EVALUATION OF THE FIDELITY OF FULL AND APPROXIMATED HRTFS FROM THE SONIC FIDELITY OF LOUDSPEAKERS

Elson C. Chiu, Archie M. Cullano, Garyle Y. Gordiel, Edward B. Kung and Clement Y. Ong
College of Computer Studies, De La Salle University - Manila
2401 Taft Avenue, 1004 Manila, Philippines

Abstract: The goal of a music playback system is to reproduce as closely as possible the sound experience of live music. Despite the excellent frequency response afforded by smaller, lighter transducers, headphones produce an unnatural "in the head" sound experience which many acute listeners find distracting.

Head Related Transfer Functions (HRTFs) are individualized summarization of the direction dependent acoustic filtering a free-field sound undergoes due to a person’s head, torso and pinna, varying as a function of source position and having large inter-subject variation. The Common Acoustical Pole and Zero (CAPZ) Model requires far fewer variable parameters to represent HRTFs. In this study different approximations of the CAPZ model are processed and evaluated in its ability to emulate the external sound field of loudspeakers while wearing headphones.

The subjective results show that there was no audible drop in quality when HRTFs are incorporated to the sound and that no approximation was singled out as having the best sound quality, but, it was observed that the amount of balance between the poles and zeroes had an audible effect to the listeners.

Key Words: Head Related Transfer Function; Common Acoustical Pole and Zero (CAPZ) Model; Early Reflection; Late Reverberation; Sonic Fidelity

1. INTRODUCTION

Earphones work in a way wherein each ear can only hear the sound coming from its own earpiece. This means that there is no natural way for the sound that is produced by the left earpiece to be heard by or to reach the right ear and vice versa. This creates an unnatural experience for the listener since the sounds are perceived as coming from inside the head [10]. This has a completely different experience compared to when listening to a standard stereo system using loudspeakers, which up to today is the format and setup used for high-fidelity, or what some others term as ‘audiophile’ music playback. This kind of system would be able to most closely reproduce the original fidelity of the music as it was recorded.

When using loudspeakers, the sound experiences attenuation, reflection, and diffraction etc. from outer environment before arriving at the ears [7]. Interaural Time Differences (ITDs) and Interaural Level Differences (ILDs) are important parameters for the perception of sounds originating from the horizontal plane. ITDs are described to be the time difference in the arrival times of a sound’s wave front at the left and right ears. ILDs are the difference in amplitude generated in the left and right ear by a sound. A sound is perceived to be closer to the ear at which the first wave front with the greater amplitude arrives [2]. ITDs and ILDs do not describe a unique spatial location; therefore, the ability to localize in the median plane is attributed to a monaural hearing mechanism which relies on the spectral coloration of a sound produced by the torso, head, and external ear, or pinna.

The unnaturalness of the sound that is produced by the earphones could be removed with the use of Head Related Transfer Functions (HRTFs), while the spaciousness of sound can be recovered by incorporating reverberation. HRTFs are often stored as impulse responses called Head Related Impulse Responses (HRIRs). HRIRs (or HRTFs) are unique for each individual. In fact there is a different set of HRIRs for each azimuth and elevation for the left and the right ear.

In this paper we discuss our attempt to reproduce the acoustic sound field of external loudspeakers in a room, through earphones. Our motivation stems from the general acceptance that high quality music reproduction is much more affordable on small transducers (i.e. headphones or earphones), yet these transducers produce an unnatural “in your head” listening experience that many high-fidelity music aficionados find disconcerting or annoying. A 24-bit 192KHz sound card with a good reputation and positive reviews was used as the playback system (Asus Xonar Essence ST). The signal source was a bit-perfect, 20-second clip of an audiophile reference CD (Audiophile Voices I).

2. LOUDSPEAKER SOUND FIELD EMULATION IN EARPHONES USING HRTFS

The system produces a sound file that contains spatial information given by the Head Related Transfer Functions (HRTFs) and information on the room’s acoustical environment given by the early and late reverberation. The HRTFs used in the system were taken from the Listen HRTF database website [11]. The site used human test subjects to measure the HRTFs and features the measurements of 51 subjects as of May 25, 2003. There are different HRTFs for each level of azimuth and elevation represented as Head Related Impulse Responses (HRIRs). The elevations measured for ranges from -45 to 90 degrees with 15 degree increments. For the -45 to the 45 degree elevation there are 24 azimuth positions ranging from 0 to 345 with 15 degree increments. The 60, 75 and 90 degree elevations have 12, 6, and 1 azimuth positions respectively. The positions also range from 0 to 345 degrees for the 60, 75, and 90 degrees but have 30 degree, 60 degree and 360 degree increments respectively.

HRTFs are approximated with the use of Common Acoustical Pole and Zero (CAPZ) to lessen the number of coefficients used in representing it. The common poles of different azimuths and the direction dependent zeros are reduced when the sound is approximated, however, it would result in an audible degradation to the sound. The emulation of the room’s acoustical environment is comprised of four phases: pre-processing, early reflection, late reverberation, and decorrelation. The parameters used for the algorithm are actual measurements of a room where the subjective evaluation of the sound is also held.

2.1 HRTF Module

The HRTF module incorporates necessary spatial characteristics to the original input sound in order to simulate the position of two sound sources which are 30^0 to the left and the right of the listener. For the testing of the system, Full HRTFs as well as the CAPZ approximation of these HRTFs were used. In this module, the elevation of the simulated sound sources was varied. The elevations used where 0^0, 15^0, 30^0, and 45^o. For the Full HRTF manipulation, the left channel of the original sound file is convolved with the direct sound of the left HRIR and the cross talk of the right HRIR. The right channel, on the other hand, is convolved with the direct sound of the right HRIR and the cross talk of the left HRIR. After this an ITD was incorporated to both of the crosstalk channels to simulate the late arrival time of the cross talk sound compared to the arrival time of the direct sound. The new left channel produced for the output is a combination of the left direct sound and right cross talk, while the new right channel is a combination of the right direct sound and the left cross talk. The CAPZ manipulation has the same basic concept, but instead of convolving the HRIRs to the corresponding channel, the system will be filtering the channels using the equivalent CAPZ approximation of the HRTFs used in the full HRTF manipulation. The CAPZ approximation of the full HRTFs used was first computed and once computations are done the resulting poles and zeros would then be used to create a filter. A detailed explanation on how to compute for the CAPZ values can be found in [4].

2.2 Reverberation Module

The original raw stereo sound file is processed with reverberation in parallel with HRTF. The reverberation module would process the sound by incorporating the early reflection of and the late reverberation to produce a simulated sound reflection of a room as shown in Figure 1. The parameters of the reverb module are based on RT60 (energy decay) of the test room, measured by computer/sound card combination, using Praxis software from Liberty Instruments. A calibrated condenser microphone.from the same company was used as the input device.

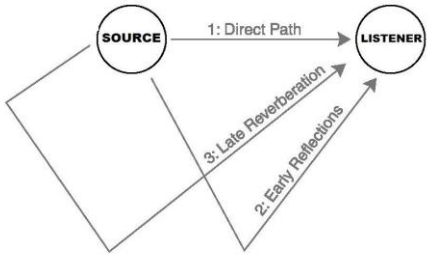

Figure 1. How a Monophonic Sound Travels in a Room and Perceived by the Listener

The raw sound file is in stereophonic format and would pass through a lowpass filter that incorporates the high frequency roll off due to room material absorption. The reverberation algorithm would require the input sound file to be converted to monophonic sound because the reverberation of a room for the left and the right channels should be applied concurrently. The sound file would then be incorporated with early reflection and late reverberation to include the emulated room characteristics. The sound would then be decorrelated to produce a stereo output sound file. This will create a difference between the left and right signals to produce a wider and more diffused sound image.

2.2.1 Early Reflection

The reflected sounds from walls and floors of a room would require a Room Impulse Response (RIR) filter that will simulate the early reflection by emulating a virtual room modeled from an actual room. The RIR would require the size of the room, the average of reflective coefficient in six frequencies, and the positions of the source and listener. Reflective coefficient or the amount of sound that is reflected by the room is inversely proportional to absorption coefficients which can easily be found in Absorption Coefficient Charts.

2.2.2 Late Reverberation

The late reverberation was simulated using an Infinite Impulse Response (IIR) filter where the original signal were filtered using six parallel comb filters, and then passed through an all pass filter. The reverberation time or RT60 allows the comb filter to control the reverberation until the sound fully decays. The RT60 depends on the materials inside the room because everything absorbs sound and is very active in 125 Hz to 4 KHz range which will eventually diminish as the frequency increases. The all pass filters with flat frequency response is used to produce the effect of multi reflections.

2.3 Mixer Module

After the reverberation and HRTF modules, the mixer module convolves the new left and right channels from the HRTF module with the reverberation for the left and right ear, respectively. These two signals are then multiplexed together to produce a WAV-compatible file suitable for playback on the computer.

3. IMPLEMENTATION

The original .wav sound file is used as an input file to the Head Related Transfer Function (HRTF) Module and the reverberation module. The HRTF module uses the Head Related Impulse Responses (HRIR) of all HRTFs, the initial pole (P) and zero (Q) order, and the Input Sound File to be used in the system. The output sound of the module is called SFAlpha for reference.

The processed sound of the HRTF module are approximated with the use of the Common Acoustical Pole and Zero (CAPZ) model to reduce the number of coefficients to represent the HRTFs. This is called SFBeta for referencing. To define the combination of poles and zeros that are used for testing, the approximated sounds' frequency responses are manually inspected. The peaks and dips in the frequency response have an audible difference during the subjective listening. An example of a combination of pole and zeros' frequency response is shown in Figure 2. There are a total of five combinations of poles and zeros that are used: 20 Poles(P)-40 Zeros(Q), 20P-230Q, 30P-50Q, 50P-100Q, and 70P-180Q. Each combination is subjectively evaluated by rating its clarity, brightness, nearness, spaciousness, and its sound quality in a scale of 0-100.

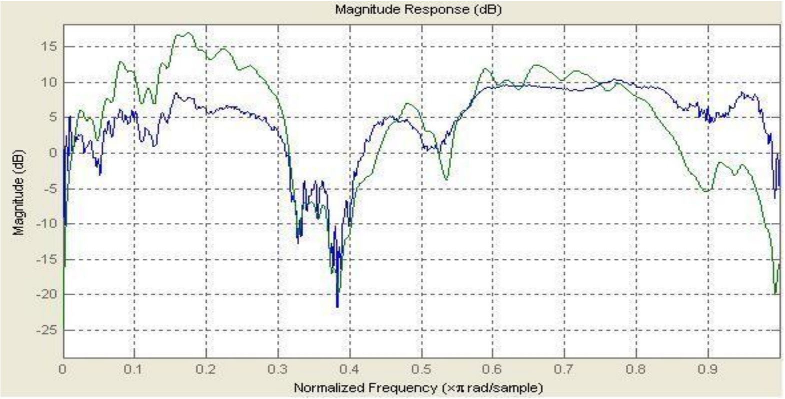

Figure 2. Frequency Response of a 70 Pole – 180 Zero CAPZ approximation (green) plotted against full HRTF frequency response (blue)

The reverberation module handles the emulation of the room's acoustical environment that affects the sound's characteristics. Figure 3 shows the room where testing was performed (and thus was used as basis for reverberation). The room dimensions are 5.47 by 3.98 by 2.78 meters, reflective coefficient of 0.948, and the positions of the loudspeaker and microphone are as shown. The RIR function used in the system is based from McGovern's Room Impulse Response Generator [6]. The pre-processed sound and RIR are convolved using Perry's High Speed Convolution [3]. The late reverberation phase uses six (6) parallel comb pass filters cascaded to one (1) all pass filter. The six comb pass filters have coprime values and the all-pass filter has 6ms delay. In this phase, it requires the RT60 characteristic of the room which is measured with the use of Praxis as its measuring tool. The sound is then decorrelated to produce a stereo output sound file to determine the left and right sound channels from the monophonic input signal. The decorrelation is based on the Mono to Stereo Upmixing using Decorrelation by Lundkvist and Ōman [5].

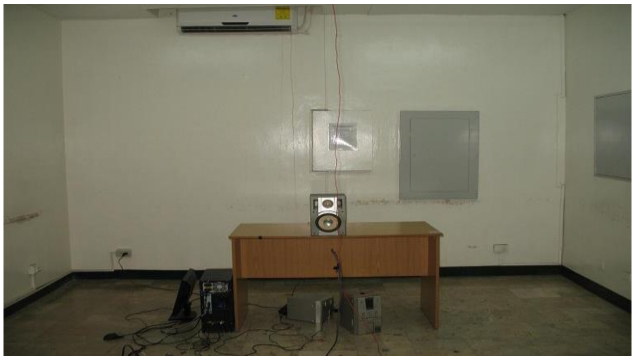

Figure 3.The Test Room

Three qualitative testing was done to provide a detailed evaluation of the sound that are incorporated with HRTFs and are done consecutively and a quantitative testing of the algorithms used to produce an externalized sound. There are a total of 10 listeners present in the tests and are same throughout the whole process.

The three qualitative testing are as follows: determining the optimal sampling rate and bit depth of the sound, subjective evaluation of the different CAPZ approximations, and to determine if the externalization of the sound is audible.

There are five stimuli used in the testing of the optimal sampling rate and bit depth. The Multi-stimulus Test with Hidden Reference and Anchor (MUSHRA) is the method of testing. One of the stimuli is the original, unprocessed .wav sound file and the other 4 stimuli are SFAlphas that are sampled at 44.1KHz with a bit depth of 16-bits and 24-bits and SFAlphas that are up sampled to 96KHz with a bit depth of 16-bits and 24-bits. Listeners are asked to rate the five stimuli from 0 to 100 according to its sound quality. They do not have prior knowledge as to which or what sound they are listening to so as to lessen the bias present in their ratings. The results are averaged to see which stimuli would have the highest rating according to its quality. The resulting combination of sampling rate and bit depth that are subjective evaluated as the best are used as the characteristics of the sound for the following tests.

There are 5 different SFBeta used, 20 Poles(P)-40 Zeros(Q), 20P-230Q, 30P-50Q, 50P-100Q, and 70P-180Q. Each approximation and an SFAlpha would have elevations of 0°, 15°, 30°, and 45°. Also in this test, MUSHRA is used as the method of testing. In total, there are 24 stimuli to be evaluated by the listeners. It is evaluated with a scale of 0-100 according to its clarity, brightness, nearness, spaciousness, and its sound quality. The approximation of 30 Poles and 50 Zeroes is evaluated as the best of the 5 approximations and are used as a reference in the next test.

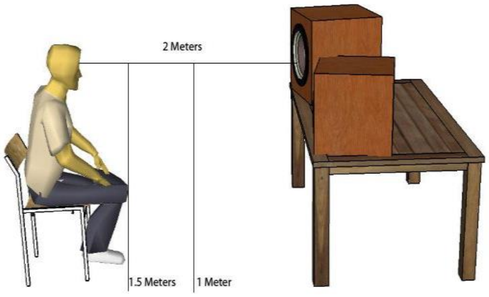

Figure 4. Illustration of the Test Setup

For determining the externalization of the sound is audible to the listeners, the ABC test method is used with a few changes on how it is implemented. The listeners are not blindfolded and are not asked to find the hidden reference from test music clips ‘B’ and ‘C’, instead, the listeners rate the similarity of test music clips ‘B’ and ‘C’ to the sound produced by sound ‘A’. It uses the sound produced by loudspeakers as its reference sound, and is represented as ‘A’. The SFBeta with an approximation of 30 Poles and 50 Zeroes is represented as ‘B’ and the SFAlpha, the full-range HRTF processed sound, is represented as ‘C’. The SFAlpha and the SFBeta are mixed with reverberation that is measured in varying distances from the speaker to amplify the effect of externalization. The distances used for the reverberation are 1 meter, 1.5 meters, and 2 meter. Figure 4 shows how the testing is done. The subject is asked to determine the similarity between music clips ‘B’ and ‘C’ from music clip ‘A’ which are held in varying distances such as 1 meter, 1.5 meters, and 2 meters with an elevation of 45°, 30°, and 15° respectively and 0° for all distances. There are a total of 12 stimuli that are rated by the test subject. The test subject is asked to listen to both ‘B’ and ‘C’ sounds and rate it according to its sound quality and its spaciousness with a rating of 0-100. This is to determine if applying reverberation to the sound would amplify the effect of sound externalization. When the test subject is finished rating its overall sound quality, the test subject is asked to listen to the sound produced by the loudspeaker and listen to music clips ‘B’ and ‘C’ again but rating it according to its similarity with respect to ‘A’ from a scale of 1.0-5.0. The scale is based on a five grade impairment scale (5.0 as excellent, 4.0 to 4.9 as good, 3.0 to 3.9 as fair, 2.0 to 2.9 as poor, or 1.0 to 1.9 as bad).

4. RESULTS AND ANALYSIS

The ratings of each listener for the four different stimuli are subtracted from the rating of the original sound file. This is to see the difference in quality that is present with the sounds. If the result is a negative number, it means that the quality of that sound is poorer compared to the original sound. After the computations for each of the test subject’s ratings, the results for each test sound file of each test subject is averaged to see the overall difference of that sound from the original sound. Figure 5 shows the overall rating of the varying SFAlphas.

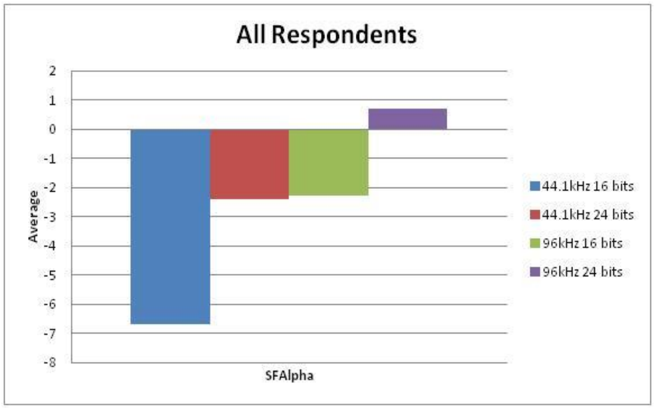

Figure 5. SFAlpha Overall Rating Against the Original Sound

The SFAlpha with a sampling rate of 96KHz and a bit depth of 24-bits are rated as the best in terms of sound quality, which verifies the fidelity of the playback system. Listeners rated the SFAlpha with a sampling rate of 44.1KHz and a bit depth of 16-bits as the worst. The results with the sampling rate of 96KHz and a bit depth of 24-bits and the sampling rate of 96KHz and a bit depth of 16-bits are perceived to have almost no audible difference with one another with having a difference in the results of only 0.1.

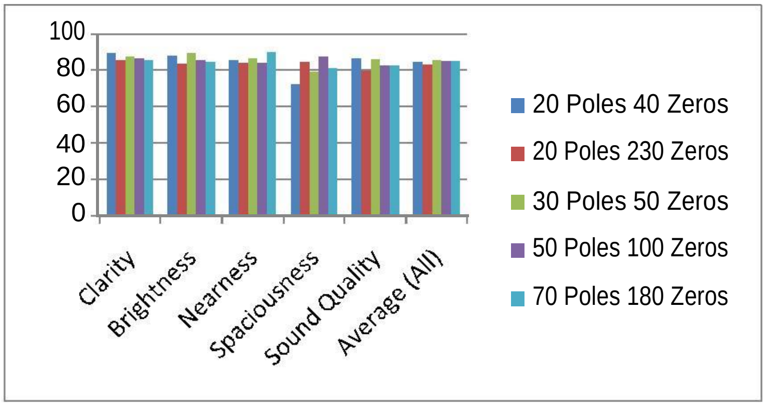

Figure 6. Subjective Evaluation of the Different CAPZ Approximation with 0° Elevation

The different approximations are evaluated separately by their elevation to see if the elevation present in the test music clips would have a perceptible audible difference. Figure 6 shows the subjective results of the SFBeta with a 0° elevation. The SFBeta with an approximation of 30 Poles and 50 Zeroes is rated overall as the best. The 70 Poles and 180 Zeroes approximation has the highest rating in terms of nearness which means that the approximation is interpreted by the listener as having the least externalized sound among the other approximations. The approximation with 20 Poles and 230 Zeroes is rated as the worst and it seems that boosting up that much zeroes without any change in the poles would have an audible degradation according to the listeners.

Tests were conducted similarly for elevations of 15, 30 and 45 degrees, with varying CAPZ approximations. Figure 7 summarizes the results.

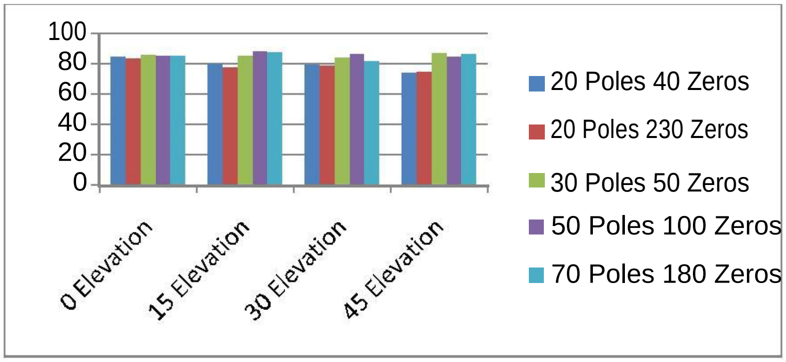

Figure 7. Overall Subjective Evaluation of the CAPZ Approximation

Figure 8 and 9 show the results of 0 degree elevation of SFAlpha and the SFBeta respectively. The SFAlpha with a reverberation parameter that is measured at 2 meters is considered as the best in terms of its overall sound quality. Having a distance of at least 2 meters from the sound source improves the overall listening experience of the test subject when using earphones as its medium of sound reproduction. The test subject's ratings with test music clips that were emulated near them were not perceived to have a good sound quality. Also, the listeners are able to determine a more spacious room when the distance between his position and the sound source increased. The approximation with a reverberation parameter that was measured at 1.5 meters is considered as the best in terms of sound quality and a distance of 2 meters from the loudspeakers is rated as the most spacious. The listeners had a hard time in perceiving the spaciousness that should be present inside the room when listening to a small distance such as 1 meter and 1.5 meters.

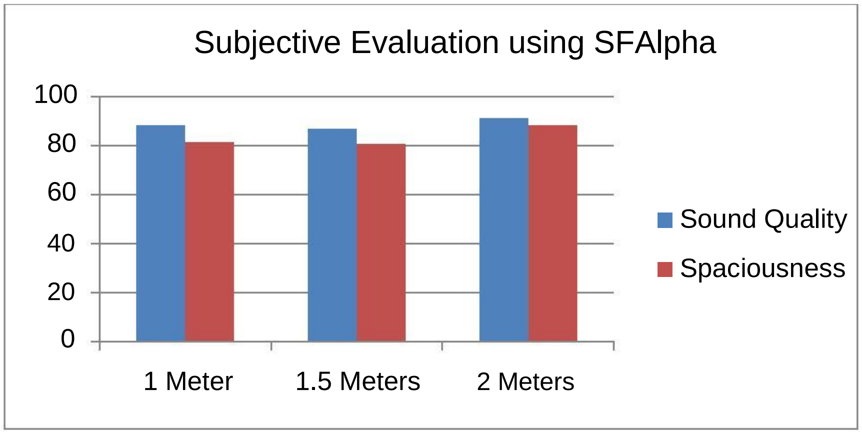

Figure 8. Full-range HRTF Sound with 0° Elevation

Tests were conducted with varying virtual elevations of 15, 30 and 45 degrees, at a distance of 1, 1.5 and 2 meters. The subjective evaluations for these tests remained similar to the results above of zero degree elevation.

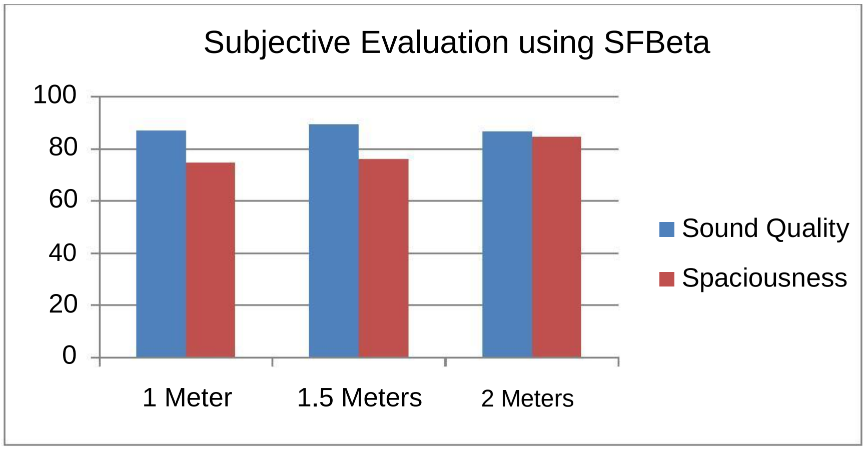

Figure 9. CAPZ Approximated Sound with 0° Elevation

For the externalization testing, the results for the SFAlpha with varying elevation and distances are shown in Figure 10. All of the ratings are around the range of 3.4 – 4.0 which means that the test music clips are perceived to be fairly good similarity between earphone versus loudspeaker reproduction. Judging by the results, when using the Full-range of HRTFs to implement externalization to the sound, having a distance of 1 meter from the loudspeaker is considered as the optimal setup for the listeners. At a distance of 2 meters, there are no notable difference between its sound quality in different elevations and are almost perceived as the same in sound quality and similarity.

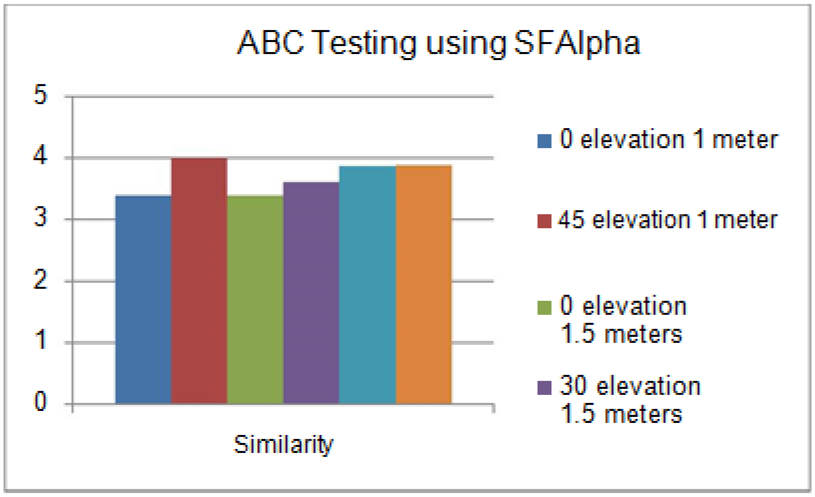

Figure 10. Similarity of Full-range HRTFs to the Original Sound

Finally, the results for the SFBeta with varying elevation and distances are shown in Figure 11. The SFBeta with an elevation of 0° with a distance of 1.5 meters is rated as the closest to the sound produced by a pair of loudspeakers, with a rating of 4.1, with regards to the results of the ABC test of SFAlpha, with a highest rating of 4.0. The range of the ABC test results are from 3.3 – 4.1 which are also perceived as having a fairly good similarity with the sound produced by a pair of loudspeakers. When SFAlpha is approximated, most of the ratings stayed the same with an exception to the 45° elevation with a distance of 1 meter and the 0° elevation with a distance of 1.5 meters.

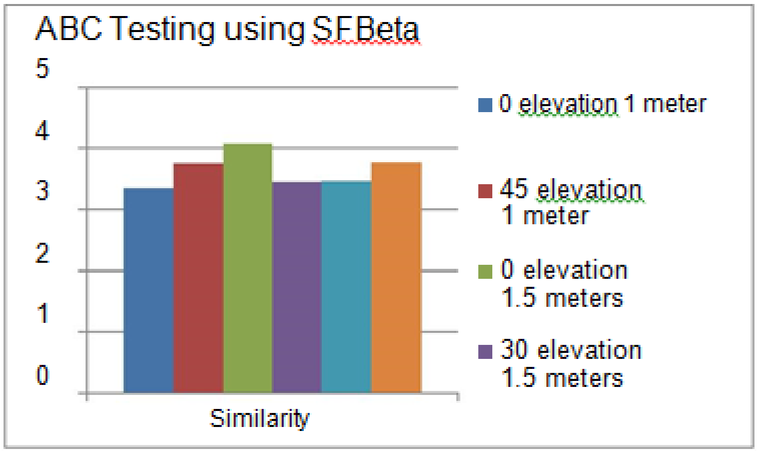

Figure 11. Similarity of CAPZ Approximation to the Original Sound

5. CONCLUSION

The results for the CAPZ approximated HRTFs with added reverberation shows that it is able to produce sounds with an average loudspeaker externalization similarity rating of 3.6. The results for the full-range HRTFs with added reverberation shows that it is rated with an average of 3.66, performing marginally better than CAPZ approximated HRTFs in terms of externalization. This indicates that the system was able to produce a sound with satisfactory, but not outstanding, externalization for the listeners. It can also be seen from the results that the CAPZ approximation does not lag far behind the full range of HRTFs in terms of sound quality and spaciousness even when reverberation is added.

6. RECOMMENDATIONS

The CAPZ-simplified HRTF provides a reasonable approximation of a full HRTF, however the fidelity of even a full HRTF, as done in this study, is limited by the fact that the HRTF used is a generalized one and not that of the actual response from each subject (person) undergoing the evaluation. Each person has their own unique HRTF characteristic, which has been shown in other research to be critical, particularly in achieving front-back localization [9]. Further, the Interaural Time Difference (ITDs) and Interaural Level Difference (ILDs) are also unique for each person. These ITDs and ILDs are well known localization cues and it has been shown that both ITDs and IIDs are important parameters for the perception of sounds originating from the horizontal plane [2]. Researching the different effects of the ITDs and the ILDs may be another area of research for the improvement of the overall externalization and localization of the output. As such, it is recommended that the HRTFs, ITD and IIDs of each individual be measured and used as basis for further fidelity tests of the CAPZ approximation.

Hybrid Reverberation Algorithm can be used in the implementation of the reverberation module, based on measuring an accurate impulse response of the actual room. This algorithm would require additional processing power compared with the simplified FIR filter algorithm implemented in this study. Further improvement of the reverberation algorithm could include a research on how to simplify the Hybrid Reverberation Algorithm while maintaining its accuracy.

7. ACKNOWLEDGEMENTS

We would like to thank Dr. Joel Ilao and Mr. Macario Cordel II for clarifying some DSP concepts; Mr. Karlo Shane Campos and Mr. Christian Echavez for suggestions in methods used testing.

8. REFERENCES

[1] Bech-Kristensen, T., Hangerman, B., Gabrielsson, A., & Lundberg, G. (1990, April 6). *Perceived sound quality of reproductions with different frequency response and sound levels.* Stolkholm: Karolinska Institute.

[2] Cheng, C., & Wakefield, G. (n.d.). *Introduction to Head-Related Transfer Functions (HRTF's) Representations of HRTF's in Time, Frequency, and Space.* Anne Arbor, Michigan, USA: University of Michigan.

[3] Giesbrecht, H., McFarland, W., & Perry, T. (2009). *Algorithmic Reverberation: Combining Moorer's Reverberator with Simulated Room IR Reflection Modeling*. Retrieved from University of Victoria.

[4] Haneda, Y., Makino, S., Kaneda, Y. & Kitawaka, N. (1999). *Common-Acoustical-Pole and Zero Modeling of Head-Related Transfer Functions*. IEEE Trans. on Speech and Audio Processing, Vol. 7, No. 2, p 188-195.

[5] Lundkvist, A. & Oman, P. (2009, March 29). *Mono to Stereo Upmixing Using Decorrelation*. Retrieved from Lulea University of Technology.

[6] McGovern, Stephen G. A Model for Room Acoustics, 2004.

[7] Moorer, J. (2009) *About This Reverberation Business*. Computer Music Journal, Vol. 3, No. 2, pp. 13-28.

[8] Robjohns, H. (2003). *Mixing on Headphones: What To Use and How To Do It.* Retrieved July 18, 2011, from http://www.soundonsound.com/sos/dec03/ articles/mixingheadphones.htm.

[9] Wanabe, K. et al. (2007) *Estimation of interaural level difference based on anthropometry and its effect on sound localization*. Acoustical Society of America.

[10] Wang, L., Yin, F., & Chen, Z. (2008). *An "Out of Head" Sound Field Enhancement System for Headphone*. Zhenjiang, China: School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116023, China.

[11] Warusfel, O. (2003). *Listen HRTF Database*. Retrieved July 19, 2011, from http://www.recherche.ircam.fr/equipes/sales/ listen/index.html,