A Qualitative Study of Acoustic Environment Simulation Using Headphones

Eng C.W.

Department of Electrical and Computer Engineering, National University of Singapore
10 Kent ridge Crescent, Singapore 119260

ABSTRACT

The focus of this research is to examine ways in which acoustic simulation can be done on headphones. In particular, ways in which MIT’s KEMAR Head Related Transfer Functions (HRTF) can be made use of to impart a spatial dimension to monophonic sound clips will be examined. The entire audio processing and examination is done on CoolEdit2000 with the Aurora audio-processing plugin.

INTRODUCTION

The aim of acoustic simulation is to immerse the audiences in an artificial cocoon of sound, to the extent that the audience believes they are in an entirely different environment. The computer simulations that allowed researchers to model and instantaneously correct their work has carried the technology to the point where they are now able to predict sound characteristics of rooms or acoustic spaces that do not exist in reality.

Much of the media exposure however, has been confined to speaker simulations – the recreation of acoustic environments with more than a couple of speakers. The most common form of this is what is known as a 5.1 system, which essentially means that the soundtracks are recorded with 5 main channels: left, center, right, left surround, right surround plus a low-frequency bass channel. Making use of a combination of 5 speakers the network is able to provide a near realistic reproduction of the desired sound environment.

Headphones, however, presented audio researchers with a unique problem. Whereas speakers can provide easily for a 3-dimensional sound field, headphones tend to give an output that sounds decidedly flat. Moreover headphones, unlike speakers that can be arranged around a person, are being located directly to the left and right side of an individual. Thus, headphones can neither give a front/back nor a top/bottom dimension to the sound field with normal recordings.

It is possible though, to compensate for this lacking with special processing, typically through making use of what is known as a person’s Head Related Transfer Function (HRTF). As 3D Audio Primer puts it, an HRTF can be thought of as a set of two audio filters (one in each ear) that contains all the listening cues that are applied to a sound as it travels from the sound’s origin (its source, or position in space) through the environments, and arrives at the listener’s ear drums.

This research attempts give a spatial dimension to monophonic clips, and consists of 3 parts:

- Part 1: Investigation into the efficacy of source localization
- Part 2: Attempt to discern what constitutes localization cues
- Part 3: Reproduction of HRTF through spectrum-flattening
PART 1 INVESTIGATION INTO THE EFFICACY OF SOURCE LOCALIZATION

MIT's Measurement Procedure

A comprehensive set of HRTF measurements of a KEMAR (Knowles Electronic Manikin for Acoustic Research) dummy head was completed by in MIT's anechoic chamber in 1994. Essentially, the KEMAR was mounted upright on a motorized turntable that could be rotated accurately to any azimuth and speaker (the sound source) was mounted on a boom stand that enabled accurate positioning of the speaker to any elevation with respect to the KEMAR. HRTF measurements can thus be taken at any angle with respect the vertical and the horizontal.

Convolution Using CoolEdit 2000

A short clip (40 seconds) of the Roland Dyen’s Tango En Skai was converted first into mono format to remove any inherent audio effects that may interfere with the experiment (such as in the case where the left channel and the right carry a different tunes.) In the CoolEdit2000 package, this process can be easily done with the sample type conversion option. A 50% mix from each channel ensures that the final clip sounds like the original, without the stereophonic effects.

The resulting single-channel mono clip is then duplicated such that a dual-channel mono clip is obtained.

Figure 1. Dual-channel mono clip

This dual channel mono clip is then convolved with the chosen set of impulse responses (in this case, 0 degrees elevation and 45 degree azimuths) through the use of the Real-time Convolver of the Aurora plugin module.
After the process, we end up with a clip that when played through headphones, sounds as if it radiates from a source in the northeast direction. The following pictorial representation illustrates the entire process of source-placement.

The process is repeated 7 more times, and the sound clip is convolved with impulse responses evenly distributed in different directions at the 0 degree elevation level. The resulting clips are then evaluated for their efficacy in sound source direction simulation.

<table>
<thead>
<tr>
<th>Azimuths</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. 0 (Front)</td>
<td>Hardly discernable</td>
</tr>
<tr>
<td>2. 45</td>
<td>Average north-east simulation</td>
</tr>
<tr>
<td>3. 90 (Left)</td>
<td>Excellent directional cues</td>
</tr>
<tr>
<td>4. 135</td>
<td>Somewhat average cues</td>
</tr>
<tr>
<td>5. 180 (Back)</td>
<td>Hardly discernable</td>
</tr>
<tr>
<td>6. 225</td>
<td>Similar to 135 degree azimuths</td>
</tr>
<tr>
<td>7. 270 (Right)</td>
<td>Excellent directional cues</td>
</tr>
<tr>
<td>8. 315</td>
<td>Similar to 45 degree azimuths</td>
</tr>
</tbody>
</table>

Table 1. Tabulated convolution results

The simulations were good in that users can now detection a source direction in the erstwhile monophonic clip. On top of that, the sound does not just appear to come from just beside the listener’s head – it seemed to originate a distance away.

In addition, as can be gleaned from the convolution results, the front/back simulations (0 and 180 degrees azimuth) were not fantastic. This perhaps resulted from the fact that the recordings were done in anechoic chambers (rooms padded to minimize echoes), which are lacking in sound cues otherwise present in real life.
PART 2  ATTEMPT TO DISCERN WHAT CONSTITUTES LOCALIZATION CUES

Although major deficiencies have been found in part 1 of the project, it has nevertheless verified the efficacy of convolution in producing an acceptable localization of sound source in space. Motivated by this, we next attempt to isolate and verify how each component of sound aids in the artificial placing of audio sources in a 3-Dimensional sound field.

Extract From 3-D Audio Primer By Aureal Corporation

The two primary localization cues are called interaural intensity difference (IID) and interaural time difference (ITD). IID refers to the fact that a sound is louder at the ear that it is closer to, because the sound’s intensity at that ear will be higher than the intensity at the other ear, which is not only further away, but usually receives a signal that has been shadowed by the listener’s head (see fig. 5). ITD means that a sound will arrive earlier at one ear than the other (unless it is located at exactly the same distance from each ear - for example directly in front). If it arrives at the left ear first, the brain knows that the sound is somewhere to the left (see fig. 6).

![Figures 5 and 6 showing Interaural Intensity Difference and Interaural Time Difference](image-url)

**Interaural Time Delay Verification**

The efficacy of Interaural Time delay will first be explored. Using CoolEdit’s silence-generation option, a generic all-pass impulse response that consists purely of interaural delay is created (fig. 7). As there are multiple peaks for each HRTF, the first notable peaks (highlighted in the screenshot by the red arrows) are chosen for the relative delay positions. Using this information, a reconstruction of the HRTF is obtained with CoolEdit. The magnitude for the resulting HRTF is fixed at 0 db so as not to affect the audio volume of the original clip.

![Figures 7 and 8 showing Peaks chosen for ITD simulation and Constructed impulse response for ITD simulation](image-url)

This constructed HRTF is then convolved with the monophonic sound clip using methods detailed in part 1.
As one might have expected, a reconstruction of the HRTF pairs that already showed distinct weakness in front/back source localization does not a better HRTF make. If anything, the reconstruction gave even less localization cues due to the absence of intensity differences.

Moreover, because only the first peak is chosen (we only want to test the initial delay), many of the reverberations are not included, giving the resulting clip an even drier feel.

One thing to note, however, is that the left/right cues are definitely present in this part of the simulation. The following screenshot shows clearly that even with absolutely no intensity differences in the left/right ear the sound certainly comes from the right.

**Interaural Intensity Difference Verification**

Verification of the Interaural intensity difference is perhaps more straightforward than the verification of ITD, if only because it can be done with simple volume control. In our case, we simply make use of CoolEdit's audio amplifier module to do the test.

By cutting the volume of the left channel by 50%, the resulting sound appear to originate from the right hand side, which is only to be expected. What is unexpected is that unlike the processing of ITD as illustrated in the previous section, simple amplification made the resultant sound seemed too close to the listener when in the original recording it felt much further.

**ITD And IID In Tandem**

In this section we will attempt to combine the effects of interaural time delay (ITD) and interaural intensity difference (IID). Essentially we are performing the same tasks as illustrated in part 2, with the addition of a magnitude dimension.

The result is as shown in the screenshot below. Note that simulation is done only on the first sound peak, and subsequent peaks (primary and secondary echoes) are largely ignored.

![Original IR](image1.png) ![Manufactured IR](image2.png)

**Figure 9. Constructed IR, ITD and IID in tandem**

Compared to the separate employment of Interaural Time Delay and Interaural Intensity Difference, their combination provided clear localization cues. Sound appeared to originate from right and at a correct distance away from the listener. However, the front/back dimension is still missing from the result.

**PART 3 REPRODUCTION OF HRTF THROUGH SPECTRUM-FLATTENING**

In the previous sections all Head Related Transfer Function (HRTF) impulse response simulations are done by the creation of entirely new impulse responses that mimics the actual
responses in the features that are to be evaluated. The creations are thus pure interaural delays or
pure volume tweaks.

In this section, we will attempt to isolate the localization cues in the frequency domain. In
doing so, we will examine the hypothesis that the imaging magic is contained entirely in the
frequency response of the filter. That is to say, magnitude changes in the HRTF will not affect
the way the brain places the sound source. To verify this claim, we have to make use of two
features in the Aurora package: the Inverse Filter function and the Flatten Spectrum function.

The inverse filter essentially finds the inverse of a given transfer function, such that it negates
the effect of the transfer function exactly. This filter finds common usage in equalizer systems
that serve to correct channel distortions during transmission. This function creates an impulse
response that is the inverse of the original impulse response BOTH in the magnitude function
and the frequency spectra.

The flatten spectrum function creates a filter that is the inverse of the original in terms of
frequency response, but is relatively flat in terms of magnitude response. Convolving the
resultant filter with the original will result in an HRTF that has a constant frequency response.

Given an HTRF impulse response, the spectrum-flattening filter obtained has a flat
magnitude response (which is desirable) and a frequency response that is the inverse of the
original IR (which is not). To work around this, we first take an inverse image of the original IR,
and then apply the Flatten Spectrum option. The gives us the same flat magnitude response, and
a frequency response that is identical with the original HRTF. This process is shown in the next
illustration.

![Original Response](image1)

![Inverted Response](image2)

![Flattened Response](image3)

Figure 10. The Inversion and Flattening process
Convolution Results

In general, part 3 presents a method that works with the impulse response at hand, instead of creating new ones with CoolEdit. In terms of sound localization, this method churns out roughly the same results as those methods detailed earlier.

The one vast improvement is the fidelity of the music, in that the convolved audio clip appears a lot more crisp in terms of tone than even the clips convolved with KEMAR HRTF directly – this being a direct result of the absence of magnitude noises (the secondary and tertiary echoes for one).

PROJECT CONCLUSION

The use of spectrum-flattening filter gave the best simulation results, as the overall fidelity of the audio clip is improved in comparison the other reconstructions. This affirms, to an extent, that the imaging magic is indeed found in the frequency response of the HRTF. Interestingly, convolving with KEMAR HRTF directly surprisingly was only second best, even though trailing bass sounds were included in the convolution.

In terms of direction, simulating sources from the left and right yielded good results. Concentration on the listener’s part is however required for identification of intermediate directions, and even with that the localization cues are often inconclusive. Much of the negative results may perhaps stem from the lack of echoes from the surroundings as the recording of HRTF is done in an anechoic chamber.

Interaural Time Delays (ITD) and Interaural Intensity Differences (IID) tweaks do provide localization cues, but they work best together. IID in particular placed the sound source too near to the listener when it was processed alone.

ACKNOWLEDGEMENTS

I would like to thank Professor Hari Krishna Garg for his guidance and unwavering support in this project, without which the research would not have materialized.

I would also like to thank Mr Robert Cain, who initiated me into the use of CoolEdit and the Aurora advanced audio processing package and imparted me his reasoning and techniques in audio simulation.

REFERENCES

Dolby Laboratories (2001), *Dolby Knowledgebase* www.dolby.com
Aureal Corporation (1998), *Primary Localization Cues, IID and ITD, 3-D Audio Primer*, www.headwize.com
Robert Cain, BSEE 1967 University of Illinois, Urbana, United States of America