A 16-KBIT/S BANDWIDTH SCALABLE AUDIO CODER BASED ON THE G.729 STANDARD

Kazuhito Koishida, Vladimir Cuperman and Allen Gersho

Signal Compression Laboratory, Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106-9560, USA

ABSTRACT

This paper proposes a bandwidth-scalable coding scheme based on the G.729 standard as a base layer coder. In the scheme, according to the channel conditions, the output speech of the decoder can be selected to be narrow-band (4-kHz bandwidth) or wideband (8-kHz bandwidth). The proposed scheme consists of two layers: base and enhancement. The base coder uses the G.729 algorithm to encode narrow-band speech. The enhancement coder is based on a fullband CELP model and it encodes wideband speech while making use of the available base layer information. Two bandwidth-scalable coders are designed: one is scalable with the 8 kb/s G.729 base coder and another with the 6.4 kb/s G.729 (Annex D) base coder. Subjective tests show that, for wideband speech, the proposed coders at 16 kb/s achieve better performance than the 16 kb/s MPEG-4 CELP with bandwidth scalability.

1. INTRODUCTION

Packetized speech communication has become increasingly important for Asynchronous Transfer Mode, Frame Relay, and Internet Protocol applications. In these networks, packet losses occur due to network congestion, maximum delay constraints, and buffer overflow. One way to avoid large degradations in speech quality with packet networks is to use scalable coding algorithms [1]–[6]. In these algorithms, the encoder generates the bit-stream in a layered manner so that the decoder can recover the reconstructed speech from a subset of the entire bit-stream. The bit-stream obtained from scalable coders consists of a base layer and one or more enhancement layers. The base layer, which is the smallest subset of the bit-stream, is generated by the base encoder and provides a minimal quality. The enhancement layers are added to the bit-stream by the enhancement encoders in such a way that the combination of the base layer and enhancement layers allows a higher quality signal to be recovered by the decoder. Two types of quality improvement are achievable by the enhancement layers: one is obtained without changing the bandwidth of the output speech [1][2], while another is achieved by increasing the bandwidth [3]–[6]. Scalability with the latter type of improvement is called bandwidth scalability. One of the most straightforward ways to implement the bandwidth scalability is to employ a subband structure [3]–[5]. However, in subband coders, an audible distortion may sometimes appear due to the subband analysis and synthesis procedures, especially at low bit rates. On the other hand, the MPEG-4 CELP (code-excited linear prediction) coder [6][7] realizes the bandwidth scalability with a fullband structure, and achieves good quality at low bit rates.

This work was supported in part by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists. This work was also supported in part by the National Science Foundation under grant MIP-9707764, the University of California MICRO program, Cisco Systems, Inc., Conexant Systems, Inc., Dialogic Corp., Fujitsu Laboratories of America, Inc., General Electric Co., Hughes Network Systems, Lernout & Hauspie Speech Products, Lockheed Martin, Lucent Technologies, Inc., Qualcomm, Inc., and Texas Instruments, Inc.

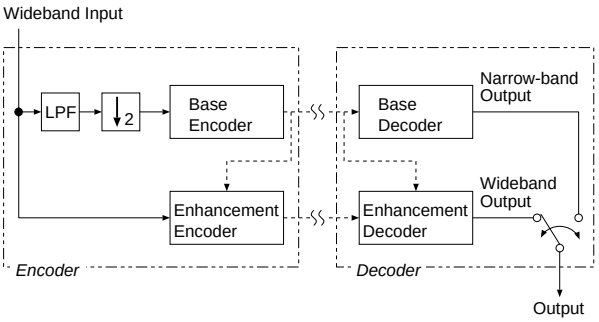

Figure 1: Overview of proposed bandwidth scalable coder.

In this paper, we propose a bandwidth-scalable coding scheme based on the G.729 standard [8][9] as a base layer coder. The proposed scalable coder consists of two layers: base and enhancement. In the base layer, the input wideband speech (sampled at 16 kHz) is down-sampled to the narrow-band speech (sampled at 8 kHz), and a base coder encodes the narrow-band speech using the G.729 algorithm. An enhancement coder in the enhancement layer is based on a fullband-type CELP model, and it encodes the wideband speech while making use of the available base layer information. In the decoder, either the narrow-band or wideband speech can be selected according to the channel conditions.

This paper is outlined as follows. Section 2 gives an overview of the proposed scalable scheme. Section 3 discusses the structure for the enhancement coder, and Section 4 describes the spectral quantization of the enhancement coder. In Section 5, the proposed bandwidth-scalable coders at 16 kb/s are evaluated through listening tests. Finally, conclusions are given in Section 6.

2. OVERVIEW OF PROPOSED SCALABLE SCHEME

Fig. 1 shows an overview of the proposed scalable coder. The proposed scheme consists of a base and an enhancement layer. In the base layer, the input wideband speech is down-sampled, low-pass filtered and fed into a base coder. The base coder encodes the narrow-band speech and generates the base layer bit-stream. In the enhancement layer, an enhancement coder directly encodes the input wideband speech and generates the enhancement layer bit-stream. The output speech of the decoder can be selected from

either narrow-band or wideband speech according to the channel conditions. When the whole bit-stream is received, the enhancement decoder reconstructs the wideband speech. If only the base layer bit-stream is available, the base decoder generates the narrow-band speech.

Two coders are employed as a base coder: G.729 at 8 kb/s [8] and G.729-D (Annex D) at 6.4 kb/s [9]. The enhancement coder is based on a CELP model and has the same frame structure as the G.729 coders, i.e, 10-ms frame with a 5-ms subframe. The enhancement coder uses a 10 ms look-ahead for 16th-order linear prediction (LP) analysis. This results in an overall algorithmic delay of 20 ms.

The wideband speech signals include almost the same information as the narrow-band speech signals in the frequency range of 0-4 kHz and, in the proposed scheme, both speech signals are encoded by the CELP model. As a result, there exists redundancy in some coding parameters, such as the line spectral pairs (LSPs) and the pitch. Therefore, the reduction of such redundancy leads to improved coding efficiency of the enhancement coder. In the following sections, we investigate the efficient use of the coding parameters of the base coder for the enhancement coder.

3. INVESTIGATION OF ENHANCEMENT CODER STRUCTURE

In this section we consider different options for the structure of the enhancement coder. Four types of structure are presented, and their coding performance is evaluated.

3.1. Enhancement Coder Structure A

We begin our investigation by considering a technique similar to that used in the MPEG-4 CELP coder, as shown in Fig. 2. In this system, the excitation signal is generated from three sources: the ACB (adaptive codebook) codevector, the FCB (fixed codebook) codevector, and the up-sampled version of the pulse codevector generated in the G.729 base coder. The up-sampled pulse codevector is obtained using the same procedure as in MPEG-4 CELP:

$$y_{up}(n) = \begin{cases} x(n/2), & \text{if } n \text{ is integer-multiple of 2} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

We note that the above procedure creates a frequency image in the 4-8 kHz band.

The excitation codebook search is done in the following way. First, the ACB is searched and its contribution is subtracted from the target vector. The contribution of the up-sampled pulse excitation is also removed from the target vector. Finally the FCB codevector is selected.

3.2. Enhancement Coder Structure B

The second structure is illustrated in Fig. 3. This system utilizes all coding parameters from the G.729 base coder and decodes the G.729 output speech without post-processing. The reconstructed wideband speech is obtained as a sum of the up-sampled G.729 output and the synthesis filter output. To obtain the up-sampled output, a low-pass filter is combined with Eq. (1) to remove the frequency image. After the up-sampled G.729 output is subtracted from the target signal, the ACB and FCB are searched in that order.

In this system, the excitation is generated not only to create the high frequency component but also to achieve quality enhancement at low frequencies. Since the up-sampled G.729 output has a certain pitch periodicity, the enhancement excitation contains relatively little periodicity.

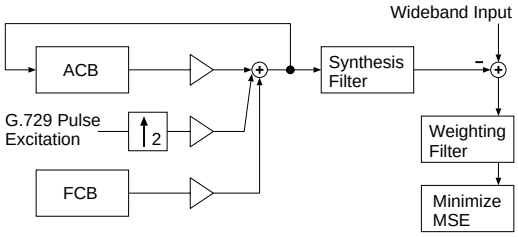

Figure 2: Enhancement coder structure A.

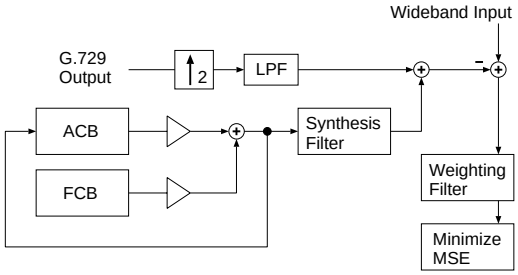

Figure 3: Enhancement coder structure B.

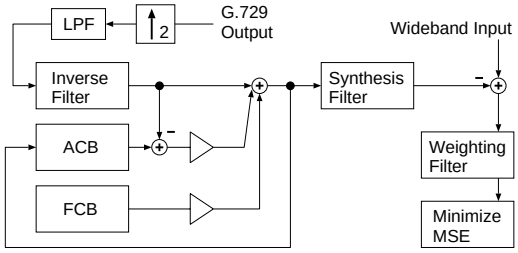

Figure 4: Enhancement coder structure C.

3.3. Enhancement Coder Structure C

In the third structure depicted in Fig. 4, the G.729 output speech without post-processing is decoded, up-sampled, low-pass filtered, and then inverse filtered. The inverse filter is defined as the inverse of the synthesis filter obtained in the enhancement coder. The inverse filtered signal is used as one of the excitation sources.

This system adopts an additional ACB (A-ACB) approach [5]. In this approach, the A-ACB codevector is obtained by subtracting the inverse filtered vector from the ACB codevector to avoid duplicating the pitch-periodic component. Consequently, the A-ACB codevector provides the periodic component which does not exist in the inverse filtered vector. The sum of the inverse filtered vector and the A-ACB codevector gives the overall pitch excitation vector.

The excitation codebook search is summarized below. First,

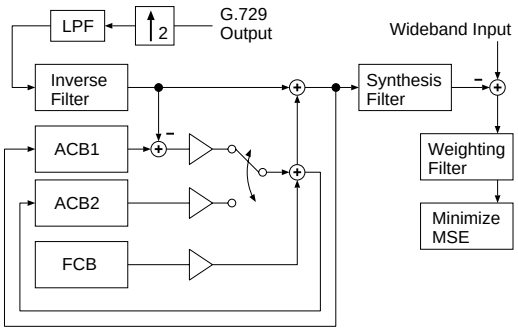

Figure 5: Enhancement coder structure D.

the contribution of the inverse filtered signal is removed from the target signal. The excitation codebook search is then performed.

3.4. Enhancement Coder Structure D

Fig. 5 shows the last structure that we have examined. This system combines the concepts of structures B and C. Here, two types of ACBs are generated and one of them is selected by a switch. It is noted that, if only the ACB2 is selected, this system provides the same output as in structure B. If the ACB1 is always chosen, the output is the same as in structure C.

The codebook search procedure is similar to that of the system C. The difference is that the switch is determined so that it selects the one of two ACBs whose codevector gives the lower distortion.

3.5. Low-pass Filter and Frame Synchronization

In the proposed scheme, the base coder encodes the down-sampled speech, and the excitation or output speech obtained in the base coder are up-sampled at the enhancement coder. Since the CELP algorithm is sensitive to waveform misalignment and phase distortions, it is important to avoid these problems during the sampling-rate conversion. The use of a linear-phase FIR filter for low-pass filtering can eliminate the phase distortions. The waveform misalignment can be removed by adjusting the filter length.

In the case of structure A, by setting the length of the low-pass filter to 10 ms, the frame of the enhancement coder is synchronized with that of the base coder [7]. On the other hand, in structures B, C and D, a different low-pass filter is incorporated for up-sampling. The input wideband speech and up-sampled G.729 output are in alignment when the length of each low-pass filter is set to 5 ms.

3.6. Performance Comparison

The proposed four types of enhancement coder were evaluated in terms of signal-to-noise ratio (SNR) and weighted SNR (WSNR). The WSNR is defined as the SNR between perceptually-weighted input and perceptually-weighted output signals. The test material was a set of sentences spoken by four females and four males. Test conditions are summarized in Table 1. Since the pitch delay of the enhancement coder is correlated with that of the base coder, the ACB index is selected from the range restricted by the G.729 pitch delay. In the structure A, the gain of the up-sampled pulse excitation of G.729 is unquantized. The 8 kb/s G.729 coder is used as a base coder in the test.

Table 1: Test conditions for objective evaluation. The number of bits for each 5 ms subframe is indicated.

	Enhancement Coder Structure	
	A, B, C	D
LSP	unquantized	
ACB	4-bit	3-bit
ACB switch	–	1-bit
ACB gain	3-bit	
FCB	21-bit	
FCB gain	unquantized	

Table 2: WSNR and SNR for four types of structure (dB).

Structure	A	B	C	D
WSNR	10.74	12.00	12.19	12.30
SNR	16.60	18.26	18.00	18.46

Table 2 shows the performance for the proposed structures. Compared to structure A, all of the other three structures provide better performance. This indicates that the use of the output speech from the base coder is more efficient than the techniques used in MPEG-4 CELP. Specifically, the results indicate that up-sampling the G.729 pulse excitation does not provide a useful component for the enhancement coder's excitation. It is also shown that structure D achieves the best performance in both SNR and WSNR. This means that it is effective to switch between two types of ACB. Based on these results, we adopt structure D for our enhancement coder.

4. LSP QUANTIZATION USING INTERFRAME AND INTRAFRAME PREDICTION

The LSP parameters show high frame-to-frame correlations for speech, and the use of interframe prediction improves the quantization performance. In addition, the LSP parameters obtained from wideband speech are also correlated with those from narrow-band speech, since the LP spectra for narrow-band and wideband speech are quite similar in the frequency range of 0-4 kHz. Hence the quantization performance can be further improved by introducing intraframe prediction where wideband LSP parameters are estimated from narrow-band LSP parameters.

In the proposed scheme, the LSP parameters of the enhancement coder are quantized with the help of both intraframe and moving-average (MA) interframe prediction. The quantized LSP parameters at time t are expressed as

$$\hat{f}_t(i) = \sum_{p=0}^{P} \alpha_p(i)\hat{l}_{t-p}(i) + \beta(i)\,\hat{f}'_t(i) \tag{2}$$

where P is the MA prediction order, $\hat{l}_t(i)$ is the quantized prediction error at time t, and $\alpha_p(i)$ and $\beta(i)$ are the interframe and intraframe predictive coefficients, respectively. The parameters $\hat{f}'_t(i)$ include the quantized LSP parameters from the G.729 base coder in the first ten parameters and zeros in the rest, i.e.,

$$\hat{f}'_t(i) = \begin{cases} \hat{f}_t^{(G.729)}(i), & i = 0, \cdots, 10 \\ 0, & i = 11, \cdots, N \end{cases} \tag{3}$$

where N is the order of LP analysis in the enhancement coder.

Table 3: Bit allocation of 8 and 9.6 kb/s enhancement coders for a 10 ms frame.

	Enhancement coder at 8 kb/s	Enhancement coder at 9.6 kb/s
LSP	18	20
ACB	3×2	3×2
ACB switch	1×2	1×2
FCB	21×2	28×2
Gain CB	6×2	6×2
Total	80	96

5. BANDWIDTH SCALABLE CODERS AT 16 KB/S

5.1. Enhancement coder

Two coders with bandwidth scalability are designed: one is scalable with the 8 kb/s G.729 base coder, and another with the 6.4 kb/s G.729-D base coder. The bit rate of the enhancement coder is set to be 8 and 9.6 kb/s for the G.729 and G.729-D base coders, respectively. In both cases, the total bit rate is 16 kb/s. The bit allocation of the enhancement coders is shown in Table 3.

Using a 25 ms Hamming window, 16th-order LP analysis is performed once per frame, and the LP parameters are quantized in the LSP domain. The excitation codebook parameters are transmitted every 5 ms subframe.

The LSP quantizer is organized as follows. The predictor codebook uses 1 bit to switch the predictive coefficients. Each entry of the predictor codebook consists of interframe and intraframe predictive coefficients. The order of the MA interframe predictor is 4. The prediction residual is encoded with a 3-stage codebook, in which (7+5+5) bits are assigned for the 8 kb/s enhancement coder and (7+6+6) bits for the 9.6 kb/s enhancement coder.

The ACB index is differentially encoded using the G.729 pitch delay. The fixed codebook is a 21-bit algebraic codebook with 4 pulses for the 8 kb/s enhancement coder, and a 28-bit algebraic codebook with 6 pulses for the 9.6 kb/s enhancement coder. The ACB and FCB gains are vector-quantized with 6 bits. The 4th-order MA prediction is applied to the FCB gain.

In the enhancement decoder, adaptive postfiltering is applied to the reconstructed speech. The adaptive postfilter is the cascade of a long-term postfilter, a short-term postfilter and a tilt compensation filter.

5.2. Subjective Evaluation

A-B comparison tests were conducted to evaluate the performance of the proposed coders. In the tests, the proposed coders were compared with a publicly available reference model MPEG-4 CELP coder at 16 kb/s. Note that the MPEG-4 encoding algorithm is not standardized so that some proprietary implementations could have better performance than the version used here as a reference. The MPEG-4 CELP used in the tests realizes a bandwidth scalability such that narrow-band speech is encoded at 6 kb/s in the base layer and wideband speech is encoded using an extra 10 kb/s in the enhancement layer. Ten people listened to 8 sentences uttered by 4 female and 4 male speakers.

The test results are provided in Table 4 and 5. Table 4 indicates that the coder scalable with the 8 kb/s G.729 provides slightly better quality than the MPEG-4 CELP coder. It is shown from Table 5 that the coder scalable with the 6.4 kb/s G.729-D outperforms the MPEG-4 CELP coder. These results also demonstrate that, when the total bit rate is fixed, there is a trade-off in performance between the base and enhancement coders. In other words, the performance of the enhancement coder improves as its bit rate increases, at the expense of the quality of the base coder.

Table 4: A-B test result for proposed coder with 8 kb/s base and 8 kb/s enhancement coders (%).

	Prefer proposed	No preference	Prefer MPEG-4
Total	41.25	26.25	32.50
Female	45.00	27.50	27.50
Male	37.50	25.00	37.50

Table 5: A-B test result for proposed coder with 6.4 kb/s base and 9.6 kb/s enhancement coders (%).

	Prefer proposed	No preference	Prefer MPEG-4
Total	45.00	27.50	27.50
Female	40.00	30.00	30.00
Male	50.00	25.00	25.00

6. CONCLUSIONS

We have proposed a bandwidth-scalable coding scheme based on the G.729 standard as a base layer coder. The proposed scheme consists of a base and an enhancement layer. The base coder encodes narrow-band speech using the G.729 algorithm, while the enhancement coder encodes wideband speech using a fullband-type CELP model. In the enhancement coder, the coding parameters are efficiently quantized using information which the coding parameters of the base coder provide. Two bandwidth-scalable coders have been designed: one is scalable with the 8 kb/s G.729 base coder and another with the 6.4 kb/s G.729-D base coder. It has been shown that, for wideband speech, both proposed coders at 16 kb/s provide better quality than the 16 kb/s MPEG-4 CELP with bandwidth scalability.

7. REFERENCES

[1] R. D. De Iacovo and D. Sereno, "Embedded CELP coding for variable bit-rate between 6.4 and 9.6 kb/s," in *Proc. ICASSP'91*, pp.681–683, 1991.

[2] A. L. Guyader and E. Boursicaut, "Embedded wideband VSELP speech coding with optimized codebooks," in *Proc. IEEE Workshop on speech coding*, pp.15–16, 1993.

[3] ITU-T Recommendation G.722, "7 kHz audio-coding within 64 kbit/s," 1988.

[4] J. Suzuki and N. Ohta, "Variable rate coding scheme for audio signal with subband and embedded coding techniques," in *Proc. ICASSP'89*, pp.188–191, 1989.

[5] A. Kataoka, S. Kurihara, S, Sasaki and S. Hayashi, "A 16-kbit/s wideband speech codec scalable with G.729," in *Proc. EUROSPEECH*, pp.1491–1494, 1997.

[6] T. Nomura, M. Iwadare, M. Serizawa and K. Ozawa, "A bitrate and bandwidth scalable CELP coder," in *Proc. ICASSP'98*, pp.341–344, 1998.

[7] ISO/JTC1, "Final draft international standard FDIS 14496-3: Coding of audiovisual objects, part 3: Audio," ISO/JTC1 SC29 WG11, 1998.

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

[9] ITU-T Recommendation G.729–Annex D, "6.4 kbit/s CS-ACELP speech coding algorithm," 1998.