LOW-COMPLEXITY BANDWIDTH EXTENSION IN MDCT DOMAIN FOR LOW-BITRATE SPEECH CODING

Kosuke Tsujino and Kei Kikuiri

Research Laboratories, NTT DOCOMO, Inc.

ABSTRACT

We propose a low-complexity Bandwidth Extension (BWE) method operating in the Modified Discrete Cosine Transform (MDCT) domain to reduce the bitrate of wideband and super-wideband speech codecs. The proposed method generates a high-frequency signal by copying the MDCT spectrum from the low frequency part, and then adjusts tonality to improve the subjective quality of the generated high-frequency signal. In combination with an MDCT-based transform codec, it requires only 64.9% of the computational complexity of MPEG-4 Spectral Band Replication (SBR). It also achieves subjective quality better than SBR for many speech samples.

Index Terms— Transform coding, Speech coding, Speech communication

1. INTRODUCTION

There is growing demand for speech communication with higher quality. To enhance the speech quality of current narrowband (NB) speech codecs [1, 2] with 300 Hz – 3.4 kHz signal bandwidth, several wideband (WB) codecs [3, 4] with 50 Hz – 7 kHz bandwidth and super-wideband (SWB) codecs [5, 6] with around 15 kHz bandwidth have been standardized recently. Although these WB and SWB codecs realize higher speech communication quality than NB speech codecs, they require higher bitrates than NB codecs. Bitrate reduction of WB and SWB codecs is therefore strongly demanded in wireless telecommunication since transmission capacity is severely limited.

Bandwidth Extension (BWE) is a promising technique for reducing the bitrate of WB and SWB codecs. BWE extends the high-frequency (HF) component of speech and audio signals from the low-frequency (LF) signal using only a small amount of auxiliary information. Thus, BWE techniques reduce the bitrates of WB and SWB encoded signals. There are two main categories of BWE methods: time domain BWE [7,8,9] and frequency domain BWE [10,11,12,13].

The time domain methods generate the HF signal by upsampling the LF signal [8,9] or by LPC-based estimation [7]. Both methods double the bandwidth of the LF signal. In contrast, the frequency domain methods can expand LF to HF in a flexible manner. The bandwidth of the generated HF signal can be wider or narrower than that of the LF signal, and frequency-dependent adjustment of the generated signal is also possible. Due to these merits, the frequency domain methods are suitable for SWB codecs, in which the extended bandwidth should be wide. Therefore, we focus on the frequency domain methods in this paper.

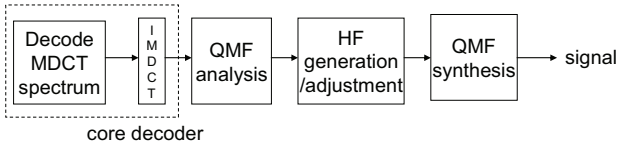

Fig. 1: SBR decoder.

Spectral Band Replication (SBR) [10,11] is the most widely used frequency domain BWE method, and is used in MPEG-4 High-Efficiency AAC (HE-AAC) [14] and AAC Enhanced Low Delay (AAC-ELD) [15]. As shown in Fig.1, a SBR decoder receives the time-domain LF signal from a core decoder, analyzes the LF signal with a complex-valued Quadrature Mirror Filter (QMF) filterbank, generates the HF signal in the QMF domain, and then adjusts the HF signal using the auxiliary information included in the bitstream. The adjustment of the HF signal is done by first adjusting the spectral envelope, and then by controlling the 'tonality', that is the ratio between the tonal and noise-like components of the signal. Due to the sophisticated tonality control made possible by inverse LPC filtering and noise injection to the QMF coefficients, SBR achieves high subjective quality for both speech and audio signals [10] and improves the coding efficiency by more than 30% [11].

While SBR is a powerful BWE technique, its computational complexity is relatively high due to the QMF filterbank. In addition, the use of QMF filterbank also introduces some additional delay. The speech codecs should offer low algorithmic delay and computational complexity, since they are used in realtime communication and are usually run on low-power handheld devices.

The QMF filterbank process can be avoided by performing BWE in the Modified Discrete Cosine Transform (MDCT) domain. If the core codec is MDCT-based, as many of SWB codecs are, BWE can be performed on the MDCT spectrum derived from the core codec, without using a dedicated filterbank to analyze the time domain signal as SBR does. Because this reduces the complexity and delay, BWE in the MDCT domain has been investigated [12, 13]. These methods can generate the HF signal and then adjust its envelope in the MDCT domain. However, they do not control the tonality of the generated HF signal, so their improvement in coding efficiency is limited.

In this paper, a novel BWE method is proposed that adjusts the envelope and tonality in the MDCT domain. Since LPC filtering as used in SBR cannot be applied with MDCT coefficients, the proposed method uses peak compression of the MDCT spectrum instead of LPC filtering.

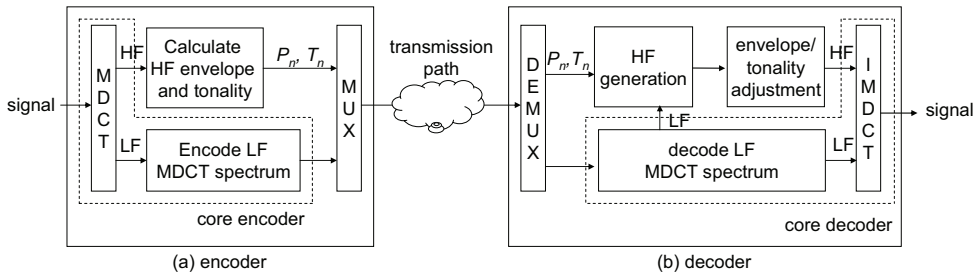

Fig. 2: Block diagram of the proposed method.

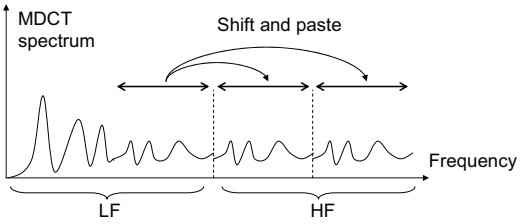

Fig. 3: HF generation.

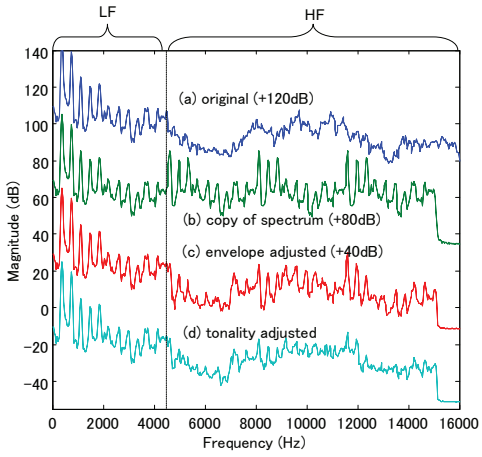

Fig. 4: Example of spectrum.

The rest of this paper is organized as follows. Section 2 describes the proposed BWE method in the MDCT domain. Section 3 assesses the subjective quality of the proposed method and SBR under the assumption that the core codec is lossless. Computational complexity, delay and the amount of the auxiliary information is also discussed in this section. Finally, this paper is concluded in Section 4.

2. BANDWIDTH EXTENSION IN MDCT DOMAIN

This section overviews the proposed method, and then details the tonality adjustment method.

2.1. Overview

The proposed BWE method consists of an encoder generating auxiliary information based on the analysis of the HF component of the input signal, and a decoder that recovers the HF signal from the LF signal and then uses the auxiliary information to adjust the generated HF signal.

The decoder of the proposed method is illustrated in Fig.2 (b). It receives the LF MDCT spectrum from the core decoder, and then generates the HF spectrum by applying the shift and copy technique to the LF spectrum as illustrated in Fig.3. An example of a generated HF spectrum is illustrated in Fig.4 (b). Next, the envelope of the generated HF signal is adjusted to approximate the shape of the original HF spectrum, yielding spectrum Fig.4 (c). For the envelope adjustment at the decoder, the encoder of the proposed method, illustrated in Fig.2 (a), splits the HF part of the MDCT spectrum into N_s subbands, and then calculates power P_n and for each subband.

We can observe that Fig.4 (c) still shows much steeper peaks and dips than the original signal Fig.4 (a), although the macroscopic shape of (c) is similar to (a). This means that the HF component of (c) has excessively high tonality. As SBR does, the proposed method adjusts the tonality of the HF component and thereby obtains spectrum Fig.4 (d), which better approximates the microscopic shape of (a) than (c). For the tonality adjustment, the encoder calculates the tonality parameter T_n for each subband, along with P_n. The tonality adjustment using T_n is detailed in the next subsection.

2.2. Tonality adjustment

The tonality adjustment is performed to the envelope-adjusted spectrum $\mathbf{M}_{H\mathrm{env}}$ according to Eq.1, where $\alpha_n (0 \le \alpha_n \le 1)$ and $\beta_n (0 \le \beta_n)$ are the parameters for controlling tonality in the n th subband, and rand is a pseudo-noise sequence with unit average power. $\mathbf{X}(i)$ denotes the i th element of vector $\mathbf{X}$.

$$\mathbf{M}_{H\mathrm{adj}}(i) = \left(\mathbf{M}_{H\mathrm{env}}(i) \Big/ \left|\mathbf{M}_{H\mathrm{env}}(i)\right|\right) \cdot \left|\mathbf{M}_{H\mathrm{env}}(i)\right|^{\alpha_n} + \beta_n \cdot \mathrm{rand} \quad (1)$$

If $\alpha_n < 1$, the peak amplitude of the MDCT spectrum is compressed, therefore $\mathbf{M}_{H\mathrm{adj}}$ has lower tonality than $\mathbf{M}_{H\mathrm{env}}$. Setting $\beta_n > 0$ also reduces the tonality of $\mathbf{M}_{H\mathrm{adj}}$, since it adds random noise to the spectrum. α_n and β_n are chosen so that $\mathrm{ton}\left(\left|\mathbf{M}_{H\mathrm{adj}}{}^n\right|\right) \approx T_n$, where $\mathbf{M}_{H\mathrm{adj}}{}^n$ is the nth subband of $\mathbf{M}_{H\mathrm{adj}}$ and the tonality measure $\mathrm{ton}(\mathbf{X})$ is defined by Eq.2; note

that we should consider one more thing to enhance the subjective quality of $\mathbf{M}_{H\text{adj}}$.

$$\text{ton}(\mathbf{X}) = \max(|\mathbf{X}|) / \text{mean}(|\mathbf{X}|) \quad (2)$$

When $\mathbf{M}_{H\text{env}}$ is highly tonal, i.e. it includes only a small amount of noise-like component, trying to control the tonality by just peak compression severely distorts the HF signal. On the other hand, excessive noise injection fills the HF signal with random noise, which also degrades the subjective quality. The proposed method reaches a balance between the two adjustment methods since it chooses parameters α_n and β_n so that the noise injection works when $\text{ton}\left(\left|\mathbf{M}_{H\text{env}}{}^{n}\right|\right)$ is too high.

Power function $\left|\mathbf{M}_{H\text{env}}{}^{n}\right|^{\alpha_n}$ in Eq.1 is computationally quite expensive. The complexity is reduced by replacing the power function by linear interpolation of predefined tables.

3. EVALUATION

3.1. Subjective quality

Subjective listening tests using Multi-Stimulus with Hidden Reference and Anchors (MUSHRA) method [16] were performed. The proposed method and SBR were applied to a lossless LF signal, in order to assess the subjective quality of just the BWE part, independent of the performance of the core codec. Codec configurations used in the listening tests are shown in Table 1. Other test conditions are described in Table 2.

The mean scores and the 95% confidence intervals for all samples are shown in Fig.5. Configuration 1 is significantly better than configuration 2. The result shows that the subjective quality of the proposed method is better than that of SBR for the speech samples used.

Table 1: Configurations used in listening tests.

Conf. #	Method	Core bandwidth	Extended bandwidth	Sample rate
1	Proposed	4.5 kHz	15 kHz	32kHz
2	SBR	4.5 kHz	15 kHz	48kHz

Table 2: Test conditions.

Frame length (proposed method)	8 ms
Test materials	4 Japanese speech files (5 to 8 seconds long)
Number of subjects	18 (all have normal hearing)

3.2. Complexity, bitrate of ancillary data, and delay

The complexity of the proposed method and SBR was evaluated by estimating the Weighted Million Operation Per Second (WMOPS) count [17] using the test materials used in the subjective tests. The estimation was intended to target only the BWE part without any assumption about the algorithm of the core codec. We should note that the input to the core codec is usually downsampled by a factor of 2 when SBR is used, as shown in Fig.6. The downsampling in SBR can influence the complexity of encoding the MDCT coefficients. Although the influence depends on the algorithm of the core codec, the bandwidth of the signal encoded by that step is the same, regardless of whether the downsampler is used or not. Therefore, a rough assumption is that the complexity of MDCT coefficient encoding is not influenced by the choice of the downsampler.

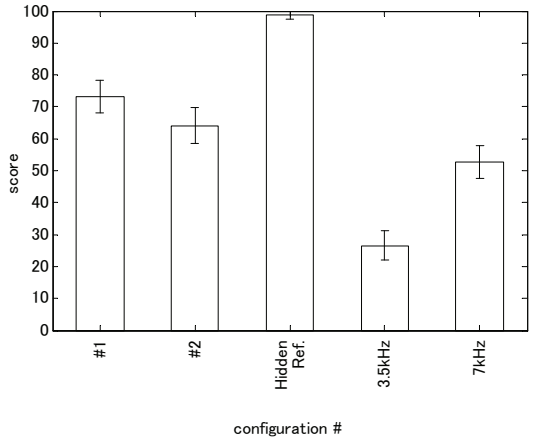

Fig. 5: Results of listening tests.

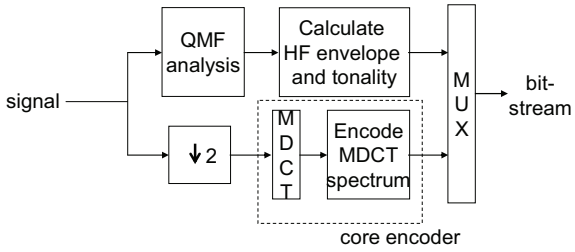

Fig. 6: SBR encoder.

The estimated WMOPS count for the proposed method and SBR is shown in Table 3. The complexity of the encode-decode chain of the proposed method is 64.9% of that of SBR, including MDCT/IMDCT and downsampling steps.

There is a low complexity version of SBR using real-valued QMF filterbank called Low Power SBR (LP-SBR) [18]. Since it reduces the complexity of the SBR decoder by 40% [18], the complexity of the encode-decode chain in that case is 82.6% of that of the normal SBR. The proposed method operates at even lower complexity than LP-SBR.

The average amount of auxiliary information needed by the proposed method and SBR is shown in Table 4. The data rate of the proposed method is almost equivalent to that of SBR.

As for delay, the proposed method introduces no additional algorithmic delay to the MDCT-based core coder, since it requires no look-ahead of future frames. If the same algorithm is used as the core codec, the proposed method achieves shorter algorithmic delay than SBR, which introduces about 20 ms delay due to the

Table 3: WMOPS count of the proposed method and SBR.

		Proposed (32kHz)	SBR (32kHz)
Encoder	Down sample	-	2.01
	MDCT	1.81	0.90
	QMF analysis	-	2.40
	Calculation of auxiliary information	0.85	1.86
	TOTAL	2.66	7.17
decoder	IMDCT	1.70	0.85
	QMF analysis	-	1.03
	HF generation and adjustment	3.86	1.13
	QMF synthesis	-	2.49
	TOTAL	5.56	5.50
encoder+decoder	TOTAL	8.22	12.67

Table 4: Amount of auxiliary information.

proposed	SBR
2.51 kbps	2.71 kbps

QMF filterbank and framing. AAC-ELD introduces a low-delay version of SBR [15], but it still introduces about 12 ms of delay.

From the evaluation results, the proposed method achieves better subjective quality with smaller complexity, almost equivalent amount of auxiliary data, and shorter algorithmic delay than SBR.

4.CONCLUSION

This paper proposed a BWE method operating in the MDCT domain. The proposed method copies MDCT spectra to generate the high-frequency signal, and then adjusts its tonality for improving the subjective quality. When applied to an MDCT-based transform codec, the proposed method achieves, compared to SBR, 35.1% lower complexity, 12 ms shorter algorithmic delay, and better subjective quality for actual speech samples. The proposed method has 17.7% lower complexity than LP-SBR. The proposed method is expected to reduce the bitrate of the SWB speech codec by at least 30 %, since it achieves better subjective quality than SBR [11]. Due to these merits, it is reasonable to adopt the proposed method as a BWE method in low-complexity MDCT-based SWB speech codecs, especially if computational resources are severely restricted.

Future work includes applying the proposed method to an actual codec, and to expand the bandwidth of the proposed method to cover music signals.

ACKNOWLEDMENTS

The authors would like to appreciate our colleague Nobuhiko Naka for constructive advices and discussions on this work.

REFERENCES

[1] 3GPP Specification TS 26.071, “AMR speech Codec; General Description.”

[2] ITU-T Recommendation G.711, “Pulse code modulation (PCM) of voice frequencies.”

[3] 3GPP Specification TS 26.171, “Adaptive Multi-Rate-Wideband (AMR-WB) speech codec; General description.”

[4] ITU-T Recommendation G.711.1, “Wideband embedded extension for G.711 pulse code modulation.”

[5] ITU-T Recommendation G.722.1, “Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss ,” annex C.

[6] ISO/IEC 14496-3:2005/Amd 1:2007, “Low delay AAC profile.”

[7] B. Geiser et al., "Bandwidth extension for hierarchical speech and audio coding in ITU-T Rec. G.729.1," *IEEE Trans. on audio, speech and language processing*, vol. 15, no.8, pp.2496-2509, 2007.

[8] J. Makinen et al., "AMR-WB+: a new audio coding standard for 3rd generation mobile audio services," *Proc. of ICASSP*, vol. 2, pp. 1109-1112, 2005.

[9] P. Jax, “Bandwidth extension for speech”, chapter 6 of *Audio bandwidth extension*, Ed. by E. Larsen and R. M. Aarts, Wiley, 2004.

[10] M. Dietz et al., “Spectral Band Replication, a novel approach in audio coding,” *Proc. of 112th AES Convention*, paper no. 5553, 2002.

[11] P. Ekstrand, "Bandwidth extension of audio signals by spectral band replication," *Proc. of IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*, pp. 53-58, 2002.

[12] M. Oshikiri et al., "A 7/10/15 kHz bandwidth scalable coder using pitch filtering based spectrum coding," *Proc. of Spring Meeting of the Acoustical Society of Japan*, pp. 327-328, 2004 (in Japanese).

[13] A. Laaksonen, "Bandwidth extension in high-quality audio coding," *Master's thesis, Helsinki university of technology*, 2005.

[14] ISO/IEC 14496-3:2001/Amd 1:2004, “Bandwidth extension.”

[15] M. Schnell et al., "Enhanced MPEG-4 low delay AAC – low bitrate high quality communication," *Proc. of AES 122nd convention*, paper number 6998, 2007.

[16] ITU-R Recommendation BS.1534.1, “Method for the subjective assessment of intermediate quality level of coding systems.”

[17] 3GPP TSG-SA4 Meeting #26 Tdoc S4-030303, “Complexity Assessment of PSS and MMS Audio Codec Candidates.”.

[18] K. S. Chong et al., "Low power spectral band replication technology for the MPEG-4 audio standard," *Proc. of International Conference on Information, Communications and Signal Processing (ICICS)*, pp.1408-1412, 2003.