



# Modified Spatial Audio Object Coding Scheme with Harmonic Extraction and Elimination Structure for Interactive Audio Service

Jihoon Park<sup>1</sup>, Kwangki Kim<sup>2</sup>, Jeongil Seo<sup>3</sup>, Minsoo Hahn<sup>1</sup>

<sup>1</sup> Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Korea

<sup>2</sup> Department of Information and Communications Engineering, Korea Advanced Institute of Science and Technology, Korea

<sup>3</sup> Electronics and Telecommunications Research Institute, Korea

batho2n@kaist.ac.kr, mulkuk3@kaist.ac.kr, seoji@etri.re.kr, mshahn@ee.kaist.ac.kr

## Abstract

An interactive audio service provides audio editing functionality to users. In the service, the users can control the wanted audio objects to make their own audio sound using a spatial audio object coding (SAOC) scheme. However, the vocal object cannot be removed perfectly from the down-mix signal in Karaoke mode of the SAOC. Thus, in this paper, a modified SAOC scheme with harmonic extraction and elimination structures are proposed. The proposed scheme perfectly removes vocal object using the harmonic information of the vocal object. Subjective and objective evaluation results show the proposed scheme is superior to the conventional ones. **Index Terms:** spatial audio object coding, interactive service, harmonic elimination

## 1. Introduction

Conventional audio service provides users with an audio signal called music which is typically stereo type and is made by properly mixing various audio objects such as vocal and several musical instruments. Because the users can only control the overall volume of the music signal, the users' demand for an alternative and advanced audio service is increased rapidly. In addition, user interactive audio services such as MUSIC 2.0, UCSing and etc. have been recently introduced in Korea [1]. In the interactive audio service, the individual audio objects and the preset information are transmitted to the users instead of the music signal. In this interactive audio service, audio signals are predetermined by the producer and generated using the audio objects and the preset information. Although the interactive audio service can satisfy the users' demands on the new audio service, it may not be practical in the network and broadcasting environments. The bit-rate is greatly increased in proportion to the number of the audio objects since an audio coder in the interactive audio service separately codes them. As a solution to the bit-rate problem of the interactive audio service, a spatial audio object coding (SAOC) scheme has been suggested [2]–[5].

The basic idea of the SAOC is that the audio objects can be represented as a down-mix signal with spatial parameters. As the SAOC only allocates the bit for the transmission of the down-mix signal and the additional side information, the bit-rate of the interactive audio service can be greatly reduced. In addition, the SAOC support Karaoke mode to users. In Karaoke mode, users can control the vocal object to make their own background music. Nevertheless, the SAOC cannot be directly used for the interactive audio service. As the audio objects reconstructed by the SAOC are not equal to the original ones, the sound quality is rather degraded. If a specific audio object is fully suppressed or played alone, the

degradation of the sound quality may be very critical and the specific audio object components remain at reconstructed signal in Karaoke mode, because the SAOC uses the sub-band processing having low frequency resolution and the audio objects are recovered from the down-mix signal. In other words, the perfect control of a particular audio object that is possible in the interactive audio service cannot be supported in the SAOC.

As a method to enhance the performance in Karaoke mode of the SAOC, the modified SAOC scheme with harmonic extraction and elimination structure is proposed in this paper. Because we know a clean vocal object, the harmonic information which is a fundamental frequency and amplitudes is well extracted in the SAOC encoder. In the SAOC decoder, vocal object is removed from the down-mix signal using transmitted spatial parameters and harmonic information.

## 2. Spatial audio object coding

The SAOC consists of the encoding and the decoding parts, as shown in Figure 1. In the encoding part, the input audio objects are represented as the down-mix signal with spatial parameters. For the calculation of spatial parameters, the input audio objects are transformed into frequency domain signals by DFT. The transformed signals are classified into parameter sub-bands to be well adapted human perception, as shown in Table 1 [1]. Table 1 presents the partition boundaries in case of partition bandwidths of equivalent rectangular bandwidth (ERB). Object level difference (OLD) used as a major spatial parameter is utilized in the SAOC. The OLD is defined as the power ratio among the input audio objects, and it is determined as

$$OLD_i(n,b) = \frac{P_i(n,b)}{\max_{1 \leq j \leq N} P_j(n,b)} \begin{cases} 1 \leq n \leq L \\ 1 \leq i \leq N \\ 1 \leq b \leq M \end{cases} \quad (1)$$

where  $P_i(n,b)$  is the estimated power of the  $i^{\text{th}}$  audio object at the sub-band  $b$  of the  $n^{\text{th}}$  frame, while  $L$ ,  $N$ , and  $M$  are the numbers of the frames, the input audio objects, and the

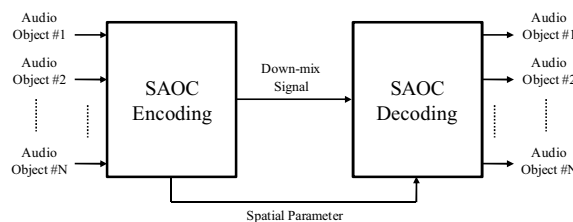


Figure 1: General structure of the SAOC.

sub-bands, respectively.

In the SAOC decoder, the background music is separated using the transmitted down-mix signal and spatial parameters. The gain factor of the each object can be calculated as

$$G_i(n,b) = \sqrt{\frac{OLD_i(n,b)}{\sum_{j=1}^N OLD_j}} \begin{cases} 1 \leq n \leq L \\ 1 \leq i \leq N \\ 1 \leq b \leq M \end{cases} \quad (2)$$

The calculated gain factor is multiplied to the transformed down-mix signal by DFT, as shown in Table 1.

$$O(n,k) = D(n,k)G_i(n,b) \begin{cases} 1 \leq n \leq L \\ 1 \leq i \leq N \\ 1 \leq b \leq M \\ A_{b-1} \leq k \leq A_b - 1 \end{cases}, \quad (3)$$

where  $O(n,k)$  and  $D(n,k)$  are the  $i^{\text{th}}$  estimated audio object signal and down-mix signal in the frequency domain, respectively.  $A_{b-1}$  and  $A_b - 1$  are the points of lower and upper boundaries for parameter sub-band  $b$ .

The SAOC should support two main application scenarios of the interactive audio service. One is the remixing music, where the users can make their own music through the amplification and attenuation of the level of the audio objects. The other is the Karaoke, where the lead vocal object is fully suppressed. For the remixing music application, the SAOC demonstrates good performance in the aspect of bit-rate and sound quality, because the bit-rate of the SAOC is slightly higher than that required for the transmission of one audio object, and the simple gain control of each audio object rarely affects the overall sound quality. However, for the Karaoke application, the SAOC shows poor performance in the aspect of removing the specific object. As the SAOC uses the sub-band processing having low frequency resolution and the audio objects are recovered from the down-mix signal, the recovered audio objects cannot be equal to the original ones.

Table 1. Sub-bands boundaries (DFT size: 2048, sampling rate: 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	318	367
$A_{24}$	$A_{25}$	$A_{26}$	$A_{27}$	$A_{28}$			
415	479	559	655	1025			

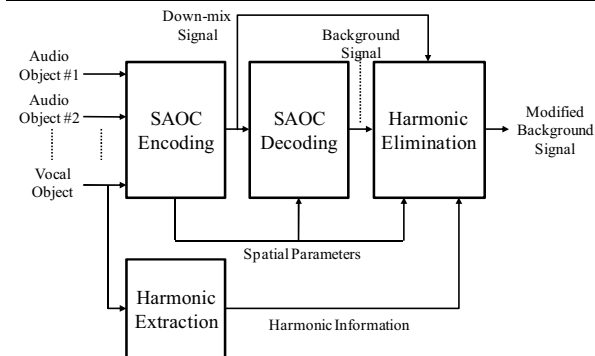


Figure 2: Proposed structure of the SAOC.

Therefore, when the specific audio object is fully suppressed or played alone, the performance of the output signal is not good.

### 3. Proposed method

To remove the vocal signal from the down-mix signal, a modified SAOC scheme with harmonic extraction and elimination structure is proposed. The structure of the proposed scheme is shown in Figure 2. SAOC encoding and decoding blocks are introduced in section 2. Harmonic extraction block extract the harmonic information from the vocal object, as explained in section 3.1. Harmonic elimination block reduces the remaining vocal signal from the separated background music, as explained in section 4.2.

#### 3.1. Harmonic extraction

The harmonic extraction extracts the transmitting harmonic information which consists of one fundamental frequency and several amplitudes. The amplitude information is calculated very simply, because it is only power spectrum magnitude at multiplying integer by the fundamental frequency in frequency domain. However, it is difficult to calculate the fundamental frequency mainly called F0. Many approaches of estimating fundamental frequency are studied [6]-[8]. In this section, the process of the fundamental frequency extraction is explained in detail, as shown in Figure 3.

Spectral whitening can flatten the rough power spectral distribution entirely or partly. It is not easy to extract the fundamental frequency without spectral whitening, because the power spectral distribution of the speech signal such as vocal object shows the large variations between low frequency and high frequency. First, the vocal object  $x(n)$  is transformed into the frequency domain signal  $X(k)$  by DFT for the spectral whitening, and center frequencies of critical sub-bands are calculated by  $c_b = 229(10^{(b+1)/21.4} - 1)$ . Then a critical sub-band  $b$  has a triangular power response  $H_b(k)$  that ranges between  $c_{b-1}$  and  $c_{b+1}$ . Next, spectral whitening coefficient  $\gamma_b$  with sub-band can be calculated as  $\gamma_b = \sigma_b^{V-1}$ , where  $\sigma_b^2$  is variance of sub-band  $b$ :

$$\sigma_b^2 = \frac{1}{K} \sum_{k=c_{b-1}}^{c_{b+1}} H_b(k) |X(k)|^2. \quad (4)$$

The spectral whitening filter coefficient  $\gamma(k)$  is obtained by linear interpolation of coefficient  $\gamma_b$  between the center frequencies of critical sub-bands. Then we obtain the spectral flattened signal as multiplying the input signal by the filter coefficient,  $Y(k) = \gamma(k)X(k)$ .

The salience function in Figure 3 is a sum of the amplitudes of a fundamental frequency candidate. In detail, the salience function  $s(\tau)$  of a pitch period candidate  $\tau$  is calculated as

$$s(\tau) = \sum_{m=1}^M \max_{k \in \kappa_{\tau,m}} |Y(k)|, \quad (5)$$

where  $\kappa_{\tau,m}$  is a calculating range of candidate  $\tau$  in frequency

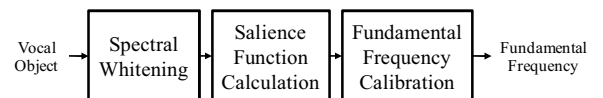


Figure 3: Process of the harmonic extraction.

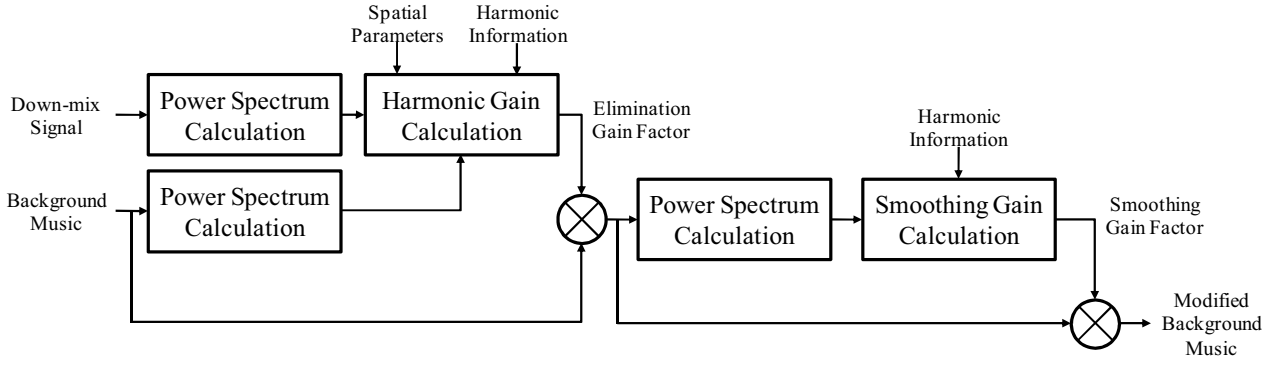


Figure 4: Proposed harmonic extraction structure.

domain. In detail,

$$\kappa_{\tau,m} = \left[ \left\langle mK / (\tau + \Delta\tau / 2) \right\rangle, \dots, \left\langle mK / (\tau - \Delta\tau / 2) \right\rangle \right], \quad (6)$$

where  $K$  is a DFT size, and the operator  $\langle \cdot \rangle$  denotes rounding off to the nearest integer. The estimated pitch period  $\hat{\tau}$  is determined as

$$\hat{\tau} = \arg \max_{\tau} s(\tau). \quad (7)$$

Finally, we estimate fundamental frequency as calculating  $f_s / \hat{\tau}$ . In general, we use the fast Fourier transform (FFT) instead of DFT for transform to the frequency domain signal; however, estimated fundamental frequency is not integer. Therefore, it is important to calibrate the estimated fundamental frequency for transmitting the precise fundamental frequency to the decoder. The calibrated fundamental frequency  $F_0$  is obtained as searching frequency bin of the power spectrum  $|X(k)|$  which has max value around the estimated fundamental frequency.

### 3.2. Harmonic elimination

Because SAOC performs the sub-band processing and the vocal objects are removed from the down-mix signal, vocal object remains in the background music of SAOC decoding output. The SAOC decoding output is calculated as

$$B(n,k) = D(n,k) \sqrt{1 - \frac{OLD_v(n,b)}{\sum_{j=1}^N OLD_j}} \begin{cases} 1 \leq n \leq L \\ 1 \leq b \leq M \\ A_{b-1} \leq k \leq A_b - 1 \end{cases}, \quad (8)$$

where  $B(n,k)$  is the background music, and  $OLD_v(n,b)$  is a OLD of vocal object. As you can see the equation (8), SAOC decoding only suppresses the spectral power of the down-mix signal. To enhance the Karaoke mode of SAOC, the proposed harmonic elimination method is described in Figure 4. The harmonic elimination method effectively eliminates the remaining vocal signal in SAOC decoding output.

A power spectrum  $|D(k)|$  is calculated from the down-mix signal, and then a harmonic gain factor  $G_H(k)$  is obtained as

$$G_H(k) = H(m) - |D(k)| G_v(n,b) \begin{cases} 1 \leq n \leq L \\ 1 \leq b \leq M \\ k = m \times F_0 \end{cases}, \quad (9)$$

where  $G_v(n,b)$  is a gain factor of the vocal object,  $F_0$  is a transmitted fundamental frequency from the harmonic extraction block, and  $H(m)$  is a transmitted magnitude of power spectrum at harmonic bins which locate at multiplying integers by the fundamental frequency in frequency domain. Then, an elimination gain factor  $G_E(m)$  for the harmonic elimination is

$$G_E(k) = \begin{cases} G_H(k) / |B(k)| & , k = m \times F_0 \\ 1 & , otherwise \end{cases}. \quad (10)$$

The background music eliminated harmonic components is obtained by weighting the background music by the elimination gain factor,  $B'(k) = G_E(k)B(k)$ . Although remaining vocal object is eliminated from the background music, the quality of background music is degraded because of the discontinuity originated from the bin eliminated harmonic components.

To modify the background music, smoothing gain factor is calculated as

$$G_S(k) = \begin{cases} \frac{\sum_{m=-1}^1 B'(k+m)}{3|B'(k)|} & , \kappa-1 \leq k \leq \kappa+1, \\ 1 & , otherwise \end{cases}, \quad (11)$$

where  $\kappa$  is a multiplying the fundamental frequency by integer,  $\kappa = m \times F_0$ . Finally, a modified background music is obtained by multiplying the background music eliminating harmonic by the smoothing gain factor,  $\tilde{B}(k) = B'(k)G_S(k)$ .

## 4. Experimental results

For the test, 5 popular Korean songs, listed in Table 2, were used. Each item composed of 4-6 audio objects, such as vocal and some musical instruments, which were sampled at 44.1 kHz with 16 bit quantization level. Analysis window for the Fourier transform is identical to [9].

We used the symmetric Kullback-Leibler distance (SKL) as objective measure between a reference signal and decoded signals. The SKL is defined as

$$SKL = \int (P(\omega) - Q(\omega)) \log \frac{P(\omega)}{Q(\omega)} d\omega, \quad (12)$$

where  $P(\omega)$  and  $Q(\omega)$  denotes the reference signal and the decoded signal, respectively. The reference signal is the original background music which consists of all objects except the vocal object. The proposed method is compared with the

other methods which are normal SAOC method and [10] based on NMF method. The NMF method is a famous method in vocal separation and has been studied until lately. The results of the performance test were shown in Table 3. As a subjective listening test, the multiple stimuli with hidden reference and anchor (MUSHRA) test was performed [11]. MUSHRA test is well known as a subjective test in audio quality test and includes the hidden reference signal and 3.5 kHz band limited anchor signal. Eight experienced listeners evaluated the background music quality of the test contents in each trail. Figure 5 shows the MUSHRA test result.

As shown in Table 3 and Figure 5, the results of the objective and the subjective test can be summarized as follows: the SAOC and proposed method have better results than the NMF method for all contents of both the objective and subjective test. And the proposed method records higher score than the SAOC method except the Snow content and which has long chorus duration, and shows the largest difference with respect to LaLaLa contest which doesn't have the chorus in the subjective test. The results of object test also show the similar trend by contents. Because the characteristics of the chorus object is similar to that of vocal one, the performance of the chorus quality is affected by the harmonic elimination of the proposed method. However, it is obvious that the proposed method guarantees the sound quality and the amount of the removed vocal object, as shown in the results.

Table 2. Test contents.

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

Table 3. The results of the objective performance test.

	NMF	SAOC	Proposed
A	2930526.4779	2409613.3432	2312958.3432
B	2851381.1351	2356486.2236	2225648.2459
C	2968441.8613	2571832.5283	2077266.6751
D	2978433.8846	2486197.4135	2296847.2456
E	2766547.3486	2411436.9634	2398461.1648
Total	14495330.7075	12235566.4720	11311181.6746

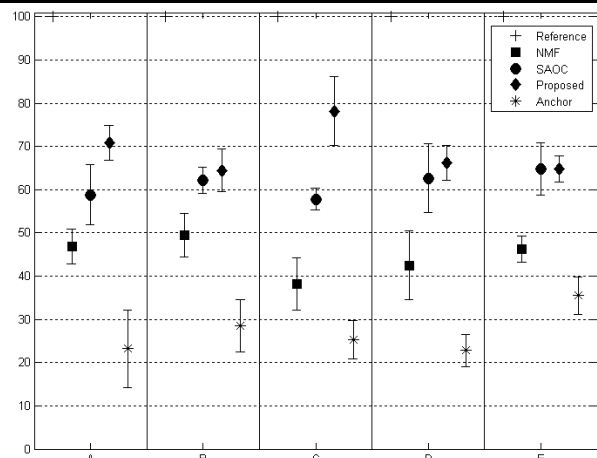


Figure 5: Subjective listening test results.

## 5. Conclusions

The SAOC is a useful technology that can support most parts of the interactive audio service with relatively low bit-rate, but is very weak to perfect gain control of a particular audio object, i.e., the vocal object for Karaoke mode. To remove the vocal object, the SAOC with harmonic extraction and elimination structure is proposed in this paper. In the SAOC decoder, the harmonic elimination removes the vocal harmonic components using the transmitted fundamental frequency and harmonic magnitudes. In particular, the proposed method has good performance with respect to removing vocal object from the down-mix signal. As a future works, to solve the bit-rate problem which is practical in the network and broadcasting, we will study with respect to reducing the bit-rate while maintain the sound quality.

## 6. References

- [1] Jang, D., Lee, T., Lee, Y. and Yoo, J., "A Personalized Preset-based Audio System for Interactive Service", 121<sup>st</sup> AES Convention, Preprint 6904, 2006.
- [2] Herre, J. and Disch, S., "New Concepts in Parametric Coding of Spatial Audio: From SAC to SAOC", 2007 International Conference on Multimedia and Expo, 1894-1897, 2007.
- [3] ISO/IEC JTC1/SC29/WG11 (MPEG), Call for Proposals on Spatial Audio Object Coding, Document N8853, 2007.
- [4] ISO/IEC JTC1/SC29/WG11 (MPEG), Study on ISO/IEC 23003-2:200X, Spatial Audio Object Coding, Document N10659, 2009.
- [5] Breebaart, J., Engdegard, J., Falch, C., Hellmuth, O., Hilpert, J., Hoelzer, A., Koppens, J., Oomen, W., Resch, B., Schuijers, E. and Terentiev, L., "Spatial Audio Object Coding (SAOC)–The Upcoming MPEG Standard on Parametric Object Based Audio Coding", 124<sup>th</sup> AES Convention, Preprint 7377, 2008.
- [6] Wu, M., Wang, D. and Brown, G. J., "A Multipitch Tracking Algorithm for Noisy Speech", IEEE Trans. Speech and Audio Proc., 11(3):229-241, 2003.
- [7] Goto, M., "A Real-time Music-scene-description System: Predominant-f0 Estimation for Detecting melody and Bass Lines in Real-world Audio Signals", Speech Com. 43(4):311-329, 2004.
- [8] Fujihara, H., Goto, M., Ogata, J., Komatani, K., Ogata, T. and Okuno, H. G., "Automatic Synchronization Between Lyrics and Music CD Recordings based on Viterbi Alignment of Segregated Vocal Signals", in IEEE International Symposium on Multimedia, 257-264, 2006.
- [9] Faller, C. and Baumgarte, F., "Binaural Cue Coding-part II: Schemes and Application", IEEE Trans. Speech and Audio Proc., 11(6):520-531, 2003..
- [10] Virtanen, T., Mesaros, A. and Ryyanen, M., "Combining Pitch-based Inference and Non-negative Spectrogram Factorization in Separating Vocals from Polyphonic Music", in ISCA Tutorial Res. Workshop Statist. Percept. Audition 2008, 17-22, 2008
- [11] ITU-R Recommendation, Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA), ITU, BS. 1543-1, 2001.