Journal of Network and Innovative Computing
ISSN 2160-2174 Volume 2 (2014) pp. 269-276
© MIR Labs, www.mirlabs.net/jnic/index.html

Modified SAOC Scheme Using Harmonic Information for KARAOKE Application

Jihoon Park

Abstract — *An interactive audio service provides audio editing functionality to users. In the service, users can control the desired audio objects to make their own audio sound using a spatial audio object coding (SAOC) scheme. However, vocal objects cannot be removed perfectly from the down-mix signal in Karaoke applications of the SAOC. Thus, in this paper, a modified SAOC scheme with harmonic extraction and elimination structures are proposed. The proposed scheme removes vocal objects using their harmonic information. In addition, statistical modeling method is applied to reduce bit rates of the information. Subjective and objective evaluation results show the proposed scheme is superior to the conventional ones which are the normal SAOC and the nonnegative matrix factorization (NMF) algorithms.*[1]

Index Terms — Spatial audio object coding, harmonic elimination, KARAOKE mode, vocal removal.

I. INTRODUCTION

Conventional audio service provides users with an audio signal called music and is made by properly mixing various audio objects such as vocal and several musical instruments. Because users can control only the overall volume of the music signal, the demand of users for an alternative and advanced audio service has been increased rapidly. In addition, user-interactive audio services such as MUSIC 2.0, UCSing, etc. have been recently introduced in Korea [1], [2]. In the interactive audio service, the individual audio objects and the preset information are transmitted to users instead of music signals. In this interactive audio service, audio signals are predetermined by the producer and generated using the audio objects and the preset information. Although the service can satisfy users' demands on new audio services, it may not be practical in communication networks and broadcasting environments. The main reason is that the bit-rate is greatly increased in proportion to the number of audio objects. As a solution to the bit-rate problem of the interactive audio service, a spatial audio coding (SAC) scheme has been suggested [3]-[9].

The SAC scheme was a new concept of multi-channel audio coding using human perception of spatial sounds. As a kind of SAC schemes, binaural cue coding (BCC) was introduced by [3]. The BCC is a scheme for a multi-channel audio based on one down-mix signal and side information. As the side information, spatial parameters such as inter-channel level difference (ICLD), inter-channel time difference (ICTD), and inter-channel correlation (ICC) are proposed in [3]. In addition, the basic idea of the SAOC introduced recently is based on the BCC scheme [5]. The SAOC scheme is that the audio objects are represented as down-mix signals with side information. As the SAOC only allocates bits for the transmission of down-mix signals and additional side information, the bit rate of the interactive audio services can be greatly reduced. In addition, the SAOC can support Karaoke/enhanced Karaoke applications to users. In the applications, users can control the vocal object to make their own background music. Nevertheless, the SAOC cannot be directly used for interactive audio services. As the audio objects reconstructed by the SAOC are not equal to the original ones, the sound quality of reconstructed audio signal is diminished than the original ones. If a specific audio object is fully suppressed or played alone, the deterioration of the sound quality may be very critical and the specific audio object components remain at the reconstructed signals in the Karaoke mode. Reasons of problems are that 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 conventional SAOC scheme. On the other hand, the decoded background music has very high quality and good performance in aspect of vocal removing in the enhanced Karaoke mode but this mode has a bit-rate problem. As a method to enhance the performance in the Karaoke mode of the SAOC using low bit rates, a modified SAOC scheme with harmonic extraction and elimination structures is proposed in this paper. Because we know a clean vocal object, the harmonic information which is a fundamental frequency with rather stable 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.

This paper's organization is as follows: Section II introduces the conventional SAOC scheme. In section III, proposed method is explained in detail and experiments and results are presented in section IV. Finally, conclusions are given in section V.

[1] J. Park is with the Center for Integrated Smart Sensors, 291 Daehak-ro, Yuseong-gu, Daejeon, 305-701, Korea (e-mail: batho2n@kaist.ac.kr)

II. Spatial Audio Object Coding

The SAOC consists of the encoder and the decoder parts, as shown in Fig. 1. In the encoder part, the input audio objects are represented as down-mix signals with spatial parameters. The decoder can obtain each object or mixed object signal using the transmitted down-mix signal and the spatial parameters.

A. Encoder

In the encoder, the input audio objects are represented as the down-mix signal with spatial parameters as shown in Fig. 2. Firstly, each object signal $x_i(n)$ is transformed into frequency domain signal $X_i(k)$ by the discrete Fourier transform (DFT) for the down-mix generation and the calculation of spatial parameters. Secondly, the transmitted down-mix signal is can be generated as

$$D(k) = \sum_{i=1}^{N} g_i X_i(k) \tag{1}$$

where N is the number of input audio object and g_i is an weighting factor for audio mixing. The down-mix signal $D(k)$ is generated simply by a weighted sum because object signals satisfy the superposition principle. Finally, the transformed signals are classified into parameter sub-bands to simulate the human perception by using TABLE I in which A_k is a starting index of $(k+1)$ th sub-band by the DTF [1]. It presents the partition boundaries according to partition bandwidths of the equivalent rectangular bandwidths (ERB).

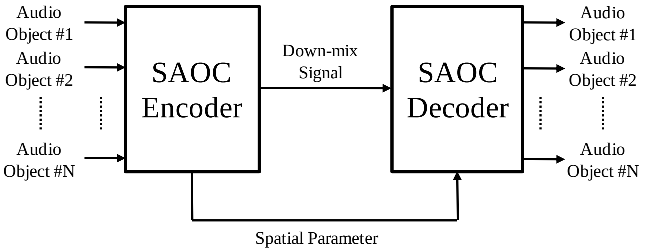


Fig. 1. General structure of SAOC coder

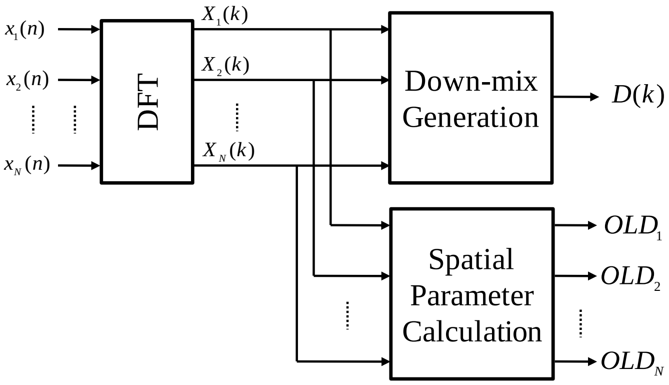


Fig. 2. Process of SAOC encoder

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

TABLE I
Sub-bands Partition Boundaries in Case of ERB
(DFT size: 2048, sampling rate 44.1kHz)

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			

$$OLD_i(b) = \frac{P_i(b)}{\max_{1\le j\le N} P_j(b)} \begin{cases} 1 \le i \le N \\ 1 \le b \le B \end{cases} \tag{2}$$

where B is the number of sub-bands. The power spectrum $P_i(b)$ estimated at the sub-band b of the i th audio object is defined by

$$P_i(b) = \sum_{k=A_{b-1}}^{A_b-1} \left|X(k)\right|^2 \tag{3}$$

where ‘$|\square|$’ is a absolute notation and A_b is a b th sub-band partition boundaries as shown in TABLE I.

B. Decoder

In the SAOC decoder, each object signal is separated using the transmitted down-mix signal and the spatial parameters by

$$O_i(k) = D(k)G_i(b) \begin{cases} 1 \le i \le N \\ 1 \le b \le B \\ A_{b-1} \le k \le A_b - 1 \end{cases}, \tag{5}$$

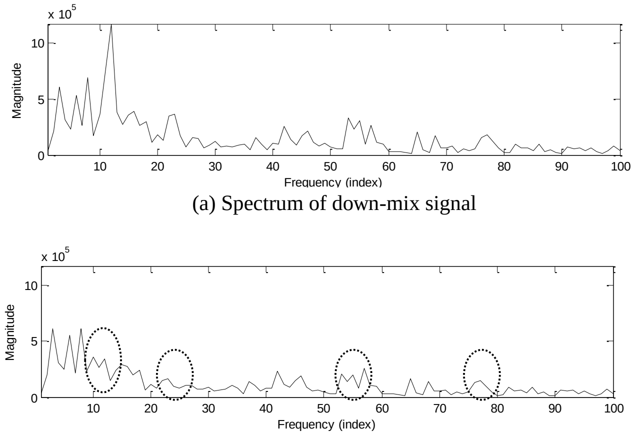
where $O_i(k)$ and $D(k)$ are the i th estimated audio object signal and down-mix signal in the frequency domain, respectively. In addition, the gain factor $G_i(b)$ of the each object can be calculated as

$$G_i(b) = \sqrt{OLD_i(b) / \sum_{j=1}^{N} OLD_j} \begin{cases} 1 \le i \le N \\ 1 \le b \le B \end{cases}. \tag{4}$$

The calculated gain factor is multiplied to the transformed down-mix signal generated by signal DFTs and TABLE I. Then separated object signals are mixed or played alone using rendering functionality according to user’s demands.

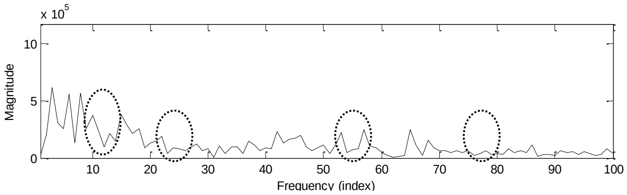
C. *Problem of Karaoke application in SAOC*

The SAOC should support two main application scenarios for interactive audio services. One is the remixing music, with which users can create their own music through the amplification and attenuation of the volume level of audio objects. The other scenario is the Karaoke/enhanced Karaoke application, where the lead vocal object is fully suppressed. For the remixing music application, the SAOC demonstrates good performance in the aspect of bit rates and sound quality because bit rates of the SAOC is slightly higher than the one required for the transmission of one audio object. Moreover, the simple gain control of each audio object rarely affects the overall sound quality. However, the SAOC shows poor performance in the aspect of removing the specific object for the Karaoke application. As the SAOC uses the sub-band processing having low resolution in the frequency domain and the audio objects are recovered from the down-mix signal, the recovered audio objects cannot be equal to the original ones. Therefore, when the specific audio object is fully suppressed or played alone, the performance of the output signal is not good. Fig. 3 shows the frequency spectra of the down-mix signal, decoded background signal by the SAOC decoder, and original background signal. It is obvious that the harmonic components of the removed vocal object remain in the decoded background signal comparing with Fig. 3. The other enhanced Karaoke application use the spatial parameter and the residual coding technique. Although the enhanced Karaoke application satisfies an audio quality and a vocal removing, high bit rates is a problem of this application. It may not be practical in communication networks and broadcasting environments because of high bitrates. Accordingly, it is necessary to do search for enhanced vocal removal methods using low bit rates.



(a) Spectrum of down-mix signal

(b) Spectrum of background signal decoded by SAOC



(c) Spectrum of original background signal

Fig. 3. Frequency spectrum comparison (dotted circle: view point of comparing harmonic remaining)

III. PROPOSED VOCAL REMOVAL METHOD

To remove vocal signals from down-mix signals, a modified SAOC scheme with harmonic extraction and elimination structure is proposed. The structure of the proposed scheme is described in Fig. 4. SAOC encoding and decoding blocks were already introduced in section II. The detailed description of each block in Fig. 4 is as follows.

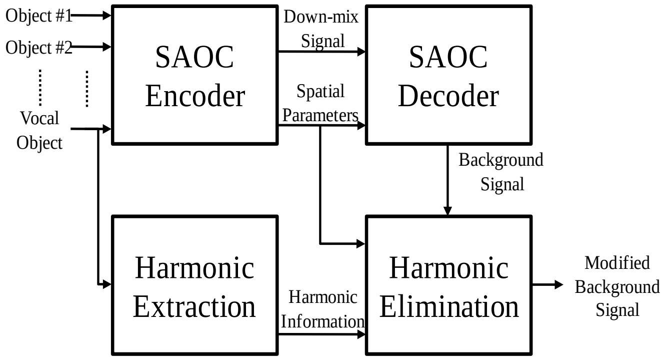


Fig. 4. Overview of proposed structure

A. *Harmonic extraction*

The harmonic extraction estimates the transmitting harmonic information which consists of one fundamental frequency and several amplitudes. The amplitude information is calculated in a simple manner, with spectrum magnitude $H(m)$ at multiplying integer m by the fundamental frequency in the frequency domain. In contrast, it is difficult to calculate the fundamental frequency mainly called F0. Many approaches of estimating fundamental frequency have been studied [10]-[13]. Fig. 5 represents the process of the fundamental frequency extraction. During the F0 estimation, spectral whitening is applied because it can flatten the rough power spectral distribution entirely or partly. Without spectral whitening, it is not easy to extract the fundamental frequency because the power spectral distribution of speech signals such as vocal objects shows the large variations between low frequency and high frequency. The detailed procedure is as follows.

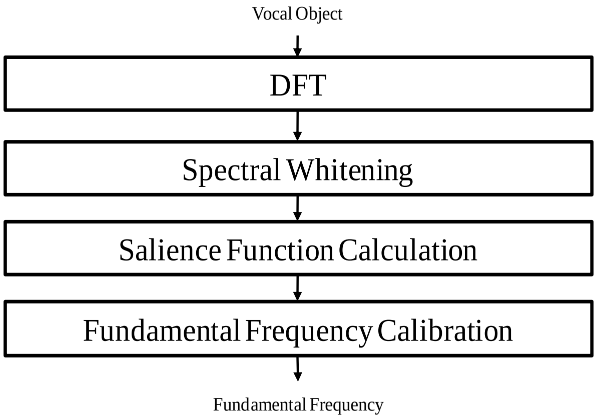

Vocal Object
DFT
Spectral Whitening
Salience Function Calculation
Fundamental Frequency Calibration
Fundamental Frequency

Fig. 5. Process of fundamental frequency extraction

Firstly, 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

$$c_b = 229(10^{(b+1)/21.4} - 1) \tag{6}$$

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 γ_b with sub-band can be calculated as $\gamma_b = \sigma_b^{\nu-1}$, where σ_b^2 is variance of sub-band b :

$$\sigma_b^2 = \frac{1}{K} \sum_{b=c_{b-1}}^{c_{b+1}} H_b(k) \left| X(k) \right|^2. \tag{7}$$

The spectral whitening filter coefficient $\gamma(k)$ is obtained by linear interpolation of coefficient γ_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 Fig. 5 is a sum of the amplitudes of a fundamental frequency candidate. In detail, the salience function $s(\tau)$ of a pitch period candidate τ is calculated as

$$s(\tau) = \sum_{m=1}^{M} \max_{k \in \kappa_{\tau,m}} \left| Y(k) \right| \tag{8}$$

where $\kappa_{\tau,m}$ is a calculating range of candidate τ in the frequency domain. Equation (9) represents the definition of $\kappa_{\tau,m}$.

$$\kappa_{\tau,m} = \left[\left\langle \frac{mK}{\tau + \Delta\tau/2} \right\rangle, \ \ldots, \ \left\langle \frac{mK}{\tau + \Delta\tau/2} \right\rangle \right] \tag{9}$$

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

$$\hat{\tau} = \arg\max_{\tau} s(\tau). \tag{10}$$

Finally, we estimate fundamental frequency by calculating $f_s / \hat{\tau}$. In general, we use the fast Fourier transform (FFT) instead of DFT to transform time-domain signals into the frequency domain; 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 F0 is obtained as searching frequency bin of the vocal power spectrum which has max value around the estimated fundamental frequency.

B. Harmonic elimination

Because the 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

$$B(k) = D(k) \sqrt{1 - \frac{OLD_v(b)}{\sum_{j=1}^{N} OLD_j}} \begin{cases} 1 \le b \le B \\ A_{b-1} \le k \le A_b - 1 \end{cases} \tag{11}$$

where $B(k)$ is the background music, and $OLD_v(b)$ is a OLD of vocal object. As it can be seen (11), 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 Fig. 6. The harmonic elimination method effectively eliminates the remaining vocal signal in the SAOC decoding output.

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

$$G_H(k) = \sqrt{H^2(m) - \left| D(k) G_v(b) \right|^2} \begin{cases} 1 \le b \le B \\ k = m \times F0 \end{cases} \tag{12}$$

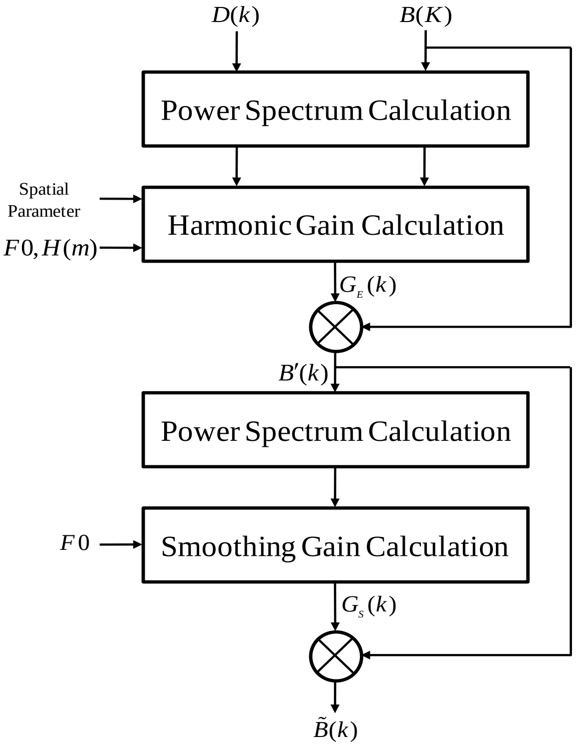

$D(k)$
$B(K)$
Power Spectrum Calculation
Spatial Parameter
$F0, H(m)$
Harmonic Gain Calculation
$G_E(k)$
$B'(k)$
Power Spectrum Calculation
$F0$
Smoothing Gain Calculation
$G_S(k)$
$\tilde{B}(k)$

Fig. 6. Process of harmonic elimination

where $G_v(b)$ is a gain factor of the vocal object, F0 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 the frequency domain. Then, an elimination gain factor $G_E(m)$ for the harmonic elimination is

$$G_E(k)=\begin{cases}\sqrt{1-\dfrac{G_H^2(k)}{B^2(k)}} & , \quad k=m\times F0 \\ 1 & , \quad otherwise\end{cases}. \tag{13}$$

The background music in which harmonic components is eliminated is obtained by weighting the background music by the elimination gain factor, $B'(k)=G_E(k)B(k)$. Although remaining vocal objects are eliminated from the background music, the quality of the background music is deteriorated because of the discontinuity originated from the eliminated frequency components. To enhance the background music, smoothing gain factor is calculated as following

$$G_S(k)=\begin{cases}\dfrac{\sum_{q=-l}^{l}\left[B'(k+q)\right]^2}{(2l+1)\left[B'(k)\right]^2} & , \quad \kappa-1\le k\le \kappa+1 \\ 1 & , \quad otherwise\end{cases} \tag{14}$$

where κ is a multiplication of the fundamental frequency by integer, $\kappa = m\times F0$ and l is a smoothing range. Finally, a modified background music is obtained by multiplying the background music eliminating harmonic information by the smoothing gain factor, $\tilde{B}(k)=B'(k)G_S(k)$.

C. Bit-rate reduction

Even though the SAOC scheme has lower bit rates than transmitting all object signals for the object control, it is important to reduce the bit rate of transmitted harmonic information in the aspect of the coder. To reduce the bit rate, we propose a new bit reduction scheme, in which an approximation of harmonic magnitudes is performed using statistical modeling. The proposed method transmits one fundamental frequency and several harmonic magnitudes. The several harmonic magnitudes can be modeled by statistical analysis because the extracted harmonic magnitudes show a decreasing trend as the frequency increases. To make the proposed algorithm operated, firstly, the harmonic magnitudes of each frame are normalized by maximum harmonic magnitude of each frame. Next, the average of the normalized magnitudes is calculated at each index. If the statistical model of harmonic magnitudes is obtained in encoder and transmitted to decoder as the header information, several harmonic magnitudes can be reconstructed using only one first harmonic magnitude of each frame. The performance of this statistical modeling will be shown minutely in section IV.

IV. Experiment and Results

A. Experimental setup and measurements

For the performance evaluation, 5 popular Korean pop songs, listed in TABLE II, were used. Each item composed of 4-6 audio objects, such as voice and other musical instruments such as guitar, piano, and drum. All objects were sampled at 44.1 kHz with 16 bit quantization level. Overlapping analysis window size and shifting size were 2048 and 1024, respectively. 2048 point FFT was executed and 20 harmonics were extracted for the harmonic elimination [3]. The OLD used 4 bit quantization and each fundamental frequency and harmonic magnitudes allocate 5 bits.

Firstly, we measured the bit-rate of side information. Secondly, we used the symmetric Kullback-Leibler distance (SKLD) and segment signal-to-noise ratio (SEGSNR) as the objective distortion measures between the reference and the test data. The SKLD was calculated by

$$SKLD = 10\log\left(\sum_{\omega\in L}\left(P(\omega)-Q(\omega)\right)\log\frac{P(\omega)}{Q(\omega)}\right) \tag{15}$$

where $P(\omega)$ and $Q(\omega)$ denote the power spectra of the reference and the decoded signal, respectively. In addition, we obtain the SEGSNR by using

$$SEGSNR = 10\log\left(\frac{\sum_{n\in L} p(n)^2}{\sum_{n\in L}\left(p(n)-q(n)\right)^2}\right) \tag{16}$$

TABLE II
TEST CONTESTS

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

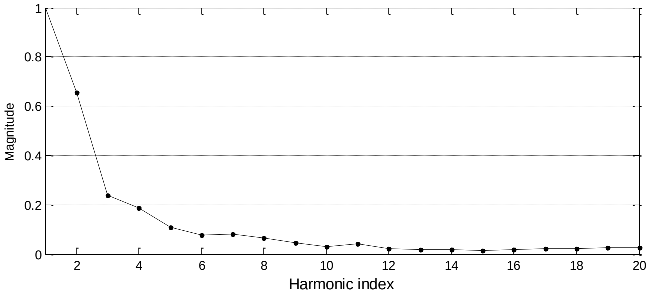


Fig. 7. Statistical modeling of harmonic magnitudes

where $p(n)$ and $q(n)$ are the sample values of the reference and the decoded signal, respectively. As a subjective listening test, the multiple stimuli with hidden reference and anchor (MUSHRA) test was performed [14]. The MUSHRA test is well known as a subjective test among audio quality test and includes the hidden reference signal and 3.5 kHz band limited anchor signal. Ten experienced listeners evaluated the background music quality of the test contests. The reference signal is an original background music which consists of all objects except the vocal object. The proposed methods are compared with other methods which are the SAOC methods and the research in [15], [16] based on the NMF method. The SAOC A and B are the normal Karaoke application and enhanced Karaoke application, respectively. The proposed method A does not use a harmonic modeling and the proposed method B applies the harmonic modeling to reduce bit rates. The example of statistical modeling is shown in Fig. 7.

TABLE III
RESULTS OF OBJECTIVE TESTS

Method	SKLD (dB)	SEGSNR(dB)
SAOC A	33.81	20.91
SAOC B	24.78	25.16
NMF	34.57	18.82
Proposed A	26.34	23.76
Proposed B	27.86	23.07

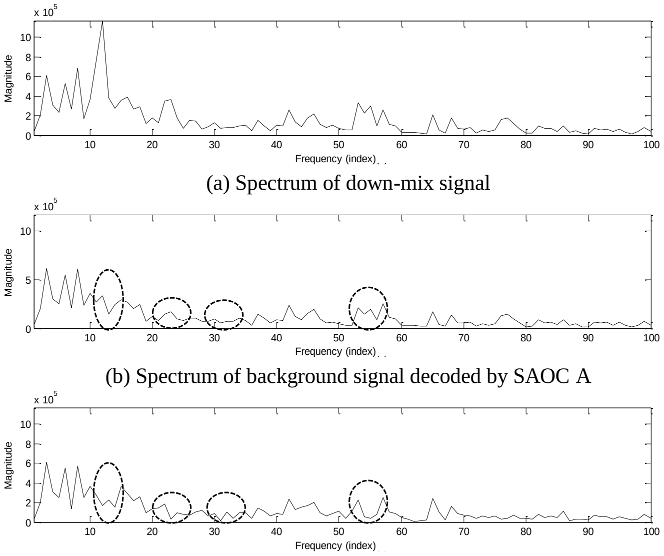
TABLE IV
NUMBER OF TRANSMITTED PARAMETER PER FRAME AND BIT-RATE ALLOCATION (EXAMPLE: 5 OBJECTS)

Method	Transmitted parameters	Bit-rate (kbps)
SAOC A	5 OLD	12.05
SAOC B	5 OLD, residual signal	42.05
NMF	1 F0	0.11
Proposed A	5 OLD, 1 F0, 21 magnitude	14.41
Proposed B	5 OLD, 1 F0	14.27

B. Experimental results

Fig. 9 represents spectrum of original background signal, decoded by the SAOC, the proposed method A, and the proposed method B. The results of the proposed method B prove that the harmonic information can be modeled well by low bit-rate allocation for the proposed methods. In overall, we can see that the proposed methods guarantee the sound quality, the sufficient amount of the vocal object removal, and the relatively low bit-rate allocation.

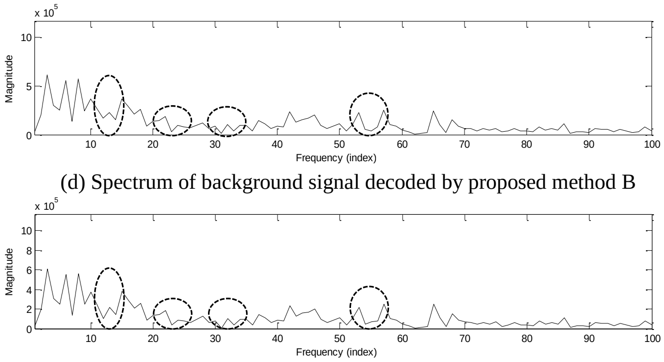
The results of the bit rates and the objective tests were listed in TABLE III and IV. As shown in TABLE III, the SKLD gain of 24.78 dB is achieved by the SAOC A while 26.34 dB and 27.86 dB are achieved by proposed method A and proposed method B, respectively. And the highest SEGSNR gain of 25.16 dB is obtained by the SAOC B. Fig. 9 shows a result of the MUSHRA test. The SAOC B also gets the best score among all test methods. As shown in TABLE III and Fig. 9, the results of the objective and the subjective tests can be summarized as follows: the SAOC methods and proposed methods shows better performance than the NMF method for all experiments of both the objective and the subjective tests. The SAOC B shows the best score among all tests and allocates the highest bit rate 42.05 kbps as shown TABLE IV.



(a) Spectrum of down-mix signal

(b) Spectrum of background signal decoded by SAOC A

(c) Spectrum of background signal decoded by proposed method A



(d) Spectrum of background signal decoded by proposed method B

(e) Spectrum of original background signal

Fig. 8. Frequency spectrum comparison (dotted circle: view point of comparing harmonic removal)

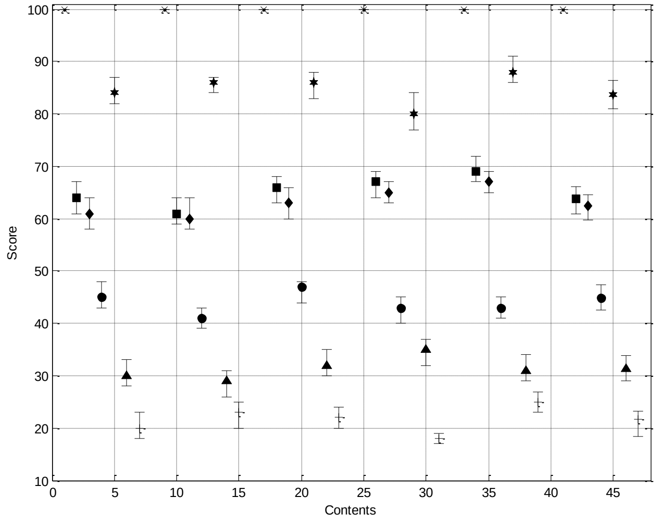


Fig. 9. Result of MUSHRA test (x-mark: hidden reference, square: proposed method A, diamond: proposed method B, circle: SAOC A, hexagram: SAOC B, triangle: NMF, minus: anchor)

And the proposed method A and B record higher and lower score than the SAOC A and the SAOC B, respectively. 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 methods. Additionally, the proposed method A and B have very similar result throughout the tests.

V. Conclusion

The SAOC is a useful technology that can support most parts of the interactive audio service with relatively low bit rate, but it is very weak to gain perfect control of a particular audio object, i.e., the vocal object for the Karaoke application. 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. Moreover, transmitted harmonic magnitudes are approximate by statistical modeling to reduce bit rates while keeping the sound quality. As a future work, we will study with respect to designing a transfer coder which operates between the encoder and the decoder for the object control like object removing, inserting, changing, etc.

Acknowledgement

“This work was supported by the Center for Integrated Smart Sensors funded by the Ministry of Science, ICT & Future Planning as Global Frontier Project" (CISS-2011-031863)

References

[1] D. Jang, T. Lee, Y. Lee, J. Yoo, “A personalized preset-based audio system for Interactive service,” 121st AES Convention, San Fransisco 2006.

[2] ISO/IEC JTC1/SC29/WG11 (MPEG), Consideration of interactive music service, Document M15390, Archamps, Apr. 2008.

[3] C. Faller and R. Baumgarte, “Binaural cue coding-part II: schemes and application,” *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, pp. 520-531, Nov. 2003.

[4] ISO/IEC 23003-1, Information technology-MPEG audio technologies - part 1: MPEG surround, 2007.

[5] J. Herre, S. Disch, “New concepts in parametric coding of spatial audio: from SAC to SAOC,” 2007 International Conference on Multimedia and Expo, pp. 1894-1897, Jul. 2007.

[6] ISO/IEC JTC1/SC29/WG11 (MPEG), Call for proposals on spatial audio object coding, Document N8853, Jan. 2007.

[7] ISO/IEC JTC1/SC29/WG11 (MPEG), Study on ISO/IEC 23003-2:200X, Spatial audio object coding, Document N10659, Maui, Apr. 2009.

[8] J. Breebaart, J. Engdegard, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, J. Koppens, W. Oomen, B. Resch, E. Schuijers, L. Terentiev, “Spatial audio object coding (SAOC) - the upcoming MPEG standard on parametric object based audio coding,” 124th AES Convention, Amsterdam 2008.

[9] ISO/IEC JTC1/SC29/WG11 (MPEG), Proposed improvement for MPEG SAOC, Document M14985, Shenzen, Oct. 2007.

[10] A.P. Klapuri, “Multiple fundamental frequency estimatiopn by summing harmonic amplitudes,” International Conference on Music Information Retrieval, pp.216-212, 2006.

[11] M. Wu, D. Wang, G. J. Brown, “A multipitch tracking algorithm for noisy speech,” *IEEE Trans. Speech and Audio Proc.*, vol. 11, no. 3, pp. 229-241, 2003.

[12] M. Goto, “A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals,” *Speech Com.*, vol.43, no. 4, pp. 311-329, 2004.

[13] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, H. G. Okuno, “Automatic synchronization between lyrics and music CD recordings based on viterbi alignment of segregated vocal signals,” in IEEE International Symposium on Multimedia, pp. 257-264, 2006.

[14] ITU-R Recommendation, Method for the subjective assessment of intermediate sound quality (MUSHRA), ITU, BS. 1543-1, Geneva, 2001.

[15] J. S. Lee, D. D. Lee, S. Choi, and D. S. Lee, “Application of nonnegative matrix factorization to dynamic positron emission tomography,” in Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation, San Diego, California, pp. 629–632, 2001.

[16] T. Virtanen, A. Mesaros, and M. Ryynanen, “Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music,” ISCA Tutorial Res. Workshop Statist. Percept. Audition 2008, pp. 17-22, 2008

BIOGRAPHIES

Jihoon Park received the B.S. degree in electronics engineering from Information and Communications University, Daejeon, South Korea, in 2005, and M.S. degree in electronic engineering from Information and Communications University, Daejeon, South Korea, in 2007. He is currently working toward Ph.D. degree at Korea Advanced Institute of Science and Technology, Daejeon, South Korea. His research interests include microphone array-based speech enhancement, HMM-based speech synthesis, VoIP, multi-channel audio coding, multi-object audio coding and ultra high definition audio coding.