IEICE TRANS. INF. & SYST., VOL.E92–D, NO.12 DECEMBER 2009

LETTER

Bandwidth-Scalable Stereo Audio Coding Based on a Layered Structure

Young Han LEE[†], **Deok Su KIM**[†], ***Nonmembers***, **Hong Kook KIM**[†a)], ***Member***, **Jongmo SUNG**[††], **Mi Suk LEE**[††], ***and*** **Hyun Joo BAE**[††], ***Nonmembers***

SUMMARY In this paper, we propose a bandwidth-scalable stereo audio coding method based on a layered structure. The proposed stereo coding method encodes super-wideband (SWB) stereo signals and is able to decode either wideband (WB) stereo signals or SWB stereo signals, depending on the network congestion. The performance of the proposed stereo coding method is then compared with that of a conventional stereo coding method that separately decodes WB or SWB stereo signals, in terms of subjective quality, algorithmic delay, and computational complexity. Experimental results show that when stereo audio signals sampled at a rate of 32 kHz are compressed to 64 kbit/s, the proposed method provides significantly better audio quality with a 64-sample shorter algorithmic delay, and comparable computational complexity.

key words: *stereo audio coding, bandwidth-scalable audio coding, parametric stereo, super-wideband stereo audio coding, layered structure*

1. Introduction

In early speech communication services, narrowband coders with a bandwidth around 3.4 kHz were commonly used, since the available network bandwidth was quite limited. These services could provide sufficient quality for comprehension, but it was generally agreed that they did not satisfy users' increasing expectations for higher sound quality. Due to advances in network technologies, however, this transmission bandwidth has recently been increased [1]. Thus, a great deal of research has been focused on extending the bandwidth of speech signals from narrowband to audio band, or extending the number of channels from mono to stereo or more [2]. In particular, bandwidth-scalable coders have been researched because they enable various bit-rates, quality, bandwidths, and channels to be provided.

Figure 1 shows a typical structure of a bandwidth-scalable coder for an IP network [3]. As shown in the figure, the bandwidth-scalable coder is composed of a baseline coder and an enhancement layer that encodes and decodes audio signals corresponding to low and high frequency bands, respectively. The two bitstreams from the baseline and the enhancement layer encoders are merged into a single bitstream which is then transmitted to a receiver. At this time, the network cannot deliver the whole bitstream due to a network condition. Hence, it is sufficient for the scalable decoder to reconstruct low-frequency signals. From the figure, it can be seen that the bitstream from either the baseline encoder or the enhancement layer encoder can be separately decoded. Such a structure makes it suitable for use in IP networks that do not guarantee high bit-rates for all times. Thus, the bandwidth-scalable coder can provide good scalability according to the available bit-rate.

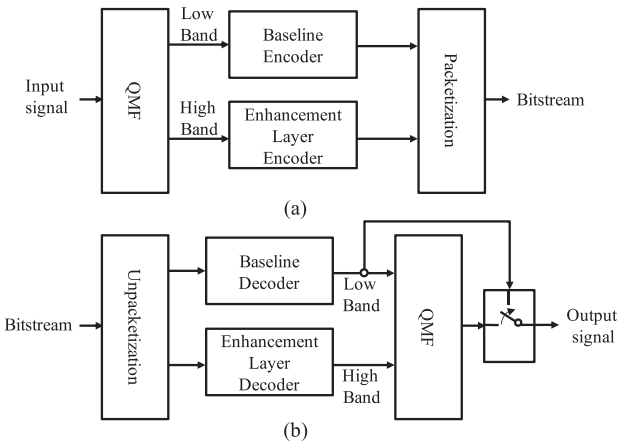

Fig. 1 Typical structures of (a) a bandwidth-scalable encoder and (b) a bandwidth-scalable decoder [3].

In addition to bandwidth scalability, there is a need to increase the number of channels to provide better audio quality, even when using a directional sound source, as is applicable in a voice conference system. Recently, there have been several stereo coding methods developed, including mid/side (M/S) coding, intensity stereo (IS) coding [4], and parametric stereo (PS) coding [5]. Among them, the PS method, standardized as a stereo coding technique for MPEG-4, has been demonstrated to provide high quality stereo signals at very low bit-rates.

However, the PS method is not suitable for bandwidth-scalable coders due to the fact that it has been developed for coding stereo signals with a single bandwidth. Because the PS method only extracts stereo coding parameters from the full bandwidth stereo signal, a down-sampler is required to adjust the bandwidth when a stereo signal, whose bandwidth is narrower than that of an input stereo signal, is insufficient for decoding. Moreover, this type of down-sampling process is apt to generate additional algorithmic delays and cause

Manuscript received March 27, 2009.
Manuscript revised July 2, 2009.

[†]The authors are with the Dept. of Information and Communications, Gwangju Institute of Science and Technology, 1 Oryong-dong, Buk-gu, Gwangju 500–712, Korea.
[††]The authors are with ETRI, Gajeong-dong, Yuseong-gu, Daejeon 305–350, Korea.
a) E-mail: hongkook@gist.ac.kr

DOI: 10.1587/transinf.E92.D.2540

distortion, ultimately resulting in quality degradation of the decoded audio signal.

In this paper, we propose a bandwidth-scalable parametric stereo audio coding method based on a layered structure to overcome such problems. The proposed method is able to reconstruct stereo signals having a lower bandwidth than the input signals—the same bit-rate as the conventional PS method. Furthermore, it will be shown that neither an additional algorithmic delay nor a significant increase in computational complexity is required.

The remainder of this paper is organized as follows. Sections 2 and 3 describe the algorithms for the conventional PS coding method and the proposed PS method, respectively, with the advantages of the proposed method over the conventional method being discussed. In Sect. 4, the performance of the proposed method is compared to that of the conventional PS method in terms of audio quality, algorithmic delay, and computational complexity. Finally, we conclude this paper in Sect. 5.

2. Parametric Stereo

The conventional PS method, standardized in MPEG-4 [5], extracts spatial parameters such as the inter-channel intensity difference (IID), inter-channel coherence (ICC), and inter-channel phase difference (IPD). Figure 2 (a) shows a block diagram of a conventional PS encoder. In the encoder, the input stereo signal is first transformed into 64 subband complex signals by means of a complex quadrature mirror filterbank (QMF) analysis†. These signals are then merged into 10 or 20 parameter bands, and the IID and ICC are extracted for each parameter band. It is noticed here that we only use IID and ICC among the three spatial parameters. This is because compared with IID and ICC, the quality improvement by IPD is the lowest whereas the bit consumption for encoding the IPD parameter is similar to those of other

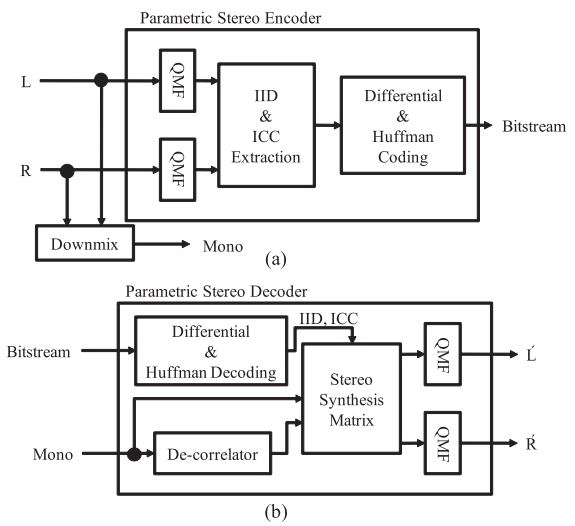

Fig. 2 Structure of a conventional parametric stereo coding method: (a) encoder and (b) decoder.

parameters [6]. For example, standard audio coders including enhanced aacPlus [7] do not use the IPD parameter.

In this paper, the IID and ICC are defined as

$$iid(b) = 10\log_{10}\left(\frac{e_l(b)}{e_r(b)}\right) \tag{1}$$

and

$$icc(b) = \begin{cases} \min\left(\sqrt{\frac{\mathrm{Re}(e_R(b))}{2\sqrt{e_l(b)e_r(b)}}}, \frac{1}{\sqrt{2}}\right), & \text{if } b < 11 \\ \min\left(\sqrt{\frac{|e_R(b)|}{2\sqrt{e_l(b)e_r(b)}}}, \frac{1}{\sqrt{2}}\right), & \text{if } b \geq 11 \end{cases} \tag{2}$$

where b is the index of the parameter band from 1 to 20, $e_R(b)$ is the correlation of the left and right channels, and $e_l(b)$ and $e_r(b)$ are the energies of the left and right channels in the parameter band b, respectively. The extracted parameters are then sequentially compressed using differential and Huffman coding. On one hand, a mono signal is generated by down-mixing the left and right channels using the equation

$$m(k,n) = \frac{l(k,n) + r(k,n)}{2} \cdot \gamma(k,n) \tag{3}$$

where $\gamma(k,n)$ is a stereo scale factor defined by

$$\gamma(k,n) = \min\left(2, \sqrt{\frac{|l(k,n)|^2 + |r(k,n)|^2}{0.5 \cdot |l(k,n) + r(k,n)|^2}}\right) \tag{4}$$

where $l(k,n)$ and $r(k,n)$ are the left and right channel signals of the k-th subband and the n-th sub-sample, respectively. Finally, the mono signal, $m(k,n)$, is compressed using a mono encoder; typically, advanced audio coding (AAC) [8] is used.

A block diagram of the conventional PS decoder is depicted in Fig. 2 (b). In the decoder, the PS parameters are first reconstructed using Huffman and differential decoding, and the reconstructed parameters are then used to make a synthesis matrix. In the decoder, the decorrelator, composed of a series of all-pass filters and time delay elements [5], is applied to the reconstructed mono signal, resulting in another mono signal that is assumed to be uncorrelated to the reconstructed mono signal. These two mono signals are subsequently combined using the synthesis matrix obtained from the PS parameters, as described above. Finally, a complex QMF synthesis is performed to reconstruct the stereo signals.

3. Layered Parametric Stereo

In this paper, we propose a bandwidth-scalable PS coding method based on a layered structure. In the proposed PS encoder shown in Fig. 3 (a), the stereo signal is first split into M different subband signals. Here, M is determined by the

†As shown in Eq. (2), ICC reflects the phase change between left and right channels. To this end, a complex QMF analysis is used here for the computation of $e_R(b)$.

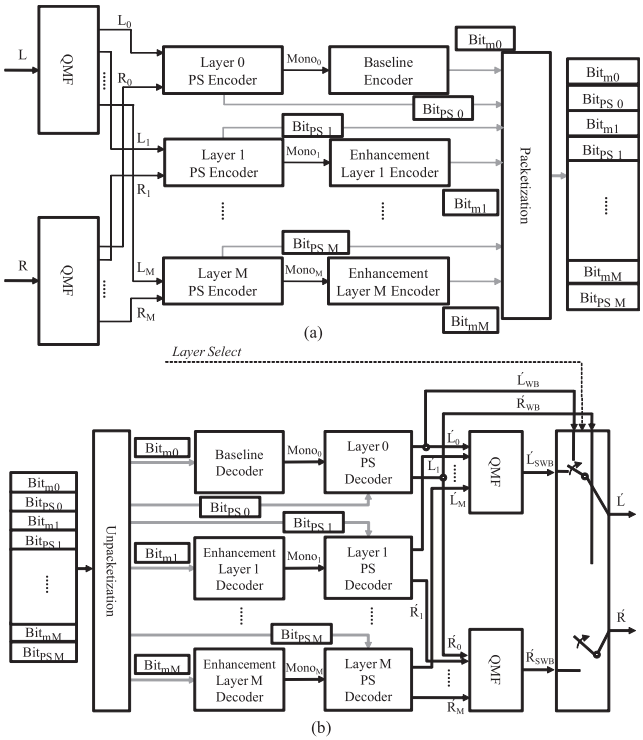

Fig. 3 Structure of the proposed parametric stereo coding method: (a) encoder and (b) decoder.

number of granules in the bandwidth scalability that should be supported by the number of filterbanks in the complex QMF analysis. Thus, stereo signals are progressively reconstructed from the first to the M-th layer, thereby becoming a layered structure, where each subband is referred to as a layer. Note that the baseline encoder corresponds to the zero-th layer in this layered structure.

Next, the PS encoding method described in Sect. 2 is applied to each layer, and stereo parameters obtained from each layer are individually quantized. In addition, a mono signal is generated from the down-mixing process defined by Eq. (3), such that the mono signal for the first layer is encoded by the baseline encoder and that each higher layer is encoded by one of the enhancement layer encoders. For example, M is set to 2 for the super-wideband (SWB) extension of G.729.1, and the baseline encoder for the SWB extension of G.729.1 is the G.729.1 encoder [2]. Using this encoder, the bitstream for the baseline encoder $\mathrm{Bit_{m1}}$, M bitstreams for the PS parameters $\mathrm{Bit_{PSi}}$ (for i = 1,$\cdots$, M), and M bitstreams for the enhancement layer encoders $\mathrm{Bit_{mi}}$ (for i = 2,$\cdots$, M) are combined into a single bitstream and then transmitted to the decoder. The bit-rate of the PS method in the layered structure is set to be the same as the conventional method. That is, the number of total parameter bands is not changed.

Figure 3 (b) displays a block diagram of the proposed PS decoder. In normal mode, we can decode all the bitstreams to obtain stereo output signals at full bandwidth. However, some exceptional cases may exist when not all the bitstream can be delivered, for example, due to network congestion, and thus only a part of the bitstream is available for decoding. In this case, if only the bitstream from the baseline and the first layer is available, the baseline decoder is applied to generate a mono signal and the first layer PS decoder is then used to extend this mono signal into a stereo signal. However, the conventional PS decoder must reconstruct stereo signals using the full bandwidth, and then uses a down-sampler to adjust the bandwidth because the bandwidth of the reconstructed stereo signal is narrower than that of the input stereo signal. In general, such down-sampling is realized using an FIR filter having a linear phase. This tends to incur an additional delay, as long as half the number of taps of the FIR filter, thereby causing the total codec delay to lengthen.

4. Performance Evaluation

In order to evaluate the performance of the proposed PS coding method, G.729.1 [9] and MDCT-based coders were used as the baseline coder and the enhancement layers, respectively, where the overall algorithmic delay by combining these coders was 49.5275 ms [10]. In addition, the sampling rates of both the input and output signals were 16 and 32 kHz for wideband (WB) and super-wideband (SWB) signals, respectively. In particular, audio signals were compressed at a rate of 64 kbit/s by using the proposed PS method as well as the conventional PS method.

A multiple stimuli with hidden reference and anchor (MUSHRA) test [11] and an AB preference test were conducted for the quality evaluation of the proposed PS coding method, and the algorithmic delay and computational complexity of the proposed PS coding method were then compared with those of the conventional PS coding method.

4.1 Quality Test

To prepare the MUSHRA test, we chose four different SWB stereo files, representing the music genres such as rock, pop, vocal, and classical. Each file was first filtered using two different low-pass filters with cut-off frequencies of 7 and 14 kHz; these filtered files were then used for anchors in the MUSHRA test. Next, each file was processed using the conventional PS coding method and the proposed PS coding method, which were denoted as fullband PS and layered PS, respectively. In addition to SWB stereo decoding, we performed WB stereo decoding using both the conventional and proposed PS coding methods under the assumption that partial bitstreams of the WB stereo signal were received, which was also referred to as SWB-WB decoding. Next, to compare the quality of the WB stereo output signal, two additional low-pass-filtered stereo signals with cut-off frequencies of 3 and 7 kHz were prepared as anchors. Finally, all the material was presented to six listeners with no auditory disorders.

Figure 4 shows the MUSHRA test results, where each column corresponds to the opinion score averaged over six

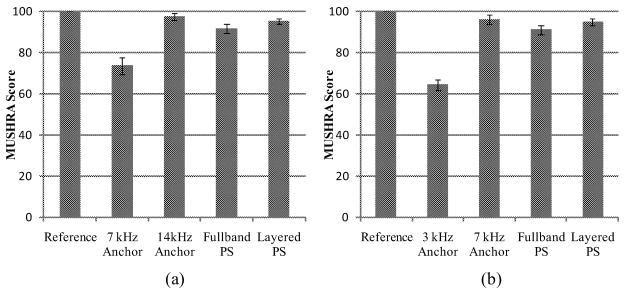

Fig. 4 MUSHRA test results: (a) SWB music signal and (b) SWB-WB music signal.

Table 1 A-B preference test result between the fullband PS and the proposed layered PS.

Genre	Preference Score(%)		
	Fullband PS	No difference	Layered PS
Classical	8.33	8.33	83.33
Pop	8.33	0.00	91.67
Rock	16.67	0.00	83.33
Vocal	16.67	0.00	83.33
Average	12.50	2.08	85.42

listeners and all audio files, and the vertical line on the top of each bar displays the 95% confidence interval (i.e., the statistical significance). It was shown from Figs. 4 (a) and 4 (b) that the proposed layered PS coding method provided better quality than the fullband PS coding method for both SWB decoding and SWB-WB decoding. This better quality was due to the frequency resolution as, especially for the low-frequency band, the layered PS had a higher resolution than the fullband PS.

Next, we evaluated the performance of the proposed layered PS coding method when the bit-rate switching occurred according to network conditions. In this experiment, we simulated such bit-rate switching by using a bitstream truncation tool defined in ITU-T Recommendation G.191 [12]. Thus, the fullband PS decoder and the proposed layered PS decoder were applied to reconstruct audio from the bitstreams, which were generated by setting bit-rate switching as 5 Hz. To measure the performance, an AB preference test was conducted, where the test conditions such as the number of participants and the number of audio files were identical to those of the MUSHRA test described above. Table 1 shows the preference test result. It could be concluded from the table that the proposed layered PS coding method provided significantly better audio quality than the fullband PS coding method in the simulated network condition.

4.2 Algorithmic Delay

Table 2 compares the algorithmic delays required for both fullband and layered PS. It was shown from the table that the algorithmic delay of the fullband PS was 2 ms due to the QMF analysis, while the algorithmic delay of the layered PS was doubled because parametric stereo was applied to the down-sampled signal. When the fullband PS decoded the WB stereo signals using stereo parameters extracted from the SWB stereo (SWB-WB) signals, both an up-sampler and a down-sampler were required for reconstructing the stereo signal, as explained in Sect. 3. Then, assuming that two 64-tap linear-phase FIR filters with a delay of 4 ms were used, the fullband PS needed a delay of 2 ms. Therefore, the overall algorithmic delay of the SWB-WB signals processed by the fullband PS was 6 ms. Conversely, the overall algorithmic delay of the SWB-WB signals for the layered PS was only 4 ms because the layered PS could reconstruct WB signals without requiring any sampling conversion. As a result, the total algorithmic delays of an audio codec using fullband and layered PS for SWB-WB signals were 55.5275 ms and 53.5275 ms, respectively, since the algorithmic delay of the baseline and enhancement encoders was 49.5275 ms, as mentioned earlier in this section.

Table 2 Comparison of algorithmic delay when audio signals are sampled at a rate of 32 kHz.

Input	Output	Fullband PS	Layered PS
SWB	SWB	2 ms	4 ms
	WB	6 ms	4 ms

Table 3 Comparison of computational complexity measured in WMOPS.

Input	Output	Tool	Fullband PS	Layered PS
SWB	SWB	Additional QMF	0.00	1.45
		SWB_PS_Encoder	6.60	6.66
		SWB_PS_Decoder	12.49	12.50
		Total	19.09	20.61
	WB	Additional QMF	0.00	1.03
		SWB_PS_Encoder	6.60	6.66
		WB_PS_Decoder	12.50	6.39
		Up-sampler	1.45	0.00
		Down-sampler	1.71	0.00
		Total	22.47	14.08

4.3 Computational Complexity

Table 3 shows the computational complexity comparison between the fullband PS and layered PS. For SWB decoding, the computational complexity of the layered PS was slightly higher than that of the fullband PS because the additional QMFs required to split the fullband into two layers in the encoder, and vice versa in the decoder. On the other hand, the computational complexity of the layered PS for WB stereo signal reconstruction was much lower than that of the fullband PS because the layered PS does not require a sampling rate conversion such as an up-sampler or a down-sampler. Moreover, it was sufficient for the proposed layered PS to only reconstruct the lower band audio signals, whereas the fullband PS should generate audio signals with the full frequency band prior to sampling rate conversion. Consequently, compared to the fullband PS, the layered PS could achieve the complexity reduction of 8.39 weighted million operations per second (WMOPS) [12].

5. Conclusion

In this paper, we proposed a bandwidth-scalable stereo audio coding method based on a layered structure. The proposed stereo coding (PS) method could encode full bandwidth stereo signals but was only able to reconstruct either the whole bandwidth stereo signal or lower bandwidth stereo signal to be decoded, depending on network congestion. To measure the performance, we implemented an audio codec, which operated at 64 kbit/s, by using the proposed PS method based on a two-layer model and then compared its performance with an audio codec by using the conventional PS method. It was shown from the experiments that the audio quality of the proposed PS method was significantly better than that of the conventional PS method with lower computational complexity and shorter algorithmic delay.

Acknowledgements

This work was supported in part by the IT R&D program of the MKE and KEIT in Korea [2008-S-011-01], Study on Acoustic Convergence Codec and its Control Technology for FMC, and by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-314-D00245).

References

[1] C. Lamblin, "Recent audio/speech coding developments in ITU-T and future trends," Proc. EUSIPCO 2008, planery lecture, 2008.
[2] ITU-T TD 298 R1, Report of Q23/16 Rapporteurs Meeting, 2008.
[3] A. Kataoka, S. Kurihara, S. Sasaki, and S. Hayashi, "A 16-kbit/s wideband speech codec scalable with G.729," Proc. Eurospeech, vol.3, pp.1491–1494, Sept. 1997.
[4] J. Breebaart and C. Faller, Spatial Audio Processing: MPEG Surround and Other Applications, Wiley, 2007.
[5] ISO/IEC 14496-3, Information technology - Coding of audio - visual objects - Part3: Audio, Dec. 2005.
[6] J. Lapierre and R. Lefebvre, "On improving parametric stereo audio coding," Proc. 120th AES Convention, Paris, France, Preprint 6804, May 2006.
[7] 3GPP TS 26.401, Enhanced aacPlus general audio codec; General Description, Dec. 2008.
[8] ISO/IEC 13818-7, Information technology – Generic coding of moving pictures and associated audio information – Part 7: Advanced Audio Coding (AAC), Dec. 1997.
[9] ITU-T Recommendation G.729.1, G.729 based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729, May 2006.
[10] B. Geiser, H. Krüger, H.W. Löllmann, P. Vary, D. Zhang, H. Wan, H.T. Li, and L.B. Zhang, "Candidate proposal for ITU-T super-wideband speech and audio coding," Proc. ICASSP, pp.4121–4124, April 2009.
[11] ITU-R Recommendation BS.1534-1, Method for the subjective assessment of intermediate quality levels of coding system, Jan. 2003.
[12] ITU-T Recommendation G.191, Software tools for speech and audio coding standardization, Sept. 2006.