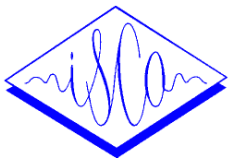

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

# VARIABLE BIT-RATE SPEECH CODING BASED ON PSI-CELP

*Hitoshi Ohmuro, Kazunori Mano, and Takehiro Moriya*

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

## ABSTRACT

This paper proposes a new variable bit-rate PSI-CELP speech coding method that switches four modes roughly corresponding to silence, unvoiced regions, voiced transient regions, and voiced stationary regions. The coder modes are determined by an open-loop procedure every two subframes (20 ms) using the feature parameters extracted from the input speech. The proposed method is based on the algorithms used in the PDC half-rate standard, but the LSP coder uses inter-frame predictive vector quantization every two subframes instead of matrix quantization every four subframes, and pitch parameters are coded by using finite-state inter-subframe prediction in voiced stationary regions to achieve good quality with fewer bits. We examined the performance using Japanese speech data which was approximately 83% active. A listening test using non-specialist subjects showed that the proposed method achieves much better quality at average bit-rate of 2.53 kbit/s over all speech data or at an average bit-rate of 2.88 kbit/s without silence than the fixed bit-rate PDC half-rate standard (3.45 kbit/s). Even when the input speech was noisy, the proposed method still achieved better quality with fewer bits than the PDC standard. This method considerably reduces the bit-rate not only in silence and unvoiced regions but also in voiced regions, so it achieves high-quality variable-low-bit-rate speech coding.

## 1 INTRODUCTION

In recent years, low bit-rate speech coding has been receiving a lot of attention for use in mobile telephone and voice storage applications. The mobile telephone system has insufficient capacity for the increasing number of users. Voice storage systems, for example voice mailing systems and multi-media communication systems, need a large voice database with small memory requirement. Low bit-rate speech coding is a promising solutions to these problems.

Pitch synchronous innovation CELP (PSI-CELP) [1][2] was selected as the half-rate codec standard [3] for the next generation mobile telephone system in Japan by a public competition. PSI-CELP has a CELP [4] structure, and is designed to improve the quality using a reasonable size of memory and reasonable computational complexity. The bit-rate is 3.45 kbit/s (the PDC standard needs an extra 2.15 kbit/s redundancy), and the subjective quality is equivalent to or better than that of the full-rate VSELP (6.7 kbit/s source + 4.5 kbit/s redundancy) [5].

This paper proposes a new variable bit-rate speech coding method based on PSI-CELP. A variable bit-rate speech coder uses a different number of bits for each input frame and thus reduces the average bit-rate. The variable bit-rate speech coder is more efficient than the fixed-rate speech coders, because silent regions hardly need any bits and stationary regions only need a much smaller number of bits to achieve sufficient quality than transient regions. Variable bit-rate speech coders are especially useful for voice storage applications. Although most conventional physical communication channels are designed for fixed-rate coders, a new communication standard called " code division multiple access (CDMA)" is also well suited for variable bit-rate coders.

A typical scheme for variable bit-rate coding previously proposed is multi-mode coding [6][7]. Cellario and Sereno proposed a seven-mode variable rate CELP (VR-CELP) [8]. The points at issue in this method are that it needs 0.3 kbit/s of mode information and that the closed-loop mode selection requires many computations. On the other hand, Paksoy and Gersho proposed variable rate phonetic segmentation (VRPS) [9]. Though this method solves previous problems, its bit reduction in voiced regions is insufficient for Japanese speech.

In this paper, we propose a four-mode variable bit-rate speech coding method based on PSI-CELP. The modes are determined by an open-loop procedure. Our method extends the efficiency of PSI-CELP by reducing the bit-rate, especially in voiced stationary regions, and achieves high quality low average bit-rate speech coding.

## 2 PSI-CELP SPEECH CODING

PSI-CELP uses the following techniques to improve the quality for voiced regions with a reasonable size of memory and real time processing.

1. Pitch synchronization of the random code vectors by the pitch period of the adaptive code vector.
2. Closed-loop switching between the adaptive codebook and the fixed codebook.
3. Two-channel conjugate structure random codebooks.
4. Vector quantization of adaptive and random code vector gains.
5. FIR type perceptual weighting filter.
6. Codebook search by delayed decision.

Figure 1 shows the PSI-CELP encoder structure and Fig. 2 shows the pitch synchronization concept of a random code vector. Pitch synchronization depends on whether the fixed or adaptive codebook is used. When the adaptive codebook is used, it is constructed by repeating the beginning of the adaptive lag length of the random code vector until the length of the vector becomes equal to the subframe length. When the fixed codebook is selected instead, on the other hand, the pitch synchronization procedure is not applied to the random codebooks.

Table 1 shows the bit allocation to the code vectors in the PDC half-rate standard.

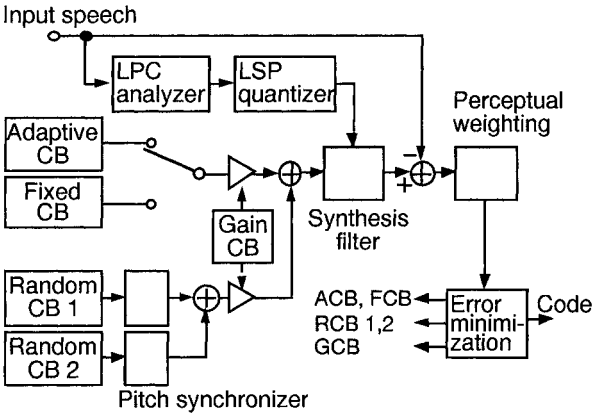


**Fig. 1** PSI-CELP encoder structure.

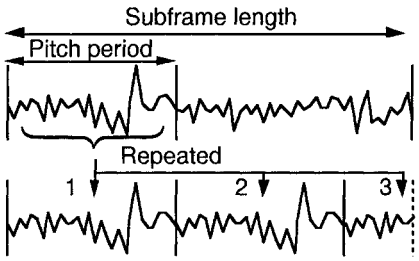


**Fig. 2** Pitch synchronization concept.

**Table 1** Bit allocations of the PDC half-rate standard. The subframe length is 10 ms, and a frame has four subframes.

| | |
|---|---|
| Power | 7 bit/frame |
| LSP | 31 bit/frame |
| Adaptive / fixed code vector | 8×4 bit/frame |
| Random code vector 1 | (4+1)×4 bit/frame |
| Random code vector 2 | (4+1)×4 bit/frame |
| Gain | 7×4 bit/frame |
| Total | 138 bit/frame (3.45 kbit/s) |

## 3 LSP QUANTIZATION FOR A VARIABLE RATE CODER

The proposed variable rate coder switches mode every two subframes (20 ms). We therefore replaced the line spectrum pair (LSP) coder of the PDC standard, which is split matrix quantization of second and fourth subframe parameters, with inter-frame predictive vector quantization (PVQ) [10][11] every two subframes. Denoting the $2n$-th-subframe quantized vector by $q_{2n}$, the $2(n\text{-}i)$-th-subframe quantized vector by $q_{2(n-1)}$ and the $2n$-th-subframe *transmitted* code vector by $x_{2n}$, $q_{2n}$ is defined by

$$q_{2n} = A_0 x_{2n} + \sum_{i=1}^{M} A_i q_{2(n-i)}, \qquad (1)$$

where $A_i$ is the diagonal matrix of predictive coefficients and $M$ is the inter-frame prediction order. We set $M$ to 2 in this study.

In addition to the LSP coding algorithm, we lengthened the analysis window from 20 ms (160 points) of the PDC standard to 32 ms (256 points). A short window causes hasty changes of LSP parameters, so the inter-frame PVQ performance becomes worse.

In this study, we allocated 15 bits per two subframes (20 ms) for LSP coding, whereas the PDC standard uses 31 bits per four subframes. The total bit-rate is therefore 0.025 kbit/s lower than the standard.

## 4 VARIABLE BIT-RATE PSI-CELP

This paper proposes a new variable bit-rate PSI-CELP speech coding that switches four modes roughly corresponding to silence, unvoiced regions, voiced stationary regions, and voiced transient regions. The mode of each subframe is determined by an open-loop procedure every two subframes.

### 4.1 Mode Selection by an Open-Loop Procedure

The coder modes are determined by the following parameters.

1. Power.
2. Maximum value of residual correlation.
3. Pitch period (open-loop).
4. Spectrum difference.
5. Previous lag value of the adaptive codebook.
6. Power of synthesized speech.

First the feature parameters described above are extracted from the input speech. If the power of the subframe is less than the threshold for mode 0, the subframe is assumed to be silence, and mode 0 is selected. Otherwise, the maximum value of the residual correlation is compared with the threshold for mode 1. If the value is less than the threshold, and if the power is less than the threshold for mode 1, the subframe is assumed to be an unvoiced region, and mode 1 is selected. Otherwise, the subframe is assumed to be a voiced region, and if all of the differences of power, pitch, and spectrum between the current subframe and the previous subframe are less than the threshold for mode 2, the subframe is assumed to be a stationary region, and mode 2 is selected. Otherwise, the subframe is assumed to be a transient region, and mode 3, which uses the highest bit-rate, is selected. Each of the above thresholds is designed so that a mode with as low a rate as possible is selected with almost no subjective degradation.

### 4.2 Bit Allocation in Each Mode

Table 2 shows the bit allocation to the code vectors in each mode. The values in the power row show that the power parameters are quantized every four subframes, and the values in the LSP row show that the LSP parameters are quantized every two subframes. Although the modes are switched every two subframes, 3 bits every 4 subframes (3/4 bits) are used for power coding only when all of the four subframes belong to mode 0, otherwise 7/4 bits are used even if a subframe belongs to mode 0. The excitation coder for mode 3 has the same structure as the PDC standard except for the bit allocations to the random code vectors. In order to follow the changes in the transient regions, more bits are used for the random code vectors than the PDC standard. Every codebook is designed only for the corresponding mode even if the number of allocated bits is the same as that in another mode.

### 4.3 Pitch Coding with "Finite-State Inter-Subframe Prediction"

In order to achieve good quality with fewer bits in stationary regions, the lag parameter of the adaptive codebook is quantized using inter-subframe prediction. Denoting the current subframe lag value by $l(i)$ and the one previous subframe lag value (which has already been

quantized) by $l(i\text{-}1)$, $l(i)$ is given by

$$l(i) = a_{jk}\, l(i\text{-}1) + b_{jk} + c_{jk}, \text{ and} \quad (2)$$

$$k: l(i\text{-}1) \in S_k, \quad (3)$$

where $a_{jk}$ is the predictive coefficient, which is typically 1.0, 2.0, or 0.5; $b_{jk}$ is the integer part of the residual; $c_{jk}$ is the fractional part of the residual; $k$ is the state index that $l(i\text{-}1)$ belongs to; and $j$ is the codebook index. Only index $j$ is transmitted in voiced stationary regions. In this study, we set the number of states to 4 and allocated 5 bits to the codebook. The fractional resolutions are defined in the finite-state codebooks, and they mainly depend on lag differences between subframes, whereas they depend on only lag length in the PDC standard. This means that the fractional resolution in a region where lag values change very slowly is higher than that in the PDC standard and that this pitch coding method can be more efficient than the standard in really 'stationary' regions.

**Table 2** Bit allocations in each mode. (bits/subframe)

| Mode | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Feature | Silence | Unvoiced | Voiced | |
| | | | Stationary | Transient |
| Power | 3 / 4 | 7 / 4 | 7 / 4 | 7 / 4 |
| LSP (*1) | 4 / 2 | 15 / 2 | 11 / 2 | 15 / 2 |
| Adaptive CV | – | – | 5 (*2) | 8 |
| Fixed CV | – | – | – | |
| Random CV 1 | 2 | 4 | 4+1 (*3) | 5+1 (*3) |
| Random CV 2 | – | – | 4+1 (*3) | 5+1 (*3) |
| Gain | 2 | 4 | 4 | 7 |
| Total (kbit/s) | 0.675 | 1.725 | 2.625 | 3.625 |
| Mode info. (kbit/s) | 0.1 | 0.1 | 0.1 | 0.1 |

*1: Inter-frame predictive VQ
*2: Finite-state inter-subframe prediction
*3: Shape + sign

## 5 CODING EXPERIMENTS

We examined the performance of our method by using Japanese telephone speech data. We used 300 short sentences for training codebooks and another 32 short sentences for testing (open data). Each sentence was 2.5-4.5 seconds long. We removed the silent regions at the beginning and the end of the sentences, so the data had much less silence than actual speech conversation. The active regions of our speech data were approximately 83% of all regions. Sample waveforms are shown in Fig. 3.

We compared the performance of the following four methods. The codebooks in the PDC standard were also trained again using our database.

Method 1: Fixed rate PDC half-rate standard (3.45 kbit/s).

Method 2: Fixed rate PSI-CELP described in section 3. (3.425 kbit/s, 32-ms window and 20-ms-frame LSP coding)

Method 3: Four-mode variable rate PSI-CELP that reduces the number of bits only in silence and unvoiced regions.

Method 4: Proposed four-mode variable bit-rate PSI-CELP.

The coding algorithm and the bit allocation for voiced regions in method 3 are the same as those in method 2.

### 5.1 Results of the Open-Loop Mode Selection

Table 3 shows the frequency with which each mode was selected by the open-loop procedure. Examples of modes selected for various actual waveforms are shown in Fig. 3.

In this experiment, mode 0 (for silence) was selected with a frequency of 16.7%. Although the frequency of silence depends on speech data, more than half of the speech in a typical conversation is usually silence. Therefore our speech data had much less silence than usual.

Mode 1 (for unvoiced regions) was selected with a frequency of 14.5%. Unvoiced consonants have a strong potential for bit reduction because the subjective quality is degraded only a little, even if random noise is used for the excitation signals. However, mode 1 was not frequently selected because a consonant rarely lasts more than 20 ms as shown in Fig. 3, and because Japanese speech does not have as many consonants as other languages.

In order to reduce the average bit-rate over all speech regions, mode 2 should be used in voiced regions as frequently as possible with almost no quality degradation, because most regions of Japanese speech are vowels. In this experiment, mode 2 was selected with a frequency of 42.6%, and mode 3 was selected with a frequency of 26.2%. As shown in Fig. 3, mode 3 is only used near the 'onset' and where the feature of the speech changes suddenly.

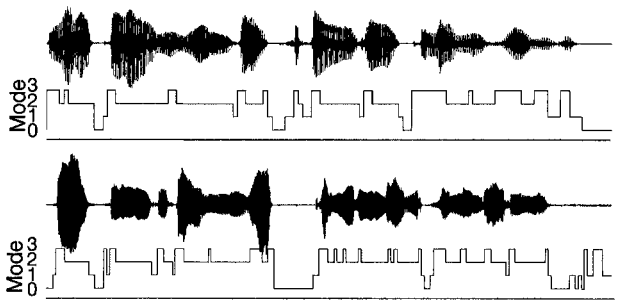


**Fig. 3** Sample waveforms and mode correspondence.

### 5.2 Performance Evaluation

In order to examine subjective quality, we carried out a listening test using 9 adult males and 15 adult females from the general public. The quality was evaluated by 5 ranks, and the mean opinion score (MOS) was converted to the equivalent Q value using MNRU references. Table 4 shows the signal-to-noise ratio (SNR) and the listening test results, where SNR does not always correspond to subjective quality especially in the case of variable rate coders.

Method 2, which is a fixed rate coder and the basis of the proposed variable rate coder, achieved almost the same quality as the PDC standard, that is, this 20-ms-frame LSP coding has the same efficiency as the 40-ms-frame matrix quantization in the standard.

Method 3, which is a variable rate coder and reduces the number of bits only in silence and unvoiced regions, achieved better subjective quality than methods 1 and 2. The switching of the codebooks according to the mode is the reason the quality of method 3 is better than that of method 2. The average bit-rate calculated only in active subframes (without silence) cannot be much reduced because Japanese does not have many unvoiced regions as described above.

The proposed variable rate method achieved 2.43+0.1 kbit/s of average bit-rate over all speech data or 2.78+0.1 kbit/s of average bit-rate calculated only in active subframes (without silence), whereas the PDC standard always uses 3.45 kbit/s. The 0.1 kbit/s represents the bit-rate for the mode information. With this proposed method, the equivalent Q value was 1.4 dB better than the PDC standard; moreover

the average bit-rate was considerably reduced.

Figure 4 shows the subjective quality of these four methods and other fixed-rate coders with various bit allocations as a function of average bit-rate over all speech data or without silence. The subjective quality of fixed-rate coder were linearly degraded as the bit-rate became lower. As a result, the proposed variable-rate coder achieved more than 4 dB better quality at the average bit-rate over all speech data and approximately 2.5 dB better quality at the average bit-rate even without silence than the fixed-rate coders.

### 5.3 Performance for Noisy Speech

In order to examine the influence of background noise on mode switching, we applied noisy speech to the coders and evaluated the quality. We added ITU-T motor noise so that the total SNR became 30 dB or 15 dB.

The open-loop mode selection results are shown in table 3. When the SNR of input speech was 30 dB, most non-active regions were coded using mode 1. However, mode 2 was still selected as frequently as when the input speech was clean. When the SNR of input speech was 15 dB, on the other hand, mode 2 was selected less frequently, and mode 3 was selected as frequently as mode 2.

Table 5 shows the average bit-rate over all speech data and the listening test results. Even when the input speech was noisy, the subjective quality coded with the proposed variable rate coder was better than that with the fixed rate coders.

## 6 CONCLUSION

This paper proposed a new variable bit-rate PSI-CELP speech coding method that switches four modes roughly corresponding to silence, unvoiced regions, voiced transient regions, and voiced stationary regions. First the coder modes are determined by an open-loop procedure every two subframes (20 ms) using the feature parameters extracted from the input speech. In our proposed method, LSPs are coded by using inter-frame predictive vector quantization, and pitch is coded by using finite-state inter-subframe prediction in voiced stationary regions. We examined the performance using Japanese speech data which was approximately 83% active, of which 26.2% were selected as voiced transient subframes and 42.6% were selected as voiced stationary subframes. The listening test using 24 non-specialist subjects showed that the proposed method achieves 1.4 dB better equivalent Q at an average bit-rate of 2.53 kbit/s over all speech data or at average bit-rate of 2.88 kbit/s without silence than the fixed bit-rate PDC half-rate standard (3.45 kbit/s). Even if the input speech was noisy, the proposed method still achieved better quality with fewer bits than the PDC standard. This method considerably reduces the bit-rate not only in silence and unvoiced regions but also in voiced stationary regions, and achieves high-quality low-bit-rate speech coding for variable bit-rate or voice storage applications.


### ACKNOWLEDGMENTS

The authors thank Dr. Nobuhiko Kitawaki for his guidance. The authors also thank all the members of the Speech Coding Group for their valuable discussions and suggestions.


### REFERENCES


[1] S. Miki, K. Mano, H. Ohmuro and T. Moriya, "Pitch Synchronous Innovation CELP (PSI-CELP)," *Proc. of EUROSPEECH* '93, pp. 261-264, 1993

[2] S. Miki *et al.*, "A Pitch Synchronous Innovation CELP (PSI-CELP) Coder for 2-4 kbit/s," *Proc. of ICASSP*-94, vol. 2, pp. 113-116, 1994

[3] T. Ohya, H. Suda and T. Miki, "Pitch Synchronous Innovation CELP (PSI-CELP) –PDC half-rate speech CODEC–," *Technical Report of IEICE*, RCS93-78, 1993

[4] M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates," *Proc. of ICASSP*-85, pp. 937-940, 1985

[5] F. Masui and M. Oguchi, "Activity of the Half-Rate Speech CODEC Algorithm Selection for the Personal Digital Cellular System," *Technical Report of IEICE*, RCS93-77, 1993

[6] M. Yong and A. Gersho, "Vector Excitation Coding with Dynamic Bit Allocation," *Proc. of GLOBECOM*, pp. 290-294, 1988

[7] T. Taniguchi, S. Unagami and R. M. Gray, "Multimode Coding: Application to CELP," *Proc. of ICASSP*-89, pp. 156-159, 1989

[8] L. Cellario and D. Sereno, "Variable Rate Speech Coding for UMTS," *IEEE Workshop on Speech Coding for Telecommun.*, pp. 1-2, 1993

[9] E. Paksoy and A. Gersho, "A Variable Rate Speech Coding Algorithm for Cellular Networks," *IEEE Workshop on Speech Coding for Telecommun.*, pp. 109-110, 1993

[10] Y. Shoham, "Vector Predictive Quantization of the Spectral Parameters for Low Rate Speech Coding," *Proc. of ICASSP*-87, pp. 2181-2184, 1987

[11] H. Ohmuro, T. Moriya, K. Mano and S. Miki, "Vector Quantization of LSP Parameters Using Moving Average Interframe Prediction," *Trans. of IEICE*, vol. J77-A, No. 2, pp. 303-313, 1994


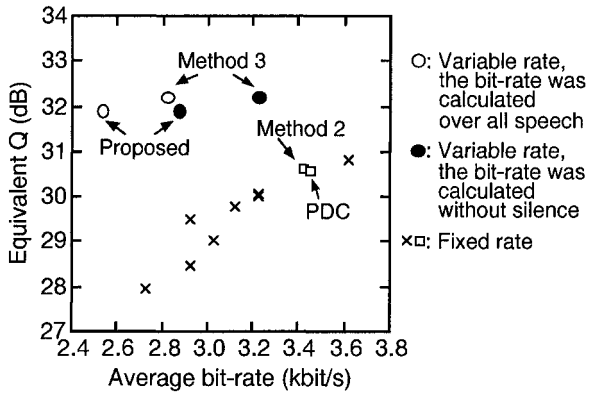


**Fig. 4** Subjective quality as a function of the average bit-rate.

**Table 3** Open-loop mode selection results.

| | Selected rate (%) | | |
|---|---|---|---|
| SNR of input speech | Clean | 30 dB | 15 dB |
| Mode 0 (Silence) | 16.7 | 0.85 | 0.0 |
| Mode 1 (Unvoiced) | 14.5 | 32.6 | 31.7 |
| Mode 2 (Voiced stationary) | 42.6 | 41.1 | 34.1 |
| Mode 3 (Voiced transient) | 26.2 | 25.5 | 34.2 |

**Table 4** Experimental results.

| | Equivalent Q (dB) | Average bit-rate (kbit/s) | | SNR (dB) |
|---|---|---|---|---|
| | | All (*1) | Active (*2) | |
| Method 1 | 30.6 | 3.450 | 3.450 | 10.33 |
| Method 2 | 30.6 | 3.425 | 3.425 | 10.42 |
| Method 3 | 32.2 | 2.722+0.1 | 3.129+0.1 | 10.30 |
| Method 4 | 31.9 | 2.433+0.1 | 2.782+0.1 | 9.89 |

*1 Average bit-rate over all speech data
*2 Average bit-rate calculated without silence subframes

**Table 5** Experimental results for noisy speech.

| | Equivalent Q (dB) | | Average bit-rate (kbit/s)* | |
|---|---|---|---|---|
| Input SNR | 30 dB | 15 dB | 30 dB | 15 dB |
| Method 1 | 20.3 | 12.0 | 3.450 | 3.450 |
| Method 2 | 20.3 | 11.9 | 3.425 | 3.425 |
| Method 3 | 23.1 | 13.8 | 2.848+0.1 | 2.885+0.1 |
| Method 4 | 22.6 | 13.0 | 2.570+0.1 | 2.681+0.1 |

* Average bit-rate over all speech data