LETTER

Efficient Residual Coding Method of Spatial Audio Object Coding with Two-Step Coding Structure for Interactive Audio Services

Byonghwa LEE[†a)], ***Student Member***, **Kwangki KIM**[††], ***and*** **Minsoo HAHN**[†b)], ***Nonmembers***

SUMMARY In interactive audio services, users can render audio objects rather freely to match their desires and the spatial audio object coding (SAOC) scheme is fairly good both in the sense of bitrate and audio quality. But rather perceptible audio quality degradation can occur when an object is suppressed or played alone. To complement this, the SAOC scheme with Two-Step Coding (SAOC-TSC) was proposed. But the bitrate of the side information increases two times compared to that of the original SAOC due to the bitrate needed for the residual coding used to enhance the audio quality. In this paper, an efficient residual coding method of the SAOC-TSC is proposed to reduce the side information bitrate without audio quality degradation or complexity increase.

key words: *interactive audio service, spatial audio object coding, residual coding*

1. Introduction

Presently, we usually enjoy music passively, i.e., it is rendered by experts like a producer. But the audio channel environment is rapidly changing to multi-directional ones and our desires for music suitable to our preference have also been steadily increasing. That is mainly why we need more reliable and higher-quality interactive audio services (IASs).

For IASs, all the component objects are supplied to users instead of the single premixed music. Hence, IASs have not only the advantage of reflecting personal preference but the disadvantage of the bitrate increase to transmit all of the object signals for object control. The bitrate increase becomes a non-negligible obstacle for successful IASs especially in mobile environments.

The SAOC has good performance with relatively low bitrate [1] and is the appropriate scheme for IASs. But the SAOC conducts sub-band processing with spatial parameters and it becomes hard to reconstruct the objects faithfully. Although the audio quality degradation does not seem significant when all of the objects are played, users can perceive noticeable degradation when a specific object is suppressed or played alone.

To cope with this, the SAOC Two-Step Coding (SAOC-TSC) utilizing the residual signal of the target object was proposed [2]. Its audio quality is becomes better but with increased complexity and 2-3 times more bitrate of the side information. In this paper, we propose an efficient residual coding method for the SAOC-TSC to reduce the side information bitrate.

Manuscript received December 1, 2015.
Manuscript revised March 11, 2016.
Manuscript publicized April 8, 2016.

[†]The authors are with Dept. of Information and Communications Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea.
[††]The author is with Digital Contents, Nazarene University, Chon-an, Korea.
a) E-mail: bhlee77@kaist.ac.kr
b) E-mail: mshahn2@kaist.ac.kr
DOI: 10.1587/transinf.2015EDL8248

2. Related Works

2.1 Spatial Audio Object Coding

The SAOC in Fig. 1 encodes needed music objects into one down-mix signal with its side information as spatial parameters. And a user can enjoy a premixed music with the down-mix signal when no SAOC decoder is available [3]. The down-mix signal is calculated as the weighted sum of the input object signals.

To obtain the spatial parameters, the input signals are transformed in frequency domain per frame by the discrete Fourier transform (DFT). The object level difference (OLD), a kind of the spatial parameter, is the power ratio between the input objects and calculated by Eq. (1)

$$OLD_i(f,b) = \frac{P_i(f,b)}{P_M(f,b)}, 1 \le i \le N, 1 \le f \le L, 1 \le b \le B \tag{1}$$

where $P_i(f,b) = \sum_{k=A_{b-1}}^{A_b-1} [S_i(f,k)]^2$ is the power of the i^{th} input object and $P_M(f,b)$ is the maximum power of the input objects at the b^{th} sub-band of the f^{th} frame. Moreover, $S_i(f,k)$ is a frequency signal of the i^{th} input object. Furthermore, A_{b-1} and $A_b - 1$ are the beginning and end points of the b^{th} sub-band, repectively. And N, L, and B are the number of the input object, the frame, and the sub-band. The sub-band frequency signals are obtained by applying human auditory characteristic according to Table 1 [4].

The signal of each object is reconstructed in the decoder by multiplying the down-mix signals with the sub-band gain as in Eq. (2)

$$\hat{S}_i(f,k) = G_i(f,b) \cdot D(f,k) \tag{2}$$

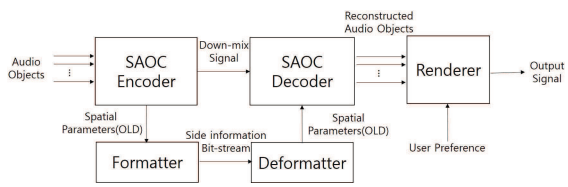

Fig. 1 Structure of SAOC

where $\hat{S}_i$ and D are the reconstructed signal of the i^{th} object and the down-mix signal in frequency domain, respectively.

The gain value of the sub-band is calculated with the OLD using Eq. (3)

$$G_i(f,b) = \sqrt{\frac{OLD_i(f,b)}{\sum_{j=1}^{N} OLD_j(f,b)}}. \quad (3)$$

2.2 SAOC Two-Step Coding [2]

The first step of the SAOC-TSC encoder is same to the SAOC encoder as shown in Fig. 2 and it generates common-object down-mix signals and OLD spatial parameters. The common-object down-mix signal is the weighted sum of the input objects except target one.

The step II encoder, the second step, makes the final down-mix signal and step II parameters. The step II parameters consist of the channel level differences (CLDs) and the residual signal. The final down-mix signal is evaluated as the weighted sum of the common-object down-mix and the target object signals. CLD values are calculated as the power ratio between the target object and the common-object down-mix signals as Eq. (4) [2], [5], [6]

$$CLD(f,b) = 10\log_{10}\frac{P_t(f,b)}{P_{dm}(f,b)} \quad (4)$$

where P_t and P_{dm} are the power of the target object and the common-object down-mix signals, respectively. The sub-band gains of the target and the common-object down-mix signals are represented with the CLD value as,

$$G_t(f,b) = \sqrt{\frac{10^{\frac{CLD(f,b)}{10}}}{1+10^{\frac{CLD(f,b)}{10}}}}, G_{dm}(f,b) = \sqrt{\frac{1}{1+10^{\frac{CLD(f,b)}{10}}}} \quad (5)$$

where G_t and G_{dm} are the gain of the target object and the common-object down-mix signals, respectively. However, the constructed signals from (2) are slightly different from original signals. The residual signal is the difference between the reconstructed and the original objects. The target object and the common-object down-mix can be represented as,

$$S_t(f,k) = G_t(f,b)\cdot D(f,k) + R(f,k)$$
$$S_{dm}(f,k) = G_{dm}(f,b)\cdot D(f,k) - R(f,k) \quad (6)$$

where D is the signal of the final down-mix in frequency domain. Finally, the residual signal can be represented using the two input signals and the CLD values as,

$$R(f,k) = \frac{G_{dm}(f,b)\cdot S_t(f,k) - G_t(f,b)\cdot S_{dm}(f,k)}{G_t(f,b)+G_{dm}(f,b)} \quad (7)$$

where $R(f,k)$ is the kth residual signal of the fth frame. Note that residual signal encoding is additionally conducted with bitrate increase to minimize the difference and it usually improves audio quality [2].

For more efficient encoding, careful effective bandwidth set-up is required because a rather high bitrate is needed to encode the whole bandwidth of the residual signal. We excluded higher frequency bands from the effective bandwidths considering poor human auditory sensitivity for those frequency ranges. Considering both the audio quality and the bitrate for residual coding, the effective bandwidth was decided on 0-5.5 kHz [2].

Line spectral frequency quantization and transform coded excitation coding are used for the residual coding after linear predictive analysis [2]. The SAOC-TSC adopts a fixed bitrate in residual coding.

The step II decoder of the SAOC-TSC as shown in Fig. 3 makes the common-object down-mix and the target signals with the transmitted final down-mix signal and the step II parameters. The step I decoder reconstructs the objects using the common-object down-mix signal from the step II decoder and the transmitted step I parameters.

2.3 Discussions on Previous Works

Each object decoded through the SAOC has some undesirable distortion caused by the other objects in the down-mix signal and discontinuities in frequency because it is reconstructed by multiplying the down-mix signals using the power ratio per sub-band as Eqs. (2) and (3). This audio quality degradation makes the SAOC almost impossible be used in high-quality IASs where a specific object needs to

Table 1 Partition boundaries of sub-band (DFT size of 2048, Sampling Rate of 44.1 kHz)

A_0	A_1	A_2	A_3	A_4	A_5	A_6	A_7
0	3	7	11	15	19	23	27
A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}
31	39	47	55	63	79	95	111
A_{16}	A_{17}	A_{18}	A_{19}	A_{20}	A_{21}	A_{22}	A_{23}
127	159	191	223	255	287	319	367
A_{24}	A_{25}	A_{26}	A_{27}	A_{28}	-	-	-
415	479	559	655	1025	-	-	-

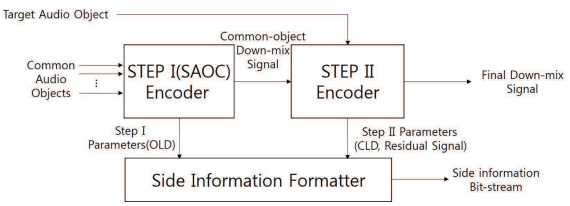

Fig. 2 SAOC-TSC encoder

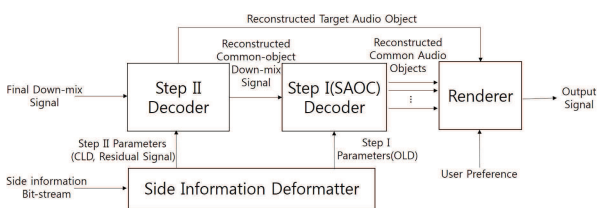

Fig. 3 SAOC-TSC decoder

be fully suppressed or played alone.

The complexity and the bitrate of the SAOC-TSC are increased about 2 times compared to that of the SAOC due to the residual coding adopted for quality enhancement. It shows improved audio quality and can be used for various services where single object elimination or playing is needed.

3. SAOC-TSC Variable Residual Coding Scheme

Residual coding of the SAOC-TSC adopts the effective bandwidth of the 0-5.5 kHz frequency regions and the fixed bitrate of 15 kbps per channel [2].

The target object and the common-object down-mix signals don't always exist both in time and effective bandwidth. Also, because the target object can be not only vocal but also instruments, it would be better to adjust the effective bandwidth along with their frequency characteristics. But the SAOC-TSC fixed residual coding scheme adopts fixed bitrate and effective bandwidth.

In this paper, we propose the SAOC-TSC variable residual coding (SAOC-TSC(VRC)) scheme. The bitrate and the effective bandwidth of the residual coding can be adjusted through an analysis of the target object and the common-object down-mix signals. In this way, the bitrate for the residual signal can be reduced. Adopted 4 residual coding modes are shown Table 2.

As explained previously, the effective bandwidth is simply obtained from the CLD because the power information of the target object and the common-object down-mix signals are included in the value. From Eq. (4), the CLD is $+\infty$ dB or $-\infty$ dB if any of two signals is zero.

Basically, the structure of the SAOC-TSC(VRC) is almost same to that of the SAOC-TSC. Only the analysis module for effective bandwidth determination is added to the step II encoder. Similarly, in the step II decoder, the decoding is performed with the transmitted CLD.

Consequently, the bitrate for the residual coding can be reduced without a noticeable complexity change.

4. Experiments and Results

4.1 Experimental Conditions

For performance evaluation, experiments were performed for the target object suppressed service. Five popular Korean songs in Table 3 were used as test items. They were CD quality stereo signals. We show the average bitrate reduction for all five items achieved by the proposed method in Table 4 and 5. To compare audio quality, we conducted the objective and subjective performance evaluations for the three target objects (vocal, rhythm, and bass).

Table 2 Mode of variable residual coding

CLD value	Effective Bandwidth	Bitrate (kbps)
$\pm\infty$ for 1^{st} to 20^{th} sub-bands	0	0
$\pm\infty$ for 13^{th} to 20^{th} sub-bands	0-1.375 kHz	10
$\pm\infty$ for 17^{th} to 20^{th} sub-bands	0-2.750 kHz	12.5
otherwise	0-5.500 kHz	15

4.2 Bit-Rate of Side Information

The average occurrence rate of the selected bitrate mode of variable residual coding for each item is shown in Table 4. As a result, for about 38 % of the occurrences, bitrate reduction can be tried.

Final bitrate reduction results are summarized in Table 5. Namely, by applying VRC to SAOC-TSC residual coding, about average 24.5 % bitrate reduction of all objects comprising the five items is achieved.

4.3 Objective and Subjective Performance Evaluation

To compare the signal distortion objectively, we used the segmental signal-to-noise ratio (SEGSNR) and the symmetric Kullback-Leibler distance (SKLD). Table 6 shows that the SAOC-TSC with residual coding has better performance than that without it and the distortion of the SAOC-TSC (FRC) is almost same to that of the SAOC-TSC (VRC).

Table 3 Performance evaluation items.

Index	Item	List of object
A	Braves	Guitar, bass, keyboard, rhythm, chorus, vocal
B	Hajiman	Guitar, bass, keyboard, rhythm, chorus, vocal
C	LaLaLa	Strings, bass, drum (rhythm), vocal
D	Snow	Guitar, bass, strings, rhythm, chorus, vocal
E	SulpunDajim	Guitar, bass, piano & brass, rhythm, chorus, vocal

Table 4 Occurrence rate of bitrate mode for residual coding (%) (the average value of all objects comprising each item)

Item	Occurrence Rate (%)			
	No signal (0 kbps)	−1.375 kHz (10 kbps)	−2.75 kHz (12.5 kbps)	−5.5 kHz (15 kbps)
A	16	16	6	62
B	20	17	5	58
C	15	13	8	65
D	22	14	2	61
E	21	11	5	63
Average	19	14	5	62

Table 5 The bitrate of the side information/residual coding per channel (the average value of all objects comprising each item)

Item	Scheme (kbps)		Reduction Rate (%)
	SAOC-TSC (FRC)	SAOC-TSC (VRC)	
A	25.33/15	21.95/11.62	13.34/22.52
B	25.07/15	21.04/10.97	16.08/26.87
C	22.54/15	19.51/11.97	13.45/20.21
D	24.54/15	20.40/10.87	16.84/27.55
E	25.87/15	22.05/11.18	14.76/25.46
Average	24.67/15	20.99/11.32	14.91/24.52

Table 6 Results of The Objective Performance Evaluation (the average evaluation value of the 5 items per target)

Target object	Item	SEGSNR (dB)	SKLD (dB)
Vocal	SAOC	18.31	27.28
	SAOC-TSC(FRC)	33.24	21.22
	SAOC-TSC(VRC)	33.24	21.22
Rhythm	SAOC	22.99	23.77
	SAOC-TSC(FRC)	37.14	20.65
	SAOC-TSC(VRC)	37.15	20.65
Bass	SAOC	26.51	23.84
	SAOC-TSC(FRC)	49.56	16.98
	SAOC-TSC(VRC)	48.83	17.19

Table 7 Configuration for MUSHRA Test

Scheme	Description
Hidden Reference	Reference signal generated with the original audio objects (the sum of the original common-objects except target object)
Anchor	3.5 kHz band-limited reference signal
SAOC	Signal decoded by the SAOC
SAOC -TSC(FRC)	Signal decoded by the SAOC and residual signal coded at fixed 15 kbps
SAOC -TSC(VRC)	Signal decoded by the SAOC and residual signal coded at variable (0/10/12.5/15 kbps)

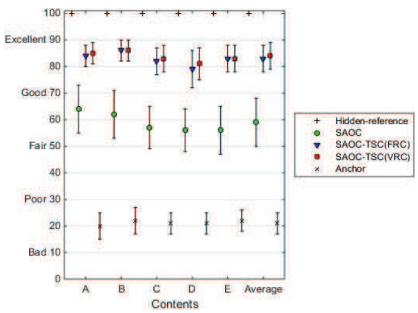

Fig. 4 Subjective test result (target object : vocal)

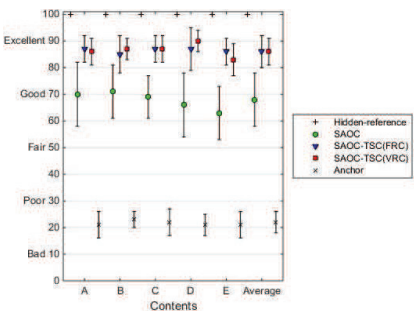

Fig. 5 Subjective test result (target object: rhythm)

For subjective performance evaluation, a multiple stimuli with hidden reference and anchor (MUSHRA) test was performed [7]. The MUSHRA test configuration is in Table 7. Listening test was performed by twelve attendants using Sennheiser HD 650 headphones.

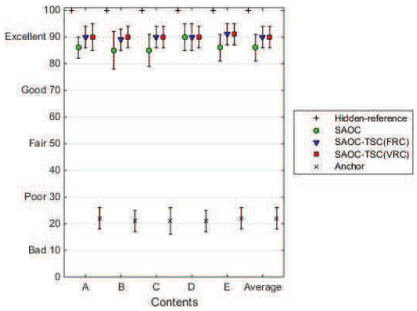

Fig. 6 Subjective test result (target object: bass)

Figures 4, 5, and 6 show the subjective test results of vocal, rhythm and bass as a suppressed target object. The results confirm that the SAOC-TSC (FRC) audio quality is almost same to that of the SAOC-TSC (VRC).

5. Conclusion

The variable residual coding technic is proposed to reduce the side information bitrate used for the residual coding in the SAOC-TSC. Our results show that it can reduce the side information bitrate without audio quality degradation and complexity increase.

Even if the reduced bitrate amount might not be so significant, our method can promote IASs in mobile environments, especially when hardware constraints are crucial due to the various applications embedded on the terminal.

References

[1] ISO/IEC 2003-2 Information Technology — MPEG AudioTechnologies — Part 2: Sptatial Audio Object Coding (SAOC), 2010.
[2] K. Kim, J. Seo, S. Beack, K. Kang, and M. Hahn, "Spatial Audio Object Codig with Two-Step Coding Structure for Interactive Audio Service," IEEE Trans. Multimedia, vol.13, no.6, pp.1208–1216, Dec. 2011.
[3] K. Kim, M. Hahn, and J. Kim, "Mastering Signal Processing in MPEG SAOC," IEICE Trans. Information &Systems, vol.E95-D, no.12, pp.3053–3059, Dec. 2012.
[4] C. Faller and R. Baumgarte, "Binaural Cue Coding-Part II: Schemes and Application," IEEE Trans. on Speech and Audio Proc., vol.11, no.6, Nov. 2003, pp.520–531.
[5] ISO/IEC 23003-1, Information Technology — MPEG AudioTechnologies — Part 1: MPEG Surround, 2007.
[6] K. Kim, S. Beack, J. Seo, D. Jang, and M. Hahn, Improved Channel Level Difference Quantization for Spatial Audio Coding, ETRI Journal, vol.29, no.1, Feb. 2007, pp.99–102.
[7] ITU-R Recommendation, Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA), ITU, BS. 1543-1, Geneva, 2001.