A 16-KBIT/S WIDEBAND SPEECH CODEC SCALABLE WITH G.729

A. Kataoka, S. Kurihara, S. Sasaki, and S. Hayashi
NTT Human Interface Labs.
3-9-11, Midori-cho, Musashino-shi, Tokyo 180, Japan
Tel. +81 422 59 4707, FAX: +81 422 60 7811, E-mail: kata@splab.hil.ntt.co.jp

ABSTRACT

A wideband speech scalable codec is proposed for improving the flexibility in telecommunication networks. This coder is scalable with G.729 (ITU 8-kbit/s standard). Its decoder can process the incoming bitstream at three bit rates (8, 12, and 16 kbit/s) and provide a choice of speech types (wideband and telephone-band). The codec has a split-band structure, where both bands are coded by analysis-by-synthesis techniques. This paper proposes two types of scalable codec: a separate one and a composite one. It also proposes a new method (an additional adaptive codebook) for predicting pitch, while maintaining scalability with the G.729 codec. Subjective testing for wideband speech showed that the quality of the proposed codec at 16-kbit/s is equivalent to that of the 64-kbit/s G.722, and at 12-kbit/s is better than that of the 48-kbit/s G.722. Testing has further demonstrated that the 8-kbit/s coder provides high quality for telephone-band speech.

1. INTRODUCTION

As use of the internet increases, services using networks are becoming a growth industry. However, networks such as LAN and ATM sometimes experience packet (or cell) loss, that is, the transmitted data disappears as a result of network congestion. Therefore, a scalable speech-coding method in ATM and LAN is desirable, so the scalable speech decoder can process a fraction of the incoming bitstream. The MPEG has already started studying the scalable methods and ITU also will study them soon [1].

Use of the personal computer is beginning to expand into tele-conferencing, it enables to handle text, figures, and voice (speech) simultaneously. In this case, since people use both hands to operate a keyboard, voice is delivered by a hands-free method, i.e. a loud speaker. To make the more natural, a loud speaker using a wideband speech (7 kHz bandwidth) is desirable because telephone-band speech has limited frequency.

This paper proposes a wideband speech codec scalable with G.729 (ITU 8-kbit/s standard) [2][3]. Its decoder can process the incoming bitstream at three bit-rates (8, 12 and 16 kbit/s) and provide two types of speech (wideband or telephone-band). This paper proposes two types of scalable codecs in a lower band: a separate one and the composite one.

After giving an overview of the proposed codec, we describe the two types of scalable codec in the lower band and the higher band codec, and propose a new method (an additional adaptive codebook) for predicting pitch, while maintaining scalability with the G.729 codec. We describe the performance of the lower band codec. Finally, we give the results of subjective quality tests for both wideband and telephone-band speech.

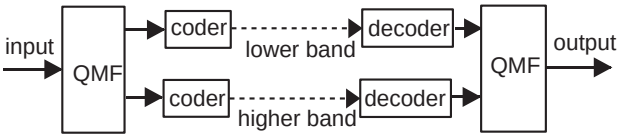

Figure 1: Proposed codec.

Table 1:Bit allocations of each type of scalable codec in the lower band.

Codec	Parameter	Separate		Composite	
		1st	2nd	1st	2nd
Additional codec	Fixed codebook	7+7	7+7	7+7	7+7
	Codebook sign	1+1	1+1	1+1	1+1
	Gain codebook	4	4	4	4
	(Sub-total)	(40)		(40)	
G.729		80		80	
Total		120		120	

2. OVERVIEW OF PROPOSED CODEC

A block diagram of the proposed codec is shown in Fig. 1. The codec has a split-band structure [4]. A quadrature mirror filter (QMF) splits the frequency band of 0 to 8 kHz into a lower sub-band (0 to 4 kHz) and a higher sub-band (4 to 8 kHz). The speech signals in each sub-band are encoded using different CELP-based coders. Each codec has the same 10-ms frame length as G.729, and each frame consists of two subframes. A 12-kbit/s scalable codec composed of a G.729 codec (8 kbit/s) and an additional 4-kbit/s codec is used in the lower band. Two types of 12-kbit/s scalable coder are described in Section 3. In the higher band, a 4-kbit/s CELP coder is used.

The bitstream consists of three parts: 8 kbit/s for the G.729 codec, 4 kbit/s for the additional codec, and 4 kbit/s for the higher band codec. If a decoder uses the whole bitstream, it reproduces high-quality wideband speech. If it uses 8 kbit/s for the G.729 codec and 4 kbit/s for the higher band codec, it reproduces good-quality wideband speech. If it uses only 8 bit/s for the G.729 codec, it provides high-quality telephone-band speech.

3. A SCALABLE CODEC IN THE LOWER BAND

The lower band codec is a scalable codec with G.729. It consists of a G.729 codec and an additional codec. We studied a separate scalable codec and a composite scalable codec. Bit allocations of each type of scalable codec in the lower band are listed in Table 1. G.729 uses 80 bits and the additional codec uses 40 bits.

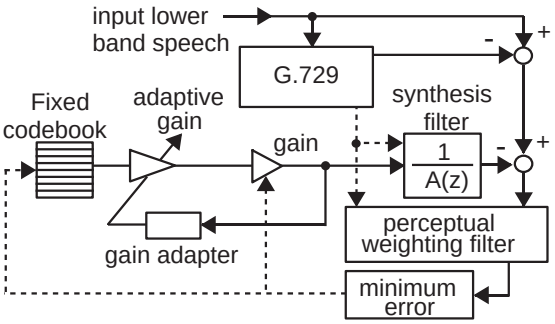

Figure 2: Separate scalable coder.

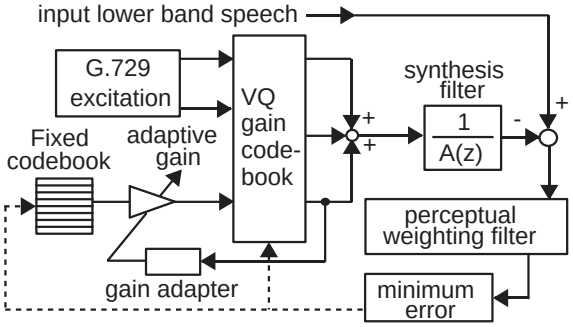

Figure 3: Composite scalable coder.

3.1. Separate Scalable Codec

The separate scalable coder shown in Fig. 2 encodes the input speech sequentially. The input speech encoded by G.729 is decoded and compared with the original speech. Then the difference is encoded by the additional coder. In this structure, G.729 gives its standard performance and the additional codec assists it.

If the additional coder has its own synthesis filter for the difference in speech, the new synthesis filter will require many bits for the LPC coefficients. Therefore, we decided it should use the synthesis filter belonging to the G.729 coder. Likewise, it uses the G.729 coder's perceptual weighting filter. The additional encoder uses an analysis-by-synthesis technique to determine the excitation vector from a fixed codebook. The fixed excitation vector is the sum of the two sub-excitation vectors [5]. The gain adapter predicts the fixed excitation gain by considering the sequence of previous fixed excitation vectors [5].

3.2. Composite Scalable Codec

Figure 3 is a block diagram of the composite scalable coder. It encodes the input speech simultaneously using two coders: the G.729 coder and an additional one. It has three excitation vectors; two from the G.729 coder and one from the additional coder. The composite coder selects excitation vectors that minimize the perceptually weighted distortion between the target vector and the synthesized speech. Although in the separate coder the G.729 and additional coders select the excitation vectors independently, in the composite coder they are selected simultaneously to achieve the best performance. The adaptive codebook is generated from the excitation vectors of only the G.729 coder to keep scalability. Furthermore, to improve the speech quality, the excitation vector of the additional coder is orthogonalized to both the pitch excitation vector and the fixed excitation vector of the G.729 coder. The

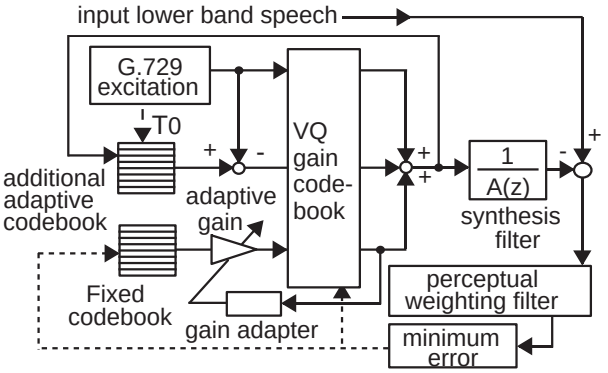

Figure 4: Composite scalable coder with additional adaptive codebook.

gains of three excitation vectors (the pitch, fixed, and the additional fixed excitation vectors) are quantized using the 3-dimensional VQ gain codebook.

When the CELP coder determines the excitation vectors from excitation sources, in each case it selects the vector that minimizes the distortion between the target vector and the synthesized speech. The target vector is generated by subtracting a zero-input vector from the input speech. The zero-input vector is the zero-input response from the previous synthesized speech. Although the G.729 part of the separate coder uses the zero-input response from the previous synthesized speech of only the G.729 local decoder, the G.729 part of the composite coder uses that from the previous synthesized speech of both the G.729 local decoder and the additional local decoder. Therefore, it gives better performance than the separate coder.

Although the standard G.729 decoder can handle data encoded by the G.729 part of the composite coder, the standard G.729 decoder does not operate at normal performance levels. This is because of the G.729 part of the composite coder selecting the excitation vectors considering the zero-input response from the previous synthesized speech of the additional coder.

4. ADDITIONAL ADAPTIVE CODEBOOK

To maintain the scalability with the G.729 coder, the adaptive codebooks of both the separate and the composite scalable coders must be generated from the excitation vectors of only the G.729 coder. Therefore, the excitation vector of the additional coder does not contribute to the adaptive codebook. The adaptive codebook (pitch prediction) is, however, important for the speech quality of the coder. Therefore, if we feed the excitation vector of the additional coder back to the adaptive codebook to maintain the scalability with the G.729 coder, speech quality can be improved. We propose a new pitch prediction method based on this idea.

Figure 4 shows a block diagram with the new pitch prediction method (additional adaptive codebook) applied to the composite coder. Bit allocations are listed in Table 2. The additional adaptive codebook is generated from the excitation vectors of both the G.729 coder and the additional coder. An additional pitch excitation vector is generated by subtracting the excitation vector of the G.729 coder from the excitation vector of the additional adaptive codebook. Therefore, the sum of the excitation vector of the G.729 coder and the additional pitch excitation vector gives the overall pitch excitation vector.

Table 2: Bit allocations of the composite scalable codec with additional adaptive codebook in lower band.

Codec	Parameter	1st	2nd
Additional codec	Fixed codebook	6+6	6+6
	Codebook sign	1+1	1+1
	Gain codebook	6	6
	(Sub-total)	(40)	
G.729		80	
Total		120	

Table 3: Bit allocations of the higher band codec.

Parameter	Subframe		Frame
	1st	2nd	(80 samples)
LSP	-	-	6
Fixed codebook	6+6	6+6	24
Codebook sign	1+1	1+1	4
Gain codebook	3	3	6
Total	-	-	40

The additional pitch lag is determined using a backward-adaptive method and an open-loop search. Pitch search finds the highest cross-correlation between the excitation vector of the G.729 coder and the past residual vectors in the additional adaptive codebook. It is only performed around the pitch lag T_0 of the G.729 coder. Since the excitation vector of the G.729 coder is determined using a closed-loop search, it has a high-correlation with the overall pitch excitation vector. This approach does not require the additional pitch lag to be sent to the decoder because the decoder can get the excitation vector and the pitch lag of the G.729 decoder, and select only the highest cross-correlation. Since the gains of three excitation vectors are determined using the closed-loop search, if the additional pitch excitation vector is not suitable for the pitch excitation vector, the gain codebook gives it a very small value.

5. HIGHER BAND CODER

The waveform in the higher band looks like a random signal, and the long-term correlation between two adjacent pitch periods is not very high. Although a conventional CELP coder uses pitch prediction (an adaptive codebook), the higher band coder does not use it. Therefore, this coder has only one excitation source: a fixed codebook.

The coder uses an analysis-by-synthesis technique to determine the excitation vectors from the codebook. Bit allocations of the higher band coder are listed in Table 3. The fixed-excitation vector is the sum of the two sub-excitation vectors [5]. The gain adapter predicts the fixed excitation gain by considering the sequence of previous fixed excitation vectors [5].

The higher band LSP quantizer uses the same structure as the G.729 except it has a fixed MA coefficient. In our study, we evaluated that the prediction gain versus the order of LPC coefficients at different bits (unquantized, 8 bits, and 6 bits). When the LSP quantizer used 8 bits, the prediction gain was improved until the 6th order of LPC. However, it saturated at higher orders. When the LSP quantizer used 6 bits, the 8th order of LPC gave the best performance. As the number of bits increased, performance improved. However, the difference between the prediction gains using 6 bits and 8 bits was not large. Based on informal listening tests, it is desirable that a fixed codebook uses many bits we chose the 6-bits LSP quantizer at the 8th order LPC.

Table 4: Performance for each type of coder in lower sub-band.(AAC: additional adaptive codebook)

G.729 part in each type of coder		
Type	Seg SNR (dB)	CD (dB)
Separate (G.729)	15.57	3.04
Composite	15.02	3.03
Composite with AAC	14.97	3.05
each type of total coder in lower sub-band		
Separate	15.92	1.78
Composite	17.83	1.87
Composite with AAC	18.20	1.86

6. PERFORMANCE OF EACH TYPE OF SCALABLE CODER IN LOWER BAND

Each type of scalable coder was evaluated using two objective measures (segmental SNR and cepstrum distance). The results are listed in Table 4. Each type of scalable coder consists of a G.729 coder (8 kbit/s) and an additional 4-kbit/s coder. The G.729 part in each type of coder shows a large cepstrum distance (CD), which is caused by the difference in frequency response. Wideband speech has bandwidth of about 7 kHz. The QMF splits the wideband speech into two sub-bands, the lower sub-band has 4-kHz bandwidth, while the telephone-band speech has a bandwidth of about 3.4-kHz. Since the G.729 coder is designed for telephone-band speech, the LSP parameter is not suitable for 4-kHz bandwidth speech. Therefore, each type of coder has to handle the frequency response from 3.4 to 4 kHz using the additional coder. With each type of total coder, this additional coder improves the CD.

As regards segmental SNR compared with the G.729 part, the separate coder achieves improvement of 0.4 dB, the composite coder 2.2 dB, and the composite coder with the additional adaptive codebook 2.6 dB. The G.729 part of the composite coder only degrades the performance by 0.6 dB compared with the normal G.729 coder, even if elects the excitation vectors considering the zero-input response from the previous synthesized speech of the additional coder.

7. EVALUATION OF SPEECH QUALITY

We evaluated the quality of each type of scalable codec using the mean opinion score (MOS). The wideband speech and the telephone-band speech were evaluated separately. Speech samples were obtained from ten male and ten female speakers for each test. The number of listeners was 24. Eight SNR values (0 to 35 dB in 5 dB steps) of the modulated noise reference unit (MNRU), called Q values, were included in the test.

The equivalent Q values of each type of scalable codec wideband speech are for shown in Figure 5(a). The separate scalable codec (denoted **A**) has equivalent subjective quality to the 56-kbit/s G.722. The composite scalable codec (**B**) scored better than the separate one. The composite scalable codec with additional adaptive codebook (**C**) was rated equivalent to the 64-kbit/s G.722. This is because the additional adaptive codebook enhances the pitch.

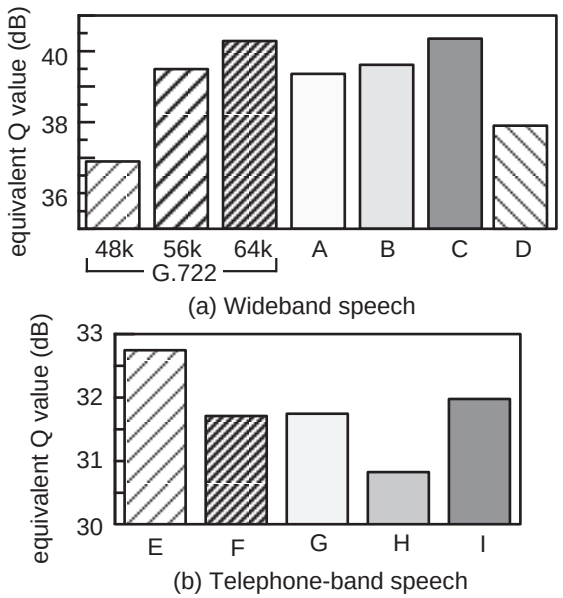

Figure 5: Subjective test results. (a) Wideband speech (b) Telephone-band speech (A to C: 16-kbit/s codecs for wideband speech, A: separate, B: composite, C: composite with additional adaptive codebook (AAC), D: 12-kbit/s codec for wideband speech, E: G.729 coder for telephone-band speech, F to I: 8-kbit/s codecs for lower band speech, F: G.729 part in separate, G: G.729 part in composite, H: G.729 part in composite with AAC, I: H + low-pass filter)

In the 12-kbit/s wideband speech codec, the codec in the lower band is only the G.729 codec, that is, the 8-kbit/s G.729 codec and 4-kbit/s higher band codec. The 12-kbit/s codec (**D**) scored worse than the 56-kbit/s G.722, but better than the 48-kbit/s G.722.

The equivalent Q value of the G.729 part in each type of scalable codec for telephone-band speech is shown in Figure 5(b). **E** is the G.729 codec under 3.4-kHz bandwidth speech, and **B** to **E** are G.729 parts in each type of scalable codec under the lower sub-band speech (4-kHz bandwidth).

The G.729 part in the separate type (**F**) scored slightly worse than the G.729 codec in 3.4-kHz bandwidth speech (**E**). This is because the G.729 codec is designed for telephone-band speech, so the coded speech from 3.4 to 4.0 kHz sounds noisy. But the degradation is only about 1 dB and the quality is still good.

Although the G.729 part in the composite codec (**G**) has the same quality as the separate codec, the G.729 part in the composite codec with the additional adaptive codebook (**H**) has worse quality than the separate codec. However, when the coded speech of the composite codec with the additional adaptive codebook is passed through a low-pass filter (cut off frequency 3.4 kHz), its speech quality is improved. This is denoted **I**. Its quality is better than that of the G.729 part in the separate type (**F**). The LPF cuts all noise between 3.4 and 4.0 kHz.

8. CONCLUSION

This paper proposed a scalable codec for wideband speech. This codec is scalable with G.729 (ITU 8-kbit/s standard). The decoder can process the same bitstream at three different bit-rates (8, 12, and 16 kbit/s) and provide a choice of speech type (wideband or telephone-band). The codec has a split-band structure, where both bands are coded using analysis-by-synthesis techniques. This paper proposed two types of scalable codec in a lower-band: a separate one and the composite one. It also proposed a new method (additional adaptive codebook) for predicting pitch, while maintaining scalability with the G.729 codec. A lower band codec consists of the G.729 and the additional codec. The separate coder encodes the input speech sequentially. The composite coder encodes the input speech simultaneously by using both the G.729 coder and the additional coder.

Subjective testing showed that the quality of the proposed codec at 16 kbit/s is equivalent to that of the 64-kbit/s G.722, and at 12 kbit/s is better than that of the 48-kbit/s G.722 for wideband speech. Testing has further demonstrated that the 8-kbit/s codec provides high quality for telephone-band speech.

Acknowledgements

We wish to thank Dr. Nobuhiko Kitawaki and Takao Kaneko for guiding our research.

9. REFERENCES

[1] “Draft text for new and continued questions for WP 2/15,” ITU-T Study Group 15 TD 48-E, June 1996.

[2] ITU-T Recommendation G.729 Coding of speech at 8-kbit/s using conjugate structure algebraic code-excited linear prediction (CS-ACELP), COM 15-152-E, 1995.

[3] A. Kataoka et al., “ITU-T 8-kbit/s Standard Speech Codec for Personal Communication Services,” Proceedings of Int. Conf. on Universal Personal Communications, pp. 818-822, 1995.

[4] Rosario D. de I. et al., “Some Experiments of 7-kHz Audio Coding at 16 kbit/s,” *Proc. ICASSP 89*, S4.19, pp. 192-195, 1989.

[5] A. Kataoka et al., “An 8-kb/s Conjugate Structure CELP (CS-CELP) Speech Coder,” *IEEE Trans. Speech and Audio Processing*., Vol. 4, Num. 6, pp. 401-411, 1996.