BANDWIDTH EXTENSION OF A NARROWBAND SPEECH CODER FOR MUSIC STREAMING SERVICES OVER IP NETWORKS

Young Han Lee and Hong Kook Kim

Dept. of Information and Communications
Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea

{cpumaker, hongkook}@gist.ac.kr

ABSTRACT

In this paper, we propose a bandwidth extension (BWE) algorithm for a low-bit-rate narrowband CELP coder using a spectral envelope sharing approach to develop a wideband speech coder. The developed wideband speech coder, referred to here as the BWE coder, is constructed using an embedded structure by adding an enhancement layer to the narrowband CELP coder. To minimize the bit-rate increase caused by the enhancement layer, the proposed BWE coder shares the spectral envelope and excitation parameters both with the narrowband CELP coder and the enhancement layer. In this paper, we choose G.729EV layer 2 as the baseline narrowband speech coder, and mel-frequency cepstral coefficients (MFCCs) are used to reconstruct the higher frequency components at the enhancement layer. By doing this, the bit-rate of the proposed BWE coder is found to be 12.7 kbit/s, just 0.7 kbit/s higher than that of G.729EV layer 2. It is also demonstrated from a MUSHRA test with audio signals from four different music genres, that the BWE coder gives better quality than G.729EV layer 2 and comparable quality to G.729EV layer 3, corresponding to an overall bit-rate reduction of 1.3 kbit/s.

Index Terms: bandwidth extension, wideband speech coding, G.729EV, IP networks, spectral envelope sharing

1. INTRODUCTION

Voice over internet protocol (VoIP) services have been receiving a great deal of attention in recent years, with most of them being provided for free, or at least a very reasonable price, as compared with legacy public switched telephone network (PSTN) services. In IP networks, a VoIP service is provided through speech coders such as ITU-T G.729 [1], ITU-T G.723.1 [2], ITU-T G.728 [3], and the Internet low bit-rate codec (iLBC) [4]. In the early days of this service, these speech coders were sufficient for efficiently processing speech signals, though it was generally agreed that they did not satisfy users' increasing expectations of higher sound quality, especially when music signals were used such as for a color ring-back-tone service [5]. However, in order to launch an improved VoIP service, several limitations of existing services should be overcome.

One of the limitations is the bandwidth of signals processed by existing VoIP speech codecs; speech codecs in VoIP systems compress input signals with a low bit-rate, around 5.3 to 16 kbit/s, under the assumption that the input signals are bandlimited up to 3.4 kHz, whereas audio signals of good quality generally require a bandwidth higher than 4 kHz. In order to improve the audio quality in the VoIP system, a coder should be able to deal with a wideband signal, i.e. of up to around 7 kHz. Another limitation degrading service quality is the bit-rate, since in IP networks a higher bit-rate results in a larger packet size, causing the decoded audio to be clipped or skipped due to packet loss [7]. This degradation implies that audio codecs are not suitable for VoIP codecs although audio codecs are able to deal with an audio bandwidth up to around 20 kHz, their bit-rates are typically too higher for use in IP networks. For example, G.722, a standard wideband coder used in VoIP, provides a wideband signal up to 7 kHz with bit-rates of 48, 56, and 64 kbit/s. These higher bit-rates of G.722 cause quality degradation in IP networks. For this reason, in order to improve the quality of music streaming services, an algorithm for extending the bandwidth with a smaller bit-rate increase is required.

In this paper, we propose a BWE algorithm for a narrowband speech coder to improve the audio quality for music streaming services over IP networks. The proposed BWE coder is based on the proposed BWE algorithm using an embedded structure using a baseline coder followed by an enhancement layer. Here, G.729EV layer 2 is selected as the baseline coder to show the effectiveness of the proposed BWE coder, since G.729EV has a bandwidth extension algorithm by using G.729EV layer 2, which is called G.729EV layer 3. By doing this, we can compare the performance of the proposed algorithm with that of the bandwidth extension algorithm in G.729EV, i.e. G.729EV layer 3. Additionally, to minimize the bit-rate increase due to the enhancement layer, the enhancement layer of the proposed coder is designed to share a spectral envelope and excitation parameters with both the baseline coder and the enhancement layer. In other words, mel-frequency cepstral coefficients (MFCCs) [8] are used to represent the spectral envelope, and are subsequently converted to linear prediction coefficients (LPCs) for both the baseline coder and the enhancement layer. Moreover, the excitation signals for the enhancement layer are expanded from those decoded by the baseline coder, incurring no additional bit increase for the enhancement layer. As a result, the proposed BWE coder operates at 12.7 kbit/s, which is only 0.7 kbit/s higher than the baseline coder. The performance of the proposed BWE coder is then evaluated by both an objective test and a subjective test. As an objective test, the spectra of the decoded wideband signal from the proposed BWE coder are compared with those from the G.729EV layer 2. In addition, a multiple stimuli with hidden reference and anchor (MUSHRA) test [9] is used for a subjective listening test.

The rest of this paper is organized as follows. Following the Introduction, the structures and algorithms of the proposed BWE encoder and decoder are described in Sections 2 and 3, respectively. In Section 4, the quality of the proposed BWE coder is evaluated through a comparison of the spectra of the decoded wideband signal and a MUSHRA test. Finally, we conclude the paper in Section 5.

1-4244-1222-6/07/$25.00 ©2007 IEEE

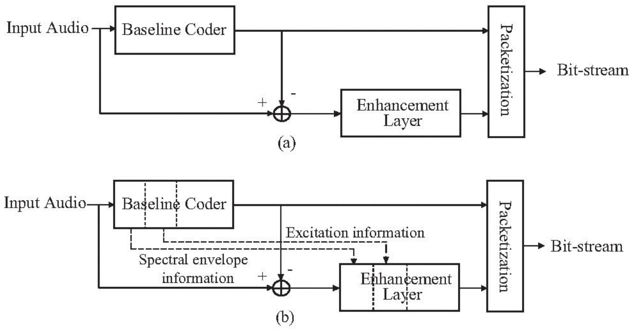

Fig. 1. Comparison of embedded encoding structures between (a) a conventional BWE coder, and (b) the proposed BWE coder.

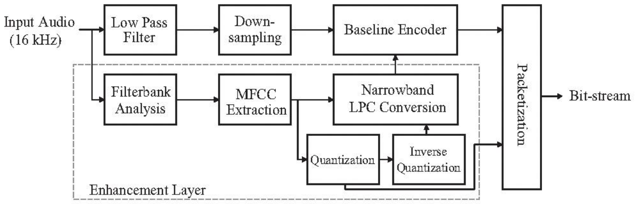

Fig. 2. Block diagram of the proposed BWE encoder.

2. PROPOSED BWE ENCODER

In this section, we describe the proposed BWE coder for music delivery services. The proposed BWE coder is based on an embedded structure using a baseline coder followed by an enhancement layer. Fig. 1(a) shows the basic principle of a conventional embedded coder, where the enhancement layer is devised to increase the bandwidth without regarding the baseline coder [10]. In other words, no parameters between the baseline coder and the enhancement layer are related. On the other hand, the structure of the proposed BWE coder shown in Fig. 1(b) shares information regarding the spectral envelope and excitation of the enhancement layer with the baseline coder. By using this structure, we can develop the enhancement layer with a smaller bit increase than the conventional structure shown in Fig. 1(a).

Fig. 2 shows a block diagram of the proposed BWE encoder. In the figure, G.729EV layer 2 is used as a baseline coder for the proposed BWE coder, where G.729EV layer 2 is a CELP-based coder and operates at 12.0 kbit/s with a frame size of 10 ms. Due to the fact that G.729EV layer 2 can only process narrowband signals, it is a viable candidate for the baseline coder. As described above, we first modify G.729EV layer 2 to incorporate the shared parts with the enhancement layer.

One of the shared parts is the representation of the spectral envelope. In the proposed BWE coder, MFCCs are used as spectral envelope parameters because the majority of other conventional speech coders, including G.729EV layer 2, use LPCs, as shown in Fig. 3. In the proposed coder, a filterbank analysis is applied to audio signals sampled at 16 kHz. The filterbank analysis begins with DC removal followed by pre-emphasis filtering. After that, a Hamming window is applied to the pre-emphasized audio signals. In order to obtain MFCCs, a 512-point fast Fourier transform (FFT) is first performed to compute the magnitude spectrum, and then the magnitude spectra are filtered by 23 triangular-shaped mel-filterbanks.

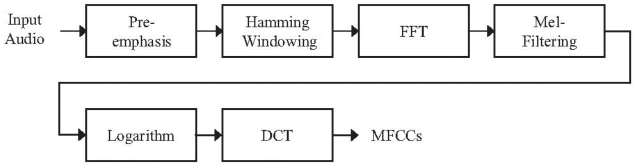

Fig. 3. MFCC extraction procedure.

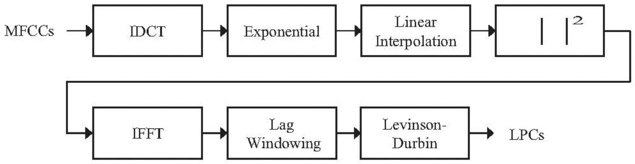

Fig. 4. Procedure for the MFCC to LPC conversion.

Table 1. Bit allocation for the proposed BWE coder.

Parameter	1st Subframe	2nd Subframe	Total (10 ms)
MFCC	-		25
ACB index	8	5	13
Pitch parity	1	-	1
FCB index	13	13	26
FCB sign	4	4	8
Gain VQ	7	7	14
2nd FCB index	13	13	26
2nd FCB sign	4	4	8
2nd FCB gain	3	2	5
FEC	-	1	1
Total			127

Each mel-filterbank energy is subsequently transformed into a logarithmic scale. Finally, a discrete cosine transform (DCT) is applied to obtain 23 MFCCs, and only 12 MFCCs out of 23 MFCCs are quantized with 25 bits/frame [11].

On the other hand, 10 LPCs are needed for the baseline coder. To do this, an inverse DCT is applied to the 12 quantized MFCCs, and 23 mel-filterbank energies are estimated, as shown in Fig. 4. Next, the power density spectrum ranging from 0 to 8 kHz is estimated by linearly interpolating these 23 filterbank energies. Note that in order to obtain the LPCs used for the baseline encoder, we only employ half of the power density spectrum, resulting in a frequency range from 0 to 4 kHz. Then, a 256-point inverse FFT is applied to compute the autocorrelation coefficients, which are subsequently smoothed by applying a lag window. Finally, we can obtain the 10 LPCs using the Levinson-Durbin recursion. These 10 LPCs are used in the baseline coder to model narrowband signals.

The bit allocation for the proposed BWE coder is shown in Table 1. It can be seen that the main difference between the bit allocation for the proposed BWE coder and that of G.729EV layer 2 is that 25 bits/frame are assigned to the MFCCs, but 18 bits are used for LSF quantization in G.729EV layer 2. This means that we need an additional 7 bits/frame for the proposed BWE coder. Note that the numbers of bits assigned for adaptive codebook indices, fixed codebook indices, and gain parameters, codebook indices, and gains remain as in G.729EV layer 2.

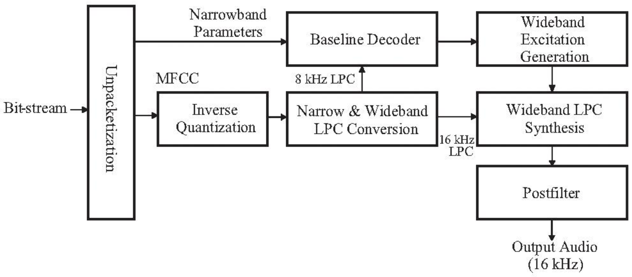

Fig. 5. Block diagram of the proposed BWE decoder.

3. PROPOSED BWE DECODER

Fig. 5 shows a block diagram of the proposed BWE decoder. As can be seen, it mainly consists of two parts: the baseline decoder and the enhancement layer decoder. The enhancement layer decoder decodes 12 MFCCs from the transmitted bit-stream and converts them into 10 narrowband LPCs and 16 wideband LPCs. Then, it generates wideband excitation from the excitation decoded by the baseline decoder. Finally, audio signals of about 8 kHz are obtained by filtering the wideband excitation. The 10 narrowband LPCs are obtained from MFCCs by using the identical procedure described in Section 2, whereas we can obtain the 16 wideband LPCs from MFCCs during the conversion procedure by using the power density spectrum ranging from 0 to 8 kHz.

In order to synthesize high frequency signals, wideband excitation is needed. A detailed block diagram of wideband excitation generation is shown in Fig. 6. To complete this generation, the narrowband excitation from the baseline decoder is first interpolated as

$$e_{i,16}(k) = \begin{cases} e_8(k/2) & k = 0, 2, \cdots, 2N-2 \\ 0 & k = 1, 3, \cdots, 2N-1 \end{cases} \tag{1}$$

where N is the number of samples per frame in the baseline decoder, and $e_8(k)$ and $e_{16}(k)$ are the k-th samples of the narrowband excitation and the interpolated excitation, respectively. Next, half-wave rectification is performed on the interpolated excitation to generate high frequency components such that

$$e_{r,16}(k) = \begin{cases} e_{i,16}(k) & \mathit{if} \quad e_{i,16}(k) > 0 \\ 0 & \mathit{otherwise} \end{cases}, \quad 0 \le k < 2N \tag{2}$$

where $e_{r,16}(k)$ is the k-th sample of the half wave rectified excitation. In order to compensate for the reduced dynamic range due to the baseline coder, the half wave rectified excitation is emphasized through the pre-emphasis filter, $1 - 0.9z^{-1}$. A high pass filter with a cutoff frequency of 4 kHz is then applied to the pre-emphasized excitation, enabling the wideband excitation, $\{e_{16}(k)\}$, to be obtained. Finally, in order to obtain the decoded speech signals, $\{e_8(k)\}$ and $\{e_{16}(k)\}$, are passed through the filters constructed by the narrowband LPCs and the wideband LPCs, respectively. Then, $\{s_8(k)\}$ filtered from $\{e_8(k)\}$ is interpolated by a factor of 2, and subsequently pre-emphasized to increase the dynamic range of the highband spectrum. That is,

$$s_{p,8}(k) = s_{i,8}(k) - \beta\ s_{i,8}(k), \quad 0 \le k < 2N \tag{3}$$

where $s_{i,8}(k)$ and $s_{p,8}(k)$ are the k-th samples of the interpolated and pre-emphasized speech signals from the baseline decoder, respectively, and β is set to 0.2. Note that $s_{i,8}(k)$ and $s_{p,8}(k)$ are sampled at a rate of 16 kHz.

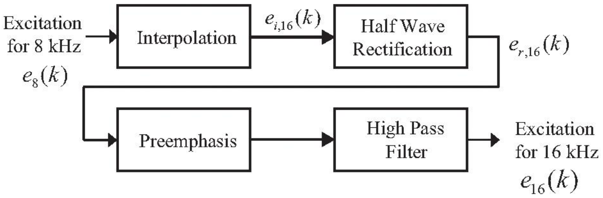

Fig. 6. Wideband excitation generation procedure for the enhancement layer.

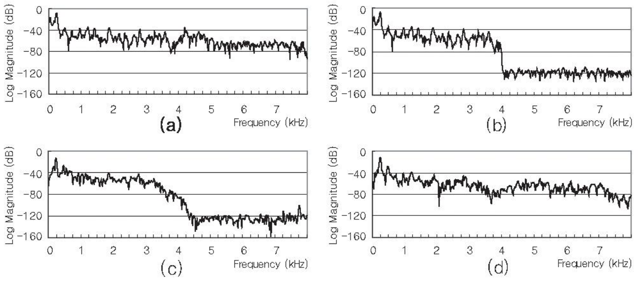

Fig. 7. Performance of the proposed BWE coder by comparing the spectra from (a) the original audio signal sampled at 16 kHz, (b) the signal used as the input for the baseline coder, (c) the signal by the baseline decoder, and (d) the audio signal decoded by the proposed BWE coder.

Finally, we add two sets of speech signals, $\{s_{p,8}(k)\}$ and $\{s_{16}(k)\}$, to obtain the resultant speech signals. In this case, there are two factors that have to be considered: one is a weighting factor between $\{s_{p,8}(k)\}$ and $\{s_{16}(k)\}$, and the other is the delay D that is occurred due to decimation at the proposed BWE encoder in the creation of the narrowband speech signals from the input speech signals. As a result, the decoded speech signals, $\{\hat{s}_{16}(k)\}$ can be obtained by

$$\hat{s}_{16}(k) = w_{16} \cdot s_{16}(k) + w_8 \cdot s_{p,8}(k + D), \qquad 0 \le k < 2N \tag{4}$$

where w_8 and w_{16} are the weighting factors for the decoded signals from the baseline decoder and the enhancement layer, respectively. From exhaustive experiments, it is found that w_8 = 1.2 and w_{16} = 0.5 provide the best speech quality. Also, the delay is set as D = 48.

4. PERFORMANCE EVALUATION

A performance comparison between the proposed BWE coder and G.729EV layer 3 was performed in two ways: a spectrum comparison and a MUSHRA test.

Fig. 7 shows the performance comparison in the spectrum domain. Here, the 70th track in the sound quality assessment material (SQAM) [12] was used as the original signal. Because SQAM audio files were recorded with stereo at a sampling rate of 44.1 kHz, each file was down-sampled from 44.1 kHz to 16 kHz and we only used the right-channel signals. The spectrum of the input audio signal is displayed in Fig. 7(a). During the process of encoding, the baseline coder requires audio signals sampled at 8 kHz, as depicted in Fig. 7(b). Figs. 7(c) and 7(d) show the spectrum of the decoded audio signal from the baseline coder and that of the proposed BWE coder, respectively. From the figure, it can be seen that the spectrum of the audio signal from the proposed BWE coder is very close to that of the original audio.

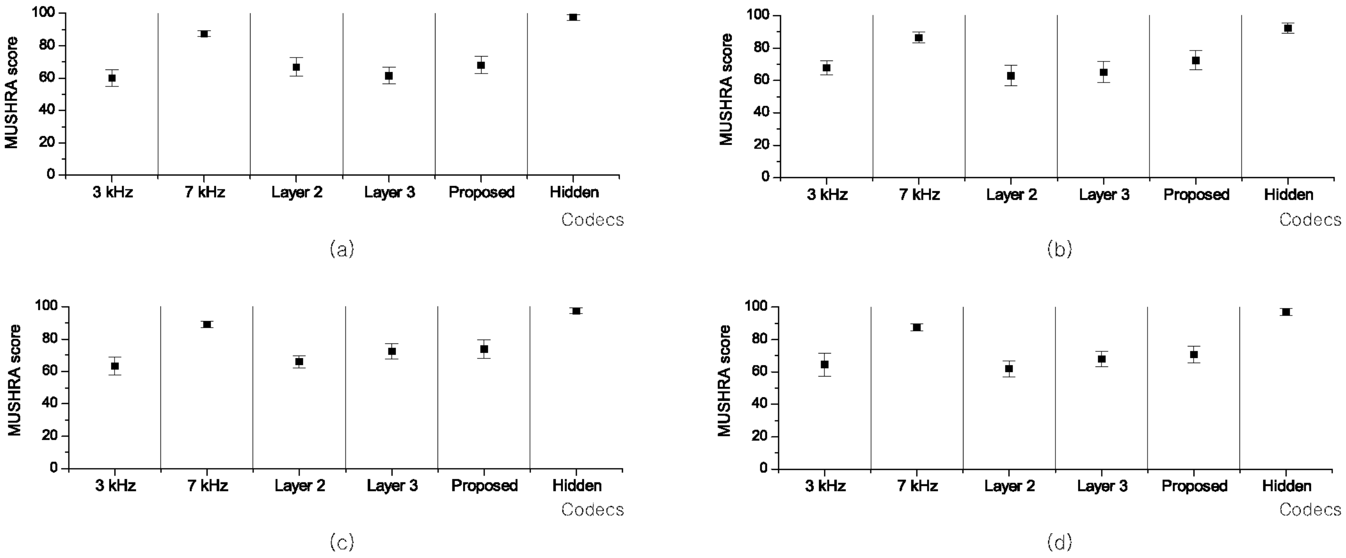

Fig. 8. MUSHRA test results for four different music genres; (a) pop, (b) classic, (c) hip hop, and (d) rock.

To further compare the quality in a subjective way, we performed a MUSHRA test with G.729EV layer 2, G.729EV layer 3, and the proposed BWE coder. To this end, we chose audio files from four different genres such as pop, classic, hip hop, and rock. Each genre consisted of five audio files, resulting in a total of 20 audio files. Here, the audio files were also prepared from SQAM. Six people with no known auditory disease participated in this test.

Fig. 8 shows the MUSHRA test results. There are three anchor signals denoted by 3 kHz, 7 kHz, and Hidden, where '3 kHz' and '7 kHz' are anchor signals low-pass-filtered with a cutoff frequency of 3 kHz and 7 kHz, respectively, and 'Hidden' means an original audio signal unknown to the participants. In addition, there are three processed audio signals by G.729EV layer 2, G.729EV layer 3 and the proposed BWE coder that are denoted in the figure by Layer 2, Layer 3, and Proposed, respectively. It was shown from the figure that the proposed BWE coder gave better quality than G.729EV layer 2. In addition, it was also shown that the proposed BWE coder achieved a comparable quality to G.729EV layer 3 for all four music genres. Especially in the classical genre, the proposed BWE coder gave better quality than G.729EV layer 3, although it operated at a bit-rate that was 1.3 kbit/s lower than the bit-rate of G.729EV layer 3.

5. CONCLUSION

In this paper, we proposed a bandwidth extension of a narrowband CELP coder for music streaming services over IP networks. The proposed BWE coder was based on an embedded structure using G.729EV layer 2 as the baseline coder. The proposed coder was designed to extend the bandwidth with a minimal bit-rate increase by sharing the spectral envelope parameters of the enhancement layer with those of the baseline coder by using MFCCs. In addition, the wideband excitation for the enhancement layer was generated by using the excitation decoded by the baseline coder. As a result, the increase in bit-rate of the proposed coder was only 0.7 kbit/s. The performance of the proposed BWE coder was then evaluated based upon a spectrum comparison and a MUSHRA test. The evaluation confirmed that the proposed bandwidth extension coder provided significantly better quality than when G.729EV layer 2 was used as the baseline coder, and gave comparable quality to G.729EV layer 3 at a bit-rate of 14 kbit/s. As a result, we could conclude that the proposed bandwidth extension could reduced the bit-rate by about 1.3 kbit/s with a comparable quality to the bandwidth extension of G.729EV.

6. REFERENCES

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

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

[3] ITU-T Recommendation G.728, *Coding of speech at 16 kbit/s using low-delay code excited linear prediction*, Sept. 1992.

[4] IETF RFC 3951, *Internet low bit rate codec specification*, Dec. 2004.

[5] Y. H. Lee, H. K. Kim, J. Yu, S. Park, D. H. Lee, D. Woo, "Performance comparison of audio codecs for high quality color ring-back-tone services over CDMA," in *Proc. SPIE Multimedia Systems and Application IX*, pp. 639105-1-8, Oct. 2006.

[6] D. Y. Pan, "Digital audio compression," *Digital Technical Journal*, vol. 5, no. 2, pp. 1-14, 1993.

[7] B. Goode, "Voice over internet protocol," *Proc. IEEE*, vol. 90, no. 9, pp. 1495-1517, July 2002.

[8] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustic, Speech and Signal Processing*, vol. 28, no. 4, pp. 357-366, Aug. 1980.

[9] ITU-R Recommendation BS.1534-1, *Method for the subjective assessment of intermediate quality levels of coding system*, Jan. 2003.

[10] A. Kataoka, S. Kurihara, S. Sasaki, and S. Hayashi, "A 16-kbit/s wideband speech codec scalable with G.729," in *Proc. Eurospeech*, vol. 3, pp. 1491-1494, Sept. 1997.

[11] G. H. Lee, J. S. Yoon, and H. K. Kim, "A MFCC-based CELP speech coder for server-based speech recognition in network environments," in *Proc. Eurospeech*, pp. 3169-3172, Sept. 2005.

[12] EBU Tech Document 3253, *Sound quality assessment material (SQAM)*, 1988.