Contemporary Engineering Sciences, Vol. 9, 2016, no. 1, 11 - 19
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2016.512315

A Study on Complexity Reduction of Binaural Decoding in Multi-channel Audio Coding for Realistic Audio Service

Kwangki Kim

Korea Nazarene University, Department of Digital Contents
Wolbong-ro 48, Cheonan Chungnam, Korea

Abstract

In this paper, we proposed the simplified binaural decoding method for reducing the complexity of the binaural decoding. In the proposed simplified binaural decoding the high frequency components of the HRTF (head related transfer function) coefficients are excluded and the binaural decoding process in the high frequency regions is simplified. From the experimental results, it is confirmed that the proposed method greatly reduces the complexity of the binaural decoding in the frequency domain as 40 % and shows the statistically same sound quality compared to the binaural decoding in the frequency domain.

Keywords: binaural decoding, multi-channel audio, spatial cue, down-mix signal, backward compatibility

1 Introduction

Recently, with increase of realistic 3D videos such as 3DTV, UHDTV (Ultra High Definition TV) and 3D movies, a realistic audio sound is getting more important in the area of audio service. The realistic audio sound can be generated by not stereo audio signals but more than 5.1 channel audio signals, and audio signals with more channels can make more realistic and immersive audio sound. But, as the data rate of multi-channel audio signals increases in proportion to the number of the audio channel, the multi-channel audio signals cannot be directly

provided through the wired and wireless network system. To solve the high bit-rate problem of the multi-channel audio signals, a spatial cue based multi-channel audio coding such as BCC (binaural cue coding), MPEG Surround, and SSLCC (sound source location coefficient coding) has been proposed and developed [1-4]. As the spatial cue based multi-channel audio coding represents the multi-channel audio signals as a down-mix signal and additional side information, the data rate of the multi-channel audio signals can be significantly reduced. So, the multi-channel audio signals can be efficiently delivered to users through the network system. Generally, the spatial cue based multi-channel audio coding has a unique functionality, called backward compatibility. With the backward compatibility, users can enjoy the down-mix signal using their stereo playback system if they do not have a multi-channel audio coder or they just want to play the down-mix signal [5]. But, as the down-mix signal cannot realize the 3D audio sound generated by the multi-channel audio signals, the backward compatibility of the spatial cue based multi-channel audio coding should be enhanced. From this reason, the binaural decoding can be applied to enhance the backward compatibility of the spatial cue based multi-channel audio coding by adding the multi-channel audio effect to the down-mix signal. The binaural decoding generates the binaural stereo sound by convolving the multi-channel audio signal with HRTF (head related transfer function) coefficients. Basically, the binaural decoding has very high complexity due to the linear convolution process in time domain. So, the binaural decoding has a limitation that the real time implementation of the binaural decoding is impossible. For the real time implementation of the binaural decoding, the binaural decoding in the SSLCC by convolving the HRTF coefficients and the multi-channel audio signals in the SSLCC synthesis domain, i.e., frequency domain was proposed [6]. Although the binaural decoding in the frequency domain successfully reduced the complexity, the binaural decoding of the SSLCC still has the rather high complexity. In this paper, we proposed a simplified binaural decoding to consist of envelope and phase modifications in the frequency domain.

2 Overview of the SSLCC

A structure of the SSLCC is depicted in Fig.1. The SSLCC encoder represents input multi-channel audio signals as the down-mix signal with additional side information. The SSLCC decoder recovers the multi-channel audio signals using the transmitted down-mix signal and the side information. A detailed process of the SSLCC encoder is shown in Fig. 2. Firstly, the input multi-channel audio signals are transformed into the frequency domain by the discrete time Fourier transform (DFT) and then they are inputted to the SSLCC analyzer for extracting the spatial parameters. Virtual source location information (VSLI) is used as the spatial parameters and it indicates a spatial image in the free space to be generated by the multi-channel audio signals. The extracted spatial parameters are quantized for the transmission. In addition, the multi-channel audio signals are summed for generating the down-mix signal.

A detailed process of the SSLCC decoder is shown in Fig. 3. Firstly, the down-mix signal is transformed into the frequency domain and the received spatial parameters are dequantized. Then, the down-mix signal and the dequantized spatial parameters are inputted into the SSLCC synthesizer for recovering the multi-channel audio signals in the frequency domain. The reconstructed multi-channel audio signals in the frequency domain are transformed into the output signals in the time domain by the inverse DFT. The detailed description of the SSLCC analysis and the synthesis can be found in [3], [4].

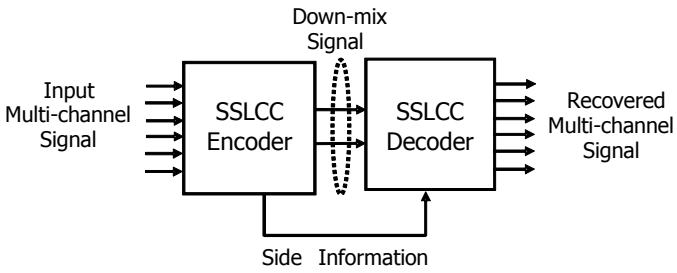

Fig. 1. Basic structure of SSLCC

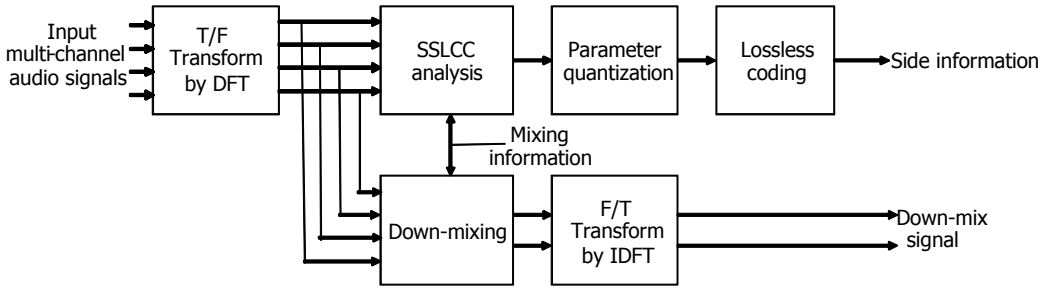

Fig. 2. Procedure of SSLCC encoder

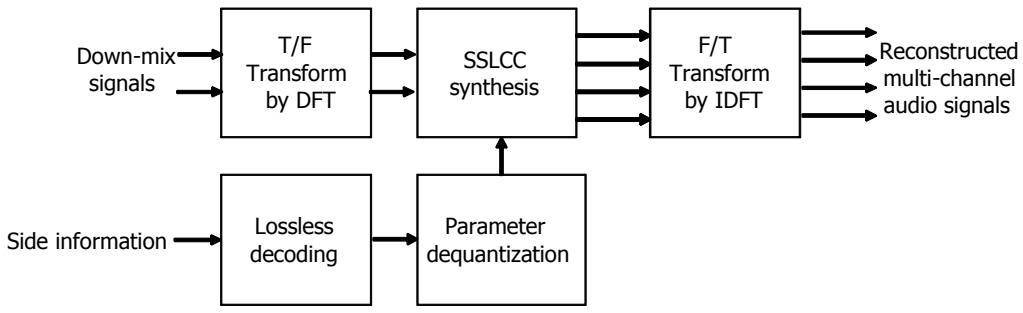

Fig. 3. Procedure of SSLCC decoder

3 Binaural Decoding in the SSLCC

Since the binaural decoding in the multi-channel audio coding has high computational loads of the linear convolution between the multi-channel audio

signals and the HRTF coefficients in the time domain, the binaural decoding cannot avoid the complexity problem and it cannot be implemented in the real time. To resolve the complexity problem, the simplified binaural decoding performed in the frequency domain was proposed in [6] and it is shown in Fig. 4. The HRTF coefficients are transformed into the frequency domain by the DFT and they are stored in the memory. The gain factors of the multi-channel audio signals are estimated using the side information in the frequency domain and they are convolving with the HRTF coefficients in the frequency domain.

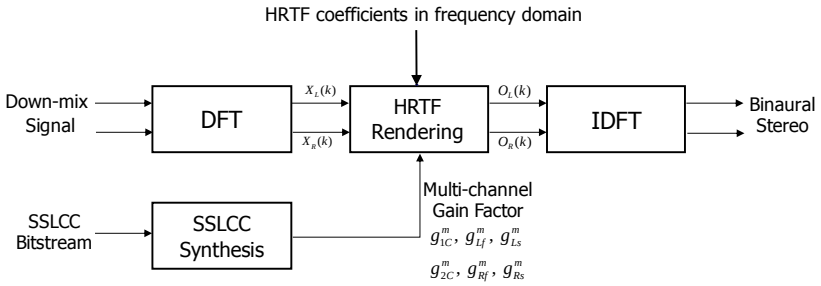

Fig. 4. Binaural decoding in SSLCC (Lf: left front, Ls: left surround, Rf: right front, Rs: right surround, C: center)

Using the down-mix signal in frequency domain, $X_L(k), X_R(k)$, the calculated he multi-channel audio signals in the frequency domain, $g_{1C}(k), g_{2C}(k), g_{Lf}(k), g_{Ls}(k), g_{Rf}(k), g_{Rs}(k)$, and the stored HRTF coefficients in frequency domain, $H_C^L(k), H_C^R(k), H_{Lf}^L(k), H_{Lf}^R(k), H_{Ls}^L(k), H_{Ls}^R(k), H_{Rf}^L(k), H_{Rf}^R(k), H_{Rs}^L(k), H_{Rs}^R(k)$, the binaural rendering can be performed as

$$\begin{aligned} H_{LL}(k) &= g_{1C}(k)\cdot H_C^L(k) + g_{Lf}(k)\cdot H_{Lf}^L(k) + g_{Ls}(k)\cdot H_{Ls}^L(k) \\ H_{RL}(k) &= g_{2C}(k)\cdot H_C^L(k) + g_{Rf}(k)\cdot H_{Rf}^L(k) + g_{Rs}(k)\cdot H_{Rs}^L(k) \\ H_{LR}(k) &= g_{1C}(k)\cdot H_C^R(k) + g_{Lf}(k)\cdot H_{Lf}^R(k) + g_{Ls}(k)\cdot H_{Ls}^R(k) \\ H_{RR}(k) &= g_{2C}(k)\cdot H_C^R(k) + g_{Rf}(k)\cdot H_{Rf}^R(k) + g_{Rs}(k)\cdot H_{Rs}^R(k) \end{aligned} \quad (1)$$

where $H_{LL}(k)$ and $H_{LR}(k)$ are elements for left and right binaural output by center, left front, and left surround channels, respectively, while $H_{RL}(k)$ and $H_{RR}(k)$ are HRTF rendering elements for elements for left and right binaural output by center, right front, and right surround channels, respectively. Here, k indicates the frequency index. At last, the binaural output signals can be obtained as

$$\begin{bmatrix} O_L(k) \\ O_R(k) \end{bmatrix} = \begin{bmatrix} H_{LL}(k) & H_{RL}(k) \\ H_{LR}(k) & H_{RR}(k) \end{bmatrix} \cdot \begin{bmatrix} X_L(k) \\ X_R(k) \end{bmatrix} \quad (2)$$

where $O_L(k)$ and $O_R(k)$ are the left and right binaural output signals, respectively.

4 Proposed Simplified Binaural Decoding in the SSLCC

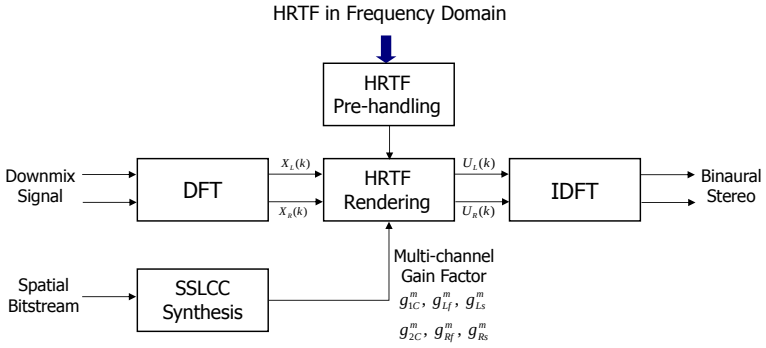

Fig. 5. Overall structure of the simplified binaural decoding in the SSLCC

The proposed simplified binaural decoding in the SSLCC consists of the envelope and the phase modifications in the frequency domain. The HRTF coefficients are pre-handled in the frequency domain to reflect human hearing property that the human hearing is insensitive to high frequency regions [7]. Therefore, the high frequency components of the HRTF coefficients can be excluded and the binaural decoding process in the high frequency regions can be skipped or simplified. Fig. 5 shows the overall structure of the proposed simplified binaural decoding in the SSLCC.

In the proposed simplified binaural decoding method, the HRTF coefficients are pre-transformed into those in the frequency domain and amplitude and phase information are calculated using them. Then, the amplitude information of the HRTF coefficients is totally stored and the phase information of the HRTF coefficients to be below 3.5 kHz are selectively stored. As the human hearing is sensitive to the phase information of the low frequency regions while being insensitive to those of the high frequency regions, we can exclude the phase information of the high frequency components of the HRTF coefficients and the HRTF rendering of the phase information in the high frequency regions can be skipped.

Using the pre-handled and stored HRTF coefficients in the frequency domain, the HRTF rendering, i.e. the envelope and phase modification, can be simply performed using the modified (1) and (2). At first, (1) is divided into the following (3) and (4).

$$\left.\begin{aligned} H_{LL}(k) &= g_{1C}(k)\cdot H^L_C(k) + g_{Lf}(k)\cdot H^L_{Lf}(k) + g_{Ls}(k)\cdot H^L_{Ls}(k) \\ H_{RL}(k) &= g_{2C}(k)\cdot H^L_C(k) + g_{Rf}(k)\cdot H^L_{Rf}(k) + g_{Rs}(k)\cdot H^L_{Rs}(k) \\ H_{LR}(k) &= g_{1C}(k)\cdot H^R_C(k) + g_{Lf}(k)\cdot H^R_{Lf}(k) + g_{Ls}(k)\cdot H^R_{Ls}(k) \\ H_{RR}(k) &= g_{2C}(k)\cdot H^R_C(k) + g_{Rf}(k)\cdot H^R_{Rf}(k) + g_{Rs}(k)\cdot H^R_{Rs}(k) \end{aligned}\right\} \; for\ 0 \le k \le L \tag{3}$$

$$\left.\begin{aligned} H_{LL}(k) &= g_{1C}(k)\cdot\left|H_C^L(k)\right| + g_{Lf}(k)\cdot\left|H_{Lf}^L(k)\right| + g_{Ls}(k)\cdot\left|H_{Ls}^L(k)\right| \\ H_{RL}(k) &= g_{2C}(k)\cdot\left|H_C^L(k)\right| + g_{Rf}(k)\cdot\left|H_{Rf}^L(k)\right| + g_{Rs}(k)\cdot\left|H_{Rs}^L(k)\right| \\ H_{LR}(k) &= g_{1C}(k)\cdot\left|H_C^R(k)\right| + g_{Lf}(k)\cdot\left|H_{Lf}^R(k)\right| + g_{Ls}(k)\cdot\left|H_{Ls}^R(k)\right| \\ H_{RR}(k) &= g_{2C}(k)\cdot\left|H_C^R(k)\right| + g_{Rf}(k)\cdot\left|H_{Rf}^R(k)\right| + g_{Rs}(k)\cdot\left|H_{Rs}^R(k)\right| \end{aligned}\right\} \text{for } L+1 \le k \le N-1 \quad (4)$$

Here, L is the frequency bin index of 3.5 kHz and N is the frame size. (3) is the HRTF rendering for the frequency regions to be below 3.5 kHz while (4) is the HRFT rendering for the high frequency regions to be beyond 3.5 kHz. Therefore, for the low frequency regions, both the envelope and the pahse information are used for the binaural decoding. Whereas, for the high frequency regions, only the envelope information is used for the binaural decoding.

5 Experimental Results

Table 1. Complexity comparison

Classification	By DFT	By convolution	Reduction
Decoded multi-channel signals to binaural output (in time domain)	5 x N log2N	10 x (N x N multiplications + N x N summations)	100 %
HRTF rendering in frequency domain	2 x 2N log22N	2 x (28N multiplications + 28N summations)	about 10 %
HRTF rendering in frequency domain with pre-handled HRTF (spectral envelope shaping)	2 x 2N log22N	28N multiplications + 28N summations	about 5 %
HRTF rendering in frequency domain with pre-handled HRTF (spectral envelope shaping+phase modification)	2 x 2N log22N	1.15 x (28N multiplications + 28N summations)	about 6 %

To validate the performance of the proposed simplified binaural decoding, we checked the complexity of various binaural decoding methods and performed a subjective listening test. Firstly, Table 1 shows the complexity comparison results. The HRTF rendering in the frequency domain can reduce the complexity of the typical binaural decoding in the time domain as 90 %. In addition, the proposed simplified HRTF rendering can reduce the complexity of the HRTF rendering in the frequency domain as 40 %.

For the subjective test, three multi-channel audio contents were used and they are listed in Table 2 [8]. The items were sampled at 44.1 kHz with 16 bit resolution and have the duration of 20 seconds. An MUSHRA test was performed [9] and four systems were used for the test and they are listed in Table 3.

Table 2. Test materials

Material	Description
ARL_applause	Ambience
Chostakovitch	Music (back: direct)
Fountain_music	Pathological

Table 3. System under test

Classification	Description
REF	Reference signal generated with the original signals
DFT	HRTF rendering in frequency domain
ENV	HRTF rendering in frequency domain with pre-handled HRTF and only envelope modification
ENV+PHA	Proposed simplified HRTF rendering. ENV and phase modification for the low frequency regions to be below 3.5 kHz

Fig. 6 shows the subjective listening test results. For all test items, 'DFT' and 'ENV+PHA' shows the good sound quality while 'ENV' has very poor sound quality. Although 'ENV+PHA' is slightly low absolute score than 'DFT', 'DFT' and 'ENV+PHA' have the statistically same sound quality. From the experimental results, it is confirmed that the proposed simplified binaural decoding method can successfully reduce the complexity while maintaining the good sound quality.

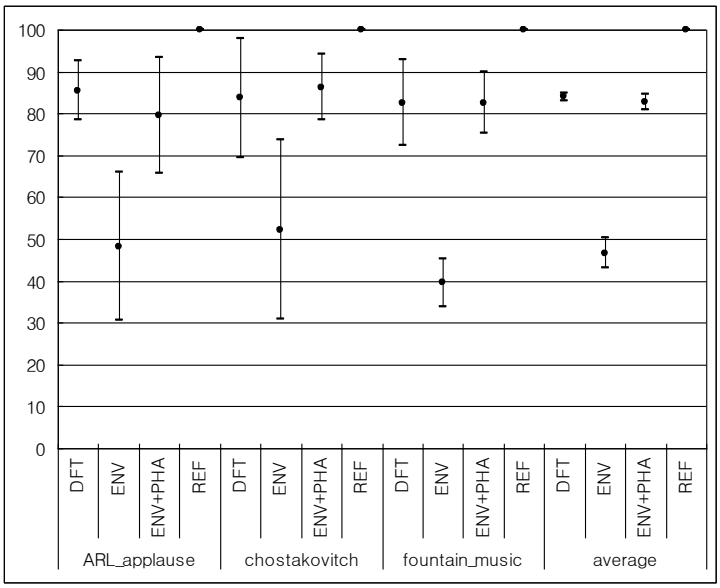

Fig. 6. Subjective listening test results

6 Conclusion

In this paper, we proposed the simplified binaural decoding method for reducing the complexity of the binaural decoding. In the proposed simplified binaural

decoding the high frequency components of the HRTF coefficients are excluded and the binaural decoding process in the high frequency regions is simplified. From the experimental results, it is confirmed that the proposed method greatly reduces the complexity of the binaural decoding in the frequency domain as 40 % and shows the statistically same sound quality compared to the binaural decoding in the frequency domain. As the future work, the binaural decoding method for more than 5.1 channel audio signals, i.e. ultra multi-channel audio environment, will be studied.

Acknowledgements. This study was funded by the research fund of Korea Nazarene University in 2015.

References

[1] ISO/IEC 23003-1, Information Technology–MPEG Audio Technologies – Part 1: MPEG Surround, (2007).

[2] C. Faller and F. Baumgarte, Binaural cue coding – Part II: Schemes and Applications, *IEEE Trans. Speech Audio Processing*, **11** (2003), no. 6, 520-531. http://dx.doi.org/10.1109/tsa.2003.818108

[3] Han-gil Moon, Jeong-il Seo, et al., A multi-channel audio compression method with virtual source location information for MPEG-4 SAC, *IEEE Transactions on. Consumer Electronics*, **51** (2005), no. 4, 1253- 1259. http://dx.doi.org/10.1109/tce.2005.1561852

[4] Seungkwon Beack, Jeongil Seo, et al., Angle-Based Virtual Source Location Representation for Spatial Audio Coding, *ETRI Journal*, **28** (2006), no. 2, 219-222. http://dx.doi.org/10.4218/etrij.06.0205.0079

[5] Kwangki Kim, Minsoo Hahn and Jinsul Kim, Mastering Signal Processing in MPEG SAOC, *IEICE Transactions on Information and Systems*, **E95.D** (2012), no. 12, 3053-3059. http://dx.doi.org/10.1587/transinf.e95.d.3053

[6] Kwangki Kim and Jinsul Kim, Binaural decoding for efficient multi-channel audio service in network environment, *2014 IEEE 11th Consumer Communications and Networking Conference*, (2014), 525-526. http://dx.doi.org/10.1109/ccnc.2014.6994429

[7] E. Zwicker and H. Fastl, *Psychoacoustics*, Springer-Verlag, Berlin, Heidelberg, 1999. http://dx.doi.org/10.1007/978-3-662-09562-1

[8] ISO/IEC JTC1/SC29/WG11 (MPEG), Procedures for the Evaluation of Spatial Audio Coding Systems, Document N6691, Redmond, 2004.

[9] ITU-R Recommendation, Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA), ITU, BS. 1543-1, Geneva, 2001.

Received: December 15, 2015; Published: December 29, 2015