Sensorimotor Coupling in Virtual Audio: Psychoacoustic Studies on the Minimum Audible Angle with 6-DoF Movement

Olli S. Rummukainen, Emanuël A.P. Habets

International Audio Laboratories Erlangen, 91058 Erlangen, Germany, Email: olli.rummukainen@iis.fraunhofer.de*

Introduction

Virtual reality systems with 6-degrees-of-freedom (6-DoF) tracking enable sensory information to be rendered in real time in response to the listener's motor actions. The minimum audible angle (MAA) has been studied with a stationary listener and a stationary or a moving sound source. The studies presented here focus on a scenario where the angle is induced by listener self-translation in relation to a stationary sound source. The self-translation minimum audible angle (ST-MAA) is shown to be 3.3° in the horizontal plane in front of the listener. Furthermore, in contrast to stationary listener MAA, the ST-MAA is shown to be unaffected by an additional visual cue.

The MAA in azimuth for a stationary listener and a sound source has been established in multiple studies to be approximately 1° in the frontal listening area and to degrade gradually moving away from the median plane [1]. In these studies, the test participant is typically seated in an anechoic chamber, and their movement is physically limited or otherwise discouraged. The sound event is fixed to a stationary loudspeaker, or in some studies to a moving boom to study the minimum audible movement angles (MAMA; [2]), where a single sound source is dynamically moved across space at a certain distance. The angles found in these studies depend heavily on velocity, frequency content, and listener training, and are found to be on average 2 to 3 times larger than MAA for stationary stimuli [3].

In contrast to MAA and MAMA studies, a natural way for humans to observe the world is an active process where motor functions support sensory information processing. Dynamic cues resulting from head rotation have been shown to resolve front-back confusions in binaural sound reproduction [4]. Further dynamic cues due to listener translation in a sound field are less studied, but some results are presented stating that the dynamic setting eases the requirements for having individualized head-related transfer functions (HRTFs) in binaural audio reproduction [5] and that the motion parallax and acoustic time to target are informative about the relative motion between observer and source [6]. Recently, active self-translation has been shown to improve auditory depth perception via the acoustic parallax phenomenon [7].

For a stationary listener, dynamic visual capture is a phenomenon where visual motion can elicit subjective motion of a stationary sound source. The motion of the sound source is perceived in the direction of the movement of the visual target [8]. A visual distractor, moving in the opposite direction from a moving sound event, reduces the perception of the direction of auditory movement to chance levels, but the detection of sound source movement is not degraded [9]. Finally, the visual capture of sound has been shown to result in larger MAAs compared to audio only conditions [10].

This study explores the absolute perception of sound event stationarity in a dynamic 6-DoF setting. Binaural reproduction is utilized in the experiment. The goal is to estimate a self-translation minimum audible angle (ST-MAA) and to compare this to a source-translation induced minimum audible movement angle, where the dynamic binaural cues elicited by the two types of translation are identical. Additionally, the ST-MAA is estimated under audio-visual conditions where the auditory and visual cues are either matching or mismatching and the potential discrepancy results from listener self-translation.

Experiment I: Audio-only ST-MAA

Experiment I establishes an estimate for self-translation induced minimum audible angle through two different two-alternative forced choice (2AFC) discrimination tasks where either the listener or the source translates across space.

Participants

In total 24 people participated in Experiment I. They were screened for hearing impairments by a standard pure-tone audiometry and all provided a written informed consent to participate in the study. Out of the 24 participants five were excluded from the final analysis due to missing a control condition, which will be defined in the Procedure section. The remaining participant pool is composed of 4 females and 15 males with average age of 30.1 years (SD = 6.0).

Stimuli

Pink noise was rendered to headphones by a parametric binaural renderer based on a spherical head model. The rendering has been shown to result in spatial resolution comparable with loudspeaker-based MAA experiments [11]. To introduce onset localization cues, the pink noise was pulsed with a pulse duration of 100 ms with an interval of 300 ms. The HTC Vive headset displayed a virtual landscape and provided the real-time 6-DoF position data of the participant.

The spatial resolution was examined by rendering the

*A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

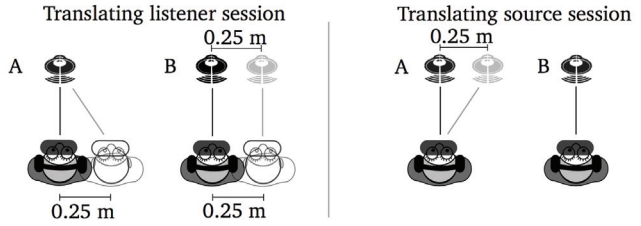

Figure 1: The audio rendering principle in the translating listener and the translating source sessions. Both include two conditions, which result in matching binaural signals between the sessions. Condition A in both sessions results in a perception of a dynamic auditory event that either reacts to self-movement or moves itself and is consistent with the visual cue, whereas Condition B results in a perception of a static auditory event located at the center of the head and audio-visual mismatch.

sound events to distances from 1 m to 10 m from the listener. Relative distance to the source was used as a proxy to reduce the effect of listener translation on the rendered signals' localization cues. The participant's head was tracked in 6 DoF and the rendering adapted in real-time to positional and rotational changes. Signal level was constant regardless of distance to avoid the possible degradation of angular localization cues due to reduced loudness.

The visual scene on the head-mounted display showed a sky-box rendered at infinite distance. There were vertical pillars denoting the end-points of the lateral translation range and the direction to which the participant should face. There was a pillar marking the position of the participant within the range and additionally a virtual carpet denoting the area where the participant was allowed to move. The visual scene was designed to remove any real-world visual cues about the size of the space and to help the participants to imagine distant sound events.

Procedure

Translating listener session presented the participants with a 2AFC task where the goal was to find the sound event that was stationary in the virtual reality instead of following the participant's translations. The task was implemented with a ±0.25 m lateral translation range with the sound event rendered at the center of the range at distances from 1 m to 10 m with a one-meter interval. Consequently, the angle range decreases with increasing distance. The allowed lateral movement range was displayed visually in the HMD and the participant received continuous visual feedback of their location within the range. The participant was in a standing position and either slightly swayed laterally or took small steps sideways. As the participant translated within the range, the sound event was either rendered to be stationary in the virtual world (Condition A) by updating the ILD, ITD, and spectral cues correspondingly, or it was rendered always at the lateral location of the participant's head (Condition B) with ILD = 0 and ITD = 0 irrespective of the listener's absolute lateral position, which resulted in a perception of an internalized or centrally-located auditory event. In both conditions head rotations were rendered naturally and only the self-translation resulted in differences in rendering between the conditions. The conditions are presented schematically in Figure 1 (Left panel).

Translating source session was the opposite case from the translating listener session. Here the participant was seated, and the sound event was either translating or stationary with a ±0.25 m translation range at distances from 1 m to 10 m. The participants were instructed to minimize their head movements, but the head was not fixed. The source translation was a periodic oscillation between the range end-points. The task was a similar 2AFC discrimination task where the participant was required to detect which event was translating. The session is depicted in Figure 1 (Right panel). The two opposed sessions produced similar audio signals to the ear canals, with the only difference being the participant self-translation or the lack thereof.

The participant controlled the playback via a hand-held controller, which they could use to switch between the two conditions as many times as desired. The only way to discriminate the two sound events was to translate laterally within the given range (±0.25 m) and listen to both options. The time to complete each trial was not limited. The system provided visual feedback after each trial to indicate whether the response was correct.

In both conditions, the trial at each distance was repeated four times by each participant resulting in 40 trials in each session. The order of the session was counterbalanced, and the order of trials was pseudo-random to reduce learning effects. There were four practice trials in both session with a visual cue of the sound event location. The visual cue was a green sphere rendered stereoscopically at eye level and at the same distance as the sound event. The practice trials spanned the distance range. During practice, it was made sure that every participant could perceive the difference at 1 m distance. Later in the analysis the 1 m condition was used as a control condition and missing it in either session was a reason for excluding the participant.

Experiment II: Audio-visual ST-MAA

This experiment replicates the previous experiment with an added visual cue.

Participants

We conducted a main experiment and a follow-up study. In total 26 people (5 female, 21 male) participated in the main experiment. Their average age is 26.1 years (SD = 2.2 years). One participant was excluded from the final analysis based on a control criterion, resulting in 25 participants. The follow-up study had 24 participants (14 female, 10 male) with the average age of 28.7 years (SD = 11.1 years). In this study four participants were excluded due to missing the control criterion, resulting in total 20 participants for the data analysis. None of the participants in the follow-up study took part in the main experiment.

Stimuli

Auditory stimulus was identical to Experiment I. The visual scene and interface elements matched the ones in Experiment I otherwise, but here the sound object was visually depicted as a sphere with a 10 cm radius. The color of the sphere changed based on the selected condition (orange or blue) and its size was rendered realistically according to distance. The sphere's position was easily detectable even at the furthest distance. Depending on the condition, the visual cue either matched the sound event or there was a mismatch between the visual and auditory cues.

Procedure

Translating listener session presented the participants with a 2AFC task where the goal was to find the sound event that was stationary in the virtual reality matching the visual cue instead of following the participant's translations. The visual cue matched the Condition A in Figure 1 (Left panel). Switching of the condition was allowed only within ±5 cm from the center line to deny the possibility to investigate the conditions at either maximum of the range.

Translating source session had the visual cue translating between the range maximums, corresponding to Condition A in Figure 1 (Right panel). The participant was required to detect which sound event was translating and thus matching the visual cue. The condition switch could be requested at any point in time, but the actual switch only happened when the translating visual sphere crossed the center line, where the sound source would either become stationary or start translating together with the visual sphere. This delay period was communicated to the participant by a visual indicator requesting them to hold on for the next condition.

Translating source with further distances follow-up study was added after the data from the previous two sessions appeared to not reach chance levels for the translating source case. Here, a new set of participants conducted two sessions with a source translating either with or without a visual cue. In these sessions the distance set was {1, 3, 5, 7, 9, 11, 12, 13, 14, 15} m. The data from these sessions was added to the corresponding previous datasets. The instructions, setup, and procedures were identical to the audio-visual translating source session described above and the audio-only translating source session described in Experiment I.

Results

Each participant's correct answers are counted for each distance in the two experiments and the probability to find the target sound event is modeled by a Weibull-distribution. The results of fitting the distributions to the data are displayed in Figure 2 together with average probability to find the target by distance. The threshold estimate and the confidence intervals (CI) are obtained by randomly sampling the dataset 10000 times with replacement and fitting the Weibull-distribution to each of the new datasets. The 95 % CIs are taken to be the 95th percentile of the resulting set of threshold estimates.

Table 1: MAAs in audio only and audio-visual dynamic scenarios.

	Audio only	**Audio-visual**
Source translation	1.1°	1.7°
Listener translation	3.3°	3.4°

Audio-only ST-MAA: Figure 2 shows a significant discrepancy in probabilities to differentiate the target sound event between the translating listener and translating source sessions. A threshold for 79.4 % correct [12] response level in the translating listener session was found to be 4.33 m (95 % CI 3.99 m to 5.19 m). This value with a 0.25 m lateral translation range corresponds to the minimum audible angle of 3.3° (95 % CI 2.8° to 3.6°). The audio only translating source session result is 13.12 m (95 % CI 11.85 m to 14.86 m), yielding a minimum audible angle of 1.1° (95 % CI 1.0° to 1.2°).

Audio-visual ST-MAA: The threshold distance in the audio-visual translating listener session was found to be 4.25 m (95 % CI 3.85 m to 4.84 m), which corresponds to the minimum audible angle of 3.4° (95 % CI 3.0° to 3.7°). The audio-visual translating source session result is 8.47 m (95 % CI 7.78 m to 9.31 m), giving a minimum audible angle of 1.7° (95 % CI 1.5° to 1.8°). The thresholds are collected in Table 1.

Discussion

The self-translation induced minimum audible angle, ST-MAA, was found to be substantially larger than the stationary MAA values in literature. Furthermore, comparing the translating listener and translating source audio only sessions, a significant difference in the thresholds is observed (3.3° versus 1.1°). The result is striking keeping in mind that there was no difference in the audio signals presented at the ear canals between these sessions. The difference results from listener self-translation or the lack thereof.

The audio-visual ST-MAA was found to be significantly larger than the audio-visual translating source MAA in Experiment II. This finding is in line with the audio-only ST-MAA thresholds reported in Experiment I. The audio-only ST-MAA of 3.3° found there does not differ from the audio-visual ST-MAA of 3.4°. However, in previous studies, apparent visual motion is shown to affect the perceived direction of auditory motion [8] and to increase the MAA [10]. Similarly, in Experiment II, the stationary listener MAA resulting from audio-visual source translation was found to be larger compared to the audio-only condition reported in Experiment I (1.7° versus 1.1°). Therefore, we conclude that ST-MAA is not affected by apparent visual motion in contrast to the stationary listener case, where the visual influence is significant.

Based on the results presented here, self-translation appears to impair absolute judgment of stationarity of sound events. Humans are shown to accept highly unnatural vi-

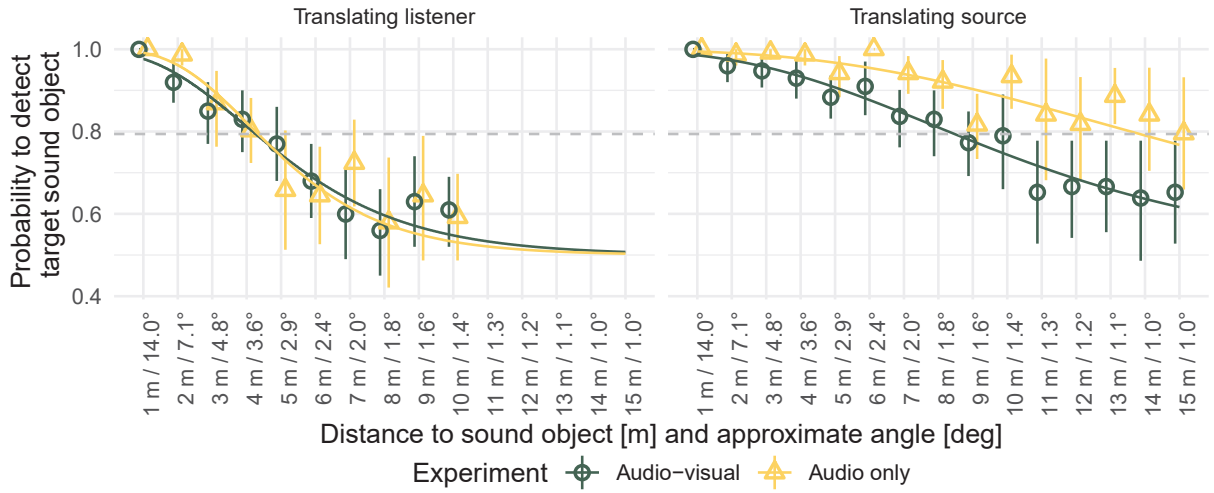

Figure 2: Psychometric functions for translating source and translating listener sessions modeled according to the Weibull-distribution. The data points are the average of each participant's average of four trials at each distance for the ±0.25 m lateral translation range. The whiskers denote the 95 % confidence interval of the mean. The grey horizontal line marks the 79.4 % correct threshold.

sual cues about spatial dimensions as long as they are consistent with self-translation [13]. A similar mechanism may be at play in the auditory system, ignoring noisy sensory data when self-translation cues are strongly in favor of a specific interpretation.

Conclusions

The classic minimum audible angle value is approximately 1° in front of the listener, which is less than a third of the angle found here for the self-translation induced minimum audible angle of 3.3°. No effect of apparent visual motion on the ST-MAA was found, which is in contrast to previous studies where the MAA has been shown to increase under corresponding conditions.

References

[1] D. R. Perrott and K. Saberi, "Minimum audible angle thresholds for sources varying in both elevation and azimuth," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1728–1731, 1990.

[2] D. R. Perrott and A. D. Musicant, "Minimum auditory movement angle: binaural localization of moving sound sources," *The Journal of the Acoustical Society of America*, vol. 62, no. 6, pp. 1463–1466, 1977.

[3] S. Carlile and J. Leung, "The Perception of Auditory Motion," *Trends in Hearing*, vol. 20, pp. 1–19, 2016.

[4] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *The Journal of the Audio Engineering Society*, vol. 49, pp. 904–916, 2001.

[5] J. M. Loomis, C. Hebert, and J. G. Cicinelli, "Active localization of virtual sounds," *The Journal of the Acoustical Society of America*, vol. 88, no. 4, pp. 1757–64, 1990.

[6] J. Speigle and J. Loomis, "Auditory distance perception by translating observers," in *Proceedings IEEE Symposium on Research Frontiers in Virtual Reality*, (San Jose (CA)), pp. 92–99, 1993.

[7] D. Genzel, M. Schutte, W. O. Brimijoin, P. R. MacNeilage, and L. Wiegrebe, "Psychophysical evidence for auditory motion parallax," *Proceedings of the National Academy of Sciences*, pp. 1–6, 2018.

[8] S. Mateeff, J. Hohnsbein, and T. Noack, "Dynamic visual capture: apparent auditory motion induced by a moving visual target," *Perception*, vol. 14, no. 6, pp. 721–727, 1985.

[9] T. Z. Strybel and A. Vatakis, "A comparison of auditory and visual apparent motion presented individually and with crossmodal moving distractors," *Perception*, vol. 33, no. 9, pp. 1033–1048, 2004.

[10] M. Stawicki, P. Majdak, and D. Baskent, "Ventriloquist illusion produced with virtual acoustic spatial cues and asynchronous audiovisual stimuli in both young and older individuals," *Multisensory Research*, vol. 32, no. 8, pp. 745–770, 2019.

[11] O. S. Rummukainen, S. J. Schlecht, and E. A. P. Habets, "Self-translation induced minimum audible angle," *The Journal of the Acoustical Society of America*, vol. 144, no. 4, pp. EL340–EL345, 2018.

[12] H. Levitt, "Transformed up-down methods in psychoacoustics," *The Journal of the Acoustical Society of America*, vol. 49, no. 2, pp. 467–477, 1971.

[13] A. Glennerster, L. Tcheang, S. J. Gilson, A. W. Fitzgibbon, and A. J. Parker, "Humans ignore motion and stereo cues in favor of a fictional stable world," *Current Biology*, vol. 16, no. 4, pp. 428–432, 2006.