MITIGATION OF BINAURAL FRONT-BACK CONFUSIONS BY BODY MOTION IN AUDIO AUGMENTED REALITY

Nick Mariette

Audio Nomad Group
School of Computer Science and Engineering
University of New South Wales
Sydney, Australia
nickm@cse.unsw.edu.au

ABSTRACT

Front-back confusions are a well-known phenomenon of spatial hearing whereby the listener incorrectly localizes a source to its mirror image position across the frontal plane. This type of localization error can occur for real and synthetically spatialised sound sources. Experiments have shown the listener can resolve front-back ambiguities by rotating their head; also that sound source movement can resolve confusions if the listener is aware of the intended direction of source movement.

The present outdoors experiment studies the mitigation of front-back confusions for synthetic binaural spatial audio interactive with body movement but not head-turns. This partly disabled mobile augmented reality system renders sound source positions relative to the world reference frame, (so the listener may walk past a stationary spatialised sound), but it renders instantaneous source bearing relative to the listener's reference frame.

Experiment participants walked past synthetic binaural sound sources with initial azimuths of ±(40°, 60°, 80°, 100°, 120° and 140°) and initial distance of 20 metres. Walk distances were chosen to result in azimuth changes of 4°, 8°, 12° and 16° between initial and final source bearings. Each factor combination resulted in a corresponding source distance change over the course of the walk. Front or back judgments of the initial source positions were recorded before and after walking. Results show statistically significant improvement of front-back localization for source azimuth changes of 12° or 16°, and source distance changes of at least 0.21 of the initial distance.

[Keywords: front-back localization, binaural, mobile audio augmented reality]

1. INTRODUCTION

Audio augmented reality (AR) applications require synthetic sound sources rendered to positions relative to the world reference frame, rather than to the listener's reference frame, as is the case for purely virtual worlds. This model simulates real-world sound behaviour, giving the listener the capability to move around a stationary sound source and perceive it from different perspectives. This mobile audio AR functionality enables many auditory display applications that utilize relationships between physical space and spatial sound. Example applications include tourist guides [1], navigation systems for visually impaired people [2], spatialised two-way communication systems [3], entertainment systems including audio AR gaming [4], and sound-art [5].

For any audio AR system, a major technical objective should be to optimize the perceptual performance afforded by the system design and component specifications. However, while necessary technologies for mobile AR are rapidly improving in speed, price, size and weight, and an increasing number of systems are being implemented, little evaluation of their usability or perceptual performance has occurred. Some evaluation examples in the literature are reviewed in an earlier paper from the present body of research that presents a novel perceptual evaluation technique for mobile audio AR [6].

Front-back confusions are a perceptual phenomenon that occurs both with real and synthetic spatial sound sources, whereby the listener is unable to discern whether a source is located in the front or rear hemisphere. The main cause of this type of localization error is ambiguity of the primary inter-aural localization cues: the inter-aural time difference (ITD) and intensity difference (IID), for which the locus of identical values forms the well-known "cone of confusion" around the inter-aural axis, rather than corresponding to one unique source direction. For the horizontal two-dimensional case, listeners incorrectly localize sources as coming from the direction at the mirror image across the inter-aural axis from the true source direction – from front to back or vice-versa.

This experiment aims to characterize the mitigation of front-back localization errors by body movement interaction with synthetic spatial audio. The expectation is that the listener will be able to discern the correct sound source location in front or behind them by combining awareness of their self-motion vector with dynamic cues of azimuth and source distance changes generated by their movement relative to world-stationary sound sources.

2. BACKGROUND

Many experiments have shown that front-back confusions exist for real spatial sound sources, yet occur more often for synthetic spatial sources [7], when degradation of localization cues may occur due to rendering technique – for example, use of non-individualized head related transfer functions. Many other experiments have shown that listeners can resolve front-back confusions by making head-movements, both with real sources or synthetic sources rendered using head tracking.

Wightman and Kistler [8] further showed that listeners can resolve front-back confusions by controlling the azimuth of a virtual sound source. Results show that a listener's awareness of the sound source's intended motion relative to their personal reference frame enables them to resolve front-back ambiguities, and that head-movements are not necessary.

Wightman and Kistler note that the necessary knowledge of intended source movement direction is presumed available in their experiment because the listener initiates and controls the movement. They also state that this condition is only possible for virtual sound sources because "in real auditory space listeners would rarely enjoy an equivalent form of control over the direction of movement of a sound source".

However, if we consider not the absolute movement of a sound source, but the relative movement between source and listener, this control is routinely possible in real auditory space because the listener can control their own position relative to stationary sound sources by walking around. The same is true of the relatively new synthetic situation of mobile audio augmented reality. The present experiment aims to evaluate resolution of front-back confusions via this un-examined mode of interaction by body motion, without the advantage of head-turn interaction.

3. EXPERIMENT

Experiment trials took place in several similar open, flat outdoors spaces, primarily a courtyard with a paved pathway running diagonally for a distance of about 40 metres (Figure 1). Participants were thirteen people (nine male; four female) of unrecorded exact age, although the range is estimated to be twenty to forty-five years, with a median age in the early thirties. Participants reported no known hearing problems.

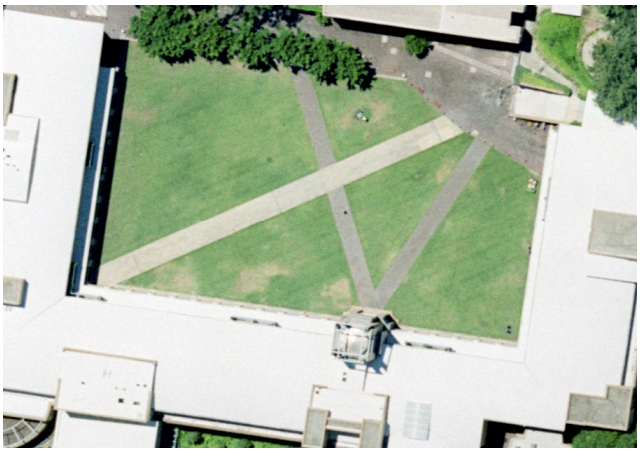

Figure 1. *Aerial view of the main experiment location.*

Participants wore and carried a system comprised of: a set of headphones; a position tracking system mounted at the centre back of the waist; and a portable computer running custom experiment software that displayed a graphical user interface, rendered sound stimuli in real-time, and logged participants' positions and responses.

The positioning system, a Honeywell DRM-III [9], combines an inertial navigation system (INS), a GPS receiver, pedometer, digital compass and barometric altimeter (that can all be individually activated/deactivated), with optional Kalman filtering and a serial RS232 interface. Stated INS position accuracy is 2-5% of distance traveled and the compass is accurate to within one degree. A feasibility study by Miller [10] using the DRM-III suggests that positioning accuracy varies significantly according to usage factors such as stride length variation. A preliminary performance test conducted for an earlier experiment [6] obtained the most accurate positioning for small distances (tens of metres) by using only the INS and digital compass, so this setting was used again in the present experiment.

Other equipment included Sennheiser HD485 headphones (an economical, open backed, circumaural design) and a Sony Vaio VGN-U71 touch-screen handheld computer with a Pentium M processor, running Windows XP Service Pack 2. The DRM-III interfaced to the Vaio with a Keyspan USB-Serial interface.

3.1. Procedure

First, to optimize DRM-III accuracy, each trial was preceded by a calibration procedure that required the participant to walk a ten-metre line at a steady pace to measure and set the stride length. After this, the trial proper began, with the participant asked to localize 48 spatialised stimuli to the front or rear hemisphere.

The experiment was self guided using custom software written in C# .NET 2.0 that interfaced to the DRM, provided a graphical user interface for the participant to supply their responses, which it logged along with position tracks, while it controlled the binaural rendering that occurred in separate software.

For each stimulus, the participant was required first to listen to the synthetic spatial sound and respond to a software prompt asking whether the sound originated in front or behind them. Then they had to walk forward in a straight line for a particular distance until the stimulus stopped, and respond to a second prompt asking whether the sound began in front or behind them. These two participant tasks are shown in Figure 2.

The sound is a synthetic spatial source rendered relative to the world reference frame, so that it seems to remain stationary relative to the real fixed surroundings as the participant walked.

Figure 2. *Experiment participant during a trial. Left: walking while listening to stimulus. Right: recording a response, with DRM-III visible at the waist.*

3.2. Instructions

Several instructions were given to each participant before they began the experiment. First, it was made clear that the sound source was intended to be stationary in the real world, so that

listeners could set up their expectation of sound source behaviour consistent with this knowledge, rather than the more familiar virtual world of computer games and portable music players in which sound is only ever positioned relative to the listener's personal reference frame. Second, participants were asked to walk at a steady pace as per the calibration procedure, and to only walk in a straight line for the whole time each stimulus played. They were also advised to keep their head facing in the direction of travel, and told that the sound was not interactive with head-turns. Participants were told to always make their best guess at a response even if the front-back localization was not easy (as expected for some stimuli), because no response option was provided for "not sure". Finally, participants were informed they should turn around between stimuli when they were nearing one end of the walkway to ensure they always had adequate space to walk forward.

3.3. Stimuli and experimental factors

Stimuli consisted of continuous noise-burst trains, real-time spatialised to a 20 metre distance, at one of twelve initial azimuth angles of ±(40°, 60°, 80°, 100°, 120°, 140°) from front centre (where positive angles are to the right), each repeated twice.

The second independent variable was the azimuth difference between initial and final source positions, which determined the required walk distance for each stimulus. Four delta azimuth values of 4°, 8°, 12° and 16° were chosen after a pilot study using angles of 5°, 10°, 15° and 20° showed scope for improved resolution. The sound source geometry is shown in Figure 3, which graphically represents the definition of initial and delta azimuth and source range values.

Note also that results analyses use the absolute value of initial azimuth, effectively giving four repetitions of each combination of initial and delta azimuth values by assuming that equal angles to the left or right of the median plane are equivalent due to head symmetry.

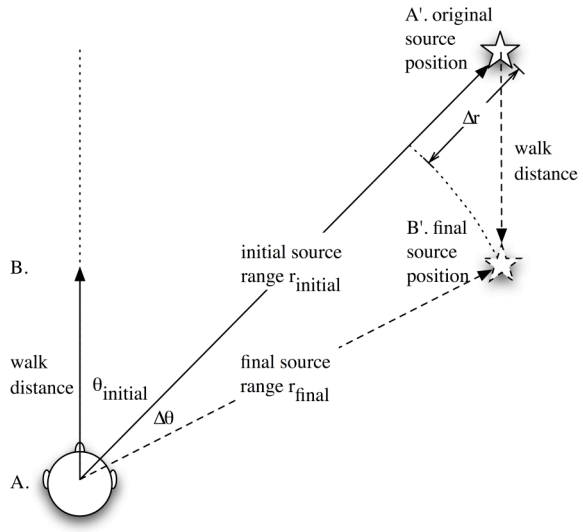

Figure 3. *Synthetic spatial sound source geometry relative to listener's frame of reference.*

The raw noise-burst train itself is a continuously looped ten-second sample of Matlab generated Gaussian white noise enveloped by a rectangular wave with duty cycle of 5ms on, 10ms off. The raw unspatialised stimulus sound pressure level was set to 75 dBA per headphone channel with the Vaio sound output set to full volume, so the level was repeatable for all participants.

3.3.1. Real-time binaural rendering

Raw noise-burst trains were rendered as synthetic spatial sound sources using a custom "patch" (process graph) developed in Pure Data (Pd), a graphical software environment for real-time digital signal processing, shown in Figure 4. The patch was a binaural adaptation of another designed by the author to spatialize multiple simultaneous sounds to a multi-channel speaker array, used originally for a mobile audio augmented reality installation on a passenger ship [11].

The raw sound source was spatialised to a virtual six-speaker array using Pulkki's vector-based amplitude panning technique (VBAP) [12], with each speaker binaurally simulated by convolving its signal with the appropriate pair of head related impulse responses (HRIRs) from subject number three chosen arbitrarily from the CIPIC database [13]. Distance was simulated only by controlling level in proportion to the inverse square of source distance.

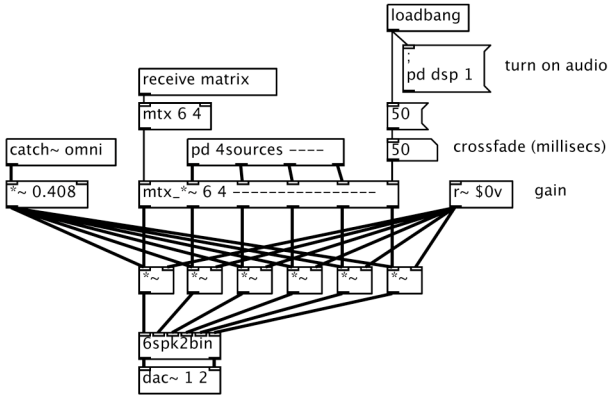

Figure 4. *Detail view of binaural rendering Pd patch.*

Control of the Pd rendering system from the custom experiment software was achieved using commands sent via the Open Sound Control (OSC) communications protocol [14]. Messages include sound file names, playback control, source positions, listener position and orientation.

This virtual speaker binaural rendering is a computationally economic system that can comfortably run on the chosen mobile computing platform. This render technique also suits audio AR because the computation load is almost constant for any number of simultaneous sound sources, because the intensive HRIR convolution load remains constant while the extra load per source consists of sound-file playback, parameter communications and a larger VBAP matrix multiplication.

4. RESULTS ANALYSIS AND DISCUSSION

For the purposes of results analysis, two measures were created to represent participants' localization performance.

First, for each stimulus, a binary *"correctness"* value is assigned according to whether a response was correct by comparing the participant's front or back localization judgment with the initial source azimuth (front for absolute angles < 90°; back for > 90°). Correctness is set to 1 for a correct response or 0 for incorrect, with a distinction made between *initial* and *final* correctness that represent responses made before and after walking.

Second, a ternary *"improvement"* value is assigned to represent the difference between initial and final correctness values, as enumerated in Table 1. Note that some information is discarded for clarity of further analysis, by setting improvement to 0 for both situations of identical initial and final correctness values. The improvement rating is considered neutral regardless of whether both judgments were correct or incorrect.

Initial Correctness	Final Correctness	Improvement
1	0	–1
1	1	0
0	0	0
0	1	1

Table 1. *"Improvement" value assignment according to all possible initial and final correctness values.*

All subsequent analysis compares population *correctness* and *improvement* rates across independent and derived experimental factors.

Independent factors are the *initial azimuth* and *delta azimuth* angles. Note that left/right head symmetry enables equal-magnitude azimuth angles to the left or right to be treated as repetitions and analyzed together as the *absolute initial azimuth* angle.

The two main derived experimental factors are *delta range* (expressed as a ratio of initial source range), and relative source-listener movement *"scenario"* (explained later). Unless otherwise stated, all plots are the result of one-way ANOVA and post-hoc multiple comparison tests using Tukey's Honestly Significant Difference (HSD) at $p<0.05$, with respect to the relevant experimental factor. Error bars in all cases represent the 95% confidence interval (CI) around each data point.

To begin, we ran a basic "reality check" on each participant's results by examining the mean final correctness versus left or right initial azimuth angles, which gave the maximum possible stimulus replication of 24 data points per factor per participant (6 azimuth values per side multiplied by 4 delta azimuth values), in turn giving the smallest possible variance. The test hypothesis is that the correctness rate should be the same for azimuth angles to the left or right. Participants' results were removed if they failed this reality check by giving significantly different correctness rates between left and right initial source azimuths. Of the thirteen original participants, three result sets were thus removed. Note however, that if all participants' results were included, the statistical significance of all presented results was not affected – only the difference between factors was decreased.

For the ten remaining participants' accumulated results, the mean final correctness is 0.72 for left azimuths and 0.74 for right azimuths, shown in Figure 5. Both left and aright azimuths have 95% confidence intervals that encompass the 0.75 value showing that the average final correctness for all stimuli is approximately half way between pure chance (correctness = 0.5) and perfection (correctness = 1.0). For comparison, the overall mean *initial* correctness is 0.60, mean *final* correctness is 0.73 and mean improvement is 0.12.

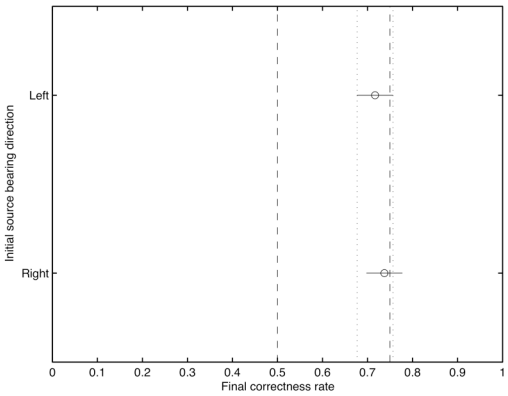

Figure 5. *Mean final correctness for left and right azimuth directions. Error bars represent 95% confidence intervals.*

Another interesting initial view of results is the overall initial correctness rates versus absolute initial azimuth. Figure 6 displays a multiple comparison test of initial correctness, with the worst performance for slightly rear azimuths, and performance significantly greater than chance (0.5 correctness) for all frontal azimuths and the far rear azimuth (±140°). However, mean initial correctness is not significantly greater than 0.75 for any angle. The worst performing azimuth of 100° might be explained by the minimal spectral cues differentiating it from the 80° azimuth that shares very similar ITD and IID cues. However, since the 80° azimuth performed significantly better than 100°, there seems to be a response bias towards the front hemisphere. This is unexpected, since without a frontal visual cue to reinforce frontal localization, ambiguous frontal sources tend to be localized behind the listener, rather than vice-versa.

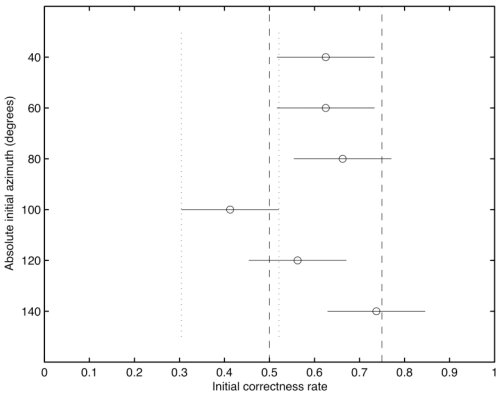

Figure 6. *Initial correctness rates versus absolute initial azimuth. Error bars represent 95% confidence intervals.*

The difference between initial and final correctness rates versus absolute initial azimuth is displayed in the improvement plot in Figure 7. Improvement rates are just significantly better than zero for rear angles of 120° and 140°, and clearly better than zero for the frontal angle of 40°. It seems the greatest mitigation of front-back localization errors occurs for the extreme frontal sound sources. Dynamic localization cues for these sources are mainly attributed to change of source range, rather than azimuth, so we might conclude that the given rendering system provides better resolution for changes of source range than for azimuth.

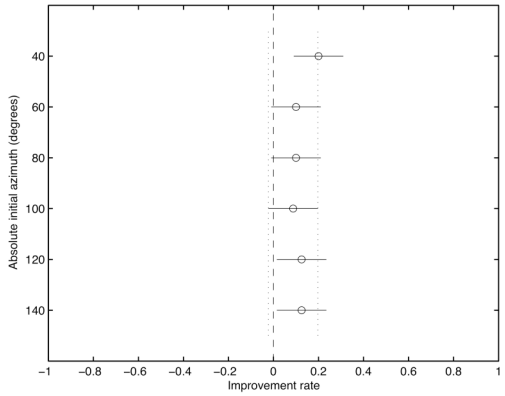

Figure 7. *Improvement versus absolute initial azimuth. Error bars represent 95% confidence intervals.*

In fact, a threshold of advantage is observed in multiple comparison plots of improvement and final correctness versus both delta azimuth, (Figure 8 and Figure 9) and the delta range ratio to initial range (Figure 10 and Figure 11).

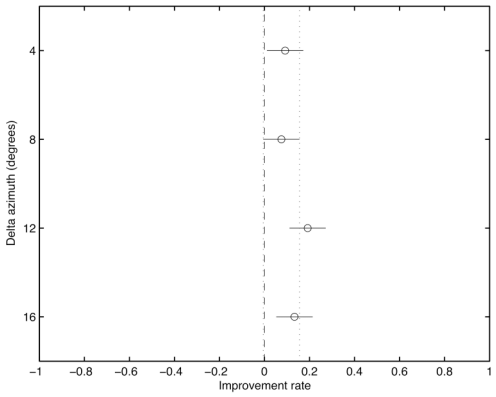

Figure 8. *Improvement rate versus delta azimuth. Error bars represent 95% confidence intervals.*

First, consider the effects of delta azimuth. The improvement plot (Figure 8) shows that an azimuth change of 12° or more gives an improvement significantly greater than zero. The final correctness plot versus delta azimuth (Figure 9) shows that this experiment produces mean final correctness above 0.75 (midway between pure chance and perfection) for delta azimuth angles between 12° and 16°. Final correctness improves monotonically with increasing delta azimuth as expected, since magnitudes of azimuth and range dynamic cues also increase monotonically.

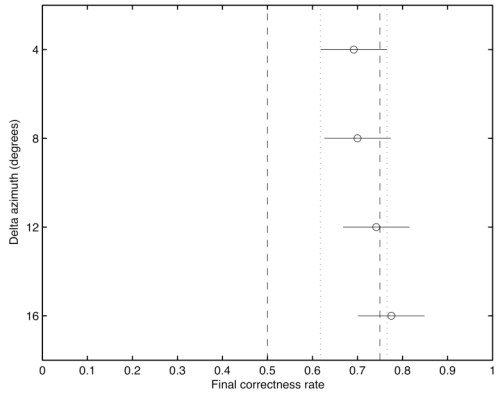

Figure 9. *Final correctness versus delta azimuth. Error bars represent 95% confidence intervals.*

Next consider the effects of delta range. For a given azimuth change, delta range values are smaller for initial azimuths towards the sides (near 90°) than for near-front or near-back angles. Also, delta range is smaller for smaller delta azimuths. Furthermore, large delta range values occur less often than smaller ones because there are fewer combinations of extreme initial azimuths with larger delta azimuth values. Note also that the only dynamic range cues generated by the present render system result are level changes proportional to the inverse square of distance. For these reasons, delta range values are presented as a ratio to the initial range, (always 20 metres), then bundled into groups containing equal proportions of all results, resulting in equal sized confidence interval error bars for each bundle, evident in Figure 10.

The improvement plot versus delta range ratio (Figure 10) shows that a delta range ratio of ≥ 0.21 of initial distance while walking results in improvement significantly greater than zero.

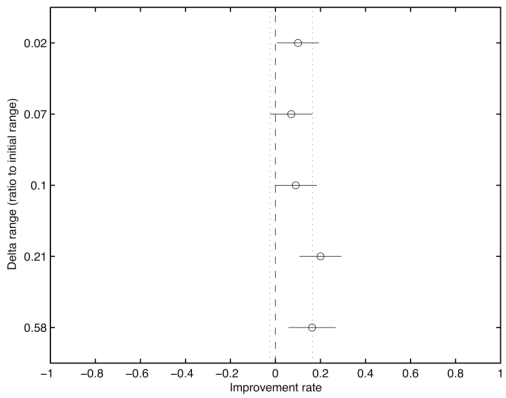

Figure 10. *Improvement rate versus delta range ratio. Error bars represent 95% confidence intervals.*

Figure 11 presents another view of the effect of dynamic range localization cues as a plot of final correctness versus delta range ratio on a linear horizontal axis. Error bars represent 95% confidence intervals and a trend line is also plotted. The trend rises above the 0.75 correctness rate (mid-way between chance and perfection) for delta range ratios above approximately 0.15 of full range.

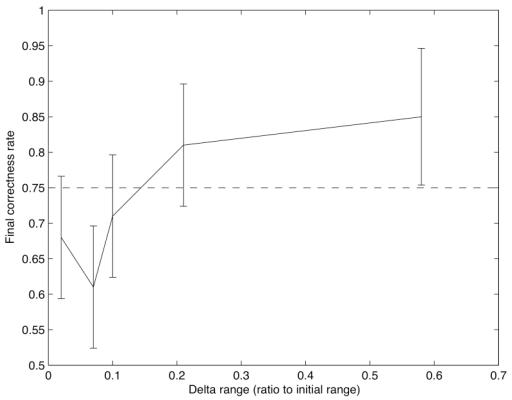

Figure 11. *Final correctness versus delta range ratio. Error bars represent 95% confidence intervals.*

The final analysis examines final correctness and improvement rates versus the "scenario" of source/listener relative movement. Scenario 1 represents the situation where the source is initially in front and the listener walks past the source until it finishes behind them. This scenario is the least common for the present values of initial and delta azimuth. Scenario 2 represents the situation of walking towards a frontal source, but never passing it – in other words, the source is *looming* towards the listener. Scenario 3 represents a source initially behind the listener, *receding* as the listener walks forward.

Figure 12 reveals that looming frontal sources (scenario 2) give significantly better localization correctness than receding rear sources (scenario 3). Apparently listeners are more likely to correctly localize a frontal source as they approach it, than to localize a rear source as it recedes. This bodes well for efficient discovery of new sound sources in an audio AR application. Possibly, if the experiment factors gave more repetitions for scenario 1 (passing sources), this would also show higher rates of correct localization, since they all begin as looming sources.

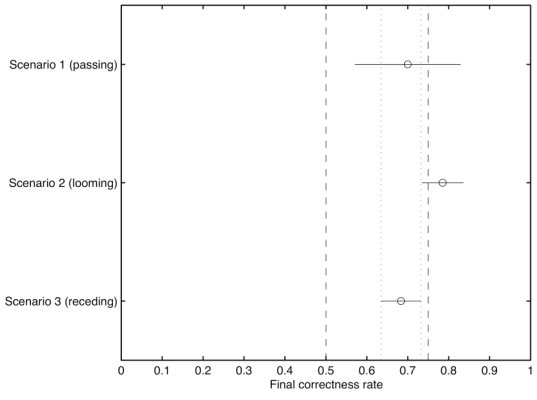

Figure 12. *Final correctness versus source-listener relative movement "scenario". Error bars represent 95% confidence intervals.*

Finally, Figure 13 displays a multiple-comparison plot of improvement versus scenario, which reinforces the out-performance of looming and receding sources over passing sources, since scenarios 2 and 3 both show improvement significantly greater than zero, while scenario 1 doesn't. However, in this representation, no significant difference exists between improvement for scenarios 2 and 3. Comparing this to the final correctness values simply represents that rear sources in this experiment have lower initial correctness rates, resulting in statistically equal improvement rates despite statistically different final correctness rates.

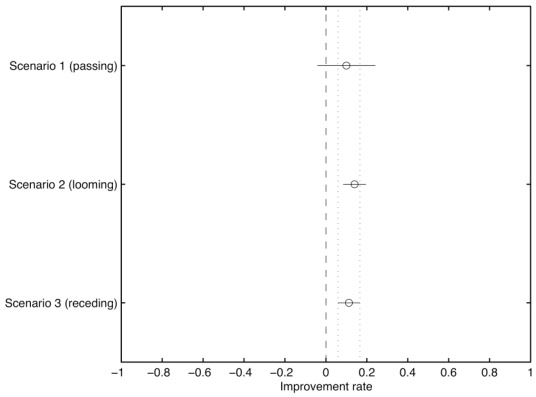

Figure 13. *Improvement versus source-listener relative movement "scenario". Error bars represent 95% confidence intervals.*

5. SUMMARY AND CONCLUSIONS

The presented experiment examines the effect of body motion interaction with synthetic spatial sound sources positioned statically in the world reference frame, as for a usual AR system without head-orientation interaction. As expected, front/back localization improves after the listener interacts with the spatialised sound by walking forward on a straight line beside the source, so it either looms towards, passes or recedes away from the listener. Dynamic localization cues of increasing source azimuth and changing source range enable the listener to constantly revise their judgment of a sound's location in front or behind them.

Delta azimuth values of 12° or 16° while walking give a mean improvement rate significantly greater than zero. Alternatively, a delta azimuth between 12° and 16° increases mean correctness after walking above the 0.75 mid-point between chance and perfect judgment. This agrees with the expected azimuth resolution of about 13° mean azimuth error for the employed render method in an earlier mobile outdoors localization experiment [6].

Delta range values of at least 0.21 of full range show improvement significantly greater than zero. Alternatively, delta range increases the mean correctness after walking above the 0.75 mid-point for ratios above 0.15 of initial range. Expressed as a gain change, the 0.21 delta range ratio equates to 4.1 dB, and the 0.15 ratio equates to 2.8 dB. Given the experiment setting in an outdoors environment with some background noise, these gain levels seem reasonable to provide adequate dynamic range cues to disambiguate front-back confusions.

Lastly, results analysis according to relative source-listener motion scenarios shows that final correctness is significantly better for looming sound sources than for receding sources.

Improvement is not significantly different between any of the scenarios, however it is significantly greater than zero for looming and receding sources, but not passing sources.

In conclusion, dynamic cue efficacy in terms of both azimuth and range changes seems to roughly match the resolution of the employed binaural render method. This suggests that higher resolution spatial sound synthesis will allow listeners to use smaller source azimuth and range changes to disambiguate front-back confusions. This also means that for a given initial source range, higher rendering resolution should enable correct localization through *smaller body movements*.

Since the geometry of dynamic localization cues (delta azimuth and delta range ratio, shown in Figure 3) scales linearly with distance, the minimum source range for an acceptable localization correctness rate will then be dictated by the position tracking resolution that limits the smallest measurable body movements.

Expressed in another way, interactions exist between the main AR system performance bottlenecks of position tracking accuracy and mobile computation power, and the weaker of the two specifications will dictate the minimum source distance that allows acceptable front/back localization performance.

6. ACKNOWLEDGMENTS

This research took place in conjunction with the Audio Nomad project, lead by Dr. Daniel Woo, Dr. Nigel Helyer and Prof. Chris Rizos, supported by an Australian Research Council Linkage Project (LP0348394) in conjunction with the Australia Council for the Arts as part of the Synapse Initiative.

7. REFERENCES

[1] Node. Node | World Leader in Location Based Media. [cited 2007 12 February 2007]; Available from: http://www.nodeexplore.com/.

[2] Loomis, J.M. "Personal Guidance System for the Visually Impaired using GPS, GIS, and VR Technologies." in VR Conference. 1993. California State University, Northridge.

[3] Kan, A., et al., Mobile Spatial Audio Communication System in Tenth Meeting of the International Conference on Auditory Display. 2004: Sydney, Australia.

[4] Röber, N. and M. Masuch, Leaving the Screen. New Perspectives in Audio-Only Gaming, in Eleventh Meeting of the International Conference on Auditory Display. 2005: Limerick, Ireland.

[5] Olivier Warusfel, G.E. "LISTEN - Augmenting everyday environments through interactive soundscapes." in IEEE VR2004. 2004.

[6] Mariette, N. "A Novel Sound Localization Experiment for Mobile Audio Augmented Reality Applications." in 16th International Conference on Artificial Reality and Tele-existence. 2006. Hangzhou, China: Springer, Berlin/Heidelberg.

[7] Wenzel, E.M., et al., "Localization using nonindividualized head-related transfer functions," Journal of the Acoustical Society of America, 1993. 94(1): p. 13.

[8] Wightman, F.L. and D.J. Kistler, "Resolution of front–back ambiguity in spatial hearing by listener and source movement," Journal of the Acoustical Society of America, 1999. 105(5): p. 13.

[9] Point Research, DRM-III OEM Dead Reckoning Module for Personnel Positioning. 2002: Fountain Valley, California.

[10] Miller, L.E., Indoor Navigation for First Responders: A Feasibility Study. 2006, National Institute of Standards and Technology.

[11] Woo, D., et al., "Syren - A Ship Based Location-Aware Audio Experience," Journal of global positioning systems. Int. Assoc. of Chinese Professionals in GPS, 2005. 4(1-2): p. 41-47.

[12] Pulkki, V., "Virtual sound source positioning using vector base amplitude panning," Journal of the Audio Engineering Society, 1997. 45(6): p. 456-466.

[13] Algazi, V.R., et al. "The CIPIC HRTF Database." in Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics. 2001. Mohonk Mountain House, New Paltz, NY.

[14] Wright, M., A. Freed, and A. Momeni. "OpenSound Control: State of the Art 2003." in Conference on New Interfaces for Musical Expression. 2003. Montreal, Canada.