4KBPS IMPROVED PITCH PREDICTION CELP SPEECH CODING WITH 20MS FRAME

Masahiro Serizawa and Kazunori Ozawa

Information Technology Research Labs., NEC Corporation
4-1-1, Miyazaki, Miyamae-ku, Kawasaki, Kanagawa 216, JAPAN

ABSTRACT

This paper proposes a new pitch prediction method for 4kbps CELP (Code Excited LPC) speech coding. In the conventional CELP speech coding, synthetic speech quality deteriorates rapidly at 4kbps, especially for female and children's speech with short pitch period. The important reason is that when the pitch period is shorter than the subframe length, simple repetition of the past excitation signal based on the estimated lag, not the true pitch prediction, is usually used in the adaptive codebook operation. The proposed pitch prediction method can carry out the true pitch prediction by utilizing the current subframe excitation codevector signal, when the pitch prediction parameters are determined. For further improvement, a split weighting method and a low-complexity harmonic and spectral perceptually-weighting method have also been developed. The informal listening test result shows that the 4kbps coder with 20msec subframe, utilizing all of the proposed improvements, achieves 0.2 MOS higher results than the coder without them.

1. INTRODUCTION

4kbps speech coding standardization has been discussed in the ITU-T study group 15 for the future FPLMTS, PSTN videophone and speech storage system applications. It requires that the speech quality in error-free condition is equal to ITU-T G.721 32kbps ADPCM, and that the algorithmic delay must be less than 20msec[21], which is applicable for tandem connection condition.

CELP[5,9,20] is one of several powerful speech coding methods. However, it is difficult to maintain high speech quality at 4kbps in a long subframe length, such as 10msec. The pitch prediction performance is significantly degraded, especially when the pitch lag is shorter than a subframe length. The long subframe can improve vector quantization efficiency, but the pitch prediction performance is deteriorated. One reason is that a pitch period varies in time in a long subframe. The other is that CELP uses an approximation in the adaptive codebook operation[20], when a pitch period is shorter than the subframe length, because the current subframe excitation signal is not available. The adaptive codebook error signal therefore includes much pitch periodicity.

To cope with this problem, pitch synchronization method [10] and comb-filtering method[11] for the excitation codebook signal have been proposed. However, these methods simply repeat or comb-filtering the excitation codebook signal to enhance the pitch periodicity. They still do not carry out the true pitch prediction.

The proposed pitch prediction method can carry out the true pitch prediction using the current subframe excitation codebook signal, when the pitch prediction parameter is determined. This method can be extended to include the pitch synchronization and comb-filter methods. To further improve the performance, a subframe-split weighting method and a low-complexity harmonic and spectral perceptually-weighting method were developed.

2. IMPROVED PITCH PREDICTION

The pitch prediction operation is represented as[19]:

$$y(n) = \beta y(n-L) + \gamma e(n), \tag{1}$$

where $e(n)$ is an excitation codevector signal selected from the excitation codebook and $y(n)$ is an excitation signal for the synthesis filter. L and β are the pitch lag and gain, respectively. In the conventional CELP, to reduce the calculation amount, the search for pitch prediction parameters β and L and the search for the excitation codevector $e(n)$ and gain γ are usually carried out sequentially. Further, when the pitch lag is shorter than a subframe length, the conventional adaptive codebook operation uses an approximation, which carries out the repetition of the past excitation signal, based on[20]:

$$y(n) = \beta y(n-kL), \tag{2}$$

where k is an integer value, such as 1, 2,... . This is because the excitation codebook signal $e(n)$ in the current subframe is not determined yet, and Eq.(1) can not be used. CELP speech quality deteriorated rapidly at 4kbps for female and children's speech due to these drawbacks.

In order to overcome these drawbacks, the proposed pitch prediction method can carry out the true pitch prediction through using Eq.(1), not the approximation based on Eq.(2), even if the pitch lag is shorter than the subframe length[2].

The procedure for the proposed algorithm is as follows: 1) First, the conventional adaptive codebook operation with a fractional lag[17], which is based on Eq.(2) when the pitch lag is shorter than a subframe length, is implemented to calculate pitch lag candidates. Excitation and gain codevector searches are then carried out. N_a candidates for pitch lags, N_e candidates for excitation codevectors and N_g candidates

0-7803-2431-5/95 $4.00 © 1995 IEEE

for gain codevectors are pre-selected, respectively. 2) The true pitch prediction based on Eq.(1), using the pre-selected pitch lag, excitation codevector and gain vector candidates in the current subframe, is executed in pitch synchronous manner, while calculating the excitation signal $y(n)$. The excitation signal $y(n)$ is filtered by the synthesis filter. The perceptually weighted error power between the input speech and the synthetic speech from the excitation signal is then calculated. 3) Operation 2) is repeated for all of the combinations of the candidates. An optimal combination of pitch lag, excitation codevector and gain codevector, which minimizes the perceptually weighted distortion, is selected, based on the delayed-decision[6].

Figure 1 is an example showing enhancement in pitch periodicity. It demonstrates that the true pitch prediction can represent pitch periodicity more accurately than the conventional one.

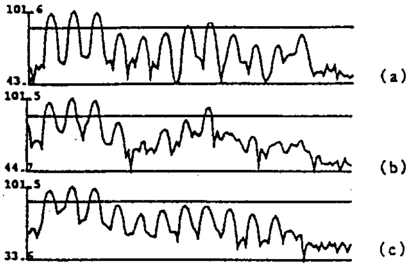

Fig.1 : Example showing enhancement in pitch periodicity, (a) the original speech, (b) speech synthesized by the conventional and (c) the proposed pitch prediction methods.

3. SPECTRAL AND HARMONIC WEIGHTING TO FURTHER IMPROVE QUALITY

3.1 Split Vector Weighting

The split vector weighting method was developed to further improve the performance, when the subframe length is fairly long[4]. Each subframe is divided into three sections. The 2nd section uses the analyzed spectral parameter set. In the 1st and 3rd sections, the spectral parameter sets, calculated by interpolating those of the adjacent subframes, are used to track spectral parameter variation more accurately.

Figure 2 shows the interpolation manner for the spectral parameter sets for the proposed split vector weighting, in comparison with the conventional manner. It is assumed that there are two subframes within one frame. The dashed line means the previous frame set and the thick solid lines mean the 1st and 2nd subframe sets in the current frame. The thin solid lines are the sets which are calculated by linear-interpolation between them. The synthetic filter coefficients are calculated in each 2nd subframe and interpolated between adjacent frames for each section in the same manner. The filter coefficient variation within subframes in the proposed method is smoother than that in the conventional method. This method therefore can estimate higher performance in the parameter-transient frame.

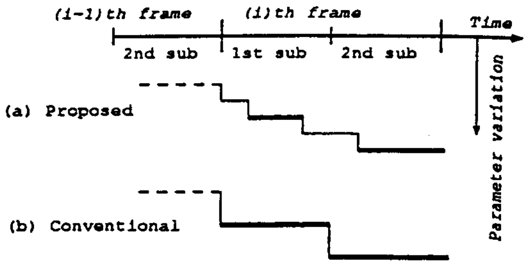

Fig.2 : Interpolation manner for spectral parameter set in split vector weighting

3.2 Harmonic Weighting

The harmonic weighting[7] is adopted to perceptually reduce pitch quantization noise. In the excitation codebook search, the conventional harmonic weighting is usually carried out by filtering the excitation codebook signal with the cascade connection of spectral and harmonic weighting filters, and thus it requires a great amount of computation. The computation amount was drastically reduced by implying the cascade filter transfer function to impulse response. The impulse response was applied to calculate the autocorrelations of weighted excitation codebook signals based on the autocorrelation method[3] with very small computation amount.

The proposed harmonic weighting cannot be valid when the pitch period is longer than the impulse response length. In general, the noise in harmonics in the synthetic speech can be perceived, especially for shorter pitch lag, for example, less than half of the subframe length (5ms). Therefore, the minimum sufficient impulse response length has to be about half of the subframe length (5ms).

4. EVALUATION BY 4KBPS CELP SPEECH CODER WITH 20MS FRAME

4.1 Experimental Condition

The proposed pitch prediction, split vector weighting, and harmonic and spectral weighting methods, are applied to a 4kbps CELP speech coder. The frame and subframe length are 20msec and 10msec, respectively. Figure 3 shows a blockdiagram of the encoder. The simulation conditions are summarized in Table 1.

Filter LSP coefficients for the 1st subframe are efficiently quantized to allocate more bits to excitation coding, by using the 1st order autoregressive predictive vector quantizer[14]. The 2-stage vector quantizer was applied, where the 2nd stage quantizer is a 2-split vector quantizer. The LSP codebook is designed based on the LBG algorithm[8]. The interpolation is carried out as mentioned in Section 3.1.

Excitation signals are quantized based on a multi-stage vector quantizer[15]. A 2-stage sparse codebook[16,18] with 8 pulses was used in modes 1,2 and 3. Mode 0 uses a 3-stage random stochastic codebook to represent various kinds of unvoiced speech[1,13]. The excitation codebooks were jointly designed by applying the Generalized Lloyd algorithm[8] to speech database.

Gains in pitch prediction and excitation codebook are jointly vector-quantized by using a 2-dimensional gain codebooks[1,13] designed by the LBG algorithm[8]. The power is vector-quantized in the μ-law region[1]. The power values at the 1st and 2nd subframes are quantized by 2-dimensional power codebook, designed based on the LBG algorithm[8].

Performance of a perceptual weighting filter on the encoder is improved through the use of non-quantized filter parameters[1,12,13].

The gain codebooks are switched in accord with each of four modes selected, based on the open-loop pitch prediction gain in each frame[1]. The excitation codebooks are switched in accord with voiced and unvoiced modes[1]. Pitch prediction is activated only in the voiced mode. LSP, excitation and gain codebooks were designed by using about 6 minute long speech signals, outside the signal used for the evaluation.

Nine Japanese speech sentence signals (30 seconds total time), uttered by two males and two females, are used for the following evaluation.

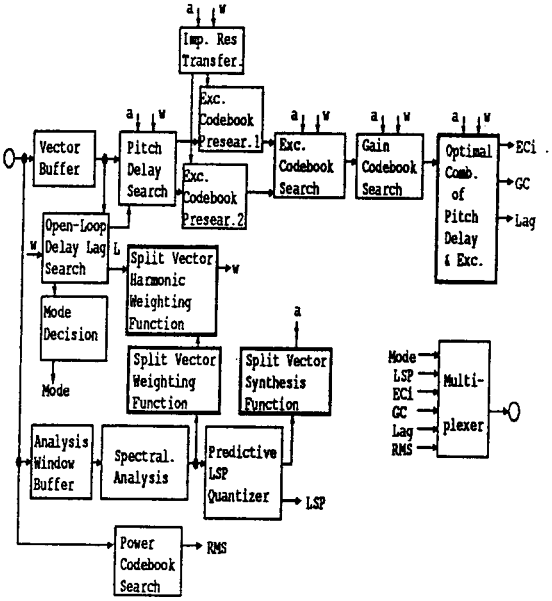

Fig.3 : 4kbps improved pitch prediction CELP with 20msec frame and 10 msec subframe

Tab. 1 : Simulation Conditions(bits per frame)

Mode	Voiced Mode	Unvoiced Mode
Mode	2	2
LSP	18	18
Power	6	6
AC	8*2	0
EC	13*2	18*2
Gain	6*2	6*2
Total	80	74

4.2 Proposed Pitch Prediction Performance

Table 2 shows the relationship between the average SNR_{seg} and the number of excitation and gain codevector candidates, N_e, N_g, in the proposed pitch prediction. In this simulation, the harmonic and spectral weighting methods described in Section 3 are adopted. The pitch lag candidate number N_a is 1. When $N_g = 30$, it can'achieve a similar SNR to the full search($N_g = 64$). The result for the conventional adaptive codebook (Conv. Pitch Pred.) is also shown. When N_g is more than 30, it achieves higher SNR_{seg} than that for the conventional adaptive codebook. Regarding excitation codevector candidate number N_e, SNR_{seg} is slightly improved as N_e is increased. An N_e value of 2 will be used in the following experiments.

Tab. 2 : SNR comparison as a function of the number of candidates ($N_a = 1$).

N_e	N_g	SNR_{seg}
2	64(full search)	9.00
2	40	9.01
2	30	8.97
2	20	8.74
2	10	7.85
3	30	9.07
5	30	9.08
6	30	9.13
2	Conv. Pitch Pred.	8.81

4.3 Proposed Split Vector Weighting Performance

The proposed split vector weighting (SVW) method was evaluated using the measure of an average SNR for subframe in each mode, which is shown in Table 3. Modes 1 and 2 were selected as the transient (Trans.) frames and mode 3 was selected as the stationary (Stat.) frames. The SVW method achieves 0.16 dB improvement, in comparison with the method without the SVW in modes 1 and 2. This indicates that the proposed method is effective in such transient subframes.

Tab. 3 : SNR comparison between with/without SVW in each mode

Mode No.	Unvoiced frame	Voiced frame		
		Trans.		Stat.
	0	1	2	3
SVW	1.83	5.81	8:90	10.82
w/o SVW	1.82	5.64	8.74	10.88

4.4 SNR and MOS Evaluation of Proposed Methods

Table 4 compares the average SNR_{seg} values for the conventional method(Basic) without incorporating the proposed improvements, the true pitch prediction(TPP), the split vector weighting (SVW), the low-complexity harmonic weighting(HW), and the proposed method, which includes all of the improvements. N_a, N_e, N_g and ϵ are 1, 2, 64 and 0.4, respectively. A SNR_{seg} improvement of 0.2dB can be seen in both TPP and SVW. The average SNR_{seg} value for HW is lower than those for the others, because the structure of the perceptual weighting differs from those of the others.

Table 5 shows MOS subjective evaluation test results in an error free condition. The proposed and basic methods are compared. G.721(32kbps ADPCM) is also evaluated as a reference. The postfilter is attached to all the methods, except G.721. Nine Japanese speakers took part in the test. The informal evaluation result shows that the proposed method can improve results by MOS of about 0.2, in comparison with the basic method results.

Tab. 4 : SNR Improvement Comparison

	SNR_{seg}
Basic Method	8.65
Basic+TPP	8.82
Basic+SVW	8.86
Basic+HW	8.47
Proposed Method	9.00

Tab. 5 : MOS Comparison

Coder	Bitrate(kbps)	MOS
ITU-T G.721 ADPCM	32	3.35
Proposed Method	4	3.07
Basic Method	4	2.85

5. CONCLUSION

This paper has proposed a new pitch prediction method for 4kbps CELP speech coding with 20msec frame. The proposed pitch prediction method can carry out the true pitch prediction using the current subframe excitation codebook signal. To further improve the performance, a split vector weighting method and a low-complexity harmonic and spectral weighting method have also been developed. The 4kbps speech coder with 20 msec frame, utilizing all of the proposed improvements, achieves 0.2 MOS higher results than the coder without them.

ACKNOWLEDGEMENTS

The authors would like to thank Mr. T. Nomura for valuable suggestions and C programs regarding LSP quantization. They also thank Mr. S. Taumi for valuable suggestions about the MOS evaluation.

REFERENCES

[1] K. Ozawa, M. Serizawa, T. Miyano, T. Nomura, M. Ikekawa and S. Taumi, “M-LCELP speech coding at 4kbps with multi-mode and multi-codebook,” IEICE Trans. Commun., Vol.E77-B, No.9, pp.1114-1121, Sept. 1994.

[2] M. Serizawa and K. Ozawa, “Improved Pitch Prediction in CELP Speech Coding,” Proc. Fall Meeting of ASJ Japan, 1-5-10, 1994 (in Japanese).

[3] I. M. Trancoso and B. S. Atal, “Efficient Search Procedures for Selecting the Optimum Innovation in Stochastic Coders,” IEEE Trans. ASSP-38,No.3, pp.385-396, 1990

[4] M. Serizawa and K. Ozawa, “Divided-vector Weighted-synthesis in CELP-type Speech Coding,” Proc. Spring Conf. of IEICE, A-283, 1994 (in Japanese).

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

[6] K. Mano and T. Moriya, “4.8kbit/s delayed decision CELP coder using tree coding,” IEEE Proc. ICASSP, pp.21-24, 1990.

[7] I. A. Gerson and M. A. Jasiuk, “Techniques for Improving the Performance of CELP Type Speech Coder,” Proc. ICASSP, pp.2185-2188, 1992.

[8] Y. Linde, A. Buzo and R. M. Gray,“An Algorithm for Vector Quantizer design, IEEE Trans. Commun., vol.COM-28, No.1, pp.84-95, 1980.

[9] I. A. Gerson and M. A. Jasiuk, “Vector sum excited linear prediction (VSELP) speech coding at 8kbps,” Proc. ICASSP, pp.461-464, 1990.

[10] S. Miki, K. Mano, H. Ohmuro and T. Moriya, “Pitch Synchronous Innovation CELP (PSI-CELP),” Proc. EUROSPEECH’93, 8.6, pp.261-264, 1993.

[11] S. Wang and A. Gersho, “Improved Excitation for Phonetically Segmented VXC Speech Coding below 4kb/s,” Proc. IEEE GLOBECOM’90, 507B.2, pp.946-950, 1990.

[12] J-H. Chen,“A robust low-delay CELP speech coder at 16kbit/s,” Proc. GLOBECOM, pp.1237-1241, 1989.

[13] T. Miyano, M. Serizawa, T. Nomura and K. Ozawa, ”A 3.6kbps LCELP coding method,” Proc. IEICE Spring Conf., SA-5-10, 1993 (in Japanese).

[14] V. Cuperman and A. Gersho, ”Adaptive differential vector coding of speech,” Proc. GLOBECOM, pp.1092-1096, 1982.

[15] T. Miyano, M. Serizawa, J. Takizawa, S. Ikeda and K. Ozawa, ”Improved 4.8kb/s CELP coding using two-stage vector quantization with multiple candidates(LCELP),” Proc. ICASSP, pp.I-321-324, 1992.

[16] W. P. LeBlanc and S. A. Mahnoud, “Structured codebook design in CELP,” Proc. 2nd Int. Mobile Satellite Conf., pp.667-672, 1990.

[17] J. S. Marques, J. M. Tribolet, I. M. Trancoso and L. B. Almeida, “Pitch prediction with fractional delays in CELP coding,” Proc. European Conf. on Speech Communication and Technology, vol.2, pp.509-512, 1989.

[18] S. Taumi, T. Nomura , M. Serizawa and K. Ozawa, “Comparative study of several excitation codebook structures in M-LCELP speech coding,” Proc. Spring Meeting of ASJ Japan, 3-8-17, 1994 (in Japanese).

[19] B. S. Atal, “Predictive coding of speech at low bit rates,” IEEE Trans. Commun., vol.COM-30, pp.600-614, Apr. 1982.

[20] W. B. Kleijn, D. J. Krasinski and R. H. Ketchum, “Improved speech quality and efficient vector quantization in SELP,” Proc. ICASSP, pp.155-158, 1988.

[21] ITU-T Study Group 15, “Provisional Terms of Reference for 4 kbit/s Speech Coding,”May 1994, Document TD 60.