EFFICIENT SPECTRUM CODING FOR SUPER-WIDEBAND SPEECH AND ITS APPLICATION TO 7/10/15 KHZ BANDWIDTH SCALABLE CODERS

Masahiro Oshikiri, Hiroyuki Ehara and Koji Yoshida

Next-Generation Mobile Communications Development Center, Matsushita Electric Industrial Co., Ltd.,
5-3, Hikarino-oka, Yokosuka, 239-0847, Japan
Email: oshikiri.masahiro@jp.panasonic.com

ABSTRACT

This paper presents an efficient spectrum coding method for super-wideband (beyond 7 kHz, e.g. 10 kHz or 15 kHz bandwidth) speech signals based on a bandwidth expansion technique. By using a 7 kHz bandwidth speech signal, its frequency band over 7 kHz is generated by the expansion technique without violation of harmonics structure of the speech signal. The bandwidth expansion is performed by pitch filtering in a frequency domain. A 7 kHz bandwidth spectrum is used as a pitch filter state, and pitch filtering is performed toward a frequency band over 7 kHz. We adopted this pitch filtering based spectrum coding (PFSC) to our proposing 7/10/15 kHz bandwidth scalable coder. The scalable coder consists of an existing standard wideband coder as a base-layer and two PFSC coders as an enhancement-layer. One PFSC coder encodes a 7-10 kHz band spectrum at 4.4 kbit/s and the other one does a 10-15 kHz band spectrum at 2.2 kbit/s. When the AMR-WB coder at 15.85 kbit/s is used as the base-layer, the total bitrate of the scalable coder is 22.45 kbit/s and the total algorithmic delay is 30 ms. We conducted degradation category rating (DCR) tests for both 10 kHz and 15 kHz bandwidth signals. The results show that the DMOS score of the proposed coder is better than that of 7 kHz bandwidth original signals in both bandwidth clean speech conditions. In addition, when G.722 at 56 kbit/s is used as the base-layer instead of the AMR-WB coder, the DMOS score of this scalable coder is close to that of 7 kHz bandwidth original signals in both bandwidth audio conditions.

1. INTRODUCTION

Although wideband (7 kHz) speech signals provide higher quality than narrowband (3.4 kHz) ones, their quality is not always sufficient for providing high-fidelity conversational communications. For such applications, super-wideband (beyond 7kHz, e.g. 10 kHz or 15 kHz bandwidth) signals are promising by our preliminary subjective evaluation tests [1].

The super-wideband signals can be encoded by an audio coding technique [2] (e.g. Advanced Audio Coding [3]) with satisfactory quality at around 64 kbit/s. However, when an audio coding technique is used in conversational applications, the applications are suffered from a long algorithmic delay because the audio coding technique needs a long frame buffering for frequency analysis. Thus, it is necessary to develop a coder to encode super-wideband signals with a reasonable algorithmic delay that is suitable for conversational applications.

Considering above background, we previously proposed a scalable speech coder at the bitrate of 23.85 kbit/s to encode 10 kHz bandwidth speech signals [1]. Since the algorithmic delay of this coder is about 30 ms, this coder is suitable for conversational applications. The scalable coder consists of the Adaptive Multirate Wideband (AMR-WB) speech coder [4] at 15.85 kbit/s as a base-layer and a transform coding using Modified Discrete Cosine Transform (MDCT) at 8 kbit/s as an enhancement-layer. However, the enhancement-layer needs a higher bitrate to maintain good speech quality when the enhancement-layer is developed to deal with 15 kHz bandwidth signals. The reason for requiring the higher bitrate is that coding efficiency of the enhancement-layer is not so high because it is based on a simple shape-gain vector quantization.

To cope with this problem, we developed an efficient spectrum coding method for the frequency band over 7 kHz. The method is based on pitch filtering in a frequency domain, and it only encodes a lag of a pitch filter and a spectral envelope of the frequency band over 7 kHz. We refer to the method as Pitch Filtering based Spectrum Coding (PFSC) in this paper.

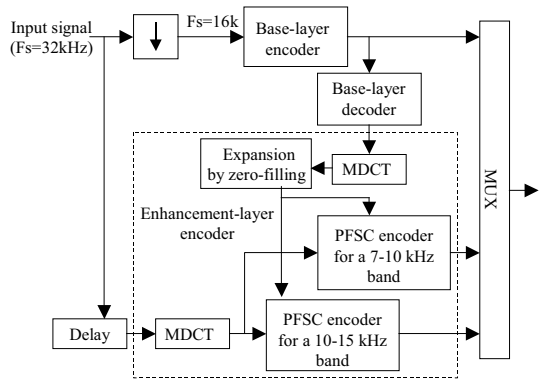
In addition, we propose a 7/10/15 kHz bandwidth scalable coder which uses an existing standard wideband coder as its base-layer and two PFSC coders as its enhancement-layer. One PFSC coder encodes a 7-10 kHz band spectrum at 4.4 kbit/s and the other one does a 10-15 kHz band spectrum at 2.2 kbit/s. When the AMR-WB coder at 15.85 kbit/s is used as the base-layer, the total bitrate of the scalable coder is 22.45 kbit/s and the total algorithmic delay is 30 ms.

The organization of this paper is as follows. In section 2, we first present the overview of the proposed scalable coder. In section 3, the description of the enhancement-layer structure and the PFSC coder are provided. Section 4 presents subjective evaluation test results. Finally, conclusions are given in section 5.

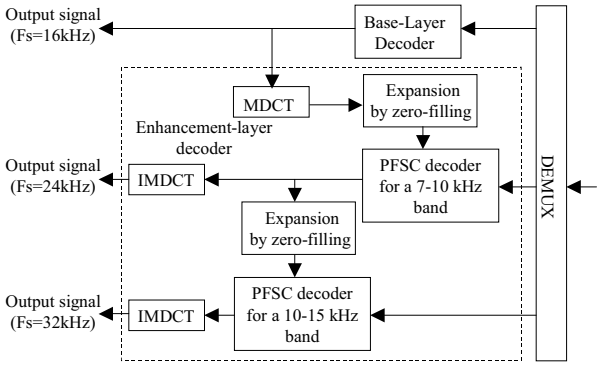
2. OVERVIEW OF THE PROPOSED SCALABLE CODER

Fig.1 shows the structure of the proposed scalable coder. Encoding and decoding procedures are performed as follows. A 32 kHz sampling input is downsampled to a 16 kHz sampling signal and the downsampled signal is encoded in a base-layer encoder. A local decoder generates a decoded signal of the base-layer. Both the decoded signal and the 32 kHz sampling original signal are supplied to an enhancement-layer encoder.

In the enhancement-layer encoder, each signal is transformed to frequency domain through MDCT. The spectrum of the decoded signal is expanded by zero-filling. This zero-filling technique is performed to obtain the spectrum of an upsampled signal. The zero-filled spectrum and the original spectrum are supplied to two PFSC encoders. One PFSC



(a) Encoder structure.



(b) Decoder structure.

Figure1: *Structure of the proposed scalable coder.*

encoder is designed for a 7-10 kHz band, and the other is designed for a 10-15 kHz band. The detailed description of the enhancement-layer encoder is presented in section 3.

In the decoder, firstly, a 16 kHz sampling signal is generated in a base-layer decoder and is supplied to an enhancement-layer decoder. The 16 kHz sampling signal is transformed to frequency domain through MDCT. After the expansion by zero-filling, the zero-filled spectrum is supplied to a PFSC decoder for a 7-10 kHz band, and the PFSC decoder generates a 10 kHz bandwidth spectrum. Similarly, the 10 kHz bandwidth spectrum is expanded by zero-filling, the zero-filled spectrum is supplied to a PFSC decoder for a 10-15 kHz band and the PFSC decoder generates a 15 kHz bandwidth spectrum. Finally, through IMDCT, the 10 kHz and the 15 kHz bandwidth spectra are transformed to a 24 kHz and a 32 kHz sampling output signals, respectively.

3. DESCRIPTION OF THE ENHANCEMENT-LAYER

Fig.2 shows a power spectrum of a Japanese vowel "o". As shown in Fig.2, harmonics structure is observed even in the frequency band over 7 kHz. By our informal listening test, the degradation of the subjective speech quality is perceived when the harmonics structure is violated in such band.

From this viewpoint, we developed PFSC that preserves the harmonics structure in the frequency band over 7 kHz. The concept of the PFSC coder is shown in Fig. 3. The basic idea is that a harmonics-structured spectrum is regarded as a temporal

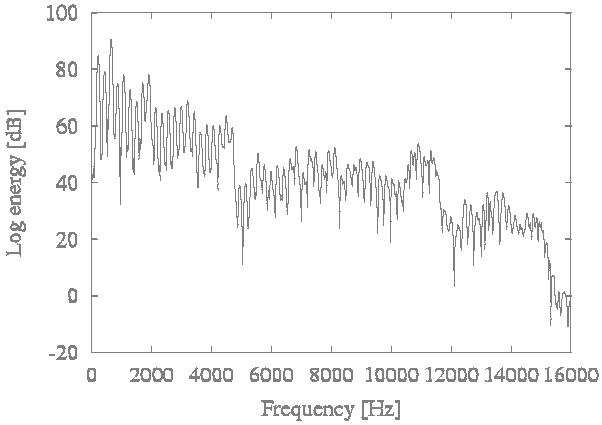


Figure 2: *Power spectrum of vowel "o" uttered by a Japanese female speaker. The spectrum is obtained by DFT analysis with a Hanning window. The analysis frame length is 25ms.*

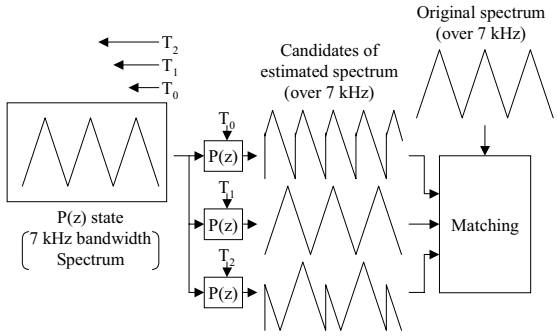


Figure 3: *Concept of the PFSC coder.*

periodic signal, and the spectrum over 7 kHz is estimated based on pitch filtering. The procedure of the PFSC coder can be described as follows: 1) a 7 kHz bandwidth spectrum obtained from the decoded signal in a base-layer is allocated to the state of the pitch filter, 2) the response of the pitch filter is regarded as a candidate of the estimated spectrum over 7 kHz, 3) the squared distortion between the candidate and the original spectrum over 7 kHz is calculated, and the lag of the pitch filter that minimizes the squared distortion is determined, 4) the optimum lag and a spectral envelope over 7 kHz are encoded. This approach can realize an efficient coding of periodic signal in a frequency domain (i.e. harmonics structure) at a low bitrate.

Furthermore, the enhancement-layer has two features to mitigate the increase of algorithmic delay. One is that it adopts a short analysis frame length for MDCT. Although MDCT is a useful method to assure the perfect reconstruction, it causes the algorithmic delay equivalent to half a length of an analysis frame. Since we chose a 10 ms analysis frame for MDCT, the algorithmic delay of this process becomes only 5 ms.

The other feature is that the enhancement-layer exploits the upsampling technique which expands the dimension of the MDCT coefficients of a 16 kHz sampling signal by using zero-filling. The zero-filled MDCT coefficients can be regarded as the MDCT coefficients of an upsampled signal. This means an upsampling process in a time domain is unnecessary for our approach. Generally, the time domain upsampling process needs an anti-aliasing filter, which causes to increase the algorithmic

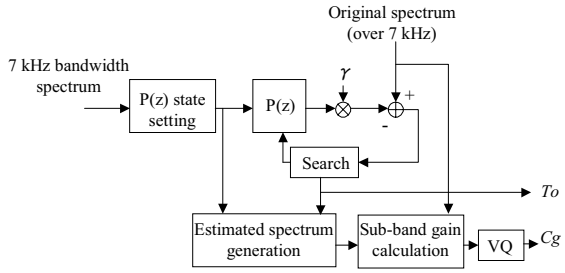


Figure 4: *Structure of the PFSC coder.*

delay. Therefore, our approach prevents the increase of the algorithmic delay caused by the time domain upsampling process.

3.1. Pitch filtering based spectrum coding

In this subsection, the coding procedure of the PFSC coder is presented. Fig.4 shows the structure of the proposed PFSC coder.

An all-pole pitch filter is generally defined as follows:

$$P(z) = \frac{1}{1 - \sum_{-M}^{M} \beta_i z^{-T+i}} \qquad (1)$$

where T and β_i are a pitch lag and pitch filter coefficients, respectively. The number of taps of this filter is 2M+1. A 7 kHz bandwidth spectrum, *S(k)*, is allocated to the state of the pitch filter.

When the input signal, *C(k)*, is given, the response of the pitch filter is described as:

$$Y(T, \beta_i, k) = S(k) = C(k) + \sum_{i=-M}^{M} \beta_i \cdot S(k - T + i), \quad (2)$$

and we regard the response of the pitch filter as a candidate of a spectrum in frequency band over 7 kHz. Here, we used an one-tap pitch filter (M=0, β_0=1.0) and a zero-vector as the input signal *C(k)*.

The squared distortion, E, between the candidate and the spectrum obtained from original signal is defined as:

$$E = \sum_{k=FL}^{FH-1} \left(O(k) - \gamma \cdot Y(T,k)\right)^2 \qquad (3)$$

where *O(k)* is the original spectrum, *Y(T, k)* is the response of the one-tap pitch filter with a pitch lag of T, γ is an optimum gain, and *FL* and *FH* are the minimum and the maximum frequency of the frequency band over 7 kHz, respectively.

By taking $\partial E/\partial \gamma = 0$ and solving for γ, we obtain the optimum pitch lag, T_o, that minimizes the squared error E by searching among all candidates of pitch lag. The optimum pitch lag is encoded and transmitted to the decoder side.

After the determination of the optimum pitch lag, a spectral envelope in frequency band over 7 kHz is calculated and encoded. In order to obtain the accurate spectral envelope, the frequency band over 7 kHz is divided into plural sub-bands, and a gain is calculated in each sub-band. The j-th sub-band gain, *g(j)*, can be defined using the original spectrum *O(k)* and the response of the pitch filter with the decoded pitch lag of $\hat{T}_o$ as:

$$g(j) = \sqrt{\sum_{k=FL(j)}^{FH(j)-1} O(k)^2 \Big/ \sum_{k=FL(j)}^{FH(j)-1} Y(\hat{T}_o, k)^2} \qquad (4)$$

where *FL(j)* and *FH(j)* denote the minimum and the maximum frequency of the j-th sub-band. A gain-vector G consists of the plural sub-band gain *g(j)*, and the gain-vector G is quantized using a vector quantizer. We set the number of sub-bands to 2.

In the decoder, the decoded spectrum in frequency band over 7 kHz, $\hat{S}(k)$, is generated as follows:

$$\hat{S}(k) = \hat{g}(j) \cdot Y(\hat{T}_o, k) \qquad (5)$$

where $Y(\hat{T}_o, k)$ denotes the estimated spectrum generated by the decoded optimum pitch lag $\hat{T}_o$, and $\hat{g}(j)$ is the decoded sub-band gain.

3.2. Bit allocation

The PFSC coders for a 7-10 kHz band and a 10-15 kHz band are respectively performed in each enhancement-layer's frame. The 7-10 kHz band is further divided into two bands and the PFSC coder is performed in each divided band. The bit allocation of the enhancement-layer is shown in Table 1.

Table 1: *Bit allocation of the enhancement-layer. The frame length of the enhancement-layer is 5 ms.*

	7-10kHz band	10-15kHz band
Pitch lag	3 bit x 2	3 bit
Spectral envelope	8 bit x 2	8 bit
Total (bitrate)	22 bit (4.4 kbit/s)	11 bit (2.2 kbit/s)

4. SUBJECTIVE EVALUATION TESTS

We conducted six Degradation Category Rating (DCR) tests [5] to assess the performance of the proposed scalable coder. The processed signals with the bandwidth of 10 kHz and 15 kHz are separately tested in different sets of the DCR tests, and the following three conditions are tested in each set. The conditions are "clean speech", "speech with background music (BGM) (SNR=15dB)" and "audio" conditions. For "clean speech" and "speech with BGM" conditions, we used four different speech utterances, which are spoken by two male and two female Japanese speakers. For "audio" condition, we used four different audio sources, which are two orchestras, a pop music and an instrumental. Sixteen non-expert listeners participated in each DCR test. Test results are shown in Table 2 and Table 3. In these tables, the 95 % confidence intervals are shown in parentheses.

The previously proposed coder in Table 2 denotes the scalable coder developed in our earlier work [1]. This coder consists of the AMR-WB at 15.85 kbit/s as a base-layer and a transform coding using MDCT at 8 kbit/s as an enhancement-layer. The proposed coder #1 is a scalable coder using the AMR-WB coder at 15.85 kbit/s as its base-layer and two PFSC coders as its enhancement-layer. As is described in section 3, one PFSC coder is designed with the bitrate of 4.4 kbit/s for the 7-10 kHz band, and the other is designed with the bitrate of 2.2 kbit/s for the 10-15 kHz band. The total bitrate of this coder is 22.45 kbit/s and the total algorithmic delay is 30 ms. The proposed coder #2 is another scalable coder using G.722 [6] at 56 kbit/s as its base-layer. The enhancement-layer of the proposed coder #2 is identical to that of the proposed coder #1.

Table 2 states the DCR test results for 10 kHz super-wideband signals. The DMOS score of the proposed coder #1 at 20.25 kbit/s is equivalent to that of the previously proposed

coder at 23.85 kbit/s in "clean speech" condition. In "speech with BGM" condition, the DMOS score of the proposed coder #1 is close to that of the previously proposed coder, even though the bitrate of the proposed coder #1 in enhancement-layer is reduced to nearly half. Thus, it is confirmed that the PFSC coder achieves efficient coding of the frequency band over 7 kHz at a low bitrate.

In "clean speech" condition, the proposed coder #1 at 20.25 kbit/s improves the DMOS by about 1.1 compared with the AMR-WB coder at 15.85 kbit/s that is a base-layer of the proposed coder #1. This means the significant improvement is achieved by the enhancement-layer using the PFSC coder. Moreover, the DMOS score of the proposed coder #1 is better than that of all standard 7 kHz band coders (G.722, AMR-WB), and it also exceeds the DMOS score of the original 7 kHz bandwidth signals. In "speech with BGM" condition, the proposed coder #1 improves the DMOS by about 0.9 compared with the AMR-WB coder at 15.85 kbit/s, and the DMOS score is close to that of G.722 at 56 kbit/s. These results show that the proposed coder #1 has good performance to encode the 10 kHz bandwidth speech signals.

However, in "audio" condition, the DMOS score of the proposed coder #1 is poor. By contrast, the proposed coder #2 improves the DMOS by about 0.5 compared with G.722 at 56 kbit/s. It is also confirmed that the DMOS score of the proposed coder #2 is close to that of the 7 kHz bandwidth original signals. These results suggest that the performance of the proposed coder is strongly depends on the performance of its 7 kHz bandwidth coding in base-layer.

Table 3 states the DCR test results for the 15 kHz super-wideband signals. It shows a similar tendency to Table 2, except the DMOS of the proposed coder #1 is a little worse in "speech with BGM" condition. It may suggest that we need to allocate more bits to a 10-15 kHz band in order to improve the quality of the signals including non-speech components.

Table 2: *DCR test results of 10 kHz super-wideband signals. The original 10 kHz bandwidth signals are used as reference signals of the DCR tests.*

	Clean speech	Speech + BGM	Audio
Original (10kHz)	4.55 (0.18)	4.59 (0.15)	4.38 (0.18)
Original (7kHz)	3.34 (0.31)	3.52 (0.26)	3.56 (0.25)
Original (3.6kHz)	2.19 (0.27)	2.11 (0.24)	2.23 (0.26)
G.722 (64kbit/s)	3.22 (0.28)	3.73 (0.24)	3.03 (0.26)
G.722 (56kbit/s)	3.06 (0.28)	3.52 (0.25)	2.94 (0.25)
AMR-WB(23.85kbit/s)	2.83 (0.25)	2.86 (0.23)	1.86 (0.18)
AMR-WB(15.85kbit/s)	2.66 (0.27)	2.44 (0.24)	1.42 (0.15)
Previously proposed coder (23.85kbit/s)	3.66 (0.28)	3.55 (0.26)	1.63 (0.19)
Proposed coder #1 (20.25kbit/s)	3.80 (0.25)	3.31 (0.28)	1.44 (0.17)
Proposed coder #2 (60.4kbit/s)	3.83 (0.25)	3.94 (0.26)	3.48 (0.28)

Table 3: *DCR test results of 15 kHz super-wideband signals. The original 15 kHz bandwidth signals are used as reference signals of the DCR tests.*

	Clean speech	Speech + BGM	Audio
Original (15kHz)	4.67 (0.15)	4.61 (0.17)	4.64 (0.17)
Original (10kHz)	4.16 (0.19)	4.28 (0.20)	4.61 (0.17)
Original (7kHz)	3.42 (0.25)	3.59 (0.24)	4.09 (0.21)
Original (3.6kHz)	1.86 (0.23)	2.06 (0.24)	2.34 (0.26)
G.722 (64kbit/s)	3.28 (0.22)	3.08 (0.26)	3.52 (0.25)
G.722 (56kbit/s)	2.81 (0.23)	3.17 (0.27)	3.48 (0.25)
AMR-WB(23.85kbit/s)	3.03 (0.26)	2.80 (0.22)	2.69 (0.29)
AMR-WB(15.85kbit/s)	2.52 (0.25)	2.20 (0.24)	1.61 (0.17)
Proposed coder #1 (22.45kbit/s)	3.88 (0.29)	2.86 (0.33)	1.55 (0.19)
Proposed coder #2 (62.6kbit/s)	3.73 (0.26)	3.94 (0.24)	3.80 (0.30)

5. CONCLUSIONS

This paper has presented pitch filtering based spectrum coding (PFSC). The PFSC coder estimates the spectrum in frequency band over 7 kHz without violating the harmonics structure of voiced speech signals, and it only encodes lag information for the pitch filtering and the envelope information of spectrum over 7 kHz.

We applied the PFSC coder to our proposing 7/10/15 kHz bandwidth scalable coder. The scalable coder uses an existing standard wideband coder as its base-layer and two PFSC coders as its enhancement-layer. One PFSC coder is designed at 4.4 kbit/s for a 7-10 kHz band and the other is designed at 2.2 kbit/s for a 10-15 kHz band. When the AMR-WB coder at 15.85 kbit/s is used as the base-layer, the total bitrate of the scalable coder is 22.45 kbit/s, and the total algorithmic delay is 30 ms.

Our DCR test results indicate that the DMOS scores of the proposed scalable coder are better than that of the 7 kHz bandwidth original signals in both the 10 kHz and the 15 kHz bandwidth clean speech conditions. Furthermore, when we use G.722 at 56 kbit/s as the base-layer, the performance of the scalable coder is significantly improved and the scalable coder maintains good subjective quality even in audio conditions.

6. REFERENCES

[1] M. Oshikiri, H. Ehara and K. Yoshida., "A Scalable Coder Designed for 10-kHz Bandwidth Speech," Proc. of the IEEE, Speech Coding Workshop, pp.111-113, Oct. 2002.

[2] T. Painter and A. Spanias, "Perceptual Coding of Digital Audio," Proc. of the IEEE, Vol. 88, No.4, pp.451-513, April 2000.

[3] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson and Y. Oikawa, "ISO/IEC MPEG-2 Advanced Audio Coding," J. Audio Eng. Soc., Vol.45, No.10, pp.789-813, Oct. 1997.

[4] B.Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola and K. Jarvinen, "The Adaptive Multirate Wideband Speech Codec (AMR-WB)," IEEE Trans. SAP, Vol. 10, No. 8, pp.620-636, Nov. 2002.

[5] "Methods for Subjective Determination of Transmission Quality," ITU-T Recommendation P.800.

[6] "7 kHz Audio coding within 64 kbit/s," ITU-T Recommendation G.722.