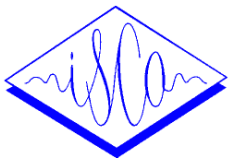

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

MEDIUM-DELAY 8 KBIT/S SPEECH CODER BASED ON CONDITIONAL PITCH PREDICTION

Takehiro Moriya

NTT Human Interface Laboratories.
Musashino-shi, Tokyo, 180 Japan

ABSTRACT

A medium bit-rate (8 kbit/s), medium delay (10 msec one-way), and high-quality speech coder is designed. The coder uses a conditional pitch predictor in the framework of the backward adaptive CELP (Code Excited Linear Prediction) coder. This scheme transmits only 3 to 5 bits to select from the pitch period candidates pruned by backward pitch analysis. It also uses block-wise backward adaptive short-term LPC analysis and backward adaptive gain quantization. In coding experiments, Signal to Noise Ratio (SNR) and subjective quality were superior to schemes with conventional forward pitch prediction or without pitch prediction. Although the quality of the proposed coder is slightly inferior to that of the conventional forward CELP (more than 50 ms delay) at the same bit-rate, it can outperform conventional CELP if delayed decision of the excitation vector is introduced by paying a computational cost.

1 INTRODUCTION

Medium bit-rate speech coding has been receiving much attention for use in various communication systems. In North America, Japan and Europe, a lot of activities have been carried out for the digital cellular radio system at around 8 kbit/s. Real-time operating high-quality coders have been built using DSP chips[1, 2]. In all cases, however, echo cancellers must be used since the coding delays in these systems are from 50 to 100 ms. Without echo control devices, communication may be largely disrupted due to echoes reflected from the hybrid circuit of the receiving telephones.

Let me review the relationship between the coding delay and the quality of the speech communication. Although echo canceller can reduce the echo signal, they are undesirable for maintaining the overall communication quality. For example, if speakers speak simultaneously at both ends, an echo canceller can not operate appropriately. In this case, it is better to cease operation rather than to produce an annoying noise. Another example is in the application to digital cordless phones or personal communication systems[3]. Because a conventional 2-wire line is replaced by a 4-wire radio channel which transmits the bit-stream of a speech coder, these network systems may have several hybrid circuits to connect 2-wire and 4-wire transmission system. Network may also have other equipment which causes transmission delay. In these cases, it is very complicated to control the echo if the speech coder has a long delay. So, the delay between the speech coder and decoder should be as small as possible for applying to the complicated communication networks. The coder for voice storage, of course, does not care about coding delay. Also, if the complete telephone network becomes fully digitalized, the coding delay of less than 100 ms can be neglected.

According to the above discussion of coding delay, low to medium bit-rate, low-delay speech coders are very important for improving the overall communication quality. Conventional speech waveform coding schemes, however, were classified into either low-bit long-delay coders or high-bit short-delay coders. CELP[4, 5] and VSELP[1] based on frame-wise processing belong to the formar category. Delta modulation and ADPCM based on sample-by-sample processing belong to the latter category. There was a technical gap between these two categories. It is very difficult to produce a high quality coder by sample-by-sample processing, even at 16 kbit/s.

In 1989, LD-CELP (Low Delay Code Excited Linear Prediction) coder[6] demonstrated that high-quality speech can be obtained at 16 kbit/s with a coding delay of less than 2 ms, by using very short frames containing only a few samples. This success has stimulated interest in finding coder structures that are able to operate at lower bit rates with short coding delay.

This paper reports on some preliminary results with an 8 kbit/s coder with a coding delay of around 10 ms. A prototype design is described in the following sections, and the performance of the coder is compared with the conventional coding schemes.

2 CODER DESIGN

2.1 Outline

The proposing coder is based on the CELP coder as shown in Fig. 1. This scheme includes the three technical components:

- backward linear prediction
- conditional pitch prediction
- backward adaptive gain quantization

In backward prediction, parameters are estimated only on the basis of the reconstructed signal. Since the reconstructed signal is available in the both encoder and decoder, information transmission is not needed.

Table 1: Comparison of the proposed coder with the CELP and LD-CELP, (F:Forward, B:Backward)

parameters	CELP	Proposed	LD-CELP
short-term LPC	F	B	B
LPC order	10	16	50
long-term delay	F	F/B	-
long-term gain	F	F	-
excitation gain	F	F/B	F/B
samples/frame	144	22	5
bit rate [kbit/s]	4 - 8	8	16

Transmitting parameters are pitch delay (preselected by backward analysis), pitch gain, excitation gain (backward adapted), and the excitation shape code. The decoder uses these information to generate the speech. The encoder selects the best code to provide the perceptually-weighted minimum distortion between the input and locally decoded signals. The parameters listed in Table 1 summarize the differences between the proposing coder, forward CELP, and LD-CELP

The proposed coding system uses the excitation gain quantizer with the same adaptation rule used in ADPCM. That is, the gain is quantized according to the step size determined from the code and the gain in the previous frame[7].

2.2 Short-term prediction

There are two ways to achieve the backward linear prediction. One is to use a recursive adaptation algorithm that updates parameters by sample-by-sample as in ADPCM[7]. The other way, used in LD-CELP and also in the proposed coder, is to extract parameters from the windowed reconstructed signal[6].

In this coding scheme, two LPC analyses are needed at the encoder as shown in Fig. 1. One analysis finds the coefficients of the synthesis filter, which are used both in the encoder and decoder. The p-th order all-pole filter is represented as $1/B(z)$.

$$B(z) = 1 - \sum_{i=1}^{p} b_i z^{-i} \tag{1}$$

The other LPC analysis is for the perceptual weighting filter, which is used only in the encoder[8]. The q-th order all-pole filter is represented by $1/A(z)$, and the weighting filter is $H(z)$ with the noise shaping factor γ_1 and γ_2.

$$A(z) = 1 - \sum_{i=1}^{q} a_i z^{-i} \tag{2}$$

$$H(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \tag{3}$$

$B(z)$ is estimated only from the reconstruction signal. On the other hand, $A(z)$ is estimated by the current speech signal. Neither LPC parameter sets are sent from encoder to decoder.

In a coder that transmits the LPC parameters, it is easy to use $B(z)$ for the perceptual weighting filter as well as the synthesis filter. In the backward adaptive coding scheme, however, $A(z)$ and $B(z)$ are different especially when the coder operate at low bit rates. Quality is significantly improved due to the introduction of $A(z)$ as well as $B(z)$, since $1/B(z)$ cannot shape the quantization noise. Even if the filter $H(z)/B(z)$ is used at the encoder, computational complexity does not increase comparing to the synthesis filter of $1/B(z)$, because the zero-state synthesis recursive filter can be replaced by a FIR filter whose tap-length is same as the number of samples in a frame.

Since the LPC parameters are not transmitted, there is no limit on the number of parameters used. To avoid interference with the pitch prediction, however, the order of the LPC is limited to 16. There were no significant differences between the different LPC analysis procedures such as the stabilized covariance method[8], weighted covariance method, and autocorrelation method. Several window shapes, such as the half Hamming window, recursive window[9], exponential window were tried. Window shape was not found to

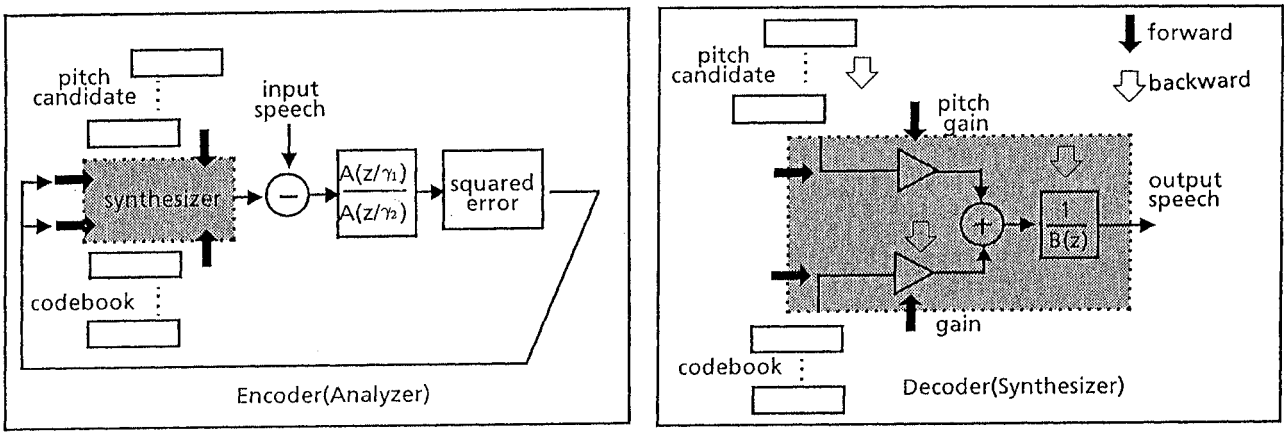

Fig. 1 Configuration of the proposed encoder and decoder.

be critical except exponential window which is significantly inferior to the others.

Ideally, synthesis needs $A(z)$ instead of $B(z)$. Therefore some part of spectral information of $A(z)$ seems to be useful for the decoder. When some encoding and quantization schemes for the difference between $B(z)$ and $A(z)$ were applied, the performance was not improved due to the increase in bit-rate for the side information.

2.3 Excitation codebook

The excitation vector is selected from a random codebook in conventional CELP. A structured codebook can improve the coder performance in terms of quality, complexity, and robustness against channel errors. With backward adaptive prediction, a structured or trained codebook can play a more important role. This is because the excitation signal should have some variations in frequency response to compensate for the synthesis filter, since the synthesis filter is different from the ideal LPC synthesis filter. The power spectrum of the synthesis filter whose coefficients are derived from the quantized speech tends to have larger value than that of actual one in the valley part, especially at higher frequencies. Therefore a simple low-passed noise codebook can provide better SNR and quality than the white noise codebook.

The proposed coder uses a trained codebook which is generated by generalized-Lloyd algorithm within a closed loop of the encoding process. This means the distortion measure for both finding the code and calculating the centroids is identical to the one used at the encoding process. Trained codebook can improve the SNR and quality even more compared to either the white noise or low-passed codebook.

3 CONDITIONAL PITCH PREDICTION

3.1 Description of the scheme

Long-term or pitch period prediction plays an important role for improving the speech quality. Although LD-CELP at 16 kbit/s does not require an explicit pitch prediction, it is very difficult to obtain a high-quality speech at 8 kbit/s without pitch prediction. In forward CELP, parameters such as pitch delay and tap gain are updated in every 5 ms subframe. In the proposed coder, frame lengths should be less than 3 ms. In this case, smaller number of bits is allowed to represent the pitch information. On the other hand, we can make use of the nature that the shorter the frame length, the more correlation the pitch parameter will have.

To reduce both the transmission bit-rate and the computational complexity without loosing performance, conditional pitch prediction is introduced. This scheme consists of forward closed loop estimation based on the backward analysis. The following two steps have been used for finding the pitch lag:

- open-loop backward analysis selects M (out of N) lag candidates, and
- closed-loop forward analysis finds the best lag out of M candidates.

Open-loop analysis is based on the cross-correlation of the residual signal. Closed-loop analysis, on the other hand, finds the pitch lag candidates by matching the synthesized signal and speech. Backward analysis uses only the reconstructed signal, while forward analysis uses current speech frame. The first step is, therefore, commonly performed at the encoder and the decoder. This means that only $\log_2(M)$, instead of $\log_2(N)$, bits have to be sent from the encoder to the decoder. This prototype uses 128 for N and 8 for M. So, 4 bits are saved for representing the pitch delay. On top of the conditional prediction, non-integer delay single-tap predictor[10] is also used only for the selected candidates.

3.2 Performance evaluation

Performance improvement due to the conditional pitch prediction is checked by comparing with other schemes. The results are shown in Table 2. The SNR values are averaged over two male and two female Japanese short utterances, all of which are outside training sequence of the excitation codebook. Note that bit-rate is fixed at 8 kbit/s by setting the vector dimension equal to the number of bit T per frame or vector. In all conditions, LPC analysis is performed by the autocorelation methods with the half Hamming window. The perceptual weighting filter with $q = 16, \gamma_1 = 0.9$, and $\gamma_2 = 0.4$ is used. Pitch period is set to be longer than the frame length. The compared schemes are summarized below.

A: Excitation vector consists only of the codebook vector. While the others(from **B** to **E**) use 16th order LPC analysis, 60th order analysis is employed. This structure is the same as that of LD-CELP.

B: Excitation of the current frame is represented by the sum of LPC residual signal with a distance of pitch period

Table 2: Comparison of pitch prediction

	b_d [bit]	b_p [bit]	b_c [bit]	T [bit]	SNR [dB]	SNR_{SEG} [dB]
A	0	0	8	11	12.1	13.1
	0	0	11	14	12.4	13.2
B	7	3	8	21	13.5	13.9
	7	3	11	24	14.1	14.4
C	9	3	8	23	13.5	14.3
	9	3	11	26	13.8	14.7
D	3	3	8	17	13.7	14.1
	3	3	11	20	14.2	14.8
E	5	3	8	19	14.0	14.5
	5	3	11	22	14.8	15.1

A : No pitch prediction
B : Forward prediction
C : Forward prediction (non-integer delay)
D : Conditional prediction
E : Conditional prediction (non-interger delay)
b_g : Number of bits for gain(3 bits for all)
b_p : Number of bits for pitch delay
b_d : Number of bits for pitch gain
b_c : Number of bits for random code
T : Total number of bits per frame

and the random codebook vector. Pitch period and gain are determined in a closed loop, so that the synthesized speech is close to the input. This structure is the same as that of CELP.

C: Pitch period is 4 times as precise as that of **B**.

D: Conditional pitch prediction described before.

E: Pitch delay is 4 times as precise as that of **D**. Selection of the candidates from past sequence is identical to the system **D**. Non-integer delay is only applied at the final candidate selection.

According to these results, conditional predictor can improve the SNR by around 1.5 dB comparing to the coder without explicit pitch predictor. Furthermore, non-integer delay also improves the SNR. Improvement of subjective quality was also confirmed by a informal listening test.

4 DISCUSSION

Since the low-delay coder can only observe data within a short window, it may be difficult to find the optimum code in a global scope. The scope can be extended if delayed decision of the excitation vector is used. According to preliminary tests, delayed decision of one frame excitation with 8 candidates can significantly improve SNR and the subjective quality. In this case, both the SNR and the subjective quality was even better than those of the conventional CELP at a similar bit rate and with a significantly longer coding delay.

Due to the delayed decision, the coding delay is increased by 2.75 ms only at the encoder. This is not serious, however, because the delay may be three times larger than this if the frame length is increased as much. The most serious drawback of the delayed decision may be the increased computational complexity. Complexity is increased more than 8 times (in the case of 8 candidates) and this amount is large enough to prevent real-time operation.

In addition, gain adaptation of the proposed coder cannot follow abrupt waveform change, which we can often see at the beginning of consonants. As far as the observed waveform is concerned, the performance of the proposed coder is poorer than that of forward CELP. According to the informal listening test, however, this degradation can hardly be detected.

5 CONCLUSION

A medium bit-rate (8kbit/s), medium delay (10 msec one-way) and high quality speech coder has been designed. This coder is based on the combination of forward and backward prediction in the framework of the CELP coder.

The proposed coder uses conditional pitch prediction, that is, forward pitch parameterization based on backward pitch analysis. In addition, block-wise backward adaptive short-term LPC, backward gain adaptation, and a trained codebook for excitation are used. The proposed conditional pitch prediction is superior to conventional pitch prediction schemes. Non-integer delay pitch prediction can also be used with the conditional pitch prediction.

The quality of the proposed coder (coding delay of 8.25 msec) is slightly lower than that of conventional forward CELP at the similar bit-rate and with a coding delay of more than 50 ms. The proposed coder, however, outperforms the conventional CELP if delayed decision of the excitation vector is introduced by paying a computational cost. This shows the possibility of reducing the coding delay of an 8 kbit/s speech coder without losing quality.

In order to apply the coding scheme in communication systems, the computational complexity should be reduced. Investigation of channel error will also be necessary for the cellular radio application.

ACKNOWLEDGEMENT

The author thanks Dr. Kiyohiro Shikano and Dr. Sadaoki Furui for their research guidance. This work is closely related to the investigations at AT & T Bell Labs where the author stayed in 1989. The author also wishes to thank Dr. Bishnu Atal and Dr. Peter Kroon for their valuable suggestions.

REFERENCES

[1] I. Gerson and M. Jasiuk: "Vector Sum Excited Linear Prediction (VSELP) Speech Coding at 8 KBS", *Proc. ICASSP'90*, 33.S9.3, 1990.

[2] T. Ohya, H. Suda, S. Uebayashi, T. Miki and T. Moriya: "Revised TC-WVQ Speech Coder for Mobile Communication System", submitted this conference.

[3] T. Hattori, *et al.*: "Personal Communication", *Proc. ICC'90*, pp.1351-1357, 1990.

[4] M. R. Schröeder and B. S. Atal: "Code-excited linear prediction (CELP): high-quality speech at very low bit rates", *Proc. ICASSP'85*, pp.937-940, 1985.

[5] P. Kroon and B. S. Atal: "Quantization Procedures for the Excitation in CELP coders," *Proc. ICASSP'87*, pp. 1649-1652, 1987.

[6] J. Chen: "High Quality 16kb/s Speech Coding with a One-Way Delay Less Than 2 ms", *Proc. ICASSP'90*, 33.S9.1, 1990.

[7] N. S. Jayant and P. Noll: *Digital Coding of Waveforms*, Prentice-Hall, 1984.

[8] B. S. Atal and M. R. Schröeder: "Predictive Coding of Speech Signals and Subjective Criteria," *IEEE Tans.* ASSP-27(3), pp. 247-254, Jun. 1979.

[9] T. P. Barnwell: "Recursive Windowing for Generating Autocorrelation Coefficients for LPC analysis," *IEEE Trans.* ASSP-29(9), pp. 1062-1066, 1981.

[10] P. Kroon and B. S. Atal: "Pitch Predictors with High Temporal Resolution," *Proc. ICASSP'90*, 46.S12.7, 1990.