Variable Subband Analysis for High Quality Spatial Audio Object Coding

Kyungryeol Koo[1], Kwangki Kim[1], Jeongil Seo[2], Kyeongok Kang[2], Minsoo Hahn[1]
[1]Speech and Audio Information Laboratory
Information and Communications University
[2]Broading Media Research Group
Electronics and Telecommunications Research Institute
[1]{dnsaram, kkkim, mshahn} @ icu.ac.kr, {seoji, kokang} @ etri.re.kr

Abstract — **Spatial Audio Object Coding (SAOC) handles a number of audio objects to provide a user with active audio services. It represents all objects as a stereo downmixed signal with some side information, and the bitrate can be significantly reduced compared to that of the conventional audio coders. In spite of the advantage of bitrate reduction, the SAOC has the severe problem such as the degradation of sound quality. To solve the problem while minimizing increase of the bitrate, variable subband analysis and the revised SAOC structure called two-step structure were proposed in this paper. Through a subjective listening test, it was confirmed that the proposed method improves sound quality by a small bitrate increase.**

Keywords — **Spatial Audio Object Coding, Variable subband analysis, Two-step structure**

1. Introduction

Audio coding standards such as Advanced Audio Coding (AAC) [1] and MPEG Surround [2, 3] have been developed by an MPEG audio subgroup. AAC is the high quality audio coding technique for mono or stereo channel signals while MPEG Surround is suitable for multi-channel audio coding. Like these techniques, the conventional audio coders generally have focused on the channel-based audio signals. Therefore, if it needs to transmit various audio objects including mono, stereo or multi-channel signals, the high bitrate problem cannot be avoidable when the conventional audio coding schemes are used. In order to cope with this problem, the alternative audio coding scheme called Spatial Audio Object Coding (SAOC) was proposed in [4].

The SAOC simultaneously deals with a number of audio objects while the conventional audio coding techniques compress each object, separately. It represents one or more audio objects as one stereo downmixed signal and side information. Hence, the bitrate can be significantly reduced compared to that of the conventional audio coders. Moreover, the SAOC can flexibly control the audio objects according to the user interaction as it adopts the mixer/renderer which provides the functionality such as panning, attenuation, and suppression.

This SAOC system has the severe degradation of sound quality because of a downmix process based on the limited number of subbands. Accordingly, many subbands are essentially needed to make up for that problem. However, if the number of subbands is carelessly increased, side information is also significantly increased. In this paper, we propose the variable subband analysis which can provide finer subband structures while minimizing increase of bitrate.

2. Spatial Audio Object Coding

2.1 Overview

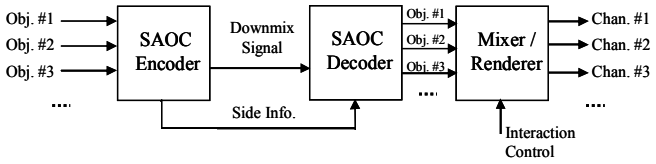

Figure 1. Block diagram of the SAOC

As shown in Figure 1, the overall structure of the SAOC is similar to the MPEG Surround. One or more input audio objects are represented as a stereo downmixed signal and some side information in the SAOC encoder. They are transmitted to the SAOC decoder and each object is reconstructed using transmitted downmixed signal and side information. After the decoding procedure, recovered audio objects are fed to the mixer/renderer component and desired sound scenes or mixing signals are produced according to the user preference.

2.2 SAOC encoder

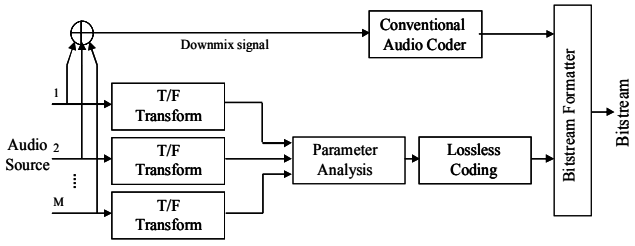

Figure 2. Block diagram of the SAOC encoder

The SAOC encoder consists of two parts. The first part is the process for downmixing input audio objects and the other one is the process for extracting spatial parameters. Figure 2 depicts a structure of the SAOC encoder. As shown in the figure, input signals are transformed to the frequency domain by the Fast Fourier Transform (FFT) and parameters are calculated based on the spatial hearing. The downmixed signal

is encoded by a conventional audio coder such as the AAC while side information is compressed by a lossless coding technique such as the Huffman coding.

Input objects are downmixed to one stereo object through several one-to-two (OTT) modules and two-to-three (TTT) modules. Here, the inter channel level difference (ICLD) [5] is mainly used as the spatial parameter. It is defined as the spectral power ratio between two or three input signals. The ICLDs of the OTT and the TTT module are estimated using following equations, respectively.

$$ICLD(i) = 10 \times log_{10} \frac{P_l(i)}{P_l(i) + P_r(i)} \quad (1)$$

$$ICLD_1(i) = 10 \times log_{10} \frac{P_l(i)}{P_c(i)/\sqrt{2} + P_l(i)} \quad (2)$$

$$ICLD_2(i) = 10 \times log_{10} \frac{P_r(i)}{P_c(i)/\sqrt{2} + P_r(i)} \quad (3)$$

$P_c(i)$, $P_l(i)$ and $P_r(i)$ are the spectral powers of input signals in the *i*th subband. Equation (1) indicates the ICLD of the OTT module, and Equation (2) and (3) show two ICLDs which are calculated between three input signals in the TTT module.

2.3 SAOC decoder

As shown in Figure 3, the SAOC decoder has the reverse structure of the SAOC encoder. A downmixed signal is decoded by a conventional audio decoder and spatial parameters are recovered by a lossless decoder. The downmixed signal is transformed to the frequency domain by FFT, and the spectra of input objects are reconstructed by using the spatial parameters. Finally, output audio objects are generated after the reconstructed spectra are transformed to the time domain.

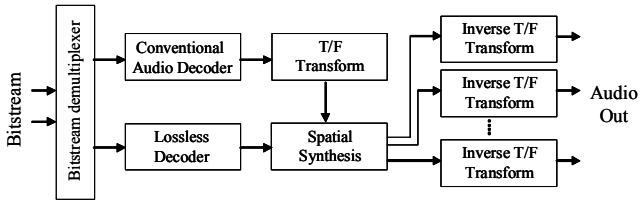

Figure 3. Block diagram of the SAOC decoder

2.4 Mixer / Renderer

To provide user interaction audio services, the SAOC adopts the mixer and the renderer after decoding objects. They have the functionalities such as attenuation and suppression of the sound level, or the sound scene control. Namely, a user can change position of the virtual source or loudness of each object in the soundtrack as shown in Figure 4. Each output channel (stereo or multi-channel) makes a phantom image by panning objects so that a user can be conscious of the reformed sound scene.

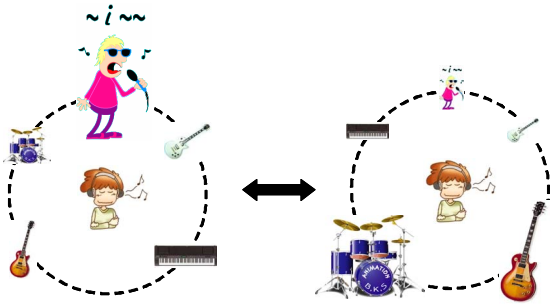

Figure 4. Sound scene control by the user's preference

3. Proposed Method

3.1 Motivation

Because of downmixing various kinds of signals to the one stereo signal in the SAOC encoder, their characteristics are mixed up each other. Hence, it is hard to reconstruct original input objects perfectly. The main reason is that spatial parameters used in the SAOC encoder are estimated based on a limited number of subbands, not every frequency bin. If each subband has low resolution, sound quality degradation is inevitable. Therefore, a large number of subbands are essentially needed to calculate more accurate parameters

On the other hand, sound quality degradation is exposed seriously when we apply sound suppression mode or solo representation mode of the target object, while it does not affect intensively in the case of sound scene panning. So, we focus on the case of sound suppression which causes severe audio quality problem.

3.2 Two-step structure

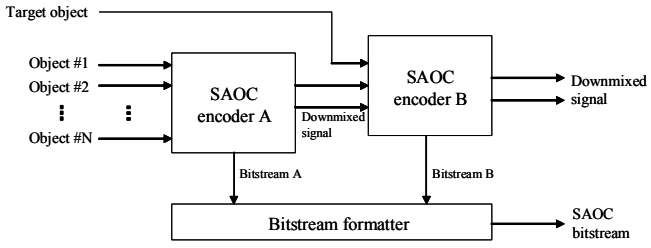

Figure 5. Block diagram of two-step structure of SAOC encoder

Figure 5 shows a block diagram of the two-step structure of the SAOC encoder. SAOC encoder A represents the conventional SAOC encoder using 28 subbands structure while SAOC encoder B represents the modified SAOC encoder which has the variable subband structure. Input audio objects except the target object are downmixed by the conventional SAOC encoder and the target object is handled by the modified SAOC encoder.

For example, in case of soundtrack objects, the target object can be a vocal sound and other objects can be background musical instruments such as a guitar, a bass, a drum, and so on. They are separately encoded by two SAOC encoders, and the

bitstream formatter makes the final bitstream of the SAOC encoder.

3.3 Variable subband analysis

To make up for sound quality degradation of the conventional SAOC system, increase of the number of subbands is needed to calculate more accurate spatial parameters. However, if the number of bands is carelessly increased, side information becomes too much in proportional to increase of the subbands. We propose the variable subband analysis which efficiently provides the high resolution subband structure adjusting the number of subbands flexibly. To minimize increase of bitrate, this method is only applied when the target object is encoded in the two-step structure.

The variable subband construction method is as follows. A basic frame of the variable subband is equal to the 28 subbands structure used in the conventional SAOC. Then, if the spectral variation between two signals in the particular subband is larger than pre-defined threshold, the subband is divided more finely.

$$avrg_b = \frac{\sum_{k=A(b)}^{A(b+1)-1} \log P_1(k) - \log P_2(k)}{A(b+1) - A(b)} \quad (4)$$

$$var_b = \frac{\sum_{n=A(b)}^{A(b+1)-1} \left(\log P_1(k) - \log P_2(k) - avrg_b\right)^2}{A(b+1) - A(b)} \quad (5)$$

$$S(b) = 1 \quad if\ var_b > \theta,\quad 8 \le b < 18$$
$$= 0 \quad otherwise \quad (6)$$

A(b) is a partition boundary of the *b*th subband in the frequency domain. $avrg_b$ means an average power difference between two signals in the *b*th subband, and var_b represents the variance of power difference. Then, if var_b is larger than the pre-defined threshold θ, *S(b)* which means whether the *b*th subband is need to be divided more finely, is set to 1. Here, if one subband is determined to be divided into smaller subbands, their resolution should be about 86 Hz according to the 2 ERB (Equivalent Rectangular Bandwidth). Also, we only consider from 8th subband to 18th subband (from 688 Hz to 6.2 kHz) as refer to the psychoacoustics.

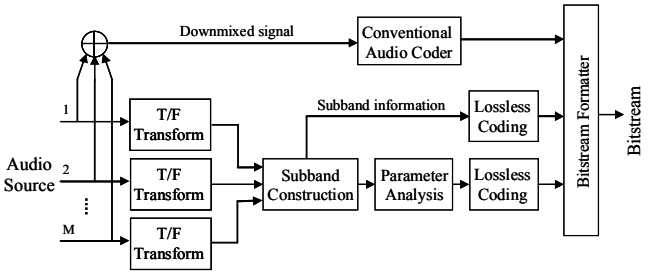

Figure 6. Block diagram of the revised SAOC encoder.

The proposed SAOC encoder with the variable subband structure is depicted in Figure 6. The subband construction part is added to the conventional SAOC encoder structure. In that part, the subband structure is constructed according to spectral variation, and one parameter describing subband structure is estimated.

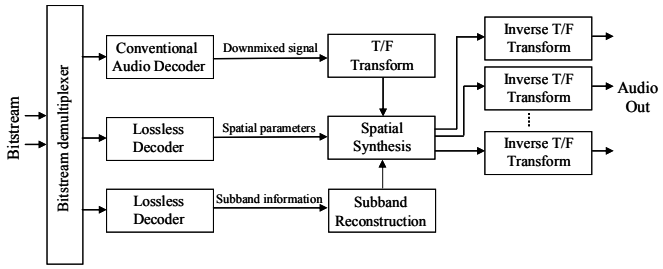

Figure 7. Block diagram of the revised SAOC decoder.

Figure 7 shows the proposed SAOC decoder which handles the variable subband structure. Most parts are same to the conventional SAOC decoder, and subband reconstruction part is only added to the revised scheme. In the subband reconstruction part, the variable subband structure is recovered by using subband information parameter passed through the lossless decoder.

4. Experiment and Result

4.1 Experiment environment

Sound suppression of the target object was mainly evaluated for confirming serious audio quality degradation. The experiments were conducted by 10 listeners and compared overall sound quality of the audio objects decoded by various methods given in Table 1. In the proposed method, threshold value was heuristically set to 10^2.1.

Table 1 shows methods used in the experiment. The 28 subbands structure is a conventional method, and the 79 subbands structure consists of subbands with 86 Hz resolution (4 frequency bins) until 18th band. The 79 subbands structure has high bitrate and high audio quality performance. The purpose of the anchor is to make the scale be closer to an absolute scale in the subjective test. Table 2 shows test materials used in the experiment.

Table 1. Experiment methods

Hidden reference	Original object
Anchor	Bad sound quality by a low pass filter
28 subbands	Conventional method
79 subbands	High quality and high bitrate mode
Proposed	Variable subband structure

Table 2. Test contents

Content 1	Ballad	6 objects
Content 2	Rock	6 objects
Content 3	K-pop 1	10 objects
Content 4	K-pop 2	12 objects
Content 5	K-pop 3	9 objects

To verify performance of the proposed SAOC subjective tests were conducted by MUSHRA test [6]. MUSHRA test

means MUltiple Stimuli with Hidden Reference and Anchor which is subjective evaluation method of audio quality. It is defined by ITU-R recommendation. The main advantage of this test is that it requires fewer participants to obtain statistically significant results. In MUSHRA test, the listener is presented with the reference, a certain number of test samples, a hidden version of the reference and one or more anchors.

4.2 Result

Figure 8 shows the experiment result by MUSHRA test and Table 3 gives the average number of subbands used in each method. As depicted in the figure, the proposed method shows similar sound quality with the 79 subbands structure method in spite of using fewer numbers of subbands. That means it effectively provides the high resolution subband structure.

In Table 3, the numbers of subbands used in the proposed method are presented. Although it uses about 40 subbands, the total increased number of subbands is not too much because of the two-step structure. The average increase rate of the total number of subbands is about 11%.

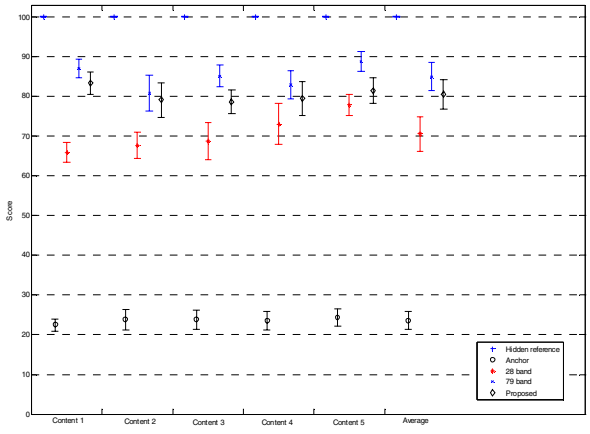

Figure 8. MUSHRA test result.

Table 3. The number of subbands used in the proposed method

Contents	Variable subbands	28 subbands (Total)	Variable subbands (Total)	Increase rate
Content 1	48.7	168	209.4	24%
Content 2	45.6	168	203.2	20%
Content 3	47.6	252	291.2	15%
Content 4	32.3	224	232.6	3.8%
Content 5	30.1	308	312.2	1.4%
Average	40.8	224	249.4	11%

4. Conclusions

SAOC is a technique to compress various audio objects by a downmixed signal and some side information. In the SAOC encoder, input objects are represented as a stereo downmixed signal, and the spatial parameters are estimated through the OTT and the TTT modules. In the SAOC decoder, each object is reconstructed from the transmitted downmixed signal and the spatial parameters. The reconstructed objects are passed to the mixer/renderer and then desired mixing signals or sound scenes can be generated according to the user interaction.

Although it has very low bitrate compared to the conventional audio coders, it shows poor audio quality because of downmix process and the limited number of subbands. In order to improve the audio quality, the two-step structure and variable subband analysis were proposed in this paper. The two-step structure consists of the two SAOC systems which are the conventional system and the proposed system to minimize the increase of bitrate. Input objects except the target object are compressed by the conventional SAOC encoder and the target object is encoded by the proposed SAOC which has variable subband method. Variable subband analysis flexibly provides the high resolution subband structure according to the spectral variation of input signals. This method is only applied to the target object in the two-step structure. Subjective test has been conducted for performance evaluation of the proposed method, and it was confirmed that the proposed method can improve the sound quality.

ACKNOWLEDGEMENTS

This work was supported by the IT R&D program of MIC/IITA. [2007-S-004-01, Development of Glassless Single-User 3D Broadcasting Technologies]

REFERENCES

[1] ISO/IEC, "Generic coding of moving pictures and associated audio information - Part 7: Advanced Audio Coding," ISO/IEC 13818-7 International Standard, 1997.

[2] J. Herre, C. Faller, S. Disch, C. Ertel, J. Hilpert, A. Hoelzer, K. Linzmeier, C. Spenger, and P. Kroon, "Spatial Audio Coding: Next generation efficient and compatible coding of multi-channel audio," 117th Conv. Aud. Eng. Soc., 2004.

[3] J. Breebaart, J. Herre, C. Faller, J. Roden, F. Myburg, S. Disch, H. Purnhagen, G. Hotho, M. Neusinger, K. Kjorling, and W. Oomen, "MPEG spatial audio coding / MPEG surround: Overview and current status," 119th Conv. Aud. Eng. Soc., 2005.

[4] Kyungryeol Koo, Kwangki Kim, Minsoo Hahn, "Spatial Audio Object Coding for User Interactive Audio Service," Proc. ITC-CSCC 2007, Vol 2, pp. 947

[5] C. Faller and F. Baumgarte, "Efficient Cue Coding Applied to Stereo and Multi-channel Audio Compression," Proc. AES 112th Conv., 2002.

[6] "Method for the subjective assessment of intermediate quality level of coding systems", ITU Recommendation, BS.1534-1