An Improved 2.4kbps Class-Dependent CELP Speech Coder

Yingbo Jiang and Vladimir Cuperman
School of Engineering Science, Simon Fraser University,
Burnaby, BC, V5A 1S6, Canada
Tel: (604) 291-4371, Fax: (604) 291-4951
email: yingbo@cs.sfu.ca, vladimir@cs.sfu.ca

Abstract

This paper presents an improved 2.4kbps class-dependent CELP speech coder. The improved coder is based on our previous efforts on a 2.4kbps CELP coder. New classification schemes in both open-loop and closed-loop are used. An extra transition class is used in the closed-loop classification. We developed a predictive LSP VQ to reduce LPC bit rate to as low as 18 bits/frame(30 ms) while maintaining low spectral distortion. Experimental results show that the quality of synthesized speech is improved. We also discuss the results obtained from class-dependent weighting filters, and a class-dependent postfilter.

1 Introduction

The current frontier of speech coding has moved to bit rates of 2.4kbps and below which has important applications such as the second generation "half-rate" mobile communication.

This paper presents our effort to improve the speech quality of a class-dependent CELP based 2.4kbps coder [2]. Our goal is to exploit the CELP structure fully, and try to get the best quality under this structure.

The system block diagram is shown in Figure 1. At the input, the original speech is divided into blocks (frames) of 240 samples. The frame-based classifier divides frames of speech into three categories: voiced, unvoiced, and transition. Each frame is divided into several subframes. The coefficients of a 10th order linear prediction (LP) filter are computed once per frame and quantized by a predictive multi-stage vector quantizer.

The reconstructed speech is obtained by passing an excitation vector through the synthesis filter. The excitation vector is computed for each subframe using a search procedure through two codebooks: an adaptive codebook and a stochastic codebook. The voiced class uses only an adaptive codebook to obtain the excitation vector, while the unvoiced class uses only a stochastic codebook. The transition classes uses both codebooks. The search procedure is based on computing the reconstructed speech for each codebook entry and then choosing the entry which provides the best reconstructed speech according to a perceptually weighted MSE criterion. In order to reduce the bit rate of the coder, we use delta-pitch coding to encode the pitch values in consecutive voiced subframes.

This paper includes 5 sections. Section 2 describes the classifiers, and Section 3 describes in detail the predictive multi-stage LSP vector quantizer. The optimum bit allocations are given in Section 4, together with the results of class-dependent perceptual weighting and post-filtering. Finally, conclusions are presented in Section 5.

2 Frame Classifiers

The coder is capable of operating using either a closed-loop or an open-loop frame classification.

The open-loop frame classifier is based on thresholding. The algorithm analyzes the speech on a frame basis and derives several parameters from the speech source. A class decision is made by applying experimentally derived thresholds to the parameters. The parameters considered in making class decisions include the normalized autocorrelation coefficient at the pitch lag, short term energy, zero-crossing rate, low-band energy, and normalized short-term autocorrelation coefficients. All these parameters have an inherent ability to discriminate between certain phonetic classes. However, there is considerable overlap between classes for any one of these parameters resulting in limited accuracy if only one parameter is considered alone. In the open-loop classification, each frame of speech is classified as either "voiced", "unvoiced" or "transition".

The closed-loop classifier encodes each speech frame once for each class defined, and computes the resulting mean square error weighted by the perceptual weighting filter. The class with minimum weighted MSE is chosen.

0-7803-2486-2/95 $4.00

In the closed-loop classification, each frame of speech is classified as either "voiced", "unvoiced" or one of two "transition" classes. The purpose of two transition classes is to provide for more diversity in speech waveform. experimental results show that it improves SegSNR and subjective quality of the reconstructed speech.

3 Predictive LSP-VQ

At 2.4kbps, the linear prediction parameters consume a large fraction of the total bit rate. Hence considerable efforts have been invested in finding efficient ways to represent these parameters. LSP parameters have some interesting properties that make them more amenable to efficient encoding than the LPC coefficients. The different LSP vector elements within a speech frame are correlated. This correlation is referred to as the intra-frame correlation of LSP parameters. In addition to this correlation, the LSP vectors from adjacent speech frames are also correlated (inter-frame correlation) [3] [4].

In order to exploit both the intra-frame and inter-frame correlation of the LSP parameters, we employed predictive LSP VQ in the coder.

By utilizing the correlations, the LSP vector for the current frame can be predicted from the previous one:

$$X(n) = A\hat{X}(n-1) + R(n) \quad (1)$$

where $X(n)$ is the LSP vector for frame n, $\hat{X}(n-1)$ is the quantized LSP vector for frame $n-1$, and $R(n)$ is the residual vector of frame n. A is a p by p prediction matrix and p is the dimension of LSP vector.

In our experiments, it was found that there existed strong correlation between adjacent vector elements having the same index, but weak correlation between elements with different indices. Based on this observation, we use a diagonal matrix A. The values we used for A are:

$$A = diag(.816\ .779\ .764\ .776\ .817\ .777\ .765\ .761\ .742\ .829) \quad (2)$$

The LSP parameters have non-zero means, and therefore, the elements of the residual vector $R(n)$ are also of non-zero mean. We obtained matrix A by minimizing the variance of each element of the $R(n)$ instead of minimizing the usual prediction error. It can be shown that this procedure is equivalent to doing the prediction after subtracting the means of individual LSP parameters.

In our original coder, 8-stage (3 bits/stage) MSVQ codebooks were designed to quantize LSPs [5]. The predictive LSP coding result shows that several bits/frame can be saved in quantization of LPC (originally 24 bits/frame) without perceptively degrading the quality of speech. Table 1 shows the detailed spectral distortions using different rates.

	Bits/Frame (30ms)	Spectral Distortion(dB))
8-stage Direct LSP-VQ	24	1.23
8-stage Predictive LSP-VQ	24	1.07
7-stage Predictive LSP-VQ	21	1.32
6-stage Predictive LSP-VQ	18	1.47
5-stage Predictive LSP-VQ	15	1.60
4-stage Predictive LSP-VQ	12	1.84

Table 1: Spectral Distortion Versus Bit Rates

4 Optimal Bit Allocations

The results of bit allocation optimization for different classes are shown in table 2.

We employ predictive LSP-VQ in the optimal bit allocations. Informal listening tests show 6 stage (3 bits/stage) MSVQ codebooks yield best speech quality at an overall bit rate constrained around 2.4kbps. In Table 2, all classes share the same predictive LSP-VQ codebooks.

The system operates with either the open-loop or the closed-loop classifier. The open-loop classifier uses the first 3 classes in Table 2, while the closed-loop classifier uses all four classes. Experimentally, the closed-loop classifier with 3 classes outperformed its open-loop counterpart. The closed-loop system with 4 classes outperformed the closed-loop system with 3 classes.

The number of taps in the adaptive codebook strongly influences the performance of the low rate CELP coder. With an unquantized adaptive codebook gain vector, higher number of taps yields better performance in both SegSNR and subjective quality. With limited quantization bits for the adaptive codebook gain in the system, three, or five tap gains are still preferable to a single tap gain. From our observation of the synthesized speech residual, it was found that the system with more taps was better able to match the excitation waveform by using a linear combination of past residual excitation wave-shapes. At

Parameter	Voiced	Unvoiced	Transition 1	Transition 2
Frame Size	240 (30ms)	240 (30ms)	240 (30ms)	240 (30ms)
Subframe Size	60	60	120	80
STP Bits	18	18	18	18
ACB Bits	7-3-3-3	-	7-7	7-4-2
ACB Gain Bits	9x4(5-tap)	-	6x2(3-tap)	6x3(5-tap)
SCB Bits	-	7x4	8x2	4x3
SCB Gain Bits	-	6x4	5x2	3x3
Classification Bits	2	2	2	2
Total(bits/frm)	72	72	72	72
Bits/sec	2400	2400	2400	2400

Table 2: Optimal Bit Allocation of the System

low rates, in the absence of a stochastic codebook, one single tap adaptive codebook is unable to adapt to variations in residual excitation wave-shapes even during sustained periods of voicing, resulting in worse performance.

4.1 Class-dependent Perceptual Weighting and Post-filtering

The human ear is more perceptive to the noise components in the valleys between the peaks in the spectral envelope of the speech waveform. The perceptual weighting filter (PWF) is used to attenuate the noise components in these valleys.

The perceptual weighting filter is of the form:

$$W(z) = \frac{1 - \sum_{i=1}^{k} a_i z^{-i}}{1 - \sum_{i=1}^{k} a_i \lambda^i z^{-i}} = \frac{A(z)}{A(z/\lambda)}, \qquad (3)$$

where $k = 10$ is the filter order, λ is a constant and $A(z)$ is the LP analysis filter.

In our system, postfiltering is applied to the reconstructed speech to improve the synthesized speech quality. The post filter is of the form [7]:

$$H(z) = (1 - \mu z^{-1}) \frac{A(z/\beta)}{A(z/\alpha)} \qquad (4)$$

Each speech class has a different LPC spectral envelope. Voiced speech envelopes tends to have greater spectral tilt than the unvoiced envelopes. A more important issue is that how much the amount of postfiltering applying to different class of speech is perceptually reasonable. The purpose of using class-dependent PWF and short-term postfiltering is to pick up the best parameter values for each class, and to obtain best subjective speech quality.

In our experiments, we tested several sets of the parameters(β,α, μ, and λ) for different classes. The system using class-dependent PWF and postfiltering parameters in Table 3 gave slightly better speech quality than the system with a class-independent choice of parameters ($\beta = 0.4$, $\alpha = 0.9$, $\mu = 0.7$, and $\lambda = 0.8$).

Parameter	Voiced	Transition	Unvoiced
λ	.75	.80	.95
α	.90	.80	.70
β	.40	.50	.50
μ	.60	.60	.40

Table 3: Class-dependent PWF and Postfilter Parameters

	Male	Female	Overall
2.4kb/s CELP	3.09	3.03	3.06
LPC10e	2.06	2.28	2.17

Table 4: MOS Scores of the CELP Coder

5 Conclusions

In this paper, we have described an improved 2.4kbps class-dependent CELP coder. A multi-stage predictive LSP vector quantizer is used to quantize the short-term spectrum of speech. New classification schemes are used, and bit allocations are optimized. Informal listening tests show that the new system offers perceptual quality improvement over the previous system, especially for male speakers.The system is very intelligible and preserves speaker identity. Table 4 compares the MOS scores of the CELP coder with those of LPC10e Vocoder. However, its subjective quality is not as good as the Federal standard at 4.8 kbps.

From the experimental results, the main deficiency of the synthesized speech is the background noise due to the

coarse quantization of the parameters at the low rate. Efficient encoding of the residual excitation is needed to further improve the analysis-by-synthesis CELP based coders.

References

[1] M. Schroeder, and B. Atal, " Code-excited linear prediction(CELP): High-quality speech at very low bit rates",*Proc. ICASSP*, pp.937-940, 1985.

[2] P. Lupini, H. Hassanein, and V. Cuperman, " A 2.4kb/s CELP Speech Codec with Class-dependent Structure,"*Proc. ICASSP*, 1993, pp. II-143 - II-146.

[3] M. Yong, G. Davidson, and A. Gersho, " Encoding LPC Spectral Parameters Using switched-adaptive inter-frame vector prediction", *Proc. ICASSP*, (San Francisco), pp. 402-405, 1988.

[4] N. Farvardin, and R. Laroia, " Efficient Encoding of Speech LSP Parameters Using the Discrete Cosine Transformation ",*Proc. ICASSP*, pp. 168-171, 1989

[5] B. Bhattacharya, W. LeBlanc, S. Mahmoud, and V. Cuperman, " Tree Searched multi-stage vector quantization of LPC parameters for 4 kb/s speech coding," *Proc. ICASSP*,(San Francisco), pp. 105-108, 1992.

[6] S. Wang, and A. Gersho, " Improving the Excitation for Phonetically-based vector excitation coding of speech at 3.6kbps," in Proc. ICASSP, pp.49-52, 1989.

[7] J. Chen, and A. Gersho " Real-time Vector APC Speech Coding at 4800 bps with adaptive postfiltering",*Proc. ICASSP*, pp. 2185-2188, 1987

[8] P. Jacobs, and W. Gardner," QCELP: A Variable Rate Speech Coder for CDMA Digital Cellular Systems,",*Speech and Audio Coding for Wireless and Network Applications*(B.S. Atal, V. Cuperman, and A. Gersho,eds.), Kluwer Academic Publishers, 1993.

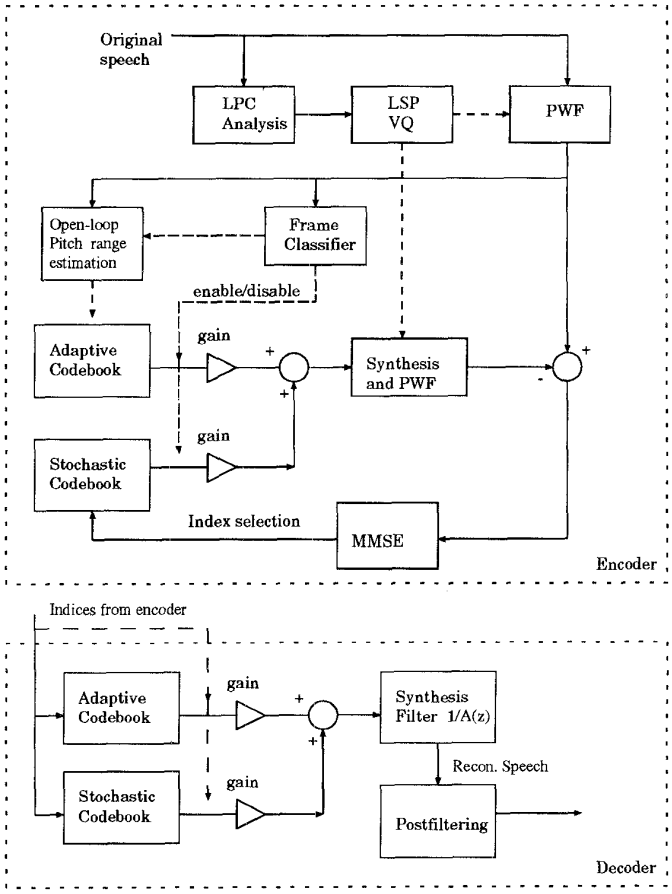

Figure 1: The Block Diagram of the Improved 2.4kbps Speech Coder