Audio Engineering Society

Conference Paper

Presented at the Conference on
Audio for Virtual and Augmented Reality
2016 September 30–October 1, Los Angeles, CA, USA

Lateral Listener Movement on the Horizontal Plane: Part 2 Sensing Motion Through Binaural Simulation in a Reverberant Environment

Matthew Boerum[1,2], Bryan Martin[1,2], Richard King[1,2] and George Massenburg[1,2]

[1] *Graduate Program in Sound Recording, McGill University, Montreal, Canada*

[2] *Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), Montreal, Canada*

Correspondence should be addressed to Matthew Boerum (matthew.boerum@mail.mcgill.ca)

ABSTRACT

In a multi-part study, first-person horizontal movement between two virtual sound source locations in an auditory virtual environment (AVE) was investigated by evaluating the *sensation of motion* as perceived by the listener. A binaural cross-fading technique simulated this movement while real binaural recordings of motion were made as a reference using a motion apparatus and mounted head and torso simulator (HATS). Trained listeners evaluated the *sensation of motion* among real and simulated conditions in two opposite environment-dependent experiments: Part 1 (semi-anechoic), Part 2 (reverberant). Results from Part 2 were proportional to Part 1, despite the presence of reflections. The simulation provided the greatest *sensation of motion* again, showing that binaural audio recordings present less *sensation of motion* than the simulation.

1 Preface

This multi-part study is focusing on the listener's auditory perception of movement between virtual sound sources from the first-person point of view in virtual reality applications. As a result, this paper serves to report only the additional work as performed in Part 2 of the study, though reference to Part 1 experiments will often occur. It is therefore suggested that the work presented in the publication on Part 1 be reviewed to gain full comprehension and background of the study's experimental procedures and data [1].

2 Introduction

Virtual reality technology provides a 360-degree, interactive immersive experience through the delivery of multisensory spatial awareness. While graphically modeled 3D environments can enhance

spatial awareness and a sense of reality through the presentation of visual information, the same is true when sound is presented in 3D.

Auditory Virtual Environments (AVEs), which model the behavior of sound within virtual spaces, present spatial awareness through acoustically generated audible events [2]. These events, or spatial audio cues are related to the sound's interaction with the acoustic space. For example, early reflections help the listener to perceive the sound with distance and direction while reverberation adds a sense of overall room dimension [3][4]. In general, early reflections occur up to 80ms after the direct sound [3][5]. As these reflections begin to regenerate and late reflections form more frequently with time, reverberation begins until the sound eventually decays to inaudible levels [6]. An illustration of this can be seen through the example of a room impulse response in Figure 1. While early reflections are known to increase localization cues [7], an extension of the early reflection period can have the inverse effect, especially with the addition of a strong reverberation period [8][9].

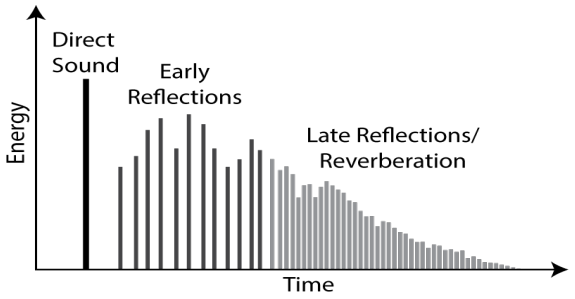

Figure 1. Example of a room impulse response.

Modeling the complex behavior of sound to accurately generate these events for precision delivery of spatial audio cues is a significant challenge for developers. In fact, once motion is involved (giving the user the ability to navigate a virtual environment in real-time), performance and quality of the spatial audio becomes compromised. This is a problem directly related to the level of complexity in the AVE and the constraints of the spatial audio processor. Moreover, poor rendering and simulation of listener movement within these AVEs can lead to audible artifacts (from miscues) in the portrayal of "realistic" 3D sound [10].

Two of the most important audible cues necessary for 3D localization and auditory motion perception are the interaural level difference (ILD) and interaural time difference (ITD). These values are determined by the change in volume and time arrival, respectively, from ear to ear for a given sound source located around the head [11]. Simply changing these values in real-time for binaural audio simulations often presents undesirable effects such as comb filtering and localization blur based on multi-positional, interpolated delays. Since sensing motion from audible cues is something we take for granted in real world situations, when motion is produced incorrectly in virtual reality, these additive miscalculations of reality become immediately apparent.

Although an abundance of binaural audio research exists, most of this work was conducted to determine sound source localization [7][12][13][14], basic sound object motion around the listener—from early physical experiments [15] to recent digital studies [16], and algorithmic modeling approaches for 3D sound reproduction [17][18]. Additional work is required to specifically focus on the performance and quality of listener motion from these binaural simulation techniques. As such, this study intends to serve as foundational research in a new approach to understanding this topic. This multi-part study is investigating how auditory listener motion is perceived when moving between two static sound sources in an AVE. The basic *sensation of motion* perceived by the listener is the measure used to evaluate the performance of binaural crossfading simulations against real world binaural motion recordings, and motionless simulations.

3 Background

3.1 Auditory Motion Processes in Virtual Reality

Simple binaural crossfading is a technique that is commonly used in virtual reality applications to simulate the listener's perspective of auditory motion while moving from one virtual sound source to another. Simulating motion through audio can be achieved through a relatively simple process of crossfading between spatial audio cues, which are captured or simulated from two positional extents on a given path.

In practice, simple changes in reverberation and early reflections are often provided as sufficient simulations of listener motion between two acoustic positions. This technique is not 3-dimensional, but is used often in game engines as a solution for motion perception and acoustic awareness. Inside the game engine, a reverb zone [19] is created as an area of graphical 3D space designed to generate specific acoustic parameters for additive reverb generation. When a listener enters a reverb zone, a

function is triggered, which applies a customized reverb to all sound sources heard within the zone by the listener. At another point in space represented by a second reverb zone, the same audio will behave differently due to the changes presented by the acoustic parameters of the new reverb zone. As a listener passes from one zone to the next, the parameters are either crossfaded, or the output of the reverb is attenuated giving the listener the perception of movement from one AVE to the next.

Likewise, with binaural audio, panning the localization of sound objects in 3D space within a single AVE can be easily performed through crossfading the positional head related transfer function (HRTF) convolution processes [20]. This method is far superior to simply changing the reverberation characteristics, but is still very basic in operation. Due to its simplicity, this technique is commonly used in virtual reality engines as an auditory motion solution. Since virtual reality content is quickly becoming a commercially accessible product, it is necessary to evaluate the performance and perception of these simple binaural crossfading processes. By understanding how simulations compare to the real world, one can better determine the performance and/or pitfalls of this technology.

4 Method

For this multi-part study, an experiment was devised to both capture and simulate auditory listener movement from real world physical motion in various acoustic environments. The goal was to evaluate basic binaural crossfading as a motion simulation process, and to analyze how reverberation affects the *sensation of motion* perceived by the listener. For all experiments in this multi-part study, the author maintained a consistent and repeatable experimental design, which controlled the methods used in the measurement procedure [1]. Therefore, the only change in Part 2 was that the measurements were conducted in a different acoustic environment, which is explained in further detail in Section 4.2.

4.1 Summary of the Replicated Measurements

The study accurately and consistently controlled a measurement process, which moved a motion apparatus along a horizontal path, to capture the binaural auditory motion between two synchronized loudspeakers. This motion was captured in both directions (left to right, and right to left). The binaural recording was performed with a head and torso simulator (HATS) mounted securely on the motion apparatus. The source audio signals used in Part 1 (stereo jazz, mono pink noise, and mono male speech) were reproduced through the loudspeakers as the stimuli for the binaural recordings. These binaural recordings became the *experimental reference* for all simulations.

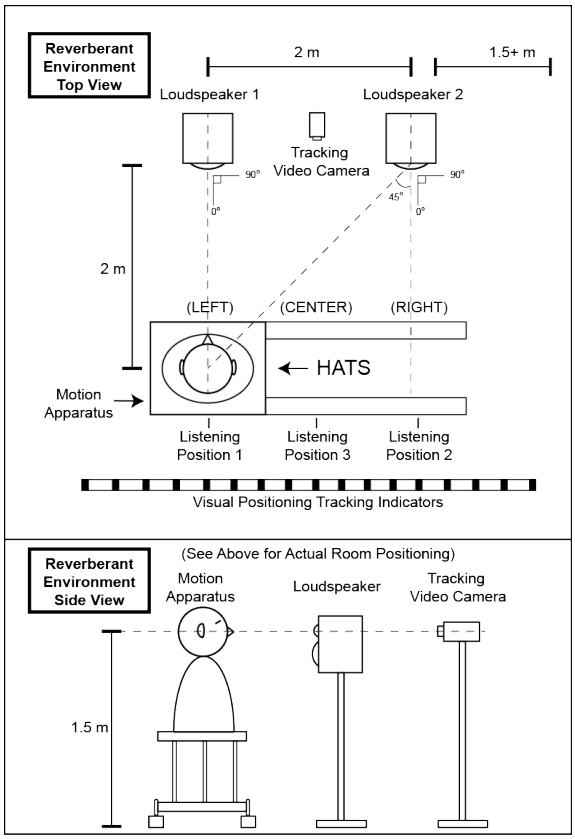

Figure 2. Layout of the reverberant environment with replicated measurement setup.

HRTFs were gathered statically at the opposite extents of the motion path at listening position 1 and listening position 2 (see Fig. 2). These measurements represented the localization cues for the two simultaneous sound source positions in space. Using motion-tracked video of the binaural motion recordings as a reference for position over time, auditory motion simulations were then created through binaural crossfading. These *experimental simulations* were synchronized in time and position to match the binaural reference recordings.

A static head related impulse response (HRIR) captured at position 3 was convolved with the stimuli and represented an audible *anchor* variable presenting no motion. Finally, first-person POV video was made for later use in the Audio/Video listening test. This captured a visual of the motion path as if perceived from the HATS' line-of-sight.

As in Part 1, angular localization positions were captured through HRIRs at five static positions along the motion path. These positions included 45 degrees left and right, 26.6 degrees left and right, and 0 degrees center. The positions would be used later to convolve with stimuli for the localization training pre-test defined in Section 5.

Figure 3. Motion apparatus setup for Part 2.

4.2Acoustic Environment

Part 1 of this study was performed in a semi-anechoic chamber in order to reduce the influence of acoustic reflections on audible motion and localization cues. This acoustically controlled environment had a total room volume of 124 m^3 and a reverb time (T30) of 90 ms.

In contrast, measurements for Part 2 were conducted in a highly reflective, reverberant space. This space, which we will term the *reverberant environment*, was an unfinished tracking room in the McGill University recording studios (Fig 3). It was temporarily used for storage during final studio construction. The non-uniform design, along with scattered storage objects and hard finished walls made the space an ideal environment for unpredictable acoustic reflections, providing complete opposition to the controlled acoustic environment used in Part 1. The reverberant environment had a total room volume of 293 m^3, nearly twice the semi-anechoic environment, and a reverb time (T30) that was 10 times the duration at 900 ms. It also had strong, equal energy early reflections at 4 ms, 14 ms, 21 ms, 25 ms and 40 ms—a potential problem for localization.

4.3POV Video Reference

The experiment also included first person point-of-view (POV) video for reference in perceptual tests. Video was included in the experiment to investigate whether fixed, visual guidance of the motion path would improve the *sensation of motion*. As visual information can be seen as a potential catalyst for change in perception, and since this experiment is studying motion perception through audio, the author chose to limit the viewing ability for testing purposes in order to control this additional variable as much as possible. It has been demonstrated often that reduced field of view in virtual environments inhibits spatial awareness [21][22][23], so by controlling the POV video to only show the front of the room from the HATS perspective, the author could prevent extreme bias from 360 degree investigation of the virtual space, thus preventing the influence of visual information from dominating the auditory investigation altogether.

5 Listening Test

Through two versions of a listening test, 59 semi-experienced subjects evaluated the audible motion examples created during the measurement stage. These subjects also performed a localization training pre-test to familiarize themselves with the listening test environment and software, and to determine the total percentage of localization accuracy among subjects when presented with the experiment's 5 positional HRTFs. As in Part 1, a threshold of the total average localization accuracy of the subjects was set to 90%. A value above this threshold would validate the ability of the experiment's HRTFs to represent accurate virtual positions along the motion path, and further validate any motion simulations generated from the use of these HRTFs. Lower values would indicate possible problems in the HRTFs gathered in the experiment.

Listening evaluations were performed in a sound proof listening suite using headphones and an exact replication of the listening test setup from Part 1, including the same Lateral Listener Movement testing application. However, this test only used audio examples from the measurements taken during Part 2 of the experiment to focus on the influence of acoustic reflections.

5.1Replicating the Testing Method

As in Part 1, subjects participated in a double blind, MUSHRA-style evaluation of the auditory motion examples [24]. Using this method based on a 100-point continuous quality scale (CSQ), the subjects were asked to rate the *sensation of motion* perceived by the presentation of the stimuli through 3 randomized examples of auditory motion (reference, simulation and anchor). These groups of examples were presented over 12 separate trials comprising 6

left-to-right passes, and 6 right-to-left passes along the motion path. The binaural auditory motion recordings served as the hidden reference; the binaurally crossfaded auditory motion simulations served as the independent variable; and the static HRIR convolutions served as the anchor (presenting no motion).

Of the 59 listening test subjects, 30 were selected at random to take an Audio Only version, while the other 29 participated in an Audio/Video version containing additional first-person video of the motion path. Subjects were assigned assessor identification numbers of "A###" for the Audio Only test and "V###" for the Audio/Video test.

6 Results

6.1Responses

In both test versions, the Shapiro-Wilk normality test showed that the sample data did not come from a normally distributed population for all conditions. Response data was therefore tested for a p value of $p < .05$ through a one-way analysis of variance on the distribution of data. All responses from the Audio Only and Audio/Video tests showed a significant difference between their pairwise comparisons of the signal by condition, $H(2) = 635.56$, $p < .05$; and $H(2) = 500.91$, $p < .05$ respectively.

When evaluating the data shown in Figure 4, one can see that for both versions of the test, the pairwise comparisons of responses by condition for all signals show that the simulation provided a greater *sensation of motion* overall to the listener while the anchor provided (by a large margin) the least *sensation of motion*, validating its use as an anchor in the perceptual tests.

Breaking down this data, one can see that when evaluating the responses by signal for individual conditions, it is clear that both the simulation and reference provided a similar *sensation of motion* when male speech was presented (see Fig. 5). Interestingly, Figure 6 shows that the presentation of male speech through the binaural motion reference reduced the *sensation of motion* significantly in comparison to the reference results recorded in Part 1. Moreover, the simulation results for male speech were basically unchanged from Part 1 to Part 2.

Again in Figure 5, one can see how the presentation of music and male speech through the binaural crossfading simulation gave similar measures of *sensation of motion*, but also how pink noise greatly enhanced this measure above all other test conditions (reference and anchor).

6.2Localization Accuracy

As in Part 1, there was only one localization training pre-test version. Therefore, all subjects' responses were pooled together to determine total localization accuracy for the entire group. Each of the five virtual positions was presented twice to each subject, giving 354 total trials per position over the entire subject group. The results of the localization training pre-test provided a mean accuracy of 90.5%, slightly above the experiment's threshold of 90%. Despite one less subject participating in Part 2 of the study than in Part 1, more localization errors occurred in Part 2 with the greatest amount of error coming from the Left Center virtual position. An illustration of these results can be seen in Table 1.

6.3 The Effect of POV Video on the Reference

The *sensation of motion* was significantly reduced when stimuli were presented through the binaural motion reference condition with the addition of the POV video. Though the results are still proportional to the Audio Only test results, the ceiling of these results was reduced by 15-20% for individual signals (Fig. 5). This result was expected as the viewing angle was fixed and head-tracked movement or investigation of the virtual space was not made possible for this experiment.

7 Analysis

The results from Part 2 for both versions of the listening tests demonstrate proportionality among data. This proportionality was present between the two test versions in Part 1 as well. In fact, Figure 6 shows that the trend exists among all data from Part 1 and Part 2.

This data validates that the *sensation of motion* is a measure that can be used to detect the perception of listener movement in an AVE. Additionally, it confirms that a binaural crossfading simulation is an acceptable process for creating a perception of listener motion in its basic form (sensation only), which is at least equal to real world motion as recorded through a dummy head. However, the data does not reflect what the measure of *sensation of motion* means in terms of directional accuracy or listener preference, but it certainly provides validation for further study in determining these factors, and to eventually understand what factors determine one's perception of audio quality in AVEs.

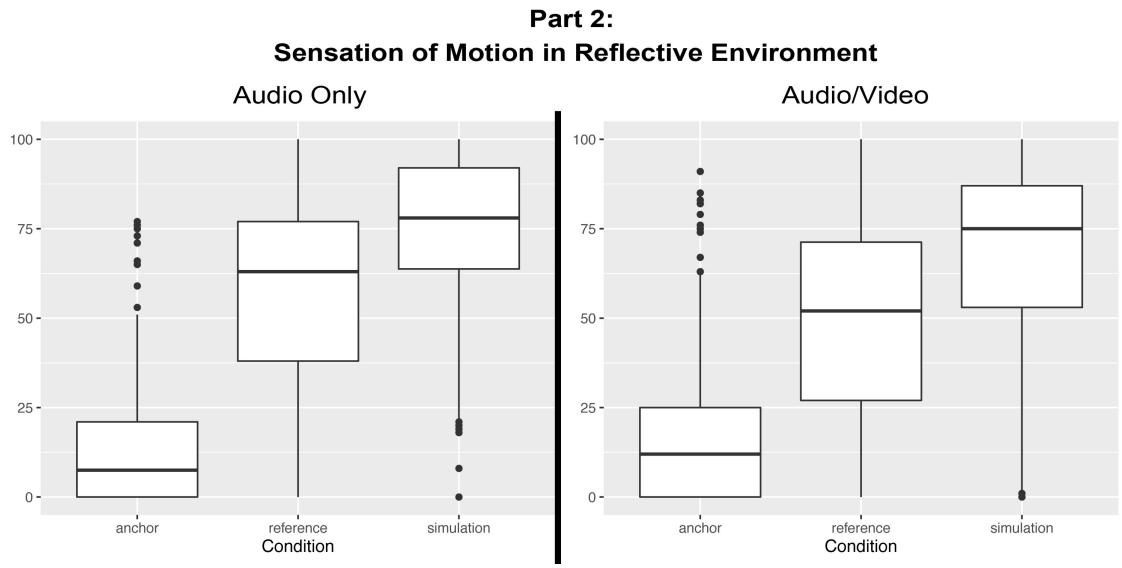

Figure 4. Responses by condition for all signals (Audio Only: Left | Audio/Video: Right).

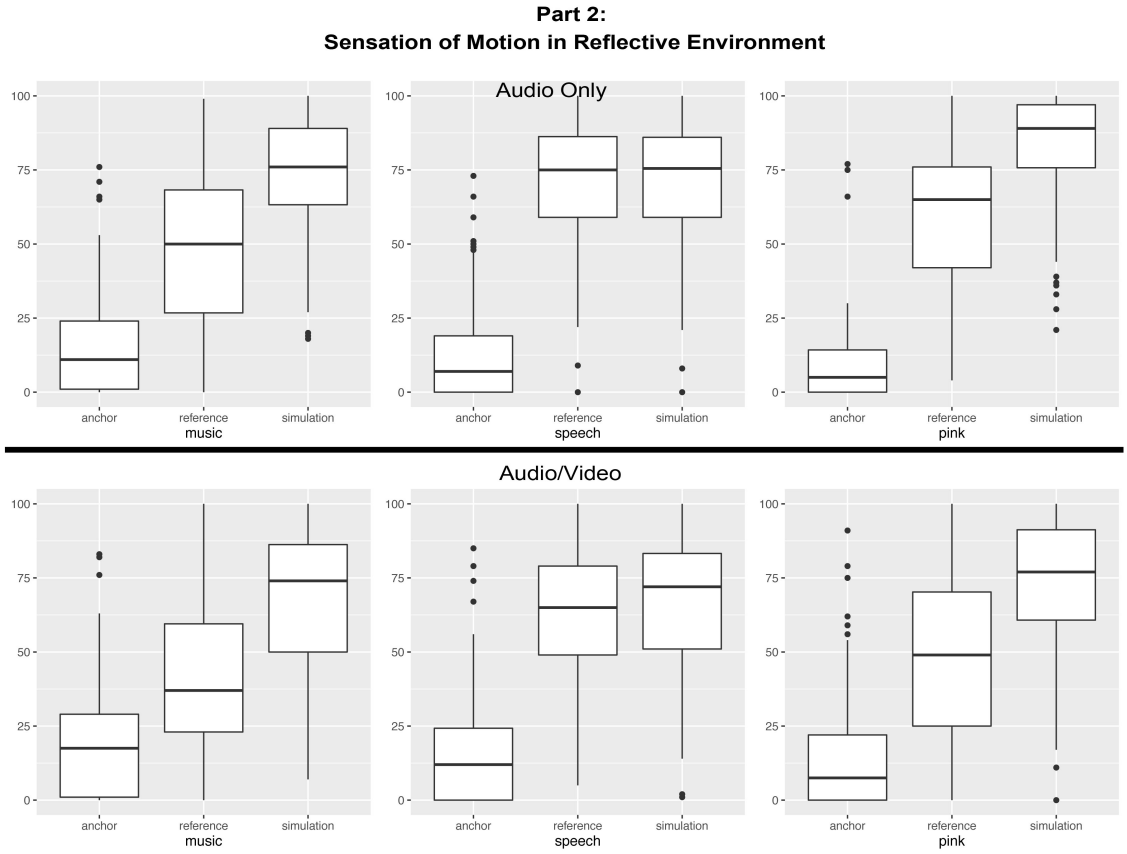

Figure 5. Responses by signal for individual conditions (Audio Only: Top | Audio/Video: Bottom).

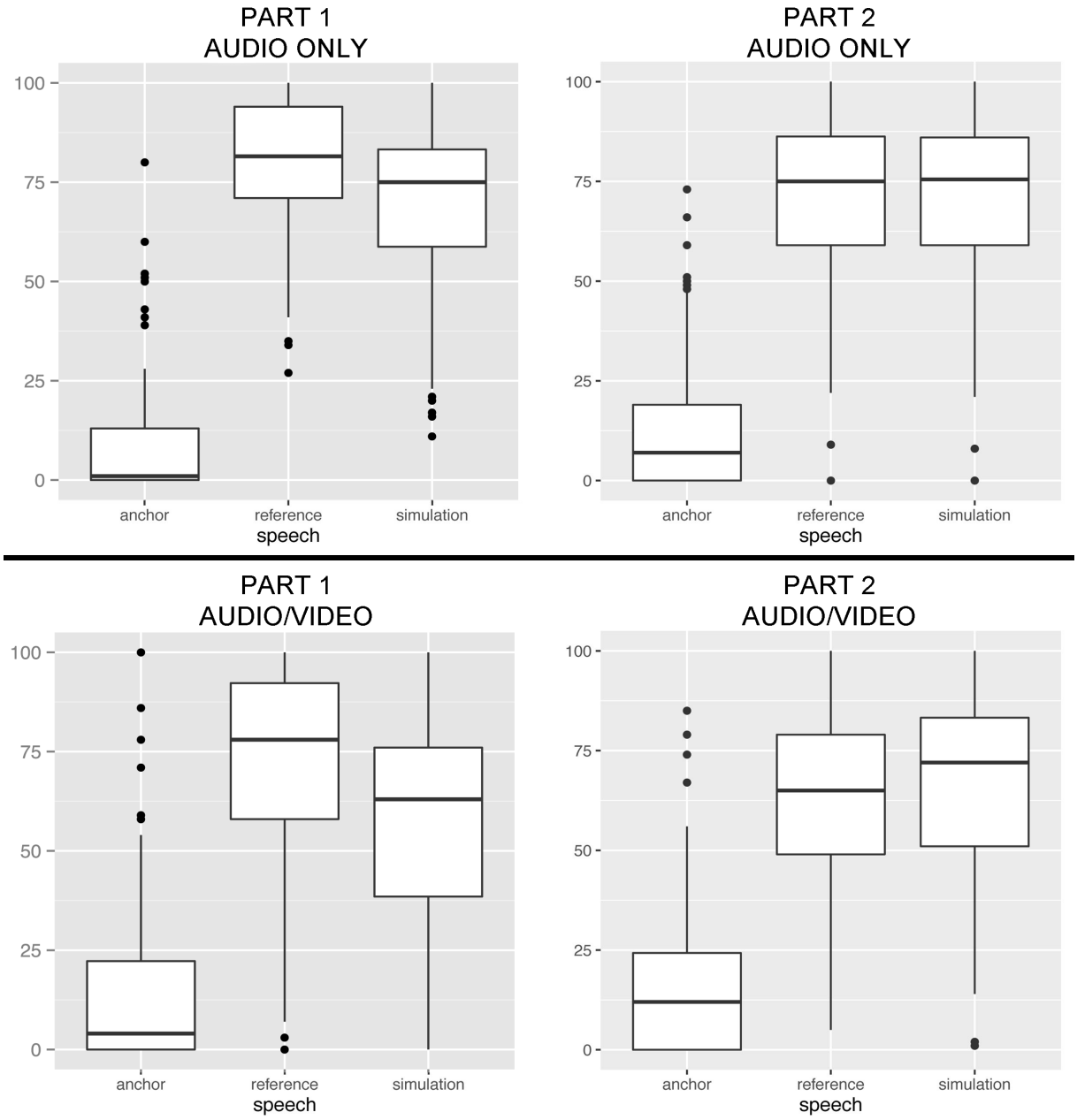

Figure 6. Comparison of Part 1 to Part 2 responses by signal for individual conditions (Audio Only: Top | Audio/Video: Bottom).

7.1 Reflections & Sensation of Motion

The presence of reflections seemed to improve the overall *sensation of motion* as presented by the binaural auditory motion simulation. Reflections seemed to account for the increased measure of responses in the static anchor as opposed to responses in Part 1. It is interesting to note, however, that the reflections reduced the *sensation of motion* as produced by the binaural motion reference from the HATS recordings. Perhaps this is due to the exact delivery of the HRTFs from the HATS during the playback of these recordings versus the continuous averaging that occurs when these HRTFs are binaurally crossfaded for the simulation. It could then be proposed that binaural crossfading is a suitable method for overcoming the problems of non-individualized HRTFs for motion perception.

7.2 Video Influence

Reflections seemed to play a big part in the Audio/Video test as the presence of video in Part 2

seemed to confuse listening subjects on the apparent size of the virtual room (according to test comments). The subjects felt that the reverberation did not match the real binaural recordings (reference), though this was naturally collected and reproduced in all examples. It is believed that the reduced visual awareness of the room in the video caused this confusion more than the reflections alone. Further investigation of visual awareness will be performed in upcoming experiments.

Sound Position	# of Trials	Accuracy	Errors	Total Accuracy
Left	354	87.6%	44	90.5%
L. Center	354	85.3%	52	
Center	354	90.1%	35	
R. Center	354	93.8%	22	
Right	354	95.5%	16	

Table 1. Localization accuracy for individual sound source positions, total localization accuracy (Part 2: Localization Pre-Test).

7.3 Localization with Reflections

The localization accuracy of 90.5% for the entire subject pool satisfies the author as validation that the HRTFs were successful at providing accurate localization. However, it should be noted that the accuracy was reduced in comparison with Part 1's results of 92.7%. The long decay and high order, high energy reflections of the reverberant environment are assumed to have played a key role in this reduced accuracy by causing significant localization blur.

7.4 Listening Test Comments

The following comments are included in this paper as reference to the results and analysis:

Assessor A120: *"B and C (reference & simulation) were equally good."*

Assessor V105: *"It's distracting how the room from the visual doesn't match the visual's reverb."*

Assessor V125: *"C (simulation) was disorienting, though it provided the most sensation of motion."*

8 Future Work

With the successful completion of Part 1 and Part 2 of the multi-part study, the author will now conclude the study with a follow-up experiment replicating the experimental measurement and procedure in virtual reality through a head-mounted display (HMD). The design of the semi-anechoic chamber and reverberant environment in virtual space will allow for listening tests to be performed with precision to exactly match the scale of the setups in the previous experiments. This experiment will also allow for the listening subjects to visually investigate the entire measurement environment, unlike the presented videos in the Audio/Video trials of Part 1 and 2. The author hopes that this future experiment will provide information on the importance of visual awareness when determining *sensation of motion* in an AVE and virtual reality environments. As stated previously, the author plans to move forward with experiments on directional accuracy in binaural crossfading simulations, as well as listener preference for specific techniques.

9 Conclusions

In Part 1 and Part 2 of this study, the simulation of auditory listener motion consistently presented a *sensation of motion* that was perceived as equal or greater than the real world reference. The presence of reflections in the experiment increased this measure for the simulation while reducing it for the real world reference, and localization accuracy. Early acoustic reflections in high order, and high energy combined with the significant reverberation period seemed to have been a catalyst for this reduction of quality. It might be argued that the continued success of the binaural simulation is due to one's innate awareness of reality vs. virtual reality. Though this research is limited specifically to self-motion perception, it could be suggested that these results show a preference for hyper-reality when dealing with virtual reality situations, especially when overly reflective environments present significant localization blur. As a result, the third and final part of this research will investigate how visual awareness of virtual rooms through an HMD impacts the listener's *sensation of motion*.

References

[1] M. Boerum et al., "Lateral Listener Movement on the Horizontal Plane: Sensing Motion Through Binaural Simulation" *AES 61st International Conference: Audio for Games*, pp. 1—10 (2016).

[2] J. Blauert, "Analysis and Synthesis of Auditory Scenes" *Communication Acoustics*, Springer-Verlag, pp. 4-7 (2009).

[3] D. Begault, "Reverberation" 3-D Sound for Virtual Reality and Multimedia, NASA, pp. 82-85 (2000).

[4] M. Mehta et al., "Design of Rooms for Speech" *Architectural Acoustics*, Prentice Hall, pp. 209 (1999).

[5] M. Vorländer, "Auralization of Spaces" *Physics Today,* vol. 62, no. 6, pp. 35—40 (2009).

[6] F. A. Everest, K. C. Pohlmann "Acoustics of Listening Rooms" *Master Handbook of Acoustics*, 5th ed., McGraw Hill, pp. 338 (2009).

[7] A. Nykänen et al., Effects on Localization Performance from Moving the Sources in Binaural Reproductions, *INTERNOISE and NOISE-CON Congress and Conference Proceedings*, Vol. 247, pp. 4023–4031 (2013).

[8] S. Devore et al., "Accurate Sound Localization in Reverberant Environments Is Mediated by Robust Encoding of Spatial Cues in the Auditory Midbrain" *Neuron,* vol. 62, pp. 123-134 (2009).

[9] W. M. Hartmann, "Listening in a Room" In R. Gilkey, T. Anderson (Ed.), *Binaural and Spatial Hearing in Real and Virtual Environments*, Psychology Press, pp. 194 (2014).

[10] S. Carlile, J. Leung, "The Perception of Auditory Motion" *Trends in Hearing*, vol. 20, pp. 1-19 (2016).

[11] J. Blauert, "Spatial Hearing with One Sound Source" *Spatial Hearing: The Psychophysics of Human Sound Localization*, Räumliches Hören, Rev. ed., pp. 37-50 (1997).

[12] G. Kearney et al., "Auditory Distance Perception with Static and Dynamic Binaural Rendering" *AES 57th International Conference*, pp. 1—8 (2015).

[13] M. Rychtarikova, T. V. d. Bogaert, G. Vermeir, and J. Wouters, "Binaural Sound Source Localization in Real and Virtual Rooms," *J. Audio Eng. Soc.*, vol. 57, no. 4, pp. 205–220 (2009).

[14] F. Chen, "Localization of 3-D Sound Presented through Headphone—Duration of Sound Presentation and Localization Accuracy," *J. Audio Eng. Soc.*, vol. 51, no. 12, pp. 1163-1171 (2003).

[15] T. Strybel, "Auditory Apparent Motion Under Binaural and Monaural Listening Conditions" *Perception & Psychophysics*, vol. 44, no. 4, pp. 371-377 (2003).

[16] C. Tsakostas and A. Floros, "Real-time Spatial Representation of Moving Sound Sources" *AES 123th Convention*, pp. 1-9 (2007).

[17] M. Matsumoto and T. Mikio, "Algorithms for Moving Sound Images" *AES 114th Convention*, pp. 1-4 (2003).

[18] V. R. Algazi and R. Duda, "Approximating the Head-Related Transfer Function Using Simple Geometric Models of the Head and Torso," *J. Acoust. Soc. Am.*, vol. 105, no. 5. pt. 1, pp. 2053-2064 (2002).

[19] Unity—Manual: Reverb Zones, Retrieved from https://docs.unity3d.com/Manual/class-AudioReverbZone.html, (August 26, 2016).

[20] F. Freeland et al, "Efficient HRTF Interpolation in 3D Moving Sound" *AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, pp. 1 – 8 (2002).

[21] D.S. Tan et al., "Physically Large Displays Improve Path Integration in 3D Virtual Navigation Tasks," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 439-446 (2004).

[22] E. Patrick et al., "Using a Large Projection Screen as an Alternative to Head-Mounted Displays for Virtual Environments," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 478-485 (2000).

[23] M.E. Grabe et al., "T.B. The role of screen size in viewer experiences of media content," *Visual Communication Quarterly*, vol. 6, pp. 4-9 (1999).

[24] "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems", *ITU-R Recommendation BS.1534-3*, International Telecom Union: Geneva, Switzerland, pp. 1-36 (2015).