Acoust. Sci. & Tech **26**, 1 (2005)

PAPER

Efficient two-stage vector quantization speech coder using wavelet coefficients of excitation signals

Seiji Hayashi*, Masahiro Suguimoto† and Erinnoviar

Department of Electronics and Computer Systems, Takushoku University, Tatemachi 815–1, Hachioji, Tokyo, 193–0985 Japan

(*Received 24 March 2004, Accepted for publication 16 September 2004*)

Abstract: An improved backward prediction coder featuring two-stage vector quantization (VQ) of shape codevectors is presented. Efficient two-stage VQ is achieved using the wavelet coefficients of excitation signals; i e., wavelet coefficients are calculated by applying a discrete wavelet transform to excitation signals, and the results are divided into an approximation group and a detail group The data lengths of both approximation and detail coefficients are half that of conventional two-stage VQ systems. Simulation results show that the proposed coder achieves a better weighted signal-to-noise ratio (WSNR) than conventional coders and, in terms of reconstructed speech quality, ranks between the FS-1016 Code Excited Linear Prediction (CELP) coder and the Vector Sum Excited Linear Predictive Coding (VSELP) coder

Keywords: Speech processing, Wavelet, Vector quantization

PACS number: 43.72 Kb, 43 72.Gy [DOI: 10 1250/ast 26.43]

1. INTRODUCTION

We proposed low-bit-rate backward-prediction coders (LBBPCs) in previous papers, introducing optimal algorithms for determining excitation gains and codebooks of excitation signals [1–3]. The proposed 3.6 kbit/s LBBPCs attained low delay of 13.5 ms but could not eliminate non-negligible noise from reconstructed speech. The coders of these studies also required a high-complexity codebook search due to a vector quantization (VQ) codebook size of 4,096.

The technique proposed newly in this paper falls basically into the category of two-stage vector quantization of excitation signals, a method used in CELP speech coders Conventional two-stage VQ, on the other hand, can be categorized primarily as (a) codebook generation using two-stage VQ of shape codevectors of excitation signals [4], and (b) codebook generation using two-stage VQ with an adaptive codebook and a shape codebook [5,6]. However, in the proposed LBBPC configuration, the effects of the adaptive codebook were not sufficiently apparent. We considered that an ideal excitation signal (for the backward prediction coder of the LBBPC) is the sum of (1) a forward LPC residual and (2) a signal that compensates and controls the backward coefficients in the backward prediction adaptor. The excitation signal therefore becomes very complicated. We did not adopt an adaptive codebook for our conventional LBBPC coder resulting in poor correlation of excitation signals between adjacent frames. In the conventional systems, two codebooks are used to attain better approximation. However, the frequency bands of the two codebooks are identical, limiting effectiveness to some extent We introduced similar techniques in our LBBPC study and investigated the resulting effectiveness. In our system, however, two codebooks for two different frequency bands were obtained by decomposing excitation signals into lower and higher frequencies. To decompose signals into subbands, we tried a method employing filter banks as well as a method using wavelet transforms. We finally adopted the latter method, wavelet transforms, for our research, for verifying improved characteristics of the codec and a possible benefit from having a diversity of wavelet transforms to choose from

In this paper, we propose an improved LBBPC by the application of wavelet techniques [7]. Conventional methods that deal with signals in subbands utilize wavelet transforms of the original speech signals. [8]. Our method, on the other hand, applies wavelet transformation to LPC residual signals. The characteristics of these wavelet transforms therefore differ greatly from those of the

*e-mail shayashi@es takushoku-u ac jp
†e-mail msuguio@es takushoku-u ac jp

original speech signals. Two-stage VQ is carried out using the wavelet coefficients of the excitation signals, i.e, a discrete wavelet transform (DWT) is applied to the excitation signals to calculate the wavelet coefficients. These coefficients are then divided into an approximation group and a detail group. The data length of coefficients in each of these groups is half that of conventional two-stage VQ systems, enabling us to decrease the codebook sizes. Furthermore, we can realize a much simpler codebook search by calculating the approximation codevectors first and then determining the detail codevectors.

In the following sections, we first briefly describe the conventional LBBPC. We then investigate the codebook search algorithm using two-stage VQ of approximation and detail coefficients and describe the design algorithm for generating the two codebooks. Finally, we evaluate the performance of the improved LBBPC and compare it with conventional LBBPCs that use conventional one-stage and two-stage VQ We also compare performance quality using other codecs for reference.

2. CONVENTIONAL LBBPC

A block diagram of the conventional LBBPC is shown in Fig 1.

This coder consists of a calculation module for excitation gain (expressed in the logarithmic domain using an ADPCM algorithm), a codebook search module, a 50th order synthesis filter and a backward prediction adapter. A brief description of the overall operation of the coder is as follows.

Step 1) Vector $s_z(n)$ is derived by subtracting $z(n)$, the zero-input response vector, from input speech vector $s(n)$. The LPC residual vector $e(n)$ can be obtained by inverse filtering of target vector $s_z(n)$

Step 2) The excitation gain, g, is defined as $\sqrt{\sum_{n=0}^{N-1} e(n)^2}$, and its logarithmic gain, g_{db}, is computed as shown in Eq. (1)

$$g_{db} = 10 \log_{10} \sum_{n=0}^{N-1} e(n)^2 \tag{1}$$

Logarithmic gain g_{db} is then quantized by the ADPCM coder to give gain code I_g (expressed by three bits). The actual values of the coefficients for ADPCM are 0 9, 0.9, 1.25 and 1.75 [9]. Gain code I_g, on the other hand, is manipulated to give $\tilde{g}_{db}$ by the ADPCM decoder. It is worth noting that $\tilde{g}_{db}$ carries excitation errors with it. Then $\tilde{g}_{db}$, the logarithmic gain value with quantization errors, is processed through the inverse logarithmic gain calculator by the following equation.

$$\tilde{g} = 10^{\frac{\tilde{g}_{db}}{20}} \tag{2}$$

Step 3) The optimal combination of σ, $\tilde{g}$ and codebook index k, is found such that Eq. (3) is minimized. In Eq. (3), $v^{(k)}(n)$ denotes the k-th codevector and the σ can be either $+1$ or -1

$$D = \sum_{n=0}^{N-1} \left\{ s_{zw}(n) - \sum_{i=0}^{n} h_w(i) \cdot \sigma \cdot \tilde{g} \cdot v^{(k)}(n-i) \right\}^2 \quad (0 \le k < M) \tag{3}$$

M Codebook size

$s_{zw}(n)$: Perceptual weighted target vector

$h_w(n)$: Perceptual weighted impulse response of synthesis filter

We used a perceptual weighting filter that is a 10th order pole-zero filter defined by transfer function $W(z)$ in Eq. (4) as

$$W(z) = \frac{1 + \sum_{i=1}^{10} \alpha_i \gamma_1^i z^{-i}}{1 + \sum_{i=1}^{10} \alpha_i \gamma_2^i z^{-i}} \tag{4}$$

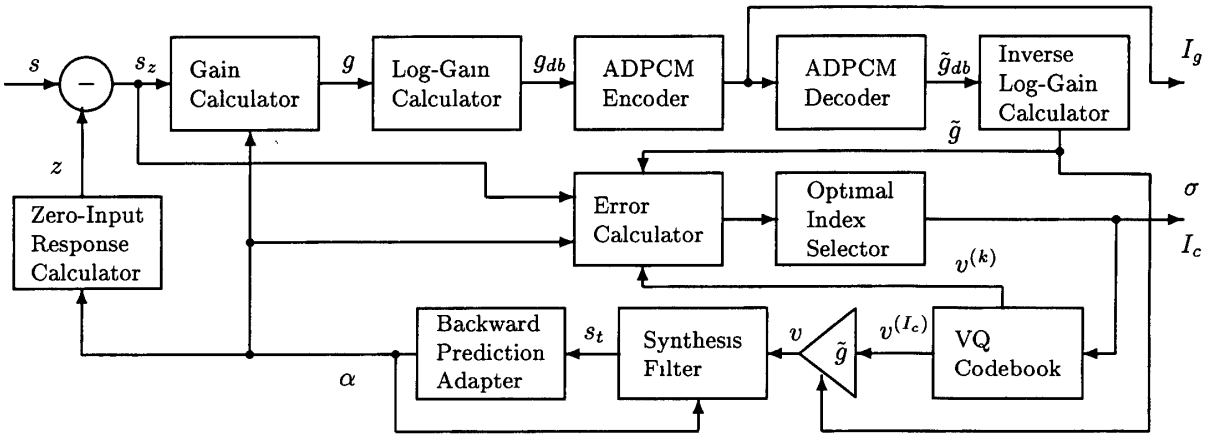

Fig. 1 Block diagram of the conventional LBBPC

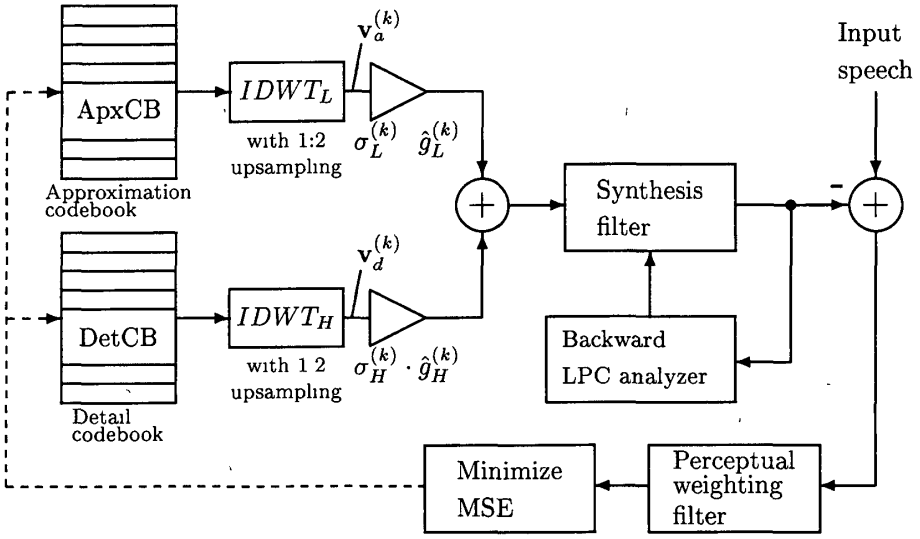

Fig. 2 Block diagram of analysis-by-synthesis procedure for determining optimal indices of wavelet coefficient vectors

where the values of γ_1 and γ_2 are 0 9 and 0.6, respectively.

The perceptual weighting filter is updated by an LPC analysis of the original input speech $s_z(n)$ but not of the previously reconstructed speech $s_t(n)$.

Step 4) The input vector for the synthesis filter is expressed by $v(n) = \sigma \cdot \tilde{g} \cdot v^{(I_c)}(n)$. This vector is processed by the synthesis filter to generate the reconstructed speech $s_t(n)$. The synthesis filter is a 50th order all-pole filter consisting of a feedback loop with a 50th order LPC predictor in the feedback branch.

Step 5) The coefficients of the synthesis filter are updated by the backward prediction adapter. The adapter takes reconstructed speech $s_t(n)$ as input and produces a set of synthesis filter coefficients as output in preparation for the next coding frame. The above steps are repeated for each frame.

In the next section, we apply our proposed two-stage VQ of wavelet coefficients to the error calculator block, optimal index selector block and VQ codebook block (shown in Fig. 1)

3. CODEBOOK SEARCH ALGORITHM

The search algorithm for determining optimal indices of wavelet coefficients for the codebooks is described here. The analysis-by-synthesis procedure used to determine optimal indices for both the approximation codebook (hereafter abbreviated as ApxCB) and the detail codebook (hereafter abbreviated as DetCB) is shown in Fig. 2.

The ApxCB codebook is for the approximation wavelet coefficient vectors, and the DetCB codebook is for the detail wavelet coefficient vectors. In other words, approximation and detail correspond to lower and higher frequencies, respectively The search algorithm is as follows

Step 1) The codebook search of ApxCB is carried out, then the optimal index I_L, gain code G_L and sign code S_L for the low frequency band are calculated so as to minimize Eq. (5)

$$D = \sum_{n=0}^{N-1} \left\{ s_{zw}(n) - \sum_{i=0}^{n} h_w(i) \cdot \sigma_L^{(k)} \cdot \tilde{g}_L^{(k)} \quad v_a^{(k)}(n-i) \right\}^2 \qquad (0 \le k < M_L) \quad (5)$$

M_L Codebook size of ApxCB

$s_{zw}(n)$: Perceptual weighted target vector

$h_w(n)$: Perceptual weighted impulse response of synthesis filter

$v_a^{(k)}(n)$: Reconstructed signal from k-th approximation code-vector of ApxCB
(optimal index I_L $\log_2 M_L$ bits)

$\sigma_L^{(k)}$: Sign of $v_a^{(k)}(n)$ (sign code S_L. 1 bit)

$\tilde{g}_L^{(k)}$: Gain of $v_a^{(k)}(n)$ after quantization/inverse-quantization
(gain code G_L: 3 bits)

The $IDWT_L$ block in Fig 2 is detailed in Fig. 3, where $\uparrow 2$ denotes an upsampling operation, and F'_L and F'_H are reconstruction filters of the approximation and the detail, respectively. The lowpass and highpass reconstruction filters (F'_L and F'_H), together

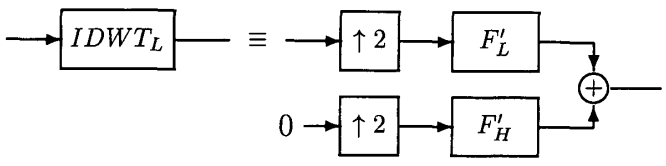

Fig. 3 Equivalent diagram of inverse DWT for reconstructed approximation

with their associated decomposition filters (F_L and F_H), are derived by a wavelet basis function. The process in $IDWT_H$, on the other hand, differs only in that a vector of zeros is fed into the approximation upsampling block of Fig. 3.

However, in the actual implementation, we computed only the upper path in Fig. 3 to decrease the computational load.

Step 2) The second target vector $s'_{zw}(n)$ is derived by subtracting the contribution of the low frequency band from $s_{zw}(n)$. This contribution is the reconstructed speech vector obtained by $v_a^{(I_L)}(n)$, $\tilde{g}^{(I_L)}$ and $\sigma_L^{(I_L)}$. As the sign $\sigma_L^{(k)}$ takes only $+1/-1$, its value is necessary to subtract the lower frequency component by taking into account the sign $\sigma_L^{(S_L)}$ from $s_{zw}(n)$. The codebook search of DetCB is then carried out. The optimal index I_H, gain code G_H and sign code S_H for the high frequency band are calculated so as to minimize Eq. (6).

$$D' = \sum_{n=0}^{N-1}\left\{ s'_{zw}(n) - \sum_{i=0}^{n} h_w(i)\cdot\sigma_H^{(k)}\cdot\tilde{g}_H^{(k)}\cdot v_d^{(k)}(n-i)\right\}^2 \quad (0 \le k < M_H) \qquad (6)$$

M_H: Codebook size of DetCB
$s'_{zw}(n)$ Second target vector for DetCB
$v_d^{(k)}(n)$: Reconstructed signal from k-th detail codevector of DetCB
(optimal index I_H: $\log_2 M_H$ bits)
$\sigma_H^{(k)}$: Sign of $v_d^{(k)}(n)$ (sign code S_H 1 bit)
$\tilde{g}_H^{(k)}$: Gain of $v_d^{(k)}(n)$ after quantization/inverse-quantization
(gain code G_H. 3 bits)

Step 3) The parameters obtained as two triplets of (I_L, G_L, S_L) and (I_H, G_H, S_H) are combined and sent to the receiver. Using the above steps, assuming $M_L = M_H = M$, we can reduce the complexity to $2M$ searches by hierarchically structuring the codebook searches for approximation and detail. On the other hand, we must calculate M^2 searches in the full search, which is impractical for a real-time codec.

4. CODEBOOK DESIGN

The algorithm for generating ApxCB and DetCB is as follows.

Step 1) Training vector $l(n)$ is generated so as to minimize Eq (7).

$$D_{MSE} = \sum_{n=0}^{N-1}\left\{ s_z(n) - \sum_{i=0}^{n} h(i)\cdot\tilde{g}\cdot l(n-i)\right\}^2 \quad (0 \le k < M_H) \qquad (7)$$

$s_z(n)$: Target vector derived by subtracting the zero input response vector from the input speech vector
$\tilde{g}$: Gain after quantization/inverse-quantization
$h(n)$: Impulse response of synthesis filter
$l(n)$: Training excitation vector

Specifically, we first compute an LPC residual vector by inverse filtering of target vector $s_z(n)$ and then calculating a norm of the LPC residual as the value of gain g Next, we encode gain g by ADPCM and obtain quantized gain $\tilde{g}$ by its decoder Consequently, we generate training vector $l(n)$ by dividing the LPC residual vector by $\tilde{g}$. It should be noted that the resulting target vector $l(n)$, in general, does not have unit norm since the estimated gain includes quantization errors.

Step 2) A discrete wavelet transformation of the training vectors is carried out using Haar to the first-level, extracting the approximation and detail coefficient vectors. Note that the coefficient vectors are only half the length of the training vector. These wavelets are the well-known Haar and Daubechies Haar represents the same wavelet as a Daubechies wavelet of the first order, and is the most basic and the simplest [7]. The procedure described in this section uses Haar for the mother wavelet. Later, in section *5.5*, we describe the procedure using Daubechies of the second order. The Haar wavelet basis function is given in Eq. (8).

$$F_L(z) = (1+z^{-1})/\sqrt{2}$$
$$F_H(z) = (1-z^{-1})/\sqrt{2} \qquad (8)$$

$F_L(z)$: Transfer function of low-pass analysis filter
$F_H(z)$ Transfer function of high-pass analysis filter

Step 3) We apply the LBG algorithm [10] to the coefficient vectors and determine approximation centroids for ApxCB and detail centroids for DetCB.

5. SIMULATION

To verify the effectiveness of the proposed two-stage vector quantization of approximation and detail, we carried out a simulation. The simulation parameters for coding and bit allocation are shown in Table 1.

5.1. Codebook Generation

Using Haar as the mother wavelet, we investigated the effectiveness of the proposed method for codebooks of different sizes. Signals comprising twelve minutes of speech by several males and females were used for training after processing with an intermediate reference system (IRS) filter [11]. We also used training excitation vectors, $l(n)$, with a power level in the frame of more than -30 dBov

Table 1 Simulation parameters

Coding conditions		
Sampling frequency		8 kHz
Frame length N		4 5 ms
Analysis window length		22 5 ms
Backward adaptation cycle		4 5 ms
Order of synthesis filter p		50
Bit allocation per frame (bits)		
	Approximation detail	
Best index I_L, I_H	8–11	8–11
Gain code G_L, G_H	3	3
Sign S_L, S_H	1	1

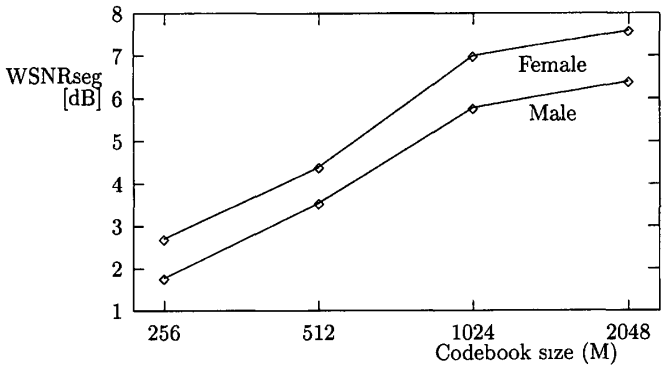

Fig. 4 $WSNR_{seg}$ characteristics for codebooks of different sizes

In step 3 of Section 4, we set a convergence condition threshold for the LBG algorithm of 0.1%. When the problem of an empty cell occurred in the training procedure, we deleted the empty cell, selected another one with more vectors than any of the others and divided it into two new cells. In this way, we maintained the same number of cells as before

We generated approximation and detail coefficient codebooks for $M = 256$, 512, 1,024 and 2,048 where M_L and M_H are set to M The experimental results of a weighted segmental SNR ($WSNR_{seg}$) are shown in Fig. 4.

In codebooks of $M \leq 512$, the required bits are 8 or 9 and the complexity is relatively very low. However, the resulting reconstructed speech quality was very poor and noise was easily recognized. For $M = 2{,}048$, the reconstructed speech quality improved greatly but at the expense of complexity. Considering the trade-off between quality and search complexity, we used a codebook size of $M =$ 1,024 (10 bits) in the following simulation. Examples of the ApxCB and DetCB coefficient vectors after training are shown in Fig 5. In the figure, each amplitude scale is set to be identical. We found that the approximation codebook ApxCB expresses the presence of long-term periodicities in voiced speech while the detail codebook DetCB expresses high frequency components. As we discuss later in the Conclusion, we could attain improved characteristics as shown in Figs. 6 and 7.

5.2. Objective Evaluation

The input speech signals used in the evaluation are non-training speeches passed through an IRS filter for 30 seconds. The $WSNR_{seg}$ results of the LBBPC coders, including the proposed two-stage VQ ($M_L = M_H = 1{,}024$), are shown in Table 2.

The LBBPC results in Table 2 include the conventional two-stage VQ (first and second codebooks are 1024 in size) and the conventional one-stage VQ (4096 in size). In conventional VQ systems, the codevectors of the codebooks are obtained by training the excitation vectors themselves. [4]. In terms of $WSNR_{seg}$, the proposed 6.2 kbit/s LBBPC (with two-stage VQ of approximation and detail coefficients) was better than the conventional LBBPC (3.6 kbit/s) with one-stage VQ by 2.14 dB, and better than the conventional LBBPC with two-stage VQ by

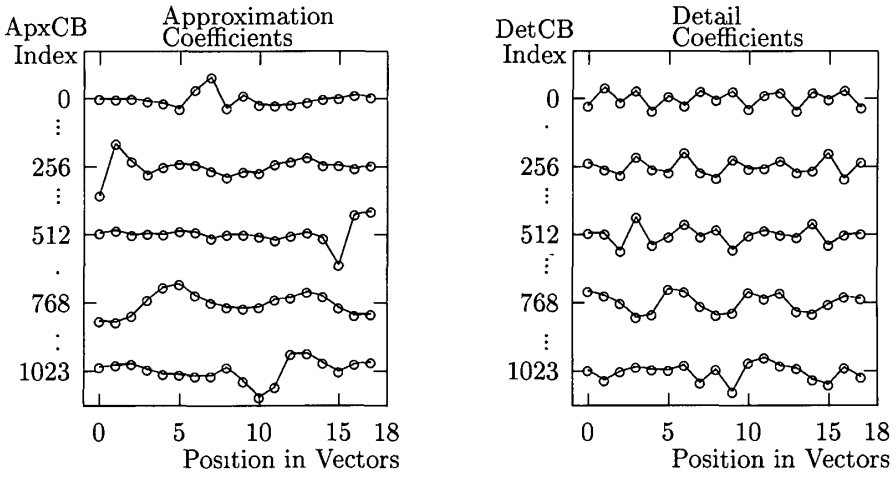

Fig. 5 Examples of coefficient vectors in ApxCB and DetCB after training ($M = 1{,}024$)

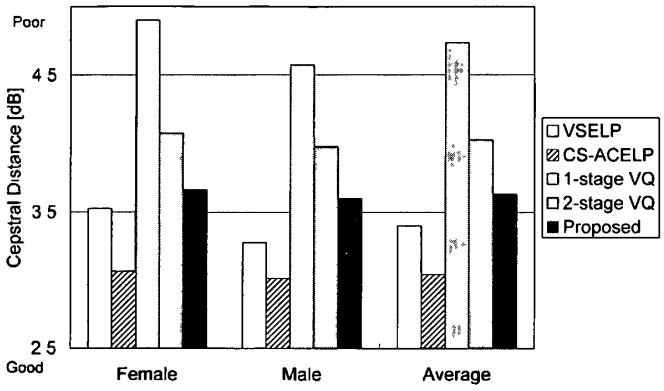

Fig. 6 Cepstrum distance characteristics

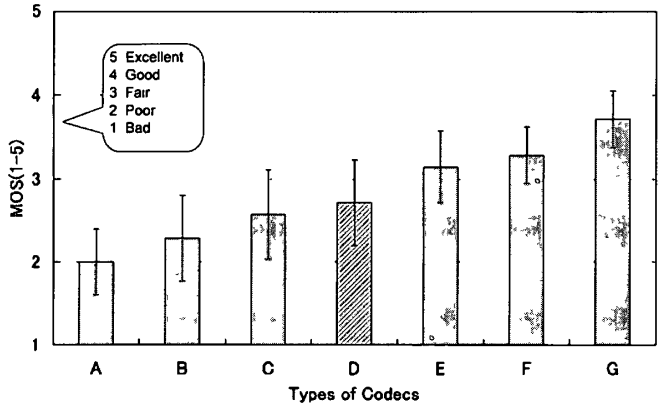

Fig. 7 Results of subjective evaluation (MOS)

0.12 dB. As another evaluation, the cepstrum distance characteristics were determined, and these results are shown in Fig. 6.

The cepstrum distance was calculated using the speech distortion cepstral distance module of the FS-1016 simulation [12] with its order set to 64. VSELP (IS-54) [11] and CS-ACELP (G.729), using 16-bit fixed-point software [13] available from ITU-T, were also evaluated. From Fig. 6 we can see that the proposed VQ could attain low spectral distortion in comparison with the conventional two-stage VQ However, it is apparent here that the quality of the proposed VQ is lower than that of VSELP.

5.3. Subjective Evaluation

To verify the subjective effectiveness of the proposed method, we performed an informal listening test based on a mean opinion score (MOS) evaluation. We compared the codec quality of the LBBPC using the proposed two-stage VQ with that of reference codecs, as shown in Table 3. As for the evaluation conditions, eight types of clean speech signals were used (4 male, 4 female, 10 seconds each) while evaluators listened with single-sided earphones The scores of 21 listeners (including some of the speech researchers) were averaged. Figure 7 shows the results of the subjective evaluation with the bars representing a 95% confidence level of error.

From Fig. 7 we can see that the proposed method (D in the figure) shows an improvement in noise reduction compared with conventional two-stage VQ (C). We introduced a codebook search that first calculates the optimal codevector of the low frequency band for the approximation codebook and then determines the optimal codevector of the high frequency band for the detail codebook. Using this approach is probably the principal reason for the improvement described above. On the other hand, while we could not directly compare the proposed system with other codecs and or with the modulated noise reference unit (MNRU) speech samples, the proposed 6.2 kbit/s coder can be said to have a reconstructed speech quality that ranks between the FS-1016 4.8 kbit/s CELP and the 8 kbit/s VSELP.

Table 2 $WSNR_{seg}$ results for LBBPC coders

	Female [dB]	Male [dB]	Bit-rate (kbits/s)
Proposed method	7 67	6 43	6 2
Conventional two-stage VQ	7 52	6 34	6 2
Conventional one-stage VQ	5 38	4 44	3 6

Table 3 Types of Codecs in Fig 7

Symbol	Coder/Reference speech	Bitrate (kbit/s)
A	MNRU Q = 15 dB (ITU-T G 191)	—
B	FS-1016 CELP	4 8
C	Conventional Two-Stage VQ	6 2
D	Proposed Method	6 2
E	MNRU Q = 20 dB	—
F	VSELP (IS54)	8
G	CS-ACELP (G 729)	8

5.4. Relationship between Codebook Specification and LBBPC Complexity

Table 4 summarizes codebook sizes, vector code lengths, required memory (to store codebooks) and complexity for three types of LBBPCs using the proposed two-stage VQ, conventional two-stage VQ and conventional one-stage VQ.

From Table 4 we can see that the proposed two-stage VQ requires codebooks half the size of, but with complexity no greater than, conventional two-stage VQ. Consequently, in comparison with conventional LBBPCs, the proposed 6.2 kbit/s coder using two-stage VQ of approximation and detail vector coefficients could attain equal or better WSNR (Table 2), better ceptstrum distance characteristics (Fig. 6) and a better MOS evaluation (Fig 7).

Table 4 Relationship between codebook specification and LBBPC complexity

	Codebook			Complexity
	Size	Code length	Required memory	
Proposed method	$2M$	$N/2$	36 kwords	0 56
Conventional two-stage VQ	$2M$	N	72 kwords	0 50
Conventional one-stage VQ	$4M$	N	144 kwords	1 00

5.5. Verification of the Propoed System with Daubechies Wavelet

We also evaluated the performance of the proposed system using Daubechies of second order as the mother wavelet However, we found that the average $WSNR_{seg}$ as well as speech quality showed only slight differences in performance compared to that of the Haar mother wavelet. The reason for this is that the influence of wavelet coefficient quantization errors was greater than the effect of differences in wavelet bases between Haar and Daubechies. This then leads to the conclusion that our two-stage VQ system is effective for both of the widely used wavelets, Haar and Daubechies.

6. CONCLUSION

We proposed an approach introducing two-stage VQ of wavelet coefficients to our conventional LBBPC speech coder to improve reconstructed speech quality. The two-stage VQ is carried out using the wavelet coefficients of excitation signals By using these wavelets, the data length of approximation and detail coefficients are each half that of conventional two-stage VQ systems enabling a decrease in the size of codebooks. Comparing our proposed coder to conventional coders that use two-stage VQ of excitation vectors, the proposed coder has smaller-sized codebooks and attains an equal or better *WSNR*. We also verified that the proposed coder attains a better mean opinion score in an evaluation comparing it to conventional coders However, future experiments with different wavelets may be mandatory to exploit these wavelets in terms of coder performance.

REFERENCES

[1] S Hayashi, M Sugimoto and T Toida, "On a low bit-rate backward prediction coder with low delay," *Proc IASTED Int Conf Modelling and Simulation-MS '94*, pp 353–356 (1994)

[2] S Hayashi and M Sugimoto, "On an improvement scheme for LBBPC speech coder," *Proc IASTED Int Conf Signal and Image Processing—SIP95—*, pp 55–59 (1995)

[3] S Hayashi, M Sugimoto and Erinnoviar, "Low bit-rate CELP speech coder with low delay," *Signal Process*, **72**, pp 97–105 (1999)

[4] B H Juang and A H Gray, "Multiple stage vector quantization for speech coding," *Proc ICASSP 82*, pp 597–600 (1982)

[5] B Kleijn, D J Krasinski and R H Kechum, "An improved speech quality and efficient vector quantization in SELP," *Proc ICASSP 88*, pp 155–158 (1988)

[6] Jian Zhang and Tian-Hu Yu, "A 4 2 kb/s low-delay speech coder with modified CELP," *Proc IEEE Signal Process Lett*, **4**, 301–303, (1997)

[7] G Strang and T Nguyen, *Wavelets and Filter Banks* (Wellesley-Cambridge Press, 1996)

[8] H Perez and F Amano, "Acoustic echo cancellation using multirate techniques," *IEICE Trans*, **E74**, 3559–3568 (1991)

[9] P Cumminskey, N S Jayant and J L Flanagan, "Adaptive quantization in differential PCM coding of speech," *Bell Syst Tech J*, **52**, 1105 (1973)

[10] Y Lindo, A Buzo and R M Gray, "An algorithm for vector quantizer design," *IEEE Trans Commun*, **COM-28**, 84–95 (1980)

[11] ITU-T Rec G 191, *Software Tools Library (STL96)* (ITU-T, 1996)

[12] C simulation source codes *celp_3 2a tar* for FS-1016 4800 bps CELP voice coder version 3 2a available by anonymous ftp

[13] ITU-T Rec G 729, *8 kbit/s CS-ACELP speech coder* (ITU-T, 1996)