4 KB/S MUTI-PULSE BASED CELP SPEECH CODING USING EXCITATION SWITCHING

Kazunori OZAWA

C&C Media Research Laboratories, NEC Corporation
4-1-1, Miyazaki, Miyamae-ku, Kawasaki, Kanagawa 216-8555, JAPAN
Email: ozawa@ccm.CL.nec.co.jp

ABSTRACT

This paper proposes an MP-CELP (Multi-Pulse-based CELP) speech coding at 4 kb/s. In MP-CELP, amplitudes or signs of multi-pulse excitation are simultaneously vector quantized (VQ). In order to improve speech quality for background noise conditions, excitation signal is switched between voiced and unvoiced speech, and the number of pulse is greatly increased for unvoiced speech by restricting pulse locations. Further, in order to improve voiced speech quality, the optimal combination among adaptive codebook lag, pulse location, sign codevector and gain codevector is selected which minimizes distortion by employing delayed-decision search. The subjective evaluation results show that speech quality for 4 kb/s MP-CELP is close to that for ITU-T G.723.1 (6.3 kb/s) and G.729 (8 kb/s) in M-IRS clean speech condition. For background noise conditions, the introduction for the excitation switching and the pulse location restriction significantly improves MOS value by 0.4. However, further improvement is still required, except for interference talker condition.

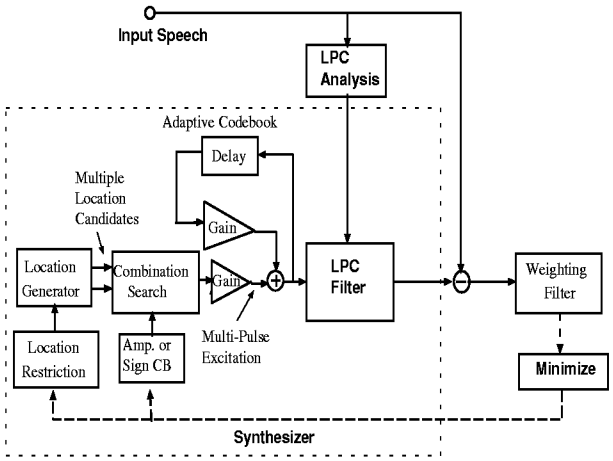

Figure 1: MP-CELP encoder operation principle.

1. INTRODUCTION

In beyond 2000, mobile multi-media communications services are expected to be provided via IMT (International Mobile Telecommunication)-2000. In that services, high-quality speech communication which is as good as wire-line quality and flexibility which can provide several kinds of source coding bit rates in accordance with transmission channel capacity and bit error rates are very important issues.

Especially, the target in which high-quality speech should be provided at 4 kb/s with moderate coding delay under both clean speech and background noise conditions is very challenging. ITU-T is now requesting candidate speech coding algorithms at 4 kb/s which can meet high-quality requirements [1].

The authors have proposed MP-CELP (Multi-Pulse-based CELP) speech coding [2], [3]. In MP-CELP, amplitudes or signs for multi-pulse excitation are simultaneously vector quantized, and combination search between multiple pulse location candidates and VQ codebook enhances performance. Speech quality for 11 kb/s MP-CELP with 10 ms frame [3] is higher than that for G.728 (16 kb/s LD-CELP) [4], and the quality for 6.7 kb/s MP-CELP with 10 ms frame is equivalent to that of G.726 (32 kb/s ADPCM) [5]. However, the speech quality is rapidly degraded when the bit rate is below 6.7 kb/s.

This paper proposes 4 kb/s speech coding with 20 ms frame, based on MP-CELP technique. In order to improve speech quality for background noise conditions, excitation signal is switched between voiced and unvoiced speech, and the number of pulse is greatly increased in unvoiced speech by restricting pulse locations. Further, in order to improve voiced speech quality, the optimal combination among adaptive codebook lag, pulse location, sign codevector and gain codevector is selected which minimizes distortion, by employing delayed-decision search.

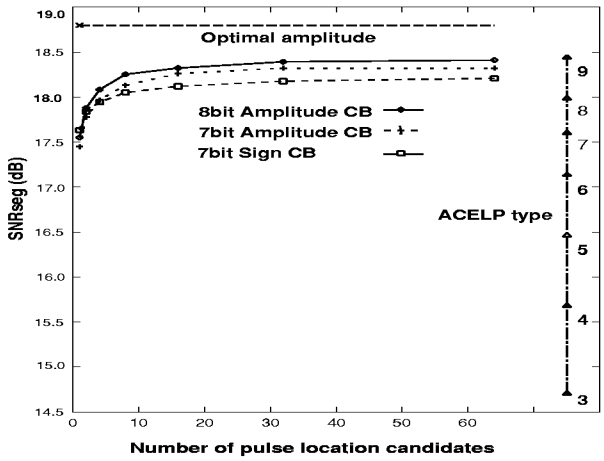

Figure 2: SNR_{seg} vs. the number of pulse location candidates in the combination search.

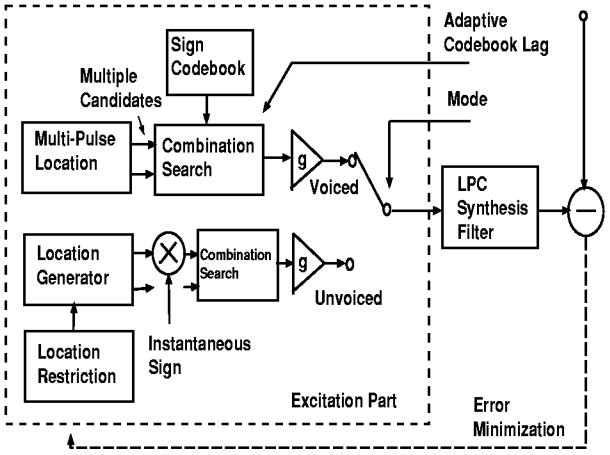

Figure 3: Blockdiagram for proposed excitation structure in 4 kb/s MP-CELP.

2. OVERVIEW OF MULTI-PULSE BASED CELP

The operation principle for MP-CELP encoder is shown in Fig. 1. In this method, excitation signal is represented by multi-pulse [6], [7]. The amplitudes or signs of all multi-pulse in a sub-frame are simultaneously vector quantized to improve the quantization performance [2], [3]. The pulse locations are restricted based on the algebraic structure whose concept is the same as that in G.729 [8] to reduce both transmission bit rate and location search complexity.

Figure 2 shows the SNR_{seg} performance against the number of the pulse location candidates N_p in the combination search, where sub-frame length is 5 ms and the number of pulse M is 7. The bit rate is 11 kb/s in this case. It is shown that the combination search improves SNR_{seg} by 0.8 dB compared to the method without the combination search (N_p=1), when the amplitude codebook is used. In the case of sign CB, SNR_{seg} improvement of 0.6 dB is achieved when the number of candidates is 16.

By using the sign CB, the excitation structure is similar to that of G.729 [8], because G.729 excitation is considered to be without the combination search.

3. 4 KB/S MP-CELP STRUCTURE

Figure 3 shows a blockdiagram of a proposed excitation structure for 4 kb/s MP-CELP encoder. Frame and sub-frame length are 20 ms and 10 ms, respectively. Mode decision is carried out in each frame using an average open-loop pitch prediction gain G over the frame.

$$G = \frac{\sum_{l=1}^{N_s} 10log_{10}(GP_l)}{N_s} \quad (1)$$

GP_l and N_s represent pitch prediction gain in the l-th sub-frame, and the number of sub-frames in a frame, respectively.

GP_l is calculated using the following equations.

$$GP_l = \frac{\sum_{n=0}^{N-1} x_w^2(n)}{A} \quad (2)$$

$$A = \sum_{n=0}^{N-1} x_w^2(n) - \frac{\left[\sum_{n=0}^{N-1} x_w(n)x_w(n-T)\right]^2}{\sum_{n=0}^{N-1} x_w^2(n-T)} \quad (3)$$

T and $x_w(n)$ show open-loop pitch lag and perceptually weighted input speech signal, respectively.

Two modes such as voiced and unvoiced modes are used in this study. Ten-th order filter LSP coefficients are vector quantized after MA (moving average) prediction [9], [10]. In adaptive codebook, lag is differentially encoded [11] in the second sub-frame in voiced frames, because lag is tempolarily correlated between sub-frames. On the other hand, lag is independently encoded in each sub-frame in unvoiced frames, because of less correlation between sub-frames.

Excitation signal is switched between voiced and unvoiced frames based on mode information. In unvoiced

Table 1: SNR_{seg} comparison at 4 kb/s

	SNR_{seg}(dB)
Method 1	12.50
Method 2	12.50
Method 3	12.73

Method 1 does not use the excitation switching.
Method 2 introduces the proposed excitation switching and pulse location restriction, but does not use the delayed-decision.
Method 3 combines the delayed-decison search with Method 2.

frames, the number of pulse is greatly increased by restricting pulse locations, based on the method reported in [5]. Instantaneous sign calculated from cross-correlation between weighted input speech and impulse reponse of weighted synthesis filter is used and transmitted as the sign for each pulse. A combination search between multiple location candidates and gain codebook is carried out to select the optimal combination. In voiced frames, the combination search is enhanced to the search among multiple adaptive codebook lag, multiple pulse location candidates, pulse sign codebook and gain codebook is carried out to select the optimal combination which minimizes distortion. This search is realized by utilizing delayed-decision [12].

In voiced frames, the number of pulse in the sub-frame is 4. The size of pulse sign codebook is 4 bits, and the number of pulse location candidates in the combination search is 16. In unvoiced frames, the number of pulse in the sub-frame is 14.

Table 1 shows SNR_{seg} performance comparison. In the Table, Method 1 does not employ excitation switching and always uses 4-pulse excitation even for unvoiced speech. Method 2 introduces the proposed excitation switching and pulse location restriction, but does not use the delayed-decision combination search. Method 3 combines the delayed-decison search with Method 2. The bit rate for all of those coders is 4 kb/s. Sixteen Japanese short sentences with flat input filter are used. From the Table, SNR_{seg} for Method 2 is the same as that for Method 1, and the excitation switching does not improve SNR_{seg} for clean speech, as is expected. Method 3 improves SNR_{seg} by 0.23 dB by incorporating the delayed-decision search in comparison with Method 2.

4. SUBJECTIVE EVALUATION

4 kb/s MP-CELP speech quality is evaluated. Simulation conditions and bit allocations are summarized in Table 2.

Two kinds of experiments were designed. Experiments 1 and 2 assess performance for clean speech conditions with

Table 2: 4 kb/s MP-CELP simulation conditions

Parameter	value
Frame (ms)	20
Sub-Frame (ms)	10
LSP VQ (bits)	16
Mode (bit)	1
Adaptive CB (bits)	8 + 5 (Voiced Frames)
Adaptive CB (bits)	8 + 8 (Unvoiced Frames)
Number of Pulses	4 x 2 (Voiced Frames)
Number of Pulse	14 x 2 (Unvoiced Frames)
Sign CB (bits)	4 x 2 (Voiced Frames)
Pulse Location (bits)	14 x 2 (Voiced Frames)
Pulse Location (bits)	17 x 2 (Unvoiced Frames)
Gain CB (bits)	7 x 2 (Voiced Frames)
Gain CB (bits)	7 x 2 (Voiced Frames)
Total/Frame (bits)	80 (Voiced Frames)
Total/Frame (bits)	79 (Unvoiced Frames)

M-IRS input filter, and performance for background noise conditions with flat input filter (car, babble and interference talker), respectively. The number of speech samples (Japanese sentence-pair) is 8 in Experiment 1 and 12 in Experiment 2. Experiments 1 and 2 use ACR and DCR evaluation methods, respectively. Two kinds of 4 kb/s MP-CELP, marked as Methods 1 and 3, which are the same as those in Table 1, are evaluated. Method 1 does not use the excitation switching. Method 3 employs both the proposed excitation switching and the delayed-decision search. In Method 3, the number of adaptive codebook lag candidates and the number of pulse location candidates are 2 and 16, respectively. ITU-T G. 726 ADPCM at 32 kb/s, G.729 and G.723.1 at 6.3 kb/s are also evaluated as reference codecs. Headphone was used as listening device. Twelve Japanese subjects took part in these Experiments.

The results are shown in Figure 4. From the comparison between Methods 1 and 3 in Experiment 1, MOS value in Method 3 is improved by 0.3. This improvement is mainly due to the delayed-decision search. Method 3 speech quality is close to those of G.723.1 at 6.3 kb/s and G.729, however, the qiality doe not reach to that for G.726 at 32 kb/s. In Experiment 2, from the comparison between Methods 1 and 3, the excitation switching and the pulse location restriction are effective for background noise conditions, and the speech quality is improved by 0.4 in MOS value, especially for car noise and babble noise conditions. The results show that the speech quality for 4 kb/s MP-CELP (Method 3) is close to those of G.723.1 at 6.3 kb/s and G.729 for interference talker condition, however further improvemt is still necessary for car noise and babble noise

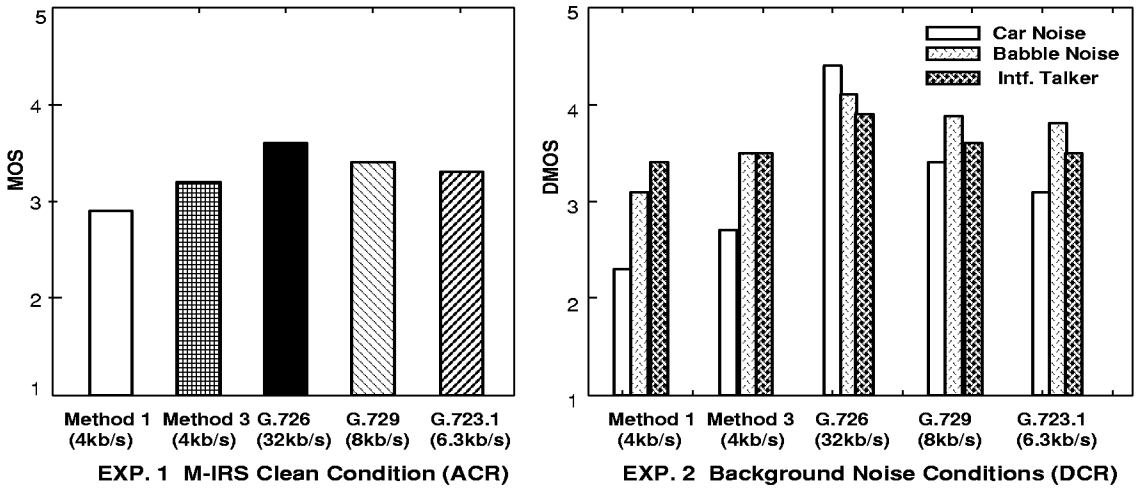

Figure 4: Subjective evaluation results.

conditions.

5. CONCLUSION

This paper proposes 4 kb/s MP-CELP speech coding. In order to improve speech quality for background noise conditions, excitation signal is switched between voiced and unvoiced speech, and the number of pulse is greatly increased for unvoiced speech by restricting pulse locations. Further, in order to improve voiced speech quality, an optimal combination among adaptive codebook lag, pulse location, sign codevector and gain vector is selected which minimizes distortion, by employing the delayed-decision search. The speech quality for 4 kb/s MP-CELP (Method 3) is close to those for ITU-T G.723.1 at 6.3 kb/s and G.729 for M-IRS clean speech condition. The excitation switching and pulse location restriction are effective for background noise conditions, and MOS value is improved by 0.4 for car and babble noise conditions. However, further improvement is still required, except for interference talker condition.

Acknowledgement

The authors would like to thank Mr. H. Kumagai and Dr. M. Serizawa for their help throughout this work.

REFERENCES

[1] ITU-T, "Subjective qualification test plan for the ITU-T 4-kbit/s speech coding algorithm," Rev.2.2.6, Feb. 1998.

[2] K. Ozawa, T. Nomura and M. Serizawa, "MP-CELP speech coding based on multipulse vector quantization and fast search," Electronics and Communication in Japan, Part 3, Vol. 80, No. 11, pp. 55-63, 1997.

[3] S. Taumi, K. Ozawa, M. Serizawa and T. Nomura,"Low-delay CELP with multi-pulse VQ and fast search for GSM EFR," Proc. ICASSP, pp. 562-565, May 1996.

[4] J-H. Chen, R. V. Cox, Y-C. Lin, N. Jayant and M. J. Melchner, "A low-delay CELP coder for the CCITT 16 kb/s speech coding standard, " IEEE J. Sel. Areas Commun. , vol. 10, no. 5 pp. 830-849, 1992.

[5] K. Ozawa and M. Serizawa, "High quality multi-pulse based CELP speech coding at 6.4 kb/s and its subjective evaluation," Proc. ICASSP, pp. 529-532, 1998.

[6] B. S. Atal and J. R. Remde,"A new model of LPC excitation for producing natural-sounding speech at low bit rates," Proc. ICASSP, pp.614-617, 1982.

[7] K. Ozawa, S. Ono and T. Araseki, "A Study on pulse search algorithms for multi-pulse excited speech coder realization," IEEE J. Sel. Areas Commun., SAC-4, pp.133-141, Feb. 1986.

[8] R. Salami, C. laflame, J-P. Adoul, A. Kataoka, S. Hayashi, C. Lamblin, D. Massaloux, S. Proust, P. Kroon and Y. Shoham, "Description of the proposed ITU-T 8 kb/s speech coding standard, " IEEE Workshop on Speech Coding for Telecommunications, pp. 3-4, Sept. 1995.

[9] H. Ohmuro, T. Moriya, K. Mano and S. Miki, "Vector quantization of LSP parameters using moving average inter-frame prediction," Trans. IEICE, vol. J77-A, No. 3, pp.303-313, March 1994 (in Japanese).

[10] T. Nomura, M. Serizawa and K. Ozawa, "Vector quantization of LSP parameters using adaptive prediction," National Meeting of ASJ, pp. 245-247, Oct. 1994 (in Japanese).

[11] K. Ozawa, M. Serizawa, T. Miyano and T. Nomura, "M-LCELP speech coding at 4 kbps," Proc. ICASSP, pp. I-269-I-272, 1994.

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