A 2.4 KBPS VARIABLE BIT RATE ADP-CELP SPEECH CODER

M. Oshikiri and M. Akamine

Kansai Research Laboratories, Toshiba Corporation,

6-26, Motoyama-Minami-Cho, 8-Chome, Higashinada-Ku, Kobe, 658 Japan

ABSTRACT

This paper presents a variable bit rate ADP-CELP (Adaptive Density Pulse Code Excited Linear Prediction) coder that selects one of four kinds of coding structure in each frame based on short time speech characteristics. To improve speech quality and reduce the average bit rate, we have developed a speech/non-speech classification method using spectrum envelope variation, which is robust for background noise. In addition, we propose an efficient pitch lag coding technique. The technique interpolates consecutive frame pitch lags and quantizes a vector of relative pitch lags consisting of variation between an estimated pitch lag and a target pitch lag in plural subframes. The average bit rate of the proposed coder was approximately 2.4 kbps for speech sources with activity factor of 60%. Our subjective testing indicates the quality of the proposed coder exceeds that of the Japanese digital cellular standard with rate of 3.45 kbps.

1. INTRODUCTION

Short-term entropy of speech varies widely [1]. A variable rate coder is able to reduce average bit rate by exploiting the nature of speech. A variety of variable rate coders has been reported [1]-[6], which are applied to speech communication systems and speech storage applications.

In variable rate coders, serious quality degradation occurs when an active speech segment is classified into a non-speech category. This is because the non-speech coding structure is generally designed at a low bit rate (e.g. 1 kbps or less) to enhance the capacity of frequency or memory resources. Several investigations on speech/non-speech classification have been reported [6]-[8]. These are based on a combination of indicators: a frame power, a prediction gain and a tilt in spectrum. The frame power is an effective indicator but it is not good in a high-level noise environment. Spectrum information, a prediction gain and a tilt in spectrum, are useful to compensate for the frame power. However, these indicators often fail to classify when spectrum characteristics of noise are similar to those of speech.

There is another problem in variable rate coders. The problem is the average bit rate considerably increases for speech sources with high activity factor. Therefore, it is important to design coding structure for active speech at low bit rates. An efficient coding technique of the pitch lag in the stationary voiced segments is needed to reduce the bit rate for active speech and several techniques have been discussed in the literature [2], [3], and [10]. However, these techniques are not efficient enough because they merely exploit the correlation of the pitch lags in a frame.

We have developed two important techniques: speech/non-speech classification method based on spectrum envelope variation (SEV) and relative pitch vector quantization (RPVQ) with interpolation of consecutive frame pitch lags. SEV is independent of a signal power and more sensitive to detect spectrum variation compared with other indicators, and therefore it can classify correctly even in the noisy environment. RPVQ exploits not only the correlation of the pitch lags in a frame but also the correlation over two consecutive frames. Therefore, the pitch lag can be represented at a very low bit rate by RPVQ.

We also propose a variable bit rate ADP-CELP coder that introduces the techniques mentioned above. The variable bit rate ADP-CELP coder selects one of four kinds of coding structure in each frame based on short time speech characteristics. Each coding structure is adapted for non-speech, unvoiced, stationary voiced and non-stationary voiced segments, and each bit rate is 0.53, 2.67, 3.17 and 4.1 kbps, respectively.

This paper is organized as follows. In section 2, we describe the speech/non-speech classification method based on SEV and evaluation results for several noise conditions. In section 3, we describe RPVQ. In section 4, we describe the structure of the variable bit rate ADP-CELP coder. In section 5, we discuss the subjective evaluation test of RPVQ and the variable bit rate ADP-CELP coder.

2. SPEECH/NON-SPEECH CLASSIFICATION

2.1 Algorithm

Spectrum envelope variation (SEV) between non-speech frames and the current frame is defined by

$$SEV = 10\sqrt{\frac{1}{M}\sum_{m=0}^{M-1}\left(\log_{10}\frac{\left|1-\sum_i \alpha_e(i)\exp(j2\pi mi/M)\right|^2}{\left|1-\sum_i \alpha_s(i)\exp(j2\pi mi/M)\right|^2}\right)} \quad (1)$$

where $\alpha_e(i)$ are averaged LPC parameters of non-speech frames and $\alpha_s(i)$ are LPC parameters of the current frame. Speech/non-speech classification is performed by SEV and power variation (PV). PV is given by

$$PV = P_s - \varepsilon \cdot P_e \qquad (2)$$

where P_e is averaged power of non-speech frames, P_s is power of the current frame and ε is a constant to keep stable classification.

Let T_{SEV} is a variable threshold depending on P_e, the classification procedure is performed according to the following rules.

If [PV > 0.0] then	speech category
else if [SEV > T_{SEV}] then	speech category
else	non-speech category

After classification, $\alpha_e(i)$ and P_e, are updated by

$$\omega_e^{new} = (1-\beta)\omega_s - \beta\omega_e \qquad (3)$$

$$P_e^{new} = (1-\gamma)P_s - \gamma P_e \qquad (4)$$

for the next frame, where $\omega_e(i)$ is a line spectrum frequency parameter corresponding to $\alpha_e(i)$, β and γ are variables for updating. The variables are determined depending on the classification result of the current frame.

2.2 Experimental result

We prepared approximately 60 seconds of speech data and correct classification data. The correct classification data is manually made. We added one of four types of background noise (car noise, babble noise, street noise, and tilt noise) to speech data to make test data used for measurements of correspondence ratio (CR). CR is expressed as

$$CR = \frac{NF_{right}}{NF_{right} + NF_{wrong}} \times 100 \qquad (5)$$

where NF_{right} is the number of frames while the speech/non-speech classification results in the correct data. Similarly, NF_{wrong} is the number of frames while the classification fails.

The result is shown in Figure 1. The symbol POW+SEV represents the proposed classification method, POW represents the conventional method using frame power and POW+K(1) represents the conventional method using frame power and a tilt of spectrum. Figure 1 shows that the proposed method improves CR compared with the conventional methods for all conditions. The improvement is especially marked at low SNRs. Therefore, we can conclude the proposed method is superior as a speech and non-speech classifier even in the noisy environment.

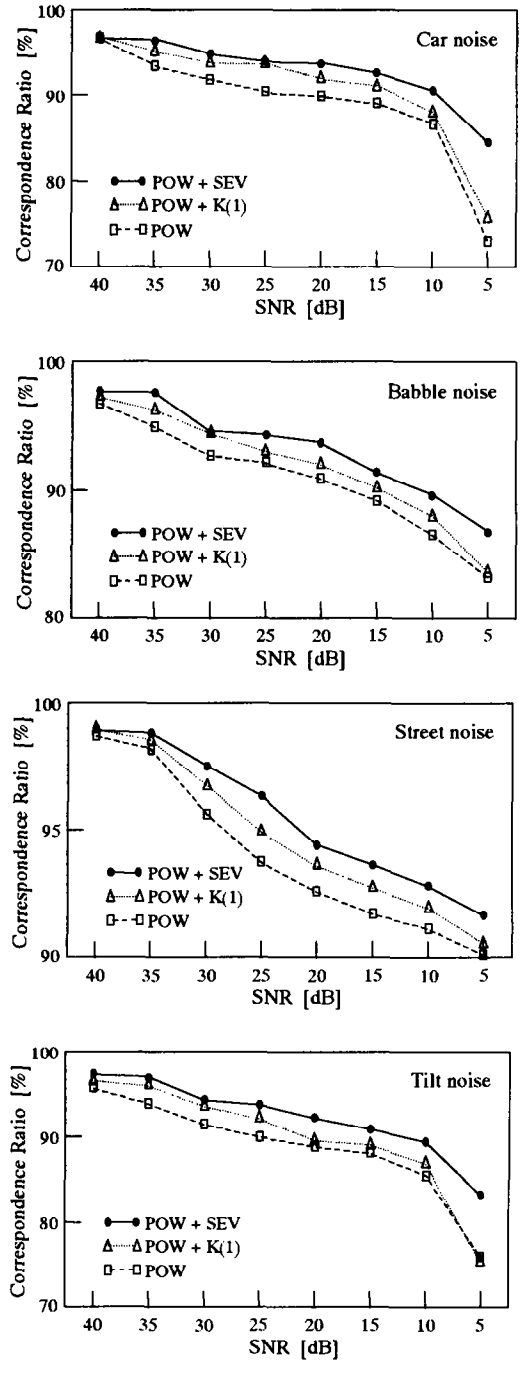

Figure 1. Signal-to-noise ratio and correspondence ratio, (a) Car noise, (b) Babble noise, (c) Street noise and (d) Tilt noise.

3. RELATIVE PITCH VECTOR QUANTIZAION

The framework of the relative pitch vector quantization (RPVQ) is shown in Figure 2. In RPVQ, subframe pitch lag is estimated by interpolation of the previous frame pitch lag and the current frame pitch lag. The pitch lag in stationary voiced segments varies smoothly, so that we can expect the estimated pitch lag is close to the target pitch lag. An error vector between the target pitch lag and the estimated pitch lag in plural subframes is quantized based on a minimum squared error criterion. In this way, RPVQ can remove not only the correlation in a frame but also the correlation over two consecutive frames. The RPVQ technique also avoids

degradation caused by the multiple errors of the pitch lag because it finds the best pitch lag around the estimated lag.

Let T(m) be a current frame pitch lag calculated by a correlation analysis method, then the pitch lag of the k-th subframe is estimated by the following interpolation.

$$STP(k) = (1-\xi(k))\cdot T(m-1) + \xi(k)\cdot T(m) \quad (6)$$

Where T(m-1) is a pitch lag of the previous frame and ξ(k) is a constant for interpolation. The adaptive codevector index of the k-th subframe is calculated by

$$\tau(j,k) = STP(k) + v(j,k) \quad (7)$$

where v(j,k) is the j-th codevector of relative pitch lags. The optimum codevector of relative pitch lags is selected to maximize the following equation.

$$C(j) = \sum_{k\in S} \frac{(r_k^T H_k P_{\tau(j,k)})^2}{\left\| H_k P_{\tau(j,k)} \right\|^2} \quad (8)$$

Where $P_{\tau(j,k)}$ is the adaptive codevector with pitch lag τ(j,k), r_k is the target vector and H_k is the impulse response matrix of the weighted synthesis filter. Evaluation of RPVQ is discussed in section 5.

4. STRUCTURE OF THE VARIABLE BIT RATE ADP-CELP CODER

The variable bit rate ADP-CELP coder consists of four kinds of coding strategy adapted for non-speech (NS), unvoiced (UV), stationary voiced (SV) and non-stationary voiced (NV) segments. In each frame, the LPC parameters quantization method and the excitation model are changed depending on the selected mode. The frame size of the coder is 30 ms and the subframe size is 7.5 ms. The bit allocation is shown in Table 1.

	NS	UV	SV	NV
LPC	9	22	22	22
ACB	-	-	5+5x2	8x4
SCB	-	8x4	9x4	12+9x3
Gain	5	6x4	5x4	7x4
Mode	2	2	2	2
Bits/frame	16	80	95	123
Bit rate (kbps)	0.53	2.67	3.17	4.10

Table 1. Bit allocations of the proposed coder.

4.1 Mode Decision Algorithm

The mode decision algorithm consists of three classifiers: a speech/non-speech classifier, a voice/unvoice classifier and a stationary/non-stationary classifier. First, the speech/non-speech classifier based on SEV determines whether the input frame belongs to the speech category or the non-speech

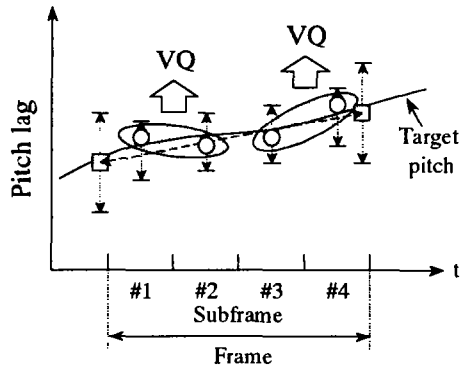

Figure 2. Framework of RPVQ.

category. If the frame belongs to the non-speech category, NS mode is selected. Otherwise, the voice/unvoice classification is carried out. The classification is based on six acoustical parameters: a frame power, a low-band power, a tilt of spectrum, a zero crossing ratio, a short-term prediction gain and a long-term prediction gain. If the frame belongs to the unvoiced category, UV mode is selected. Otherwise, the stationary/non-stationary classification is carried out based on continuity of consecutive frame pitch lags. If the variation of the pitch lag is smaller than the pre-determined threshold, SV mode is selected. Otherwise, NV mode is selected.

4.2 LPC Parameter Quantization

The LPC parameters are quantized in a line spectral frequency (LSF) representation, which is known to provide efficient quantization. The LSF parameters are predicted by a first-order AR predictor, and the residual is vector quantized. In the variable bit rate ADP-CELP coder, two kinds of vector quantization scheme are introduced. One is used in US mode. The residual is split into low-band, middle-band and upper-band vector, and each vector is quantized using a 3-bit codebook.

Another scheme is employed in UV, SV and NV mode. The residual is split into low-band and upper-band vector. Each vector is quantized by a two-stage vector quantizer. The total bits for vector quantization are 22 bits including one bit for switching a prediction coefficient.

4.3 Excitaion Model

Excitation model of each mode is differently designed in order to represent a variety of speech properties. In NS mode, normalized Gaussian source is used as excitation. Hence, the excitation parameter to be transmitted to the decoder is only a gain parameter. The gain is smoothly interpolated in the decoder. This is to avoid the degradation caused by power discontinuity.

In UV mode, the variable bit rate ADP-CELP coder eliminates an adaptive codebook but exploits a stochastic codebook because unvoiced speech is non-periodic signal. The stochastic codebook contains clipped random sequences.

In NV and SV mode, the variable bit rate ADP-CELP coder exploits both the adaptive codebook and the stochastic codebook to represent excitation signal. The NV mode is applied when the pitch lag and the gain are rapidly changed. Therefore, many bits need to be allocated in order to follow the changes. In addition, the stochastic codebook possesses ADP (Adaptive Density Pulse) structure [9], which dynamically changes the pulse interval. The ADP structure is suitable for NV mode.

The SV mode works for speech segments where the periodic speech continues stably. The relative pitch vector quantization technique described in section 3 is applied to SV mode to encode pitch lags. Here, we allocate five bits for a frame pitch lag and five bits for a vector of the relative pitch lags of two consecutive subframes. The number of total bits for pitch lag is 15 bits per frame. Moreover, the size of stochastic codebook and gain codebook can be small without any perceptual degradation.

5. SUBJECTIVE EVALUATION TEST

We carried out mean opinion score (MOS) test using six listeners to assess the performance of RPVQ and the quality of the variable bit rate ADP-CELP coder. In the subjective evaluation test, we used ten clean speech files as test data. The percentage of mode selection for the test data were 40% (NS mode), 5.2% (UV mode), 32.1% (SV mode) and 22.7% (NV mode) respectively. In this condition, the average bit rate of the variable bit rate ADP-CELP coder was 2.3 kbps.

In the subjective evaluation test, two reference coders were used. One is the same as the variable bit rate ADP-CELP coder except for the adaptive codevector search in SV mode. All candidates of the adaptive codebook (256 candidates) are fully searched in this coder. Therefore, the total number of bits for pitch lags is 32 bits. We call this reference coder REF-RPVQ. Another reference coder is the half-rate Japanese digital cellular standard (HR-PDC) with rate of 3.45 kbps. The subjective evaluation test result is shown in Table 2.

The performance of RPVQ can be seen from the comparison between the variable bit rate ADP-CELP coder and REF-RPVQ. The evaluation result indicates that MOS of proposed coder is equivalent to that of REF-RPVQ. Consequently, RPVQ is able to reduce the bits for pitch lags without any quality degradation. In this case, RPVQ can reduce the total number of bits per frame by 17 bits.

Table 2 also indicates that MOS of proposed coder is higher than that of the HR-PDC. Therefore, we can conclude that the variable bit rate ADP-CELP coder achieves lower bit rate and better quality than HR-PDC.

6. SUMMARY

We proposed a variable bit rate ADP-CELP coder. The following techniques are introduced to improve speech quality and reduce the bit rate.

- A speech/non-speech classification technique using ***spectrum envelope variation***. This indicator is independent of signal power and is sensitive to spectrum variation, so that the proposed technique can classify speech and non-speech correctly even in noisy environment.
- An efficient pitch coding technique named ***relative pitch vector quantization (RPVQ)***. This technique exploits not only the correlation of the pitch lags in a frame but also the correlation over two consecutive frames. The pitch lags can be coded by RPVQ with 15 bits in a frame without any quality degradation.

According to our subjective evaluation test, the variable bit rate ADP-CELP coder achieved lower bit rate and better quality than the half-rate Japanese digital cellular standard.

	Proposed coder	REF-RPVQ	HR-PDC	ORIGINAL
MOS	2.6	2.6	2.4	3.5

Table 2. MOS subjective evaluation test result.

REFERENCES

[1] A. Gersho and E. Paksoy, "An Overview of Variable Rate Speech Coding for Cellular Networks, " *IEEE Proc. ICWC*, pp.8.1-8.4, 1992.

[2] E. Paksoy, K. Srinivasan, and A. Gersho, "Variable rate speech coding with phonetic segmentation, " *IEEE Proc. ICASSP*, pp. II-155-158, 1993.

[3] H. Ohmuro, K. Mano, and T. Moriya, "Variable bit-rate speech coding based on PSI-CELP, " *Proc. ICSLP*, pp.2067-2070, 1994.

[4] A. DeJaco, W. Gardner, P. Jacobs, and C. Lee, "QCELP: The North American CDMA digital cellular variable rate speech coding standard, " *Proc. of the IEEE Workshop on Speech Coding for Telecommunications*, pp5-6, Oct. 1993.

[5] E. Paksoy, A. McCree, and V. Viswanathan, "A variable-rate multimodal speech coder with gain-matched analysis-by-synthesis, " *IEEE Proc. ICASSP*, pp.751-754, 1997.

[6] L. Zhang, T. Wang, and V. Cuperman, "A CELP variable rate speech codec with low average rate, " *IEEE Proc. ICASSP*, pp.735-738, 1997.

[7] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, "The voice activity detector for the pan-European digital cellular mobile telephone service, " *IEEE Proc. ICASSP*, pp.369-372, 1989.

[8] K. Itoh and M. Mizushima, "Environmental noise reduction based on speech/non-speech identification for hearing aids, " *IEEE Proc. ICASSP*, pp.419-422, 1997.

[9] M. Akamine and K. Miseki, "Adaptive density pulse excitation for low bit rate speech coding," *IEICE Trans. FUNDAMENTALS, Vol.E78-A*, pp.199-207, 1995.

[10] J. P. Campbell, Jr., V. C. Welch, and T. E. Tremain, "An expandable error-protected 4800 bps CELP coder (U.S. federal standard 4800 bps voice coder)," *IEEE Proc. ICASSP*, pp.735-738, 1989.