PSYCHOACOUSTIC EVALUATION OF DIFFERENT METHODS FOR CREATING INDIVIDUALIZED, HEADPHONE-PRESENTED VAS FROM B-FORMAT RIRS

ALAN KAN, CRAIG T. JIN∗ and ANDRÉ VAN SCHAIK

Computing and Audio Research Laboratory,
School of Electrical and Information Engineering,
University of Sydney, Australia, 2006

∗ E-mail: craig.jin@sydney.edu.au

We evaluate a new technique for synthesizing individualized binaural room impulse responses for headphone-rendered virtual auditory space (VAS) from B-format room impulse responses (RIRs) recorded with a Soundfield microphone, and a listener’s anechoic head-related impulse responses (HRIRs). Traditionally, B-format RIRs are decoded for loudspeaker playback using either Ambisonics or Spatial Impulse Response Rendering. For headphone playback, virtual loudspeakers are commonly simulated using HRIRs. However, the number and position of loudspeakers should not really be a factor in headphone playback. Hence, we present a new technique for headphone-rendered VAS which is not limited by the number and position of loudspeakers and compare its performance with traditional methods via a psychoacoustic experiment.

Keywords: Virtual auditory space; Binaural room impulse response; Soundfield microphone; Room impulse response

1. Introduction

A virtual auditory space (VAS) is an auditory display that conveys three-dimensional acoustic information to a listener such that a virtual sound source in the VAS will be perceived to be the same as that of a naturally-occurring sound source in an equivalent real-world space. A VAS can be presented to a listener using loudspeakers or headphones. For headphone-presented VAS, a binaural room impulse response (BRIR) is typically recorded at the ears of a listener for every sound source position of interest in the room or listening space. The BRIR completely characterizes the acoustical transformation of the sound signal from its source position to the listener’s ears. This transformation arises from reflections and scattering due to the room and the listener’s ears, head and physique, and
provides acoustic information to the listener about the source location and also the room’s physical characteristics. Recording BRIRs may not always be easily achieved or even possible because it clearly requires that each listener must travel to the acoustic space of interest to have the measurements taken. A possibly more flexible method would be to record the components of the acoustical transformation that arise from the room and the listener separately and then to recombine these components together again to synthesize the BRIR. This paper examines various techniques to achieve this separation and recombination of acoustic information for the synthesis of individualized VAS.

Consider now the two separate components of a BRIR. First, a head-related impulse response (HRIR), or in the frequency domain a head-related transfer function (HRTF), characterizes the directionally-dependent acoustical transformation of a sound signal from a location in the free-field to the listener’s ears. These are typically recorded for a listener in an anechoic room, i.e. a room without reflections, and therefore characterize the acoustic properties of a listener’s ear. Secondly, the acoustical transformation of a sound signal from its source location in a room to a listening position is characterized by a room impulse response (RIR) and can be recorded using a Soundfield microphone. The advantage of using a Soundfield microphone is that the directional characteristics of the RIR are encoded within its B-format signals, which consist of an omni-directional pressure signal, $W(t)$, and three orthogonal figure-of-eight, pressure-gradient signals, $X(t)$, $Y(t)$ and $Z(t)$, oriented in the directions of the Cartesian axes.

Because the methods for decoding B-format signals have traditionally been designed for loudspeaker playback, we will first review the application of B-format RIRs for loudspeaker playback and then consider common adaptations of this technique for headphone presentation which will ultimately use a listener’s recorded HRIRs. There are two primary methods for loudspeaker playback of B-format signals: Ambisonic decoding and Spatial Impulse Response Rendering (SIRR). With the Ambisonic technique, a monaural sound source signal is first filtered with the B-format RIRs to produce a vector of B-format signals, $b$. Ambisonic decoding then solves a least-mean square optimization problem based on the location of the loudspeakers to obtain a decoding matrix $M_d$. Given the decoding matrix, the vector of loudspeaker feeds, $l$, are obtained using $l = M_d b$. It should be noted that with a limited number of loudspeakers, the size of the listening area and frequencies at which the sound field can be accurately reconstructed is limited due to spatial aliasing. Above the spatial-aliasing
frequency (typically around 400 Hz), the loudspeaker gains can be modified in order to maximize the high frequency energy coming from the direction of a sound source; we will refer to this as “Ambisonic - max_re”. To improve the robustness of the sound field across a larger listening area, an additional decoding correction can be added such that the loudspeakers are played “in-phase”, that is, the decoding prevents loudspeakers from playing signals out of phase, particularly from those loudspeakers that are diametrically opposite to the sound source location. We will refer to this method of Ambisonic decoding as “Ambisonic - in-phase”.

An alternative method for loudspeaker playback using B-format RIRs is SIRR. SIRR assumes that perfect reconstruction of the original sound field is not necessary to reproduce the spatial impression of a room, but rather the same spatial impression can be generated by recreating the time-frequency features of a sound field. To achieve this, SIRR applies an energy analysis to the B-format RIRs in the time-frequency domain in order to determine the direction of arrival and the diffuseness of the energy at each time-frequency tile. The time-frequency analysis is usually performed using a short-time Fourier transform (STFT). The information derived from the energy analysis is then used to create a set of decoding filters for a loudspeaker array. A monaural source signal is then filtered with the decoding filters to generate loudspeaker signals that preserve the direction of arrival, diffuseness and spectrum of the sound field when played back over the array of loudspeakers.

In our view, the primary drawback with SIRR is that the diffuse sound field is rendered somewhat arbitrarily. One of the main contributions of this work is that we have developed a technique along the lines of SIRR that better preserves the diffuse sound field when rendered via headphones. Before describing our new method, we first review SIRR in some detail.

The SIRR energy analysis is based on the concept of sound intensity which describes the transfer of energy in a sound field. For a given time-frequency tile, the active intensity, \( I_a(k, \omega) \), and diffuseness, \( \psi(k, \omega) \), of the B-format RIR are given by:

\[
I_a(k, \omega) = \sqrt{2Z_0} \text{Re} \{ W^*(k, \omega)V(k, \omega) \}\]

and

\[
\psi(k, \omega) = 1 - \frac{\sqrt{2} \| \text{Re} \{ W^*(k, \omega)V(k, \omega) \} \|}{|W(k, \omega)|^2 + |V(k, \omega)|^2 / 2}
\]

where \( W(k, \omega) \) and \( V(k, \omega) \) are the STFT (\( k \) is the time-frame index and
\( \omega \) is the frequency variable) of \( W(t) \) and \( V(t) = X(t)e_x + Y(t)e_y + Z(t)e_z \), respectively, where \( e_x, e_y \) and \( e_z \) are the unit vectors in the directions of the Cartesian co-ordinate axes; \( * \) denotes the complex conjugation, \(|.|\) denotes the absolute value of a complex number, \(||.||\) denotes the norm of a vector, and \( Z_0 \) is the acoustic impedance of air (typically 413.2 Nsm\(^{-3}\) at 20°C).

The quantity \( \psi \) takes a value between 0 and 1. A value of \( \psi = 1 \) indicates an ideal diffuse sound field (no net transport of energy), and a value of \( \psi = 0 \) signifies the sound field consists only of a directional component. From the intensity vector, the direction of arrival of the net flow of energy, i.e. the azimuth, \( \theta(k, \omega) \), and elevation, \( \phi(k, \omega) \), can be calculated as:

\[
\theta(k, \omega) = \tan^{-1}\left(\frac{-I_y(k, \omega)}{-I_x(k, \omega)}\right), \phi(k, \omega) = \tan^{-1}\left(\frac{-I_z(k, \omega)}{-\sqrt{I_x^2(k, \omega) + I_y^2(k, \omega)}}\right)
\]

(3)

where \( I_x(k, \omega), I_y(k, \omega), I_z(k, \omega) \) are the components of the active intensity in the directions corresponding to the Cartesian co-ordinate axes.

After performing the energy analysis of the B-format RIRs as described above, an STFT representation of the decoding filters for the loudspeaker array is determined as follows. It should be noted that for each time window, zero-padding is used prior to the Fourier transform to prevent time-domain aliasing. For each time-frequency tile, the omni-directional signal, \( W(k, \omega) \), is split into directional and diffuse components according to the diffuseness estimate \( \psi(k, \omega) \). The directional component is given by: \( \sqrt{1 - \psi(k, \omega)}W(k, \omega) \) and the diffuse component by: \( \sqrt{\psi(k, \omega)}W^2(k, \omega) \). At each time-frequency tile, the directional component is distributed among the decoding filters using a vector-based amplitude panning (VBAP) technique\(^7\) while the diffuse component is added to all of the decoding filters using a technique that distributes the total diffuse energy in a decorrelated manner among all of the loudspeakers. Pulkki et al.\(^8\) suggests a decorrelation method for SIRR whereby random panning of the diffuse energy from different loudspeakers is used at the low frequencies (< 800 Hz), with a smooth transition (800 - 1200 Hz) into a phase randomization method at the high frequencies. Time-domain decoding filters for the loudspeakers are then obtained by applying an inverse STFT to the STFT representation of the decoding filters with appropriate overlap-and-add processing.

To render the signals from Ambisonics or SIRR over headphones as VAS, it is common to use a virtual loudspeaker technique in which the loudspeaker signals are filtered with the HRIRs corresponding to the direction of the loudspeaker relative to the listener and summed together to
create left and right headphone signals. In reality, however, the limitations on the number and position of the loudspeakers should not be a factor when reproducing the sound field over headphones. For example, the quality of an Ambisonic decoding varies with the order of the decoding and also with the number of loudspeakers. With too many loudspeakers, Ambisonics solves an under-determined system of equations to determine the decoding matrix and the quality of the reproduction suffers. On the other hand, with too few loudspeakers the directional resolution of sound sources will suffer. SIRR partly overcomes the problems associated with using a large number of loudspeakers in an Ambisonic decoding. It achieves this by using VBAP, but the diffuse or ambient sound can be incorrectly reproduced. In the following, we propose a new method called binaural sound field rendering (BSFR) for using B-format RIRs to generate an individualized VAS for headphone playback which is not limited by the number and position of the loudspeakers.

2. Binaural sound field rendering

BSFR is a method for synthesizing individualized BRIRs from B-format RIRs and a set of anechoic HRIRs. Fig. 1 shows the steps for BSFR. BSFR begins with exactly the same steps as SIRR and applies an energy analysis on the B-format RIRs in the STFT domain to determine the directional and diffuse components of the omni-directional signal, $W(k, \omega)$. The STFT of the desired BRIR is then determined as follows. At each time window, $W(k, \omega)$, is split into directional and diffuse components according to the diffuseness estimate $\psi(k, \omega)$. The directional component of the BRIR is then obtained by: $\sqrt{1 - \psi(k, \omega)}W(k, \omega)HRTF_{lr}(k, \omega, \theta, \phi)$ where $\psi(k, \omega)$ is the estimated diffuseness, $W(k, \omega)$ is the omni-directional channel of the Soundfield RIR, $HRTF_{lr}(k, \omega, \theta, \phi)$ is the complex-valued HRTF.
corresponding to the direction of the active intensity vector at a particular frequency bin and the subscript ‘lr’ denotes the left or right ear. The real-valued magnitude spectra of the diffuse component of the BRIR is obtained by: $\sqrt{\psi(k,\omega)W^2(k,\omega)DHRTF_{lr}}$ where $DHRTF_{lr}$ is the real-valued magnitude of the directionally-averaged or diffuse-field HRTF for the left or right ear. It is calculated separately for the left and right ears from HRTFs recorded for an evenly distributed set of sound source directions around the listener using:

$$DHRTF_{lr} = 10^{\frac{1}{20} \sum_{i=1}^{N} \log_{10}(|HRTF_{lr}(\theta_i,\phi_i)|)}$$

where $N$ is the number of HRTFs and $\theta_i$ and $\phi_i$ are the azimuth and elevation co-ordinates, respectively, corresponding to the direction of the HRTF. In order to estimate the phase of the diffuse component of the BRIR, a spectrogram inversion method was used. This method iteratively estimates the phase at a particular time window while minimizing the difference in the magnitude response between the magnitude-only diffuse field BRIR and the estimated complex-valued diffuse field BRIR. Additionally, phase continuity between time windows is maintained by taking into account the magnitude spectra from past, present and future time windows during the phase estimation process. The use of the spectrogram inversion method for synthesizing the diffuse field BRIR gives a natural sounding reproduction of the diffuse sound field without the need for decorrelation methods. The diffuse field BRIR estimated by our method is naturally decorrelated at the two ears since the diffuse-field HRTF for the left and right ears are different and hence lead to different phase estimates for the final left and right ear signals. Finally, the directional and diffuse-field parts of the BRIR are added together and the time-domain BRIR is obtained by applying an inverse STFT with appropriate overlap-and-add processing.

3. Listening Test

A listening test was conducted to evaluate the different methods mentioned above for generating headphone-rendered VAS. Subjects rated the VASs generated by these methods against a reference VAS generated from their own BRIRs. In the following, we first describe the methods employed in recording the B-format RIRs used in this listening test, and the BRIR and HRIIs of each subject. Details on how the different methods are applied to these recordings to generate test stimuli are then given. Finally, a description of the listening test will be presented.
Subject BRIRs and a B-format RIR were recorded in a room 7.52 x 12.14 x 2.72 m$^3$ in size. A Tannoy V6 loudspeaker, driven by an Ashley 4400 power amplifier, was used to provide the stimulus. The loudspeaker was located 2.7 m away from the recording position at a height of 1.5 m. A silent computer equipped with an RME Multiface sound card was used to play and record the audio signals at a 48 kHz sampling rate. Since the output transfer function of the loudspeaker did not have constant gain across frequency, a compensation filter was used so that the output transfer function of the loudspeaker was flat within 3 dB between 300 Hz and 20 kHz. A 6 s long logarithmic sine sweep from 10 Hz to 20 kHz, filtered with the compensation filter, was used as stimulus for the recordings and the impulse responses recovered from the recorded sweep via deconvolution. A Soundfield microphone was used for recording the B-format RIR. Subject BRIRs were recorded using a “blocked ear canal” method. The subjects faced the loudspeaker for the BRIR recordings.

HRIRs were recorded for each of the subjects in an anechoic chamber using the “blocked ear canal” method. HRIRs were recorded for 393 different sound source directions, around the subject’s head. HRIRs for any sound source direction were then obtained by interpolation of the 393 HRIR recordings using a spherical thin-plate spline interpolation method. 14 subjects participated in the listening test. Of the 14 subjects, 7 subjects had extensive experience, 5 subjects had some previous experience and 2 subjects had no previous experience in listening tests.

Test stimuli were generated using the four different methods described above. For Ambisonic - max$\_re$, Ambisonic - in-phase and SIRR decoding, a cubic configuration of virtual loudspeakers was used where the loudspeakers were placed at the corners of the cube. For SIRR and BSFR, 3 ms sine-squared windows with 50% overlap were used for the energy analysis. The same windows were used for the synthesis with 1.5 ms zero-padding before and after each window. The diffuse-field HRTF for BSFR was calculated by averaging the 393 recorded HRTFs for each subject separately. Additionally, a reference sound was created by filtering anechoic sound stimuli with the measured BRIRs and a low quality anchor stimuli was created by filtering the anechoic sounds with the anechoic HRIR of the subject for a sound source in front of the listener and low-passed filtered at 3.5 kHz. A total of 8 anechoic sounds were chosen for the listening test (see Table 1). In order to achieve a consistent perceived loudness across the test stimuli generated by the different methods, a loudness model was used to calculate a single gain adjustment factor for each of the test stimuli separately. The calculated
Table 1. The different sound excerpts are shown along with the name (key) with which the sound will be identified.

<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Key</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Female Speech - English voice</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Guitar Capriccio Arabe guitar</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Xylophone Sabre Dance xylo</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>Bb Trumpet Over the Rainbow</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Symphony No. 4 in E-flat orch (Bruckner)</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>The Marriage of Figaro figaro (Mozart)</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Pizzicato Polka (Strauss) strings</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Violin solo violin</td>
<td></td>
</tr>
</tbody>
</table>

Gain adjustment factor was then applied to the corresponding left and right ear sound signals of the test stimuli.

The listening test was conducted in a sound-attenuating booth to reduce external sound interference. Sound stimuli were presented using Etymotic ER-1 headphones from an RME Multiface soundcard attached to a computer located outside the booth. An adapted version of the multi-stimulus test with hidden reference and anchor (MUSHRA) paradigm was used.

In the standard MUSHRA paradigm, a subject is asked to rate how close each test stimuli, generated by the different methods, is to a reference sound using a scale from 0 to 100. The scale is divided into 5 equal intervals, where [0-19] = bad, [20-39] = poor, [40-59] = fair, [60-79] = good, and [80-100] = excellent. However, during preliminary listening tests, it was determined that making one rating for each test stimuli was too difficult since the stimuli generated by the different VAS methods differed from the reference in more than one perceptual aspect. Hence, subjects were instructed to first rate the test stimuli on three perceptual attributes separately, prior to making an overall rating of the test stimuli. The three perceptual attributes were: (1) the quality of the reverberation in the sound, that is, whether the test sound sounded like it was in the same room as the reference; (2) the quality of the sound source, that is, how similar the sound source was to the reference and whether there were noticeable timbral difference or changes in the sound source width; and (3) the position of the sound source, that is, how close the sound source was in position compared to the reference sound. Sliders were provided on a graphical user interface for the subject to make the ratings for each trial. After rating each test stimuli on the perceptual attributes, the subject was then asked to make an overall rating of the test stimuli. For the overall ratings, subjects were required to rate one of the sounds in each trial at a score of 100 and one at
4. Results

The mean overall rating given by subjects for the different VAS generation methods are shown in Fig. 2. A number of observations can be made from the overall ratings: (1) For most of the sounds, subjects gave similar scores to Ambisonic - max\textsubscript{re} and Ambisonic - in-phase decoding methods. This is to be expected since the subject always remains in the “sweet spot” when the Ambisonic reconstruction of the sound field is presented over headphones. (2) For two of the sounds, trumpet and violin, BSFR was on average rated significantly higher than the other methods; and (3) the scores for SIRR were lower compared to the other methods for most sounds.

To test the significance of the above observations, a Kruskal-Wallis non-parametric ANOVA was conducted on the mean overall ratings to test the hypothesis that there are no statistically significant differences in the ratings for the four different methods. The analysis was done for each of the test stimuli separately and the results are shown in Table 2. The analysis revealed significant differences in the ratings for all test sounds except for the strings. A post-hoc analysis (Tukey HSD) was conducted to investigate these differences and the analysis revealed the higher ratings for BSFR for
Table 2. The $\chi^2$ and p-value results of a Kruskal-Wallis non-parametric ANOVA conducted on the overall ratings for each of the test sounds. The number of degrees of freedom for all tests is 3.

<table>
<thead>
<tr>
<th></th>
<th>trumpet</th>
<th>violin</th>
<th>guitar</th>
<th>strings</th>
<th>xylo</th>
<th>figaro</th>
<th>orch</th>
<th>voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>&lt; 0.005</td>
<td>&lt; 0.005</td>
<td>&lt; 0.005</td>
<td>0.08</td>
<td>&lt; 0.005</td>
<td>&lt; 0.005</td>
<td>0.01</td>
<td>0.02</td>
</tr>
</tbody>
</table>

the trumpet and violin sounds and the low ratings for SIRR for most test stimuli to be statistically significant. The ratings for the two Ambisonic methods showed no statistically significant differences.

Some understanding of how subjects may have arrived at their overall ratings can be obtained by studying the ratings for the sound quality and position of the sound source, and the quality of the reproduced reverberation. The mean ratings for each perceptual parameter is shown in Fig. 2. It can be observed that for the trumpet and violin sounds, BSFR was, on average, rated as being better at reproducing the sound quality and position of the sound source. Also, it can be observed that SIRR was rated significantly lower for most stimuli when judged for its ability to reproduce the reverberant qualities of the sound field.

5. Discussion and Conclusions

A listening test was conducted to evaluate a number of different methods for generating an individualized, headphone-rendered VAS from B-format RIRs. The results show that there is a noticeable difference in the VAS generated from the different methods compared to the VAS generated using the subjects’ measured BRIRs. Anecdotally, subjects commented that the VAS generated by the different methods were acceptable, even reasonable, except for SIRR where most subjects commented that the reproduced sound field was too reverberant. This is due to the fact that there is no control over the amount of decorrelation applied in the decorrelation method used in SIRR. In the case of the trumpet and violin sounds, there is an improvement in the generated VAS when using BSFR. Furthermore, the BSFR method was anecdotally reported to provide a better frontal image. Subject ratings in these sounds for the three perceptual attributes indicate that there was improved position localization and timbral qualities of the sound sources when using BSFR. In summary, while the B-format RIRs do not provide complete information to synthesize perceptually-accurate BRIRs, the BSFR method provides a technique that is not limited by the position or number of loudspeakers and seems to recreate the characteristics of the sound field.
reasonably well.

References