



## A 4.8 kb/s HIGH-QUALITY SPEECH CODING USING VARIOUS TYPES OF EXCITATION SIGNALS

Kazunori Ozawa

C&C Information Technology Research Laboratories, NEC Corporation,  
4-1-1, Miyazaki, Miyamae-ku, Kawasaki 213, JAPAN

### ABSTRACT

A high-quality speech coding method (SPMEX) at 4.8 kb/s is proposed. The SPMEX selects a suitable excitation signal, based on the decision from acoustic features of speech signal in a frame. Improved pitch interpolation multi-pulse (PMPC) excitation is selected for vowel-like speech. In PMPC, multi-pulse during only one pitch period is calculated in the frame. Further, gain and phase adjusting coefficients are calculated for each of the other pitch periods in the same frame, in order to produce high-quality synthetic speech. Multi-pulse excitation (MPC) is selected for stop and transition-like speech. Stochastic codebook (SC) excitation is selected for frication-like speech. Subjective evaluation results show that 4.8 kb/s SPMEX reconstructs high-quality synthetic speech, which is equivalent to 48 kb/s  $\mu$ -law PCM.

### 1. INTRODUCTION

Demands for low bit rate speech coding methods have been rapidly increasing, especially in the areas of mobile radio communications systems. In Europe, a 13 kb/s speech coding method was recently standardized [1], and in the U.S., standardization is progressing around 8 kb/s. Taking these facts into account, a speech coding method, which can encode speech at 4.8 kb/s and below, will be required in advanced mobile radio communications and mobile satellite communications systems.

Several types of speech coding methods [2-6] have been proposed from 4.8 to 8 kb/s. In these methods, multi-pulse coding has an advantage for accurately representing pitch pulses for vowel parts of the speech, compared with other speech coding methods, such as CELP [2]. The authors proposed multi-pulse coding with pitch interpolation at 4.8 kb/s [6]. In this method, for voiced speech, a small number of multi-pulse was calculated during only one pitch period (representative period). The excitation signal for other pitch periods was reconstructed by linearly interpolating amplitudes and locations of the multi-pulse in the representative period. Synthetic speech quality of this method was generally good, but sometimes slightly degraded, due to phase or amplitude distortion caused by interpolating the multi-pulse. Further, for unvoiced speech, multi-pulse excitation is very difficult to represent frication-like signal when the number of the pulse is small.

This paper describes a new speech coding method (SPMEX: Speech Coding Using Multiple Excitation Signals) [7] based on multi-pulse coding. This method uses multiple excitation signals, in order to highly represent any kind of speech signal, such as vowels,

stops and fricatives. This method selects a suitable excitation signal, based on the decision from acoustic features of speech signal in a frame. Further, an improved pitch interpolation method, based on multi-pulse excitation, is proposed to produce high-quality synthetic speech for vowel-like speech. In this method, gain and phase adjusting coefficients for the multi-pulse in the representative period are calculated, in order to accurately represent gain and phase variation between pitch periods in a speech signal.

### 2. SPMEX ALGORITHM

#### 2-1. EXCITATION SIGNALS

Speech signal in a frame is mainly classified into three categories, as shown in Fig. 1, based on several acoustic features obtained from the speech signal. By selecting these categories, a suitable excitation signal is allocated. In order to avoid serious degradation of synthetic speech quality, such as deciding on an incorrect category by background noise, excitation signals, shown in Fig. 1, are adopted. Each of the excitation signals has a capability of representing a speech signal without significant loss in speech quality, even if excitation selection is incorrect. For vowel-like category, improved pitch interpolation multi-pulse (PMPC) excitation has been developed. For stop and transition-like category, multi-pulse (MPC) excitation is adopted, because the excitation source characteristics are changed rapidly with time. For frication-like category, stochastic codebook (SC) excitation is adopted, because the excitation source is like a noise signal and a small number of multi-pulse is difficult to represent such excitation.

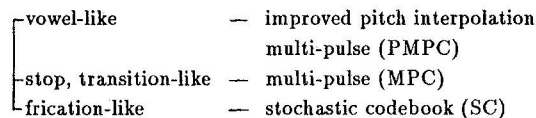


Fig.1. Types of excitation signals used in SPMEX.

#### 2-2. ACOUSTIC FEATURES

The acoustic features adopted in SPMEX to select the excitation signals are pitch prediction gain (P), short-term prediction gain (G), rms value (RMS), temporal rms value variation (RMS-D) and 1st order log-area ratio (LAR1). Appropriately selecting the excitation signal is as follows: If RMS, P and G are high and RMS-D is low, vowel-like category is selected and PMPC excitation is used. If LAR1 and G are high, MPC excitation is selected.

Periodicity of LPC prediction residual signal is not strong in this case, since LPC coefficients are strongly affected by pitch harmonics. If RMS, P, G and LARI are low and RMS-D is high, stop and transition-like category is selected, and MPC excitation is used. The rest is categorized into frication-like category and stochastic excitation is selected.

### 2-3. IMPROVED PITCH INTERPOLATION MULTI-PULSE EXCITATION

Figure 2 shows a speech synthesis model, based on improved pitch interpolation multi-pulse (PMPC) excitation. In this model, a small number of multi-pulse is calculated during only one pitch period (representative period) in the frame, in order to efficiently utilize periodicity of the excitation source. In each of the other pitch periods within the same frame, multi-pulse in the representative period is repeated while adjusting the gain and phase for the multi-pulse. Gain and phase adjustments are important to further improve synthetic speech quality. The gain and phase adjusting coefficients can be calculated, so as to minimize weighted mean-squared error between original and synthetic speech, during each of the pitch periods. The synthetic speech  $s_j(n)$  in the  $j$ -th pitch period can be written as follows:

$$S_j(n) = c_j \sum_{i=1}^K g_i h(n - m_i - T - d_j), \quad (1)$$

where  $c_j$ ,  $d_j$  and  $T$  denote gain adjusting coefficient, phase adjusting coefficient in the  $j$ -th pitch period and average pitch period in the frame, respectively.  $g_i$ ,  $m_i$  are amplitude and location of the  $i$ -th pulse in the representative period. The perceptually weighted mean-squared error in the  $j$ -th pitch period is

$$E_j = \sum_n [\{x_j(n) - s_j(n)\} * w(n)]^2 \quad (2)$$

where  $w(n)$  shows impulse response for a perceptual weighting filter [8]. Gain and phase adjusting coefficients in each of other pitch periods can be calculated by minimizing  $E_j$ . This leads to the following equation,

$$c_j = \frac{\sum_n x_w(n) s_w(n)}{\sum_n s_w(n) s_w(n)} \quad (3)$$

where

$$x_w(n) = x(n) * w(n) \quad (4)$$

$$s_w(n) = s(n) * w(n). \quad (5)$$

By inserting Eq. (3) into Eq. (2), Eq. (2) is rewritten as

$$E_j = \sum_n x_w^2(n) - \frac{\{\sum_n x_w(n) s_w(n)\}^2}{\sum_n s_w(n) s_w(n)}. \quad (6)$$

Phase adjusting coefficient  $d_j$  is obtained by maximizing the second term of Eq. (6), and then gain adjusting coefficient  $c_j$  is obtained from Eq. (3).

Figure 3 shows an example of waveforms for the process used in determining gain and phase adjusting coefficients described above. The waveform in the frame is shown in Fig.3 (a), multi-pulse in the representative pitch period is indicated in Fig.3 (b), reconstructed excitation signal in the frame by gain and phase adjusting coefficients is shown in Fig. 3 (c) and synthetic speech is indicated in Fig. 3(d).

In order to maintain high speech quality for frames, where

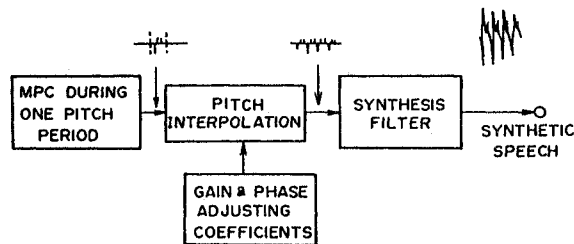


Fig.2. Speech synthesis model by improved pitch interpolation multi-pulse.

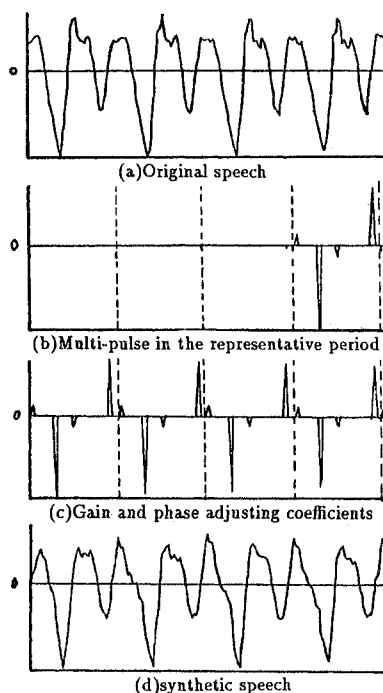


Fig.3. Waveforms showing the process for determining gain and phase adjusting coefficients in improved pitch interpolation multi-pulse.

speech characteristics gradually change in time, the representative pitch period is searched for in several pitch periods within the frame. The pitch period which minimizes the following error power  $E$  is determined as the representative period.

$$E = \sum_{n=1}^N [\{x(n) - s(n)\} * w(n)]^2 \quad (7)$$

### 2-4. MULTI-PULSE (MPC) AND STOCHASTIC (SC) EXCITATIONS

Multi-pulse excitation is used for stop and transition like-category. Amplitudes and locations of multi-pulse are calculated according to [9] in the frame.

The Gaussian stochastic codebook [2] is used as the excitation signal for frication-like category. The speech synthesis model is the

same as that for CELP. The long-term predictor order was set to 1. The coefficient of the long-term predictor, codebook index and gain are determined in 10ms sub-frame. The optimum index and gain for the codebook are calculated according to [2].

### 2-5. SPMEX CODER AND DECODER STRUCTURE

Figure 4 shows a blockdiagram of coder and decoder structure for SPMEX. In the coder side, LPC and pitch analysis are carried out and acoustic features described in 2-2 are calculated. Then, an excitation signal is selected, and parameters for the selected excitation signal are calculated. PARCOR and excitation selection information are transmitted as side information. Average pitch period in the frame, amplitudes and locations of multi-pulse in the representative period, gain and phase adjusting coefficients and location of the representative period in the frame are transmitted for PMPC excitation. Amplitudes and locations of multi-pulse are transmitted for MPC excitation. Pitch period, long-term predictor coefficient, codebook index and gain are transmitted for SC excitation.

## 3. EXPERIMENTS

### 3-1. S/N AND CD PERFORMANCES

Table 1 summarizes 4.8kb/s SPMEX simulation conditions. Eight short Japanese sentences (uttered by 3 male and 3 female speakers) were used as speech database. They were recorded through a dynamic microphone and digitized at an 8kHz sampling frequency.

Table 2 shows average segmental S/N (SNRseg) and average LPC cepstrum distance (CD). 4.8 kb/s multi-pulse coding with pitch interpolation [6] (PP2) and 4.8 kb/s pitch predictive multi-pulse coding [10] (PP1) were also evaluated as conventional methods. From the table, SNRseg of the SPMEX is 9.2dB. This value is 0.7dB better than PP1 and 0.2dB better than PP2. Differences in SNRseg are not so large between SPMEX and PP2. Differences in CD are very small among the three methods.

Figure 5 shows S/N performance plotted versus time for a male speaker. The power of short-time segments of speech is plotted over a sentence in Fig. 5 (a) and S/N of short-time segments is plotted in Fig. 5 (b). S/N for the whole sentence is 8.4 dB for SPMEX and 7.4 dB for PP2. SPMEX improves S/N in the large

Table 1. 4.8kb/s SPMEX simulation condition.

frame	20ms	<u>MPC</u>	
LPC order	10	Number of	
	38bits	multi-pulse	6
Excitation			53bits
selection	2bits	<u>CELP</u>	
<u>PMPC</u>		Pitch	2x11bits
Pitch	6bits	Codebook	2x10bits
Number of	4	Gain	2x5bits
multi-pulse	37bits		
Rep. period	2bits		
Gain and phase			
adj. coeff.	10bits		

Table 2. Segmental S/N and LPC Cepstrum distance.

	SNR(dB)	CD(dB)
4.8kb/s SPMEX	9.2	1.9
4.8kb/s PP2	9.0	2.1
4.8kb/s PP1	8.5	2.0

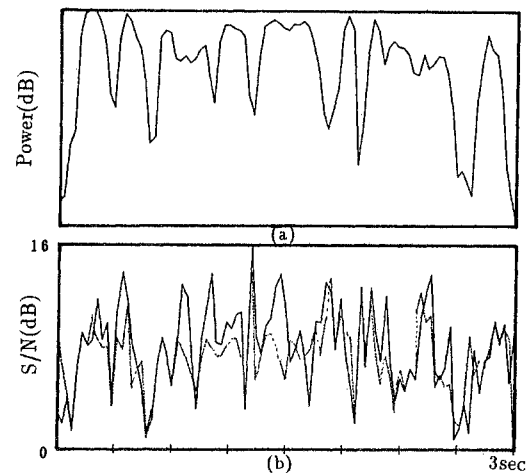


Fig.5. S/N performance vs. time. The plot (a) shows short-time power of speech, (b) shows S/N. The solid line shows S/N of SPMEX and the dotted line shows S/N of PP2.

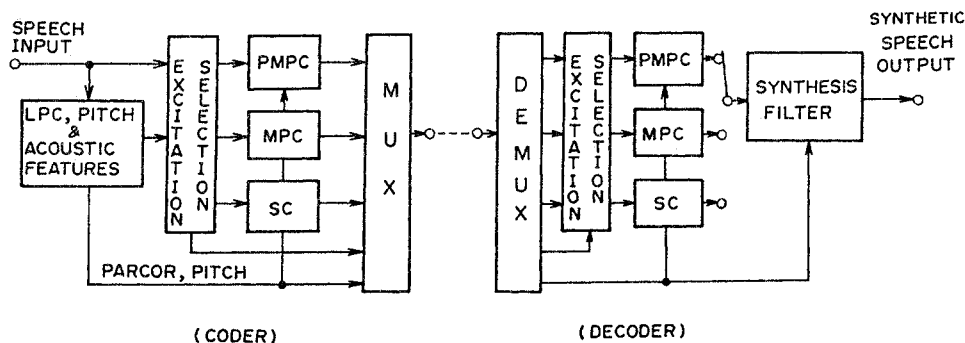


Fig.4. Coder and decoder structures of SPMEX.

speech power regions. On the other hand, S/N of PP2 is sometimes better than SPMEEX in the low speech power regions. Reasons are that the stochastic codebook excitation is often used in SPMEEX in low speech power regions, and multi-pulse excitation of PP2 gives better S/N than stochastic codebook excitation in SPMEEX in such regions. Informal listening tests showed S/N performance does not fully reveal synthetic speech quality difference. Thus, subjective evaluation of speech quality was carried out.

### 3-2. SPEECH QUALITY EVALUATION

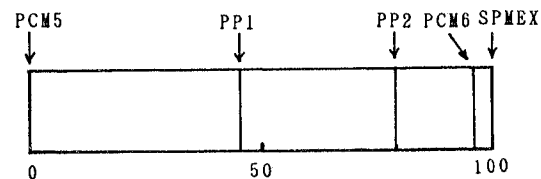
Synthetic speech quality was evaluated using a paired comparison method [11] with 5 point categories. Evaluated speech coding methods are 4.8 kb/s SPMEEX, PP2 and PP1. In addition to these methods, 48 kb/s  $\mu$ -law PCM (PCM6) and 40 kb/s  $\mu$ -law PCM (PCM5) were included in the evaluation as references. Four short Japanese sentences (uttered by 2 male and 2 female speakers) were used as speech samples. Six female listeners took part in the evaluation. Synthetic speech was presented through headphones in an anechoic quiet room.

Figure 6 shows the evaluation results. In this figure, speech coding which obtained the highest score is plotted at 100%. The one which had the lowest score is plotted at 0%. The speech quality for SPMEEX is high and judged to be equivalent to PCM6. The quality is significantly higher than PP2 and PP1. The statistical analysis results show that the difference between SPMEEX and PCM6 is not significant, but the difference is significant between SPMEEX and PP2 and PP1 individually. The evaluation results also show that SPMEEX improves speech quality for vowel-like speech, as well as for frication-like speech, although the SPMEEX S/N is lower than PP2 for frication-like speech. These improvements are due to adoption of improved pitch interpolation multipulse excitation for vowel-like speech and stochastic codebook excitation for frication-like speech.

### 4. CONCLUSION

SPMEEX uses multiple excitation signals, such as improved pitch interpolation multi-pulse, conventional multi-pulse and stochastic codebook. A suitable excitation signal is selected using acoustic features of speech in the frame. Improved pitch interpolation multi-pulse is proposed to further improve speech quality for vowel-like speech. In this method, the excitation signal in the frame is represented by multi-pulse during one pitch period. For the other pitch periods in the same frame, gain and phase adjusting coefficients are calculated. Subjective evaluation results show that 4.8 kb/s SPMEEX produces high-quality synthetic speech, which is equivalent to 48 kb/s  $\mu$ -law PCM, and which is significantly higher than 4.8 kb/s multi-pulse coding with pitch interpolation and 4.8 kb/s pitch predictive multi-pulse coding.

The author would like to thank Mr. T. Watanabe for his discussion, and Mr. H. Kumagai and Mr. E. Hanada for carrying out simulation and subjective evaluation experiments.



PP1: 4.8kb/s pitch predictive multi-pulse coding  
PP2: 4.8kb/s multi-pulse coding with pitch interpolation

Fig.6. Subjective evaluation results.

### REFERENCES

- [1] P. Vary et al., "Speech codec for the European mobile radio system," Proc. ICASSP, pp.227-230, 1988.
- [2] M. Schroeder and B. Atal, "Code-excited linear prediction: High quality speech at very low bit rates," Proc. ICASSP, pp.937-940, 1985.
- [3] Y. Yatsuzuka et al., "A variable rate coding by APC with maximum likelihood quantization for 4.8kb/s to 16kb/s," Proc. ICASSP, pp.3071-3074, 1986.
- [4] T. Moriya and M. Honda, "Transform coding of speech using a weighted vector quantizer," IEEE J. Sel. Areas, Commun., pp.425-431, 1988.
- [5] T. Taniguchi et al., "Multimode coding: Application to CELP," Proc. ICASSP, pp.156-159, 1989.
- [6] K. Ozawa and T. Araseki, "Low bit rate multi-pulse speech coder with natural speech quality," Proc. ICASSP, pp.457-460, 1986.
- [7] K. Ozawa, "A high-quality 4.8kb/s speech coding using multiple types of excitation signals," Tech. Rep. Study Group Speech, IEICE Japan, SP89-2, pp.9-16, 1989.
- [8] B. Atal and J. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," Proc. ICASSP, pp.614-617, 1982.
- [9] K. Ozawa et al., "A study on pulse search algorithms for multiple excited speech coder realization," IEEE J. Sel. Areas, Commun., pp.133-141, 1986.
- [10] K. Ozawa et al., "High quality multi-pulse speech coder with pitch prediction," Proc. ICASSP, pp.1689-1692, 1986.
- [11] H. Scheffe, "An analysis of variance for paired comparisons," Am. Statist. Assoc., J., pp.381, 1952.