Enhancement of Headphone-based Music Listening Experience through Binaural Synthesis

Kai-Chieh Huang
(kaichieh@stanford.edu)

Adviser: Julius O. Smith
Center for Computer Research in Music & Acoustics
Stanford University, Palo Alto, CA

This paper discusses the basic psychoacoustics principle behind spatial hearing and the implementation of binaural synthesis on headphones using Head Related Transfer Function (HRTF). Successful binaural rendering of spatialized virtual sound source through headphones can lead to an improvement on listening experience for entertainment systems. In particular, a more immersive music listening experience through headphones can be achieved by up-mixing stereo music tracks, spatialize each up-mixed channel through binaural rendering, and down-mixing back to stereo. A study on up/down-mixing algorithms and virtual speaker sptialization through binaural synthesis is conducted to enhance general headphone-based music listening experience. The final algorithm is evaluated by comparing normal stereo music to binaural synthesized version through listening test.

0 INTRODUCTION

As the advent of smartphone era and proliferation of portable devices in recent years, music streaming service through Apps is becoming more and more popular. This improves and expedite the way people can access to music. However, stereo playback through headphones is still the dominant presentation for providing the listening experience through these services. As a result, it will be beneficial to enhance the traditional stereo listening experience through signal processing techniques. One possible way of enhancing the immersiveness of listening music through headphones is to create virtual speakers and reproduce the spatial image of real stereo speaker setup or home theater surround sound setup [1].

Binaural synthesis is a technique widely used for generating spatialized virtual sound image through headphones from monaural sound source. It uses the concept of Head Related Transfer Function (HRTF) or binaural room transfer function (BRTF) to recreate the spatial cue of a monaural sound source coming from different location. With the improvement of computational efficiency on computers and portable devices, binaural synthesis technique is becoming relatively feasible to implement. This technique can also be used to tackle the in-the-head sound problem that typically appear in stereo music playback on headphones and provide a more externalized sound image. Applications such as gaming, mobile video, and music streaming can also apply this technology to augment the overall entertainment experience. By applying this technology, consumers can experience multi-channel surround sound without the need of a dedicated loudspeaker setup, and without disturbing other people in the same physical environment.

In the following, an introduction to the basic psychoacoustics principle behind spatial hearing and binaural synthesis is given first. Then, methods of extracting multi-channel sound tracks from stereo music is presented. Thirdly, the implementation of virtual speaker spatialization and algorithm to render binaural synthesized sound tracks into stereo headphone playback is discussed. Finally, the evaluation results of binaural synthesized music using proposed algorithm and normal stereo music are compared.

1 Binaural Synthesis Principles

The basic concept behind Binaural Synthesis is that when the sound sources coming from a certain location, it arrives at both ears at different time and has different level at different frequencies due to the reflections of head, body, pinna. The interaural time difference (ITD) and the interaural level difference (ILD) are captured in the response pair of the sound source to the left ear and the sound source to the right ear. This response pair is often refer to as Head Related Transfer Functions (HRTFs), or Head Related Impulse Responses (HRIRs) in time domain. For each location, there will have a unique HRTF pair that describes the

spatial hearing cues of that location. Thus, if we can successfully reproduce these cues with a HRTF pair to trick the ear while playing back on headphones, we can spatialize sound sources at different location.

The main problem of virtual spatialization through binaural synthesis has always been the lack of solution for individualized HRTFs. Since, every person's anthropometric features are different, individualized HRTFs are required for reproducing accurate spatial cues for each person. However, for non-ideal case, simple simulation of the ITD and ILD through general HRTF measured from KEMAR dummy head can provide sufficient intuition on the location of the sound source[2]. Currently, the MIT and CIPIC library are the two most popular open source measured HRIR database.

In addition to using HRIRs, in order to make the binaural synthesized sound more externalized and give a more realistic feeling of the synthesized sound source is coming from a location outside of headphone, we can add some synthesized reverb on top of HRIRs since people tend to use the reflected sound as a cue to determine distance and reverberation is known to associate with the perception of distance and externalization. The overall response of HRIRs and the synthesized reverberation is refer to as binaural room impulse responses (BRIRs) in time domain or binaural room transfer functions (BRTFs) in frequency domain. By using a pair of BRIRs, one for the left ear, and one for the right ear, we can convolve a single channel input with the left ear BRIR and right ear BRIR respectively to reproduce the spatial hearing cues while playing back on headphones and spatialize sound source at a certain location as shown in figure 1.

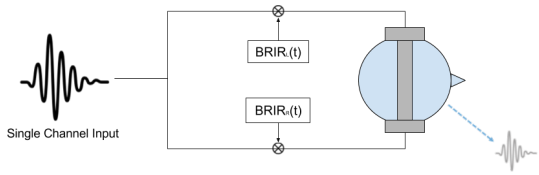

Fig. 1. Sound source spatialization process

Typically, the room reverberation is simulated by adding some discrete early reflections following by more condensed late reflections. Since only the first few early reflections are effective, we can randomly generate some early reflection at different arrival time under 30 milliseconds and at different angles, then use filters to simulate air absorption and wall absorption[3]. To simulate the reflections from different angles, we also need to convolve the reflected signal with the corresponding HRIRs. Finally, we can use Feedback Delay Network to simulate late room reflections[4]. Note that we need to maximize the externalization with as little reverberation as possible in order to preserve the naturalness of the surround sound. Furthermore, diffuse sound appears as left and right ear input signals are nearly fully correlated at low frequency with decreasing correlation as the frequency increases. Thus, it is also important to simulate the correct correlation to provide a proper distance perception[5].

2 Surround Sound Configuration

The 5.1 surround sound configuration is the most commonly used layout in home theater system and therefore is suitable to used for simulating the virtual surround rendering on headphone. For this project, we will spatialize each virtual speaker at the speaker position as in the 5.1 surround sound configuration to create a more immersive listening experience. The detail of the 5.1 surround sound configuration is explained as follows. The front left (L) and front right (R) speakers are placed at ear height, producing a 30 degrees angle as viewed from the main seating position. This delivers a wide sound stage and precise localization of individual sounds. The center channel speaker (C) is placed at 0 degree directly in front of the listener. The left surround (LS) and right surround (RS) speakers are placed at 110 degrees to each side and 2 feet higher above the listener. Finally the subwoofer or the low frequency enhancement (LFE) channel is place in the middle of the front wall. An example of the standard ITU-R BS 775 5.1 surround sound configuration is presented in figure 2.

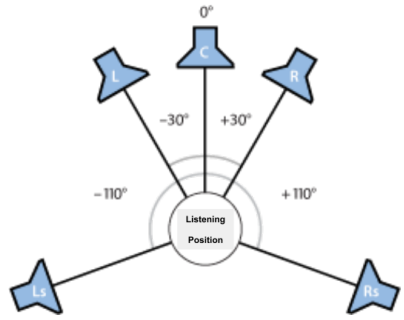

Fig. 2. 5.1 surround sound configuration

3 Up-mixing Algorithm

The typical approach for generating additional channels is by exploring the correlation between stereo channels. Four types of correlation based upmixing methods including the passive surround decoding (PSD), the least-mean-square (LMS)-based method, the PCA-based upmixing method, and the Adaptive Panning method (ADP) are discussed and evaluated in [6]. In this project, we adopted the PSD method due to its simplicity and better performance from the evaluation result of [6]. The PSD upmixing method for converting audio from a stereo format to a 5.1 channel format is described below. First, the center and surround channels are obtained by the following equations:

$$Center(n) = (x_L(n) + x_R(n)) \times \frac{1}{\sqrt{2}} \quad (1)$$

$$Surround(n) = (x_L(n) - x_R(n)) \times \frac{1}{\sqrt{2}} \quad (2)$$

Then, the center channel is low pass filtered with a cutoff frequency of 4 kHz to extract and emphasize on the voice and dialog content. A low-pass filter having a cut-off frequency of 200 Hz is also used to filter the center channel to generate the low frequency enhancement (LFE) channel that emphasizes low frequency region. On the other hand, the surround channel is first delayed by 12 ms to provide some ambiance effects and then low pass filtered at a cut-off frequency of 7 kHz to simulate the high-frequency absorption effect. Finally, the left and right surround channel are created by phase shifting this post processed surround channel with +/- 90 degrees to present spaciousness effects. In order to preserve the original spectrum of the frequency range in interest as accurate as possible, all the filters used in the upmixing process is chosen to be 10 order butterworth filter due to its maximum flat pass band advantage. The full upmixing procedure is summarized in figure 3.

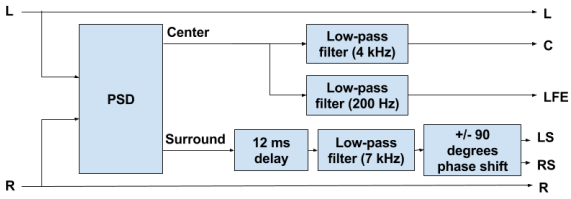

Fig. 3. Sound source spatialization process

Where the channels are labeled L (front left), R (front right), C (center), LFE (low frequency enhancement), LS (left surround), and RS (right surround). As illustrated in the figure, the L (front left) and the R (front right) 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.

4 Binaural Rendering Implementation

To provide a more realistic and immersive listening experience through binaural synthesis, it is essential to synthesize the reverberation to create a natural sense of externalization. In this section, we will discuss the implementation of generating synthesized BRIR and the down-mixing of spatialized channels. Since the findings of J. Catic et al. in [7] indicates that the fluctuation of binaural cues in the BRIRs strongly affects externalization for sound sources, the early reflections in our synthesized BRIR is chosen to be randomly vary from right 90 degrees and left 90 degrees incident to the listener. In our BRIR, 3 early reflections with right 90 degrees and 8 milliseconds delay, left 60 degrees and 15 milliseconds delay, and right 40 degrees and 23 milliseconds delay, with a gain of 0.5, 0.3, and 0.2 respectively are added after the HRIRs of the angle we tried to spatialize. As for the late reflections, a 200 milliseconds white noise is used to simulate the randomness of the late reflection. Then, the white noise is converted into spectrogram to simulate the frequency dependent decay by windowing each frequency bin with a exponential decay. The T60 for simulating the exponential decay from low frequencies to high frequencies is chosen to be from 200 milliseconds ramping down to 1 milliseconds in logarithmic scale. Finally, the late reflections is added right after the last early reflection. An example of the generation process for the left ear BRIR at azimuth right 30 degrees is provided in figure 4. Note that the synthesized BRIR for the

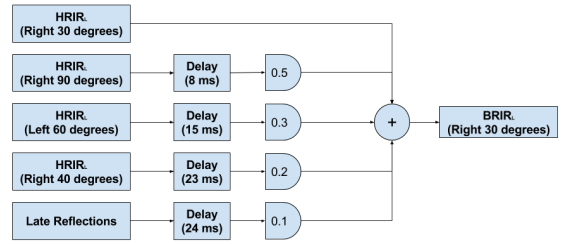

Fig. 4. BRIR synthesize process for azimuth right 30 degrees

right ear is generated with the same process by replacing all the left ear HRIRs with the right ear HRIRs. The synthesized BRIRs for azimuth right 30 degrees and its frequency response are provided in figure 5.

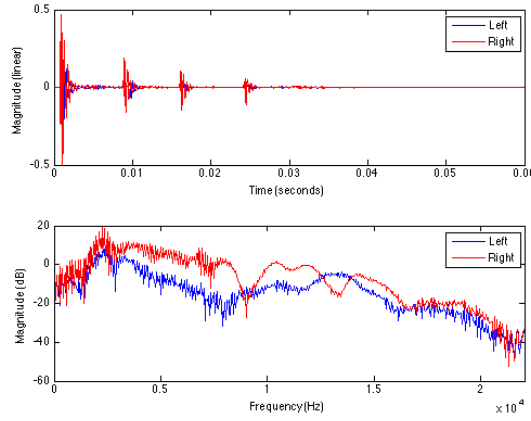

Fig. 5. Example BRIR pair of azimuth right 30 degrees

After generated the BRIRs for each of the azimuth angle for the speaker position in the 5.1 surround sound configuration, each channel (excluding the LFE is spatialized at its corresponding position and linear down-mixed to stereo. Lastly, the LFE is added equally to the left and right channel because low frequency bass is less directional. Some examples of the 5.1 binaural rendering can be found in [8], [9], and [10].

5 Evaluation

To understand the performance of our 5.1 surround binaural rendering, an AB selection listening test was conducted where the listening were asked to indicate their

preference by force ranking among the original stereo version and our 5.1 surround binaural version. In the listening test, 10 audio contents that spans a wide range of music genre including Hard Rock, British Rock, Acoustics, Blues, Jazz, and Pop music were used. The full list of audio contents used for this listening test are shown in figure 6.

Index	Content Name	Artist	Genre
1	Wind Blow	No Party For Cao Dong	Alternative Rock
2	Yellow	Coldplay	British Rock
3	Just Tonight	Pretty Reckless	Pop Rock
4	Yellow Brick Road	Angus & Julia Stone	Indie
5	How Blue Can You Be	B.B. King	Blues
6	You Shook Me All Night Long	AC/DC	Hard Rock
7	Angelina	Tommy Emmanuel	Acoustics
8	Rock Hometown	Omnipotent Youth Society	Indie Rock
9	Bold As Love	John Mayer	Pop
10	Jesus Etc.	Norah Jones	Jazz

Fig. 6. Audio contents used in listening test

There are 10 trials in total for this listening test. In each trial, one audio content was presented to the listener blindly in the original stereo version and the 5.1 surround binaural rendering version. The listener was asked to force rank their preference among these two versions. The listening test environment was modified from the web-based open source MUSHRA test [11]. The listening test environment used for this project can be found at [12]. 10 Subjects were involved in this listening test. A total of 100 data points were collected (10 subjects $\times$ 10 trials) and the percentage of the listener likes one version over the other was compared. The result of the listening test is presented in the following.

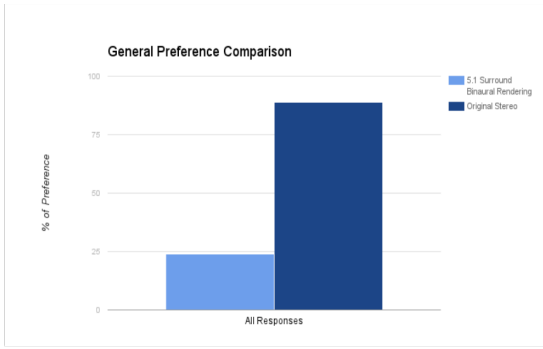

Fig. 7. Preference comparison result

Surprisingly, more of the listeners prefer the original stereo playback over our 5.1 surround sound binaural rendering version. The explanation of this observation is further derived by studying the comments that each listener provides at the end of the listening test. This inference and some future works are discussed in the next section.

6 Conclusion & Future Work

From the listening test result, we found that the 5.1 surround rendering of our algorithm is actually not as preferable as the original stereo play back. If we dig deeper into finding the reason, we found that people tend to choose their preference base on the spectral performance of the audio content over the spatial performance from the comments of the listeners. Since the frequency response of the BRIRs we used in the 5.1 surround binaural rendering are not flat, it is inevitable to shape the spectrum of the original stereo mixing through BRIR processing. Thus, further equalization method needs to be investigate to adjust the spectrum of the 5.1 surround binaural rendering. In addition, the coherence matching to simulate the correlation of left and right ear in the diffuse field is also a valuable direction to enhance the performance of the 5.1 surround binaural rendering.

Some possible future directions can also be adopted based on [13], where Grant Davidson et al. studied the factor in reverberation that maximize externalization to generate BRIRs that maximize externalization under the nature timbre constraint. More specifically, they incorporated those factors in a stochastic room model to randomly generate synthesized BRIRs and used an objective function that measures the distortion of each critical band to automatically select the one that is most spectrally accurate. In the paper, they also mention that spatialized 5.1 binaural rendering generally offer marginal improvement over stereo playback, so they suggest using a spatialized 7.1.4 down-mix, where the extra 4 virtual speaker on the top may offer potential preference gain.

7 REFERENCES

[1] Christof Faller and Jeroen Breebaart, “Binaural Reproduction of Stereo Signals Using Upmixing and Diffuse Rendering,” *J. Audio Eng. Soc*, Presented at the 131st Convention, 2011 October 2023 New York, USA.

[2] William L. Martens, “Perceptual evaluation of filters controlling source direction: Customized and generalized HRTFs for binaural synthesis,” Acoust. Sci. & Tech. 24, 5 (2003).

[3] Yougen Yuan, Zhonghua Fu, Ming Xu, Lei Xie, Qi Cong, “Externalization improvement in a real-time binaural sound image rendering system,” International conference on orange technologies.

[4] XXIA Risheng, LI Junfeng, XU Chundong, and YAN Yonghong, “A sound image externalization approach for headphone reproduction by simulating binaural room impulse responses,” Chinese Journal of Electronics, Vol.23, No.3, July 2014.

[5] Fritz Menzer and Christof Faller, “Stereo-to-Binaural Conversion Using Interaural Coherence Matching,” Presented at the 128th Convention, 2010 May 2225 London, UK.

[6] Chan Jun Chun, Yong Guk Kim, Jong Yeol Yang, and Hong Kook Kim, “Real-Time Conversion of Stereo Audio to 5.1 Channel Audio for Providing Realistic Sounds,” In-

ternational Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 2, No. 4, December 2009.

[7] J. Catic, S. Santurette, T. Dau, Role of Reverberation-Related Binaural Cues in the Externalization of Speech, J. Acoust. Soc. Am., Vol. 138, August 2015.

[8] https://goo.gl/oH10S9

[9] https://goo.gl/VKkhiH

[10] https://goo.gl/N1pfBX

[11] https://github.com/HSU-ANT/beaqlejs

[12] https://ccrma.stanford.edu/%7Ekaichieh/ABTest/

[13] Grant Davidson, Dan Darcy, Louis Fielder, Zhiwei Shuang, Richard Graff, Jeroen Breebaart, and Poppy Crum, Design and Subjective Evaluation of a Perceptually-Optimized Headphone Virtualizer, Presented at the 140th Convention, 2016 June 47 Paris, France.