Real-Time Conversion of Stereo Audio to 5.1 Channel Audio for Providing Realistic Sounds

Chan Jun Chun, Yong Guk Kim, Jong Yeol Yang, and Hong Kook Kim

Department of Information and Communications
Gwangju Institute of Science and Technology, Gwangju, Korea
{cjchun, bestkyg, jyyang, hongkook}@gist.ac.kr

Abstract

In this paper, we address issues associated with the real-time implementation of upmixing stereo audio into 5.1 channel audio in order to improve audio realism. First, we review four different upmixing methods, including a passive surround decoding method, a least-mean-square based upmixing method, a principal component analysis based upmixing method, and an adaptive panning method. After that, we implement a simulator that includes the upmixing methods and audio controls to play both stereo and upmixed 5.1 channel audio signals. Finally, we carry out a MUSHRA test to compare the quality of the upmixed 5.1 channel audio signals to that of the original stereo audio signal. It is shown from the test that the upmixed 5.1 channel audio signals generated by the four different upmixing methods are preferred to the original stereo audio signals.

Keywords: *Audio Upmixing, Stereo Audio, Multi-channel Audio, Passive Surround Decoding, Least Mean Square, Principal Component Analysis, Adaptive Panning*

1. Introduction

As technologies related to audio systems have advanced, the demand for multi-channel audio systems has increased too. Such audio systems not only provide more realistic sounds, but also offer more ambient effects than standard stereo audio systems. For example, if audio content having fewer channels than can be provided by a target system is available, the target audio system cannot take full advantage of it. Therefore, in order to utilize such audio content, it is necessary to use an upmixing method that converts mono or stereo audio formats into a multi-channel audio format suitable for such a system.

Many multi-channel audio systems currently exist with a wide range in the number of channels available. Since 5.1 channels are better for creating the effects of ambience and spaciousness than stereo channels, we need to develop upmixing methods that convert audio from a stereo format to a 5.1 channel format. Since one of the typical approaches for creating additional channels is to use a correlation property between stereo channels, we first review and implement four correlation-based methods; a passive surround decoding method [1], a least- mean-square based method [2], a principal component analysis based method [2], and an adaptive panning method [3]. We then compare the upmixed audio contents obtained by each of the four different methods to the original stereo audio contents.

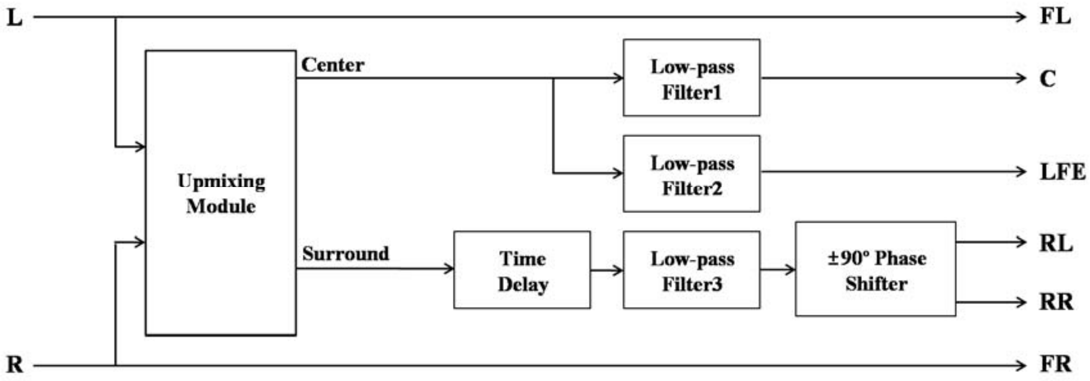


Figure 1. Procedure for upmixing a stereo audio to a 5.1 channel audio

The remainder of this paper is organized as follows. Following this introduction, we review four different upmixing methods for converting audio from a stereo format to a 5.1 channel format in Section 2. After that, we design a simulator for upmixing stereo contents using these methods in Section 3. In Section 4, the quality of the upmixed 5.1 channel audio is compared with that of the original stereo audio by performing a multiple stimuli with hidden reference and anchor (MUSHRA) test [4]. Finally, we conclude this paper in Section 5.

2. Audio Upmixing Algorithm from Stereo to 5.1 Channel Audio

In this section, we describe four different upmixing methods for converting audio from a stereo format to a 5.1 channel format. Figure 1 shows the upmixing procedure, where the channels are labeled FL (front left), FR (front right), C (center), LFE (low frequency enhancement), RL (rear left), and RR (rear right). As illustrated in the figure, the FL and the FR channels for the 5.1 channel audio format are directly obtained from the original stereo channels, while the remaining channels are generated from the center channel and the surround channels. Therefore, it is discussed in the following subsections how to derive the center channel and the surround channels by using each upmixing method.

2.1. Passive Surround Decoding Method

The passive surround decoding (PSD) method is an early passive version of the Dolby Surround Decoder [1]. In this method, the center channel is obtained by adding the original left and right channels. On the other hand, the surround channel can be derived by subtracting the right channel from the left channel. Note that in order to maintain a constant acoustic energy, the center and the surround channel are lowered by 3 dB, which is implemented by multiplying $1/\sqrt{2}$ to the center and the surround channels. That is, the center and the surround channels are obtained by using the equations of

$$Center(n) = (x_L(n) + x_R(n))/\sqrt{2}\,, \tag{1}$$

$$Surround(n) = (x_L(n) - x_R(n))/\sqrt{2}\,, \tag{2}$$

where $x_L(n)$ and $x_R(n)$ represent the left and the right samples at the time index n, respectively.

2.2. LMS-based Upmixing Method

The least-mean-square (LMS)-based upmixing method creates the center and surround channels using the LMS algorithm [2][5]. In this method, one of the original stereo channels is taken as the desired signal, $d(n)$, and the other is considered as the input, $x(n)$, of the adaptive filter. The error signal, $e(n)$, is then the difference of the output, $y(n)$, of the filter and the desired signal, $d(n)$. The output, $y(n)$, is defined as a linear combination of the input signals by using the equation of

$$y(n) = \mathbf{w}^T(n)\mathbf{x}(n) = \mathbf{w}(n)\mathbf{x}^T(n)\,, \tag{3}$$

where $\mathbf{x}(n) = [x(n)\; x(n-1) \cdots x(n-N+1)]^T$ and $\mathbf{w}(n) = [w_0\; w_1\; \cdots\; w_{N-1}]^T$. In Equation (3), $\mathbf{w}(n)$ is a coefficient vector of the N-tapped adaptive filter and is obtained based on the LMS algorithm described as

$$\mathbf{w}(n+1) = \mathbf{w}(n) + 2\mu\, e(n)\mathbf{x}(n)\,, \tag{4}$$

where μ is a constant step size which is set to 10^{-4} in this paper. In this case, $y(n)$ and $e(n)$ are considered as the signals for the center channel and the surround channel, respectively.

2.3. PCA-based Upmixing Method

The principal component analysis (PCA)-based upmixing method decomposes the original stereo channels into correlated and uncorrelated portions [2]. In order to derive the center and the surround channels, we first need to find a 2x2 covariance matrix, A, such that

$$\mathbf{A} = \begin{bmatrix} \mathrm{cov}(x_L, x_L) & \mathrm{cov}(x_L, x_R) \\ \mathrm{cov}(x_R, x_L) & \mathrm{cov}(x_R, x_R) \end{bmatrix}, \tag{5}$$

where $\mathrm{cov}(x_p, x_q)$ is the covariance of x_p and x_q where p (or q) could be the left or the right channel. The covariance matrix, A, gives two eigenvectors which are the basis vectors for a new coordinate system [6]. These eigenvectors are then used as weight vectors corresponding to the left and right channels to generate the center and the surround channels such as

$$Center(n) = c_L x_L(n) + c_R x_R(n)\,, \tag{6}$$

$$Surround(n) = s_L x_L(n) + s_R x_R(n)\,. \tag{7}$$

The eigenvector $[c_L \; c_R]$ corresponding to the greatest eigenvalue becomes the weight vector for the center channel. Thus, the other eigenvector $[s_L \; s_R]$ becomes the weight vector for the surround channel. Typically, the PCA-based method is implemented as frame-wise processing that may cause unwanted artifacts at the frame boundaries. These artifacts can be mitigated using an overlap-and-add technique [7]. For the overlap-and-add technique, analysis and synthesis windows that perfectly satisfy the overlap-and-add reconstruction condition are required. In this paper, we choose the same window for the analysis and the synthesis, which is defined as

$$w(n) = \begin{cases} \sin\left(\dfrac{\pi(n+\frac{1}{2})}{2(N-M)}\right), & 0 \le n \le N-M-1 \\ 1, & N-M \le n \le M-1 \\ \sin\left(\dfrac{\pi(N-n-\frac{1}{2})}{2(N-M)}\right), & M \le n \le N-1 \end{cases}, \tag{8}$$

where N is the length of a frame, and M is the overlap region. Here, N and M are set to 1024 and 128, respectively.

2.4. Adaptive Panning Method

The adaptive panning (ADP) method proposed in [3] generates the center and the surround channels by panning the original stereo channels. The weight vector for ADP is recursively estimated using the LMS algorithm. Let us now define $y(n)$ as a linear combination of the original stereo channels as

$$y(n) = \mathbf{w}^T(n)\mathbf{x}(n) = \mathbf{w}(n)\mathbf{x}^T(n)\,, \tag{9}$$

where $\mathbf{x}(n) = [x_L(n) \;\; x_R(n)]^T$ and $\mathbf{w}(n) = [w_L(n) \;\; w_R(n)]^T$. Two coefficients, $w_L(n)$ and $w_R(n)$, which are the elements of the weight vector corresponding to the left and the right channels, respectively, are then estimated using the LMS algorithm defined as

$$w_L(n+1) = w_L(n) - \mu y(n)[x_L(n) - w_L(n)y(n)], \qquad (10)$$

$$w_R(n+1) = w_R(n) - \mu y(n)[x_R(n) - w_R(n)y(n)], \qquad (11)$$

where μ is a constant step size and is set to 10^{-10}. Finally, the center and surround channels can be determined as

$$Center(n) = w_L(n)x_L(n) + w_R(n)x_R(n), \qquad (12)$$

$$Surround(n) = w_R(n)x_L(n) - w_L(n)x_R(n). \qquad (13)$$

2.5. Low-pass Filters

For 5.1 channel audio contents such as movie, live music, voice and dialog are usually emphasized when they are played through the center channel. Therefore, the center channel is further processed by a low-pass filter, where we design a finite-duration impulse response (FIR) low-pass whose length is 256 and cut-off frequency is 4 kHz, denoted as *Low-pass Filter1* in Figure 1. On one hand, the low frequency enhancement (LFE) channel is used to emphasize low frequency region ranged from 100 to 200 Hz. To this end, an FIR low-pass filter having a cut-off frequency of 200 Hz is designed with 256 taps, which is denoted as *Low-pass Filter2* in Figure 1. In the surround channels, a low-pass filter (*Low-pass Filter3* in Figure 1) is also used to simulate a high-frequency absorption effect. An FIR low-pass filter with 256 taps and a cut-off frequency of 7 kHz is also used in this paper.

2.6. Time Delay and ±90° Phase Shifter

The rear left and the rear right channels are intended to provide ambience and spaciousness effects. A time delay element is used to provide such ambience effects, and a ±90° phase shifter is needed to present spaciousness effects. Assuming that the distance from front loudspeakers to the wall is about 2 meters and the distance from rear loudspeakers to the wall is about 1 meter, a time delay of 12 ms is applied to the surround channels. In addition, a discrete Hilbert transform is used as a phase shifter [8]. By using FIR approximations having a constant group delay, we can implement a discrete Hilbert transform. In particular, the approximation is done using a Kaiser window which is defined as

$$h(n)=\begin{cases}\dfrac{I_0(\beta(1-[(n-n_d)/n_d]^2)^{1/2})}{I_0(\beta)}\cdot\dfrac{\sin(\pi(n-n_d)/2)}{\pi(n-n_d)/2}, & 0\le n\le M\\ 0, & otherwise\end{cases}, \tag{14}$$

where M is the order of the FIR discrete Hilbert transform, and n_d is $M/2$. In this paper, M and β are set to 31 and 2.629, respectively.

3. Design and Implementation of an Audio Upmixing Simulator

In this section, we present a simulator used to implement the four upmixing methods described in the previous section. Figure 2 shows a snapshot of the simulator that is designed based on the procedure shown in Figure 1. The simulator can play both stereo and 5.1 channel audio files. Furthermore, it enables us to have 5.1 channel audio files from stereo audio files by using one of the four different upmixing methods. The simulator mainly consists of five parts as follows.

1) File information

This part offers information about the currently loaded audio file. It displays the filename, the number of channels, the sampling rate, etc.

2) Control panel

The buttons in the control panel part control how to play the audio file. We can play, pause, stop, and open audio files. In addition, we can turn up and turn down the volume being played.

3) Output mode

Once the simulator has loaded a stereo audio file, the stereo file can be played in either a stereo audio format or an upmixed 5.1 channel format. In case of 5.1 channel audio files, the simulator only plays them in the 5.1 channel format.

4) Upmixing algorithm

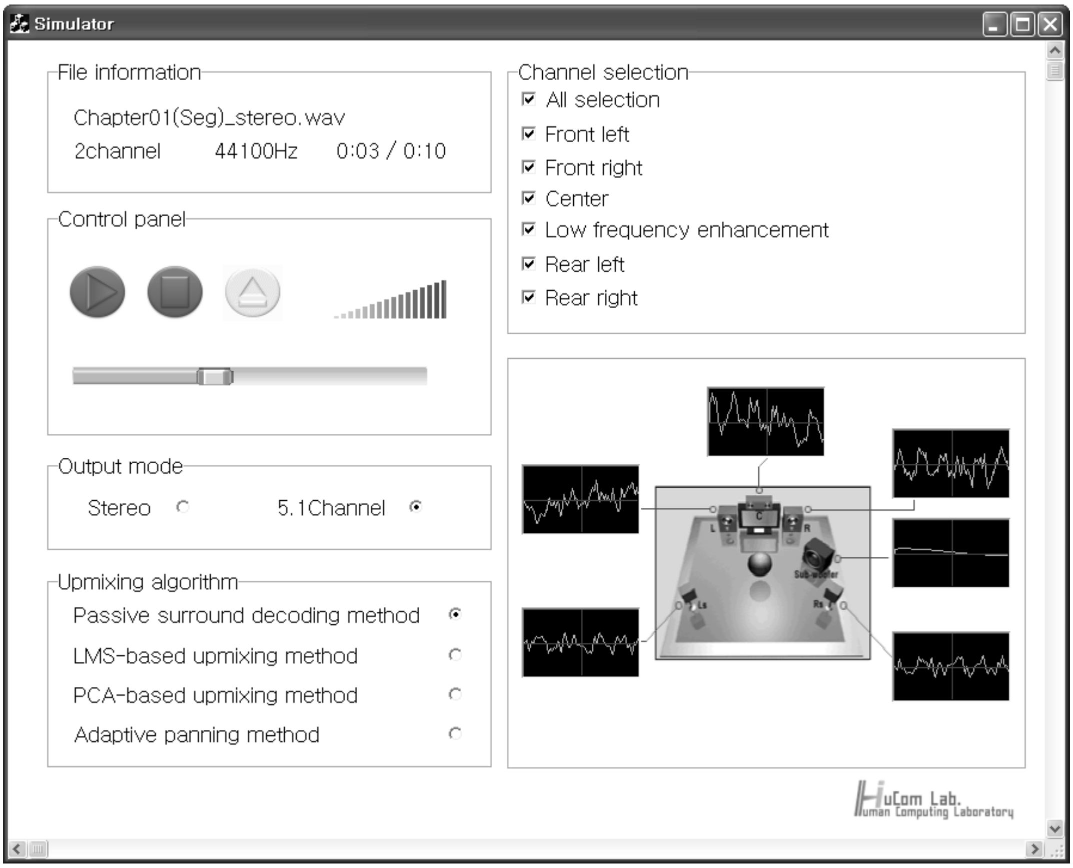
When an audio file is played after upmixing from stereo to 5.1 channel audio format, we can select one of the upmixing methods described in Section 3.

5) Channel selection

We can select the activity of each speaker.

4. Performance Evaluation

In this section, we compared the quality of the upmixed 5.1 channel signals with that of the original stereo signals. First of all, we were in full compliance with the ITU multi-channel configuration standard defined by the ITU-R Recommendation BS.775-1 [9]. A multiple stimuli with hidden reference and anchor (MUSHRA) test [4] was conducted by using five



Simulator

File information

Chapter01(Seg)_stereo.wav
2channel 44100Hz 0:03 / 0:10

Control panel

Output mode

Stereo ○ 5.1Channel ●

Upmixing algorithm

Passive surround decoding method ●
LMS-based upmixing method ○
PCA-based upmixing method ○
Adaptive panning method ○

Channel selection

☑ All selection
☑ Front left
☑ Front right
☑ Center
☑ Low frequency enhancement
☑ Rear left
☑ Rear right

Figure 2. A snapshot of the simulator for upmixing a stereo audio format to a 5.1 channel audio format in real-time

music genres such as rock, ballad, hip-hop, classical, and heavy metal. Eight people with no auditory disease participated in this experiment. For the MUSHRA test, the audio contents to be compared in the test were listed as

- Hidden reference
- 3.5 kHz low-pass filtered anchor
- 7 kHz low-pass filtered anchor
- Stereo audio content containing the front left and the front right channels of the hidden reference
- Upmixed audio content by the passive surround decoding method
- Upmixed audio content by the LMS-based method
- Upmixed audio content by the PCA-based method, and
- Upmixed audio content by the adaptive panning method.

Figure 3 shows the MUSHRA test result. As shown in the figure, the upmixed 5.1 channel audio contents by any of the upmixing methods were preferred to the stereo audio contents.

This implies that 5.1 channel audio could provide a better listening environment than stereo audio. Moreover, it was found out that the adaptive panning method outperformed the other methods.

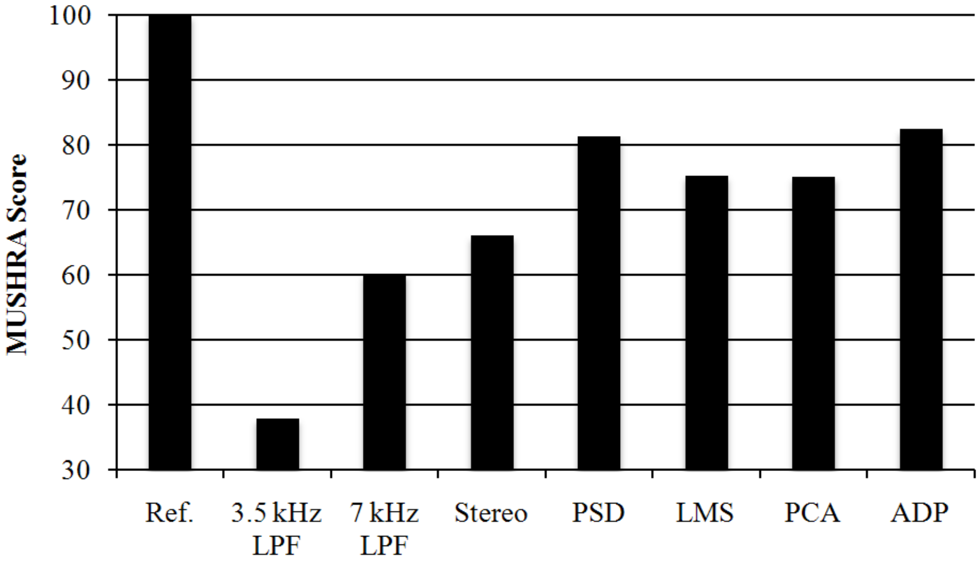


Figure 2. Comparison of MUSHRA test scores for the audio signals upmixed by different methods

5. Conclusion

In this paper, we described four different upmixing methods for converting audio from a stereo format to a 5.1 channel format based on correlation techniques. After implementing these techniques, we then designed a simulator that was able to upmix stereo audio files and play them in real time to produce better realistic sounds. In order to evaluate the performance of the upmixing algorithms and compare the upmixed 5.1 channel signals with the original stereo signals, a MUSHRA test was conducted. It was shown from the test that the 5.1 channel audio generated by any of the upmixing methods provided better audio quality than the original stereo audio and the adaptive panning method yielded the best performance among all the methods.

Acknowledgments

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (No. 2009-0057194) and by a basic research project grant provided by GIST in 2009.

References

[1] Dolby Laboratory, *Dolby Surround Pro Logic Decoder Principles of Operation*, http://www.dolby.com/professional/getting-dolby-technologies/index.html.

[2] M. R. Bai, G.-Y. Shih, and J.-R. Hong, "Upmixing and downmixing two-channel stereo audio for consumer electronics," *IEEE Trans. on Consumer Electronics*, vol. 53, no. 3, pp. 1011-1019, Aug. 2007.

[3] R. Irwan and R. M. Aarts, "Two-to-five channel sound processing," *J. Audio Eng. Soc.*, vol. 50, no. 11, pp. 914-926, Nov. 2002.

[4] ITU-R BS.1534-1, *Method for the Subjective Assessment of Intermediate Quality Levels of Coding System*, Jan. 2003.

[5] B. Widrow and S. D. Stearns, *Adaptive Signal Processing,* Prentice-Hall, 1985.

[6] I. T. Jolliffe, *Principal Component Analysis*, Springer, Heidelberg, 2002.

[7] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards, Kluwer Academic Publishers*, Dec. 2002.

[8] A. V. Oppenheim, R. W. Schafer and J. R. Buck, *Discrete-time Signal Processing,* Prentice-Hall, 1989.

[9] ITU-R BS.775-1, *Multi-Channel Stereophonic Sound System with or without Accompanying Picture*, July 1994.

Authors

Chan Jun Chun received a B.S. degree in Electronics Engineering from Korea University of Technology and Education, Korea in 2009. He is now a student for the M.S. degree at the Gwangju Institute of Science and Technology (GIST). His current research interests include 3D audio and audio upmixing.

Yong Guk Kim received a B.S. degree in Electronics and Computer Engineering from Chonnam National University, Korea in 2006, and an M.S. degree in Information and Communications Engineering from the Gwangju Institute of Science and Technology (GIST), Korea in 2008. He is now a Ph.D. student at GIST. His current research interests include 3D audio and multi-channel audio rendering.

Jong Yeol Yang received a B.S. degree in Electronics Engineering from Inha University, Korea in 2008. He is a student for the M.S. degree at the Gwangju Institute of Science and Technology (GIST), now. His current research interests include speaker recognition in noisy environments and pattern classification.

Hong Kook Kim received a B.S. degree in Control and Instrumentation Engineering from Seoul National University, Korea in 1988. He then received both M.S. and Ph.D. degrees in Electrical Engineering from the Korea Advanced Institute of Science and Technology (KAIST), Korea in 1990 and 1994, respectively. He was a senior researcher at the Samsung Advanced Institute of Technology (SAIT), Kiheung, Korea, from 1990 to 1998. During 1998-2003, he was a senior member technical staff with the Voice Enabled Services Research Lab at AT&T Labs-Research, Florham Park, NJ. Since August 2003, he has been with the Department of Information and Communications at the Gwangju Institute of Science and Technology (GIST) as a professor. His current research interests include speech recognition and coding, audio coding and 3D audio, and embedded algorithms and solutions for speech and audio processing for handheld devices.