-Algorithm and Adaptive Quantization", *IEEE Trans. Comm., Vol. COM-26, pp. 1376-1379, Sept. 1978*
4. T. Svendsen, "Tree Encoding of the LPC Residual", *Proc. Int. Conf. Acoust., Speech, Signal Proc., Vol. 1, pp. 10.11.1-10.11.4, Mar. 1984*
5. C.C. Chu, "Tree Encoding of Speech Signals at Low Bit Rates", *M.Eng Thesis, Department of Electrical Engineering, McGill University, Montreal, 1986*

[†] This point has been verified by analyses of the computational complexity and memory requirements. See [5] for more details.

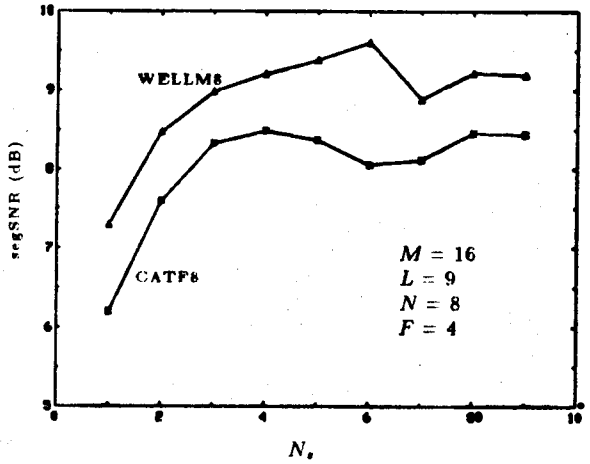


Fig. 1 Objective performance of the system as a function of N_s

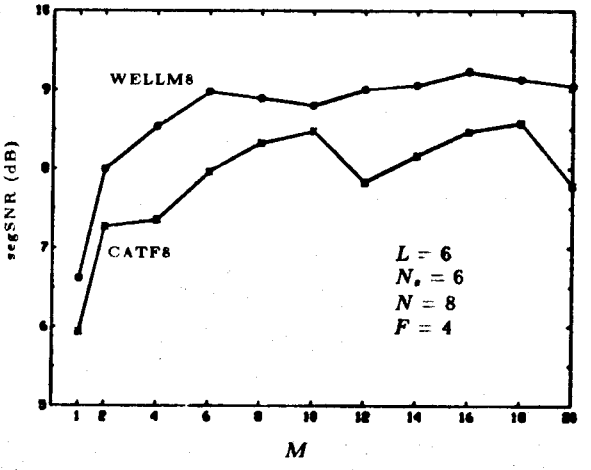


Fig. 2 Objective performance of the system as a function of M (when $M = 1$, $L = 1$)

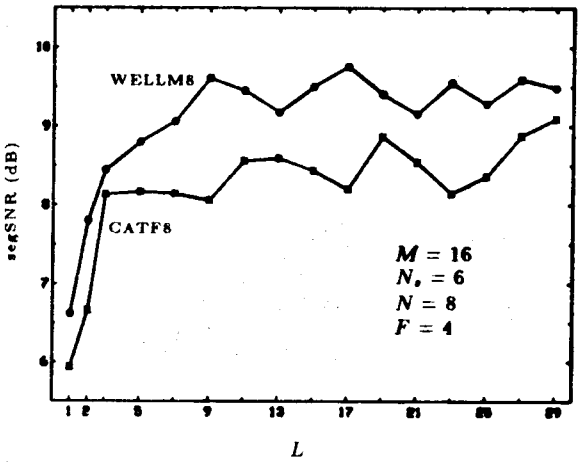
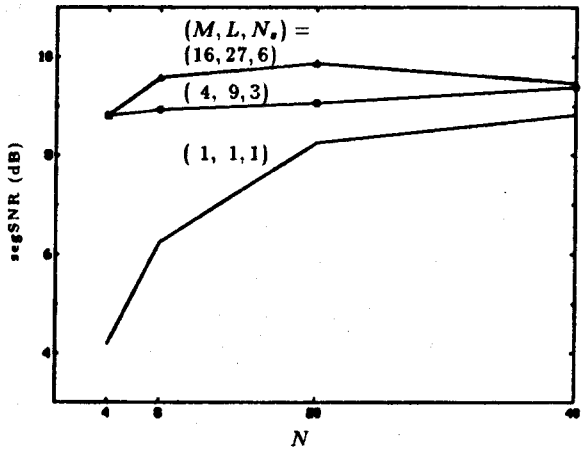
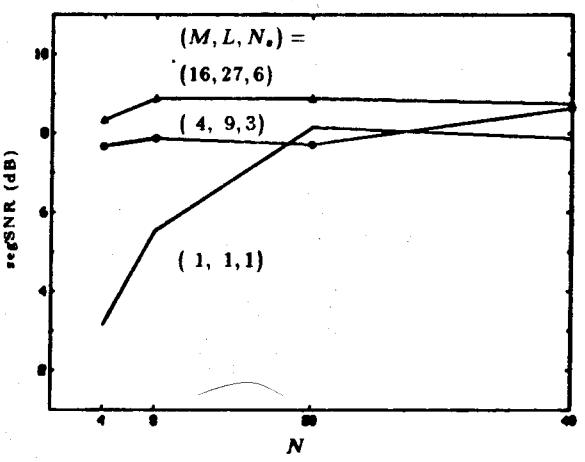


Fig. 3 Objective performance of the system as a function of L ($M \leq F^{L-1}$)



(a) $N = 4, 8, 20, 40$ and input sentence is WELLM8



(b) $N = 4, 8, 20, 40$ and input sentence is CATF8

Fig. 4 Objective performance of the system as a function of N