A BIT-RATE/BANDWIDTH SCALABLE SPEECH CODER BASED ON ITU-T G.723.1 STANDARD

Sung-Kyo Jung , Kyung-Tae Kim, and Hong-Goo Kang

Center for Information Technology (CITY)
Dept. of Electrical and Electronic Eng., Yonsei University

E-mail: {*skjung, kktae, hgkang*}*@mcsp.yonsei.ac.kr*

ABSTRACT

This paper presents a new scalable coder based on the ITU-T G.723.1 standard which is one of the most famous speech coders for VoIP applications. In order to support both bit-rate scalability and bandwidth scalability, the proposed coder adopts a split-band approach, where the input signal, sampled at 16 kHz, is decomposed into two equal frequency bands. The lower-band speech is coded with the standard coder such as the G.723.1 standard. In addition, the low-band enhancement layer for the lower-band speech improves the perceptual quality of decoded speech by employing additional coding units based on a cascaded codebook approach. The higher-band signal is encoded using an MDCT-based transform coding scheme. The proposed coder at a bit-rate of 19.4 kbit/s provides speech quality comparable to the ITU-T 24 kbit/s G.722.1 coder, while it also has interoperability with G.723.1.

I. INTRODUCTION

The recent speech coding standards provide high speech quality that is sufficient for deploying them to the current market. A future speech coding standard will be requested to provide more flexibility and interoperability with current standards. In other words, a new standard should be able to support audio bandwidth quality if channel condition allows, and it also minimizes the need for transcoding. Recently, ITU identified the application area and terms of reference (ToR) of new variable bit-rate (VBR) codecs to be developed: multi-rate source controlled VBR (MSC-VBR) and embedded VBR (EV) approaches [1]. In [2]-[4], several scalable coding schemes were proposed to support both bandwidth scalability and interoperability with the conventional standard coders such as G.729 [5] and its bit-rate extension [6]. Another candidate of a core layer for scalable coder is G.723.1 [7] because it is the most widely used coder for VoIP applications. However, it is difficult to directly use G.723.1 for this purpose due to its insufficient quality. The structure proposed in [4] is good to improve the quality in lower-band area. However, its higher-band quality might not be good because the additional bits might be generally allocated to the lower-band area where the relative energy is higher.

This paper focuses on voice coder that not only has interoperability with the traditional codec, but also supports bit-rate scalability and bandwidth scalability. A two-band split approach is introduced to have bandwidth scalability. To maintain backward compatibility with conventional standard coders, the narrow-band speech signal is coded with already standardized coder, 5.3 kbit/s G.723.1 standard. The difference of the proposed algorithm from the previous approaches [2]-[4] is the modular algorithm that separately improves the perceptual quality of the low and high band. Specifically, a low-band enhancement layer is added into the narrowband coders in a cascade form to maintain bit compatibility with the standard coders. The high-band input signal through high-pass filtering is encoded by a transform-based coding scheme to obtain high-quality sound and to provide flexibility in terms of bandwidth and bit-rate. Once every frame, the high-band speech is transformed by a modified discrete cosine transform (MDCT) [8]. The MDCT coefficients are split into subbands, and then the coefficients in each band are encoded by a gain-shape vector quantization. The proposed coder is implemented at a bit-rate of 19.4 kbit/s. The proposed coder provides comparable quality to the reference coder, 24 kbit/s G.722.1 standard [9].

II. BIT-RATE/BANDWIDTH SCALABLE WIDEBAND CODER

This section describes a proposed bandwidth-scalable coder for wideband speech signal that has a capability of interoperability with narrowband standard coders. The proposed coder splits full bandwidth by low band (0-4 kHz) and high band (4-8 kHz). A block diagram of the proposed scalable coder is shown in Fig. 1.

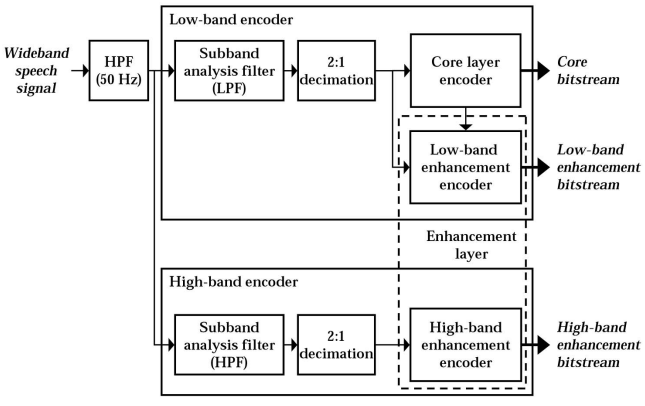


Fig. 1. Block diagram of the encoder for the proposed coder.

The distinguished characteristic of the proposed algorithm from the previous approaches [2]-[4] is using two different enhancement layers which try to improve the qualities of both the low-band speech and the high-band signal. Specifically, a low-band enhancement unit is simultaneously operated with the narrowband standard coders in a cascade manner while maintaining bitstream compatibility. This enhancement unit improves the synthesized quality by increasing the accuracy of the excitation model, such that additional bits are allocated to reduce the quantization error. The high-band enhancement layer is designed to keep the same frame structure as a core codec. The high-band input signal through high-pass filtering is encoded by a transform-based coding scheme that is flexible for generating high quality. For the transformation, an MDCT is used to reduce block edge effects [10] which are discontinuities in the reconstructed signal due to the independent processing of each block in traditional block transform. The gains and shapes of MDCT coefficients are individually quantized using scalar quantization (SQ) and vector quantization (VQ) codebooks, respectively.

2.1. Low-band enhancement layer

The purpose of this layer is to enhance the perceptual quality of the 5.3 kbit/s G.723.1 coder. Particularly, since the lower band components are more perceptually important than the higher band ones, we improve the quality of the core layer by inserting an enhancement unit to the conventional codec. The additional bits in the low-band enhancement unit are allocated to the quantization of linear predictive parameters and approximation of the excitation signal with a cascaded structure manner.

Since the LPC filter for the last subframe is quantized using 24 bits per frame and the LPC filters for other subframes are obtained by linear interpolation, the performance of the LSP quantizer is degraded in terms of spectral distortion (SD), as shown in Table 1. Additional 24 bits for the LP filter to the 2nd subframe reduce the number of outlier subframes as well as the average SD. Since we use the same quantization scheme as standard coder, no additional quantization table is required. As shown in Fig. 2, the cascaded codebook approach [11] in the fixed codebook search provides bit-rate scalability as well as improved quality. In other words, the non-periodic component of the excitation is approximated using a three-stage algebraic codebook structure. The first codebook is used in a core layer and the last two codebooks in a low-band enhancement layer. For the efficient implementation, the structure of the two additional algebraic codebook is the same as G.723.1 coder. The gain term of the additional codebook is encoded with a 3-bit scalar quantizer. The contribution of two additional fixed codebooks are quantized with 160 bits per subframe. Table 2 shows the bit allocation of the low-band enhancement layer. In order to evaluate the performance of the low-band enhancement layer, we use ITU-T P.862 perceptual evaluation of speech quality (PESQ) [12]. Table 3 compares the PESQ results of the proposed low-band coder and the reference standards. The proposed coder shows much higher quality than G.723.1, but it provides a 0.12 point lower PESQ score in comparison with G.729 annex E.

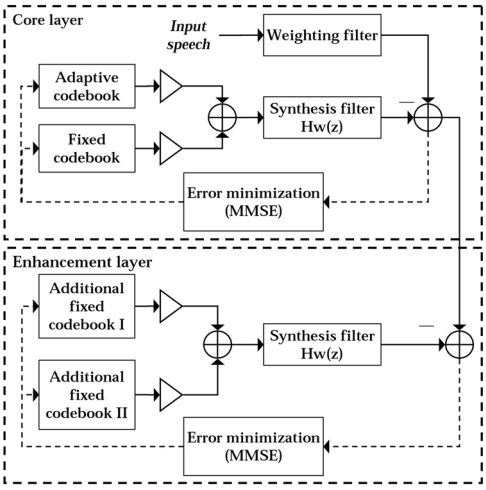


Fig. 2. Block diagram of low-band enhancement layer.

Table 1. SD results of LSP quantization schemes for the G.723.1 coder.

Bits per frame (%)	SD (dB)	Outliers (%)	
		2-4 dB	>4 dB
24 bits (last subframe)	2.23	29.38	12.01
24 bits (all subframes)	2.96	44.44	21.87
48 bits (all subframes)	2.36	40.20	11.26

Table 2. Bit allocation for the coded parameters in the low-band enhancement layer.

Coded parameters		Number of bits per frame
LPC indices		24
2nd FCB	Shape	4×17=68
	Gain	4×3=12
3rd FCB	Shape	4×17=68
	Gain	4×3=12
Total		184

Table 3. PESQ results of the proposed low-band coder and the reference coders.

Coders	PESQ		
	Female	Male	Average
5.3 kbit/s G.723.1	3.32	3.58	3.45
6.3 kbit/s G.723.1	3.46	3.72	3.59
11.4 kbit/s proposed	3.76	4.02	3.89
11.8 kbit/s G.729E	3.89	4.13	4.01

2.2. High-band enhancement layer

A block diagram of the high-band enhancement layer is shown in Fig. 3. The role of each block will be briefly described in this subsection. In the high-band enhancement layer for the proposed coder, a simplified transform-based coding scheme is employed since the high-band signal is perceptually less important

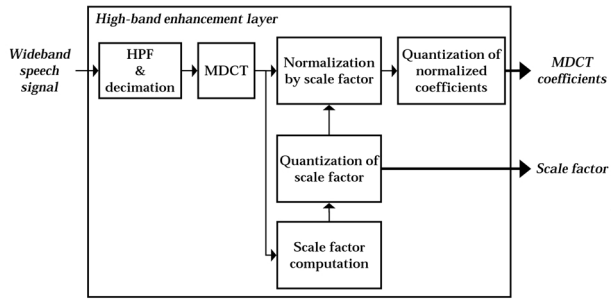


Fig. 3. Block diagram of high-band enhancement layer.

than the low-band one. The wideband speech, sampled at 16 kHz, is high-pass filtered with a bandwidth of 4-8 kHz and is decimated to 8 kHz sampling rate. An MDCT is used to transform the time-domain input signal into spectral-domain components. To transmit the transform coefficients, we decompose subbands of the transform coefficients into gains (scale factors) and shapes (normalized coefficients).

For the MDCT, windowing is used to select the portion of the input signal to analyze. The length of window is a compromise between long windows (high coding gain) and short windows (better approximation for non-stationary speech). In general, the analysis/synthesis window for MDCT is 50 % overlapped. Since every input sample is transformed twice using the 50 % window in the analysis filter, there is no reduction in the transform coding performance. To cope with this problem, special attention was paid to the design of the overlapping window shape. In [13], a possible window function was given for an adaptive block size MDCT. We employ this window function for MDCT analysis and synthesis so that the windowing procedure in MDCT analysis matches the frame structure of the core layer. In other words, no additional delay, excepting a lookahead delay of core coder, is required for an overlap operation between two consecutive blocks. Due to the long frame size of the core coder, three small analysis windows are applied on every frame in order to increase the time-domain resolution. The window size is set to 20 ms (N=160) and the overlap length is fixed to 7.5 ms.

In our coder, we decompose subbands of the transform coefficients into gains and shapes. For the transform coding of the signal, we firstly apply an N-point analysis window on the transform target signal, and then calculate the N-point MDCT. Due to the antisymmetry of the MDCT coefficients, only the first $N/2$ coefficients need to be encoded. We need not transmit the first $N/8$ MDCT coefficients corresponding to the frequency region ranging from 7000 Hz to 8000 Hz because this frequency region is outside the bandwidth of interest in wideband speech applications. Note that the term obtained by the decimation is a right-shifted version of the original component by an amount π on the unit circle [14]. The remaining $3N/8$ coefficients are partitioned into R frequency bands with a same dimension of 10 to calculate a scale factor in each region. The transform coefficients are divided into six subvectors in case of selecting G.723.1 coder as a core coder (R=6). The transform coefficients in each region are normalized by the corresponding square root energy which must be transmitted to the receiver as side information. The normalized MDCT coefficients which are divided into $2R$ frequency groups with a same dimension of 5. While the scale factors are transformed and then scalar quantized, the shape vectors are directly vector quantized. We compute the R DCT coefficients of the R scale factors. The DCT coefficients are individually predicted using intra-frame correlation and then each residual value is quantized with a 4- or 3-bit scalar quantizer. The 12 groups of the normalized coefficients are vector quantized using different bit allocation patterns as described in Table 4.

Table 4. Bit allocation for the coded parameters in the high-band enhancement layer.

Coded parameters	No. of bits per analysis	Total bits per frame
Gains	4+4+3+3+3+3=20	3×20=60
Shape vectors	7+7+6+6+5+5+5+5 +4+4+3+3=60	3×60 =180
Total	80	240

III. IMPLEMENTATION AND PERFORMANCE EVALUATION OF THE PROPOSED CODER

3.1. Bit allocation

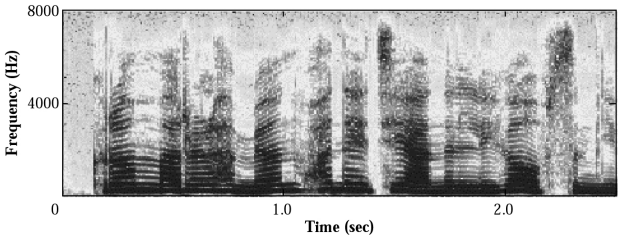
The allocated bits consist of three parts; core bitstream, low-band enhancement bitstream, and high-band enhancement bitstream. As presented in the previous section, the low-band enhancement bitstream describes detailed characteristics of the LP filter coefficients and the excitation components in lower frequency band. High-band components are transformed in the frequency domain and then the MDCT coefficients in each subband are quantized by a gain-shape VQ approach. The proposed coder employs 5.3 kbit/s G.723.1 as a core layer speech coder. The additional bit-rate for the enhancement layer is set to about 14.1 kbit/s. So, the total bitrates needed to encode wideband speech are 19.4 kbit/s. The bit allocation of the proposed scalable coder is shown in Table 5. The proposed coder has the same frame structures as the G.723.1 coder (i.e., 7.5-ms look-ahead and 30-ms frame with 7.5-ms subframe). This results in the overall algorithmic delay of 37.5 ms. Note that the enhancement layer uses a 7.5-ms look-ahead for overlap of MDCT window as previously described.

3.2. Performance evaluation

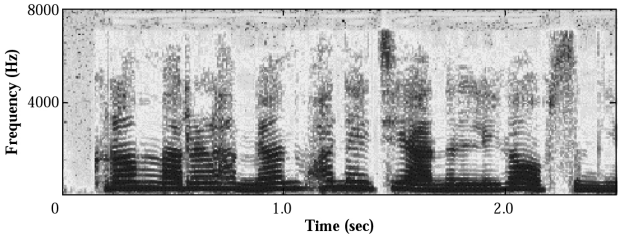
Fig 4 shows the spectrograms of the original female speech and the decoded signal. The proposed coder provides good representation of the energy distribution, but does not well describe the fine structure, especially in the high-frequency region. The main

Table 5. Bit allocation for the proposed scalable coder.

Scalable bitstream		Bit allocation	
		Number of bits	Bit-rate
Core bitstream		158 bits/frame	5.3 kbit/s
Enhancement bitstream	Low-band	184 bits/frame	6.1 kbit/s
	High-band	240 bits/frame	8.0 kbit/s
Total		582 bits/frame	19.4 kbit/s



(a)



(b)

Fig. 4. Spectrograms of original speech and the decoded signal by the proposed coder. (a) Original speech. (b) Output speech of the proposed coder at a bit-rate of 19.4 kbit/s.

Table 6. Preference test results between the proposed coder at 19.4 kbit/s and the reference coder.

Preference (%)	Female	Male	Average
Proposed coder	35.000	25.000	**30.000**
Reference coder	25.000	17.500	**21.250**
No preference	40.000	57.500	**48.750**

reason can be explained by the mean-square-criterion in the shape quantization.

To compare perceptual quality difference, we performed an informal blind preference test for a subjective evaluation involving 10 listeners. The test material is chosen from NTT multilingual database including 8 Korean sentences pronounced by four male and four female speakers each. Table 6 shows the results of the subjective test between the proposed coder and the reference coders. Results imply that the proposed coder at a bit-rate of 19.4 kbit/s provides comparable speech quality in comparison with the 24.0 kbit/s G.722.1 coder.

IV. CONCLUSION

The primary objective of this paper has been to propose a wideband scalable coder, which has backward compatibility with the conventional coders such as G.723.1 standard. To accomplish the goal, the proposed scalable coder supports both bit-rate and bandwidth scalabilities, which result in reducing signal distortion for narrowband or achieving better speech quality with high-frequency components. The proposed scalable coder, which is implemented at a bit-rate of about 19.4 kbit/s, provides synthetic speech quality comparable to the reference coder, while maintaining a bit-rate scalability.

V. ACKNOWLEDGEMENTS

This work was supported in part by the Electronics and Telecommunications Research Institute (ETRI) and the Biometrics Engineering Research Center (KOSEF).

VI. REFERENCES

[1] ITU-T SG 16 Q.9, "Variable bit rate coding of speech signals," Oct. 2002.

[2] A. McCree, "A 14 kb/s wideband speech coder with a parameteric highband model," in *Proc. Int. Conf. Acoust. Speech. Sign. Process.*, pp. 1153-1156, 2000.

[3] K.-T. Kim, S.-K. Jung, Y.-C. Park, and D. H. Youn, "A new bandwidth scalable wideband speech/audio coder," in *Proc. Int. Conf. Acoust. Speech. Sign. Process.*, pp. 657-660, 2002.

[4] K. Koishida, V. Cuperman and A. Gersho, "A 16 kbit/s bandwidth scalable audio coder based on the G.729 standard," in *Proc. Int. Conf. Acoust. Speech. Sign. Process.*, pp. 1149-1152, 2000.

[5] ITU-T Rec. G.729, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)," Mar. 1996.

[6] ITU-T Rec. G.729 Annex E, "11.8 kbit/s CS-ACELP speech coding algorithm," Sep. 1998.

[7] ITU-T Rec. G.723.1, "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s," Mar. 1996.

[8] J. P. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time-domain aliasing cancellation," in *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 1153-1161, Oct. 1986.

[9] ITU-T Rec. G.722.1, "Coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss," 1999.

[10] H. S. Malvar, *Signal processing with lapped transforms*, Artech House, 1992.

[11] S.-K. Jung, K.-T. Kim, H.-G. Kang, and D. H. Youn, "A cascaded algebraic codebook structure to improve the performance of the speech coders," in *Proc. Int. Conf. Acoust. Speech. Sign. Process.*, pp. II.173-II.176, Apr. 2003.

[12] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech coders," Feb. 2001.

[13] M. Iwadare, A. Sugiyama, F. Hazu, A. Hirano, and T. Nishitani, "A 128 kb/s hi-fi audio CODEC based on adaptive transform coding with adaptive block size MDCT," in *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 1, pp. 138-144, Jan. 1992.

[14] P. P. Vaidyanathan, *Multirate systems and filter banks*, Prentice Hall, 1993.