Rummukainen, Olli S.; Schlecht, Sebastian J.; Habets, Emanuël A.P.

No dynamic visual capture for self-translation minimum audible angle

Published in:
The Journal of the Acoustical Society of America

DOI:
10.1121/10.0001588

Published: 01/07/2020

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Rummukainen, O. S., Schlecht, S. J., & Habets, E. A. P. (2020). No dynamic visual capture for self-translation minimum audible angle. *The Journal of the Acoustical Society of America*, *148*(1), EL77-EL81. https://doi.org/10.1121/10.0001588

No dynamic visual capture for self-translation minimum audible angle

Olli S. Rummukainen, Sebastian J. Schlecht, and Emanuël A. P. Habets

Citation: The Journal of the Acoustical Society of America **148**, EL77 (2020); doi: 10.1121/10.0001588

View online: https://doi.org/10.1121/10.0001588

View Table of Contents: https://asa.scitation.org/toc/jas/148/1

Published by the Acoustical Society of America

JASA
EXPRESS LETTERS

https://doi.org/10.1121/10.0001588

No dynamic visual capture for self-translation minimum audible angle

Olli S. Rummukainen,[1,a)] Sebastian J. Schlecht,[2] and Emanuël A. P. Habets[1]

[1]*International Audio Laboratories Erlangen, A Joint Institution of the Friedrich-Alexander-University Erlangen-Nürnberg and Fraunhofer Institute for Integrated Circuits, Erlangen, Germany*

[2]*Department of Signal Processing and Acoustics and Department of Media, Aalto University, Espoo, Finland*

olli.rummukainen@iis.fraunhofer.de, sebastian.schlecht@aalto.fi, emanuel.habets@iis.fraunhofer.de

Abstract: Auditory localization is affected by visual cues. The study at hand focuses on a scenario where dynamic sound localization cues are induced by lateral listener self-translation in relation to a stationary sound source with matching or mismatching dynamic visual cues. The audio-only self-translation minimum audible angle (ST-MAA) is previously shown to be 3.3° in the horizontal plane in front of the listener. The present study found that the addition of visual cues has no significant effect on the ST-MAA.

[Editor: Qian-Jie Fu] Pages: EL77–EL81

Received: 7 April 2020 **Accepted:** 28 June 2020 **Published Online:** 21 July 2020

1. Introduction

The minimum audible angle (MAA) is the smallest difference in the direction of two sound events that can be reliably detected. The baseline horizontal MAA for broadband sounds has been established in loudspeaker-based studies to be approximately 1° in the frontal listening area (Perrott and Saberi, 1990). Our previous study (Rummukainen *et al.*, 2018) focused on gathering empirical evidence on the self-translation minimum audible angle (ST-MAA), which was found to be 3.3° in front of the listener. In this case, the sensation of sound movement is the result of listener lateral full-body self-translation instead of source movement or head rotation. In other words, we studied the MAA when a listener makes a step to the side relative to a fixed sound source. The study at hand investigates the ST-MAA under audio-visual conditions, giving rise to the ventriloquist illusion where the perceived location of the sound event is drawn towards a visual cue (Alais and Burr, 2004).

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 (Mateeff *et al.*, 1985). Dynamic visual capture is stronger in the vertical and depth orientation than in the horizontal orientation (Kitajima and Yamashita, 1999). The judgment of sound movement in these directions is more difficult and inaccurate, leading to a stronger visual capture. Coincidentally, the perceived direction of auditory apparent motion is strongly modulated by motion in vision only when the sensory events are spatially and temporally sufficiently aligned (Soto-Faraco *et al.*, 2002). A visual distractor, moving in the opposite direction from a moving sound event, can reduce the detection of auditory apparent motion direction to chance levels, but the overall detection of sound source movement is not degraded (Strybel and Vatakis, 2004). Finally, the visual capture of sound has been shown to result in larger MAAs compared to audio only conditions (Stawicki *et al.*, 2019).

This study explores the audio-visual cue integration in a dynamic six-degrees-of-freedom (6-DoF) setting that tracks the head position in three dimensions of translation and three dimensions of rotation. In the present study, the main movement direction is a side-to-side one-dimensional movement. Binaural reproduction is utilized in the experiment. The goal is to estimate a ST-MAA (Rummukainen *et al.*, 2018) under audio-visual conditions where the auditory and visual cues are either matching or mismatching and the potential discrepancy results from listener self-translation. A comparison is made to a source-translation minimum audible movement angle, where the possible audio-visual mismatch results from external object translation instead of listener translation, who remains stationary. With both origins of translation, listener, or source, the resulting dynamic binaural cues are identical. Based on previous research, we

[a)]Author to whom correspondence should be addressed.

hypothesize H_1, the ST-MAA is larger than the source-translation MAA also under audio-visual conditions and H_2, the audio-visual ST-MAA is larger than the audio-only ST-MAA.

2. Method

2.1 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, which will be defined in Sec. 2.3, 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 a total of 20 participants for the data analysis. None of the participants in the follow-up study took part in the main experiment.

2.2 Stimuli

Auditory stimulus: Pink noise, where each octave band has an equal amount of energy, was rendered binaurally to headphones. The interaural time difference (ITD) and interaural level difference (ILD) were computed dynamically from a spherical head model (Algazi *et al.*, 2001; Duda and Martens, 1998). No individualization was performed, and the radius of the head model was set to 87 mm. All sound events were located on the virtual horizontal plane. No pinna model and therefore no elevation adjustment were included. The signal was played back with an RME Fireface audio interface at 48 kHz and buffer size of 32 samples via a Beyerdynamic DT770 PRO headphone; the buffer added 0.67 ms of latency to binaural cue update. To introduce onset localization cues the pink noise was pulsed with a pulse duration of 100 ms with an interval of 300 ms. This rendering setup was shown to be able to produce MAA resolution comparable to loudspeaker-based experiments in our previous study (Rummukainen *et al.*, 2018).

To introduce experimental conditions, the spatial resolution was examined by rendering the sound events to distances from 1 to 10 m from the listener. Distance was used as a proxy for reducing the effect of listener translation on the rendered signals' localization cues. In the experiment, the listener or the source made side-to-side translations of ±0.25 m, and the target distance was varied to vary the resulting effective angular change. The signal level was kept equal at all distances to avoid the possible degradation of localization cues due to reduced loudness.

Visual stimulus: The visual scenery is shown in Fig. 1. The HTC Vive Pro head-mounted display (HMD) showed a virtual landscape and provided the real-time position data of the participant. The average total latency from movement to stimulus for the HTC Vive tracking system is 22 ms (Niehorster *et al.*, 2017). Experiment logic, interface, and stimuli were programmed, and the experiment controlled, in Max/MSP 8. Setting this study apart from Rummukainen *et al.* (2018), 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. Depending on the condition, the visual cue either matched the sound event or there was a mismatch between the visual and auditory cues.

The visual scene showed a sky-box rendered at infinite distance which did not react to positional changes. There were vertical bars denoting the end-points of the lateral movement range (black bars in Fig. 1). In the self-translation session, there was a bar marking the position of the participant within the range. Switching of the condition using a hand-held controller was

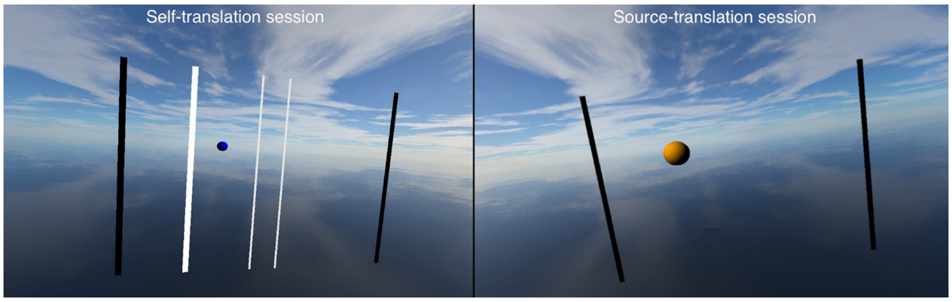

Fig. 1. (Color online) Self-translation and source-translation sessions. In the self-translation session, the visual cue was stationary at the center of the movement range and the sound event was either stationary (audio-visual match) or translating following the head of the listener (audio-visual mismatch). In the translating-source session, the visual cue was always translating between the range end-points and the sound event was either stationary (audio-visual mismatch) or translating together with the visual cue (audio-visual match). The black bars denote the maximum translation range and the thick white bar the lateral position of the participant. The thin white bars mark the area where condition switching is allowed in the self-translation session.

allowed only within ±5 cm from the center line (narrow white bars in Fig. 1, left panel) to ensure that listener judgements were made only in response to dynamic cues rather than static cues. In the source-translation session 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.

2.3 Procedure

The self-translation session presented the participants with a two-alternative forced choice (2AFC) task where the goal was to find the condition with a sound event that was stationary instead of following the participant's translations, thus matching the visual cue. The task was implemented with a ±0.25 m lateral translation range in which the participant was encouraged to make the translation. The sound event was rendered at the center of the range at distances from 1 to 10 m with a 1-m interval. 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 Fig. 2, where the visual cue always matches the condition A. 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 source-translation session was the opposite case of the self-translation session. Here the participant was seated and the sound event was either translating or stationary with a ±0.25 m translation range. The visual cue was always translating between the range end-points, corresponding to condition A in Fig. 2. The participants were instructed to minimize their head movements, but their head was not fixed. The source translation was a sinusoidal oscillation with a period of 5 s between the range end-points. This frequency with a 0.25 m radius corresponds to linear velocity of 0.31 m/s observed at the center of the range and diminishing to 0 m/s towards the range end-points. The task was a similar 2AFC discrimination task where the participant was required to detect which sound event was translating and thus matching the visual cue.

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. In all conditions, the system provided visual feedback after each trial whether the response was correct. The trial at each distance was repeated four times by each participant. 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 sessions with a text label revealing the correct condition. The practice trials spanned the distance range. In the practice session, 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.

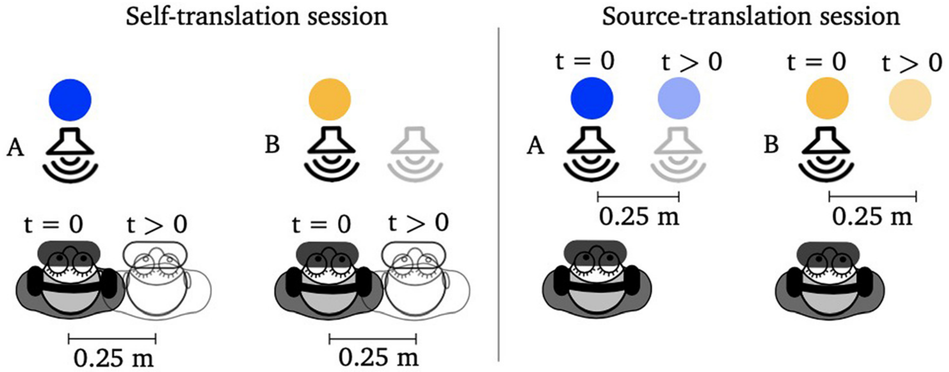

Fig. 2. (Color online) The audio rendering principle in the self-translation and the source-translation sessions. Both include two conditions, which result in matching binaural signals between the sessions. Condition A in both sessions results in a perception of a dynamic auditory event that is consistent 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.

The source-translation with further distances follow-up study was added after the data from the previous two sessions appeared not to reach chance levels for the source-translation sessions. 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 were added to the corresponding previous datasets. The instructions, setup, and procedures were identical to the audio-visual source-translation session described above and the audio-only source-translation session described in Rummukainen *et al.* (2018).

2.4 Results

Audio-visual ST-MAA: Each participant's correct answers are counted for each distance in the two sessions and the probability to find the target sound event is modeled by a Weibull-distribution. The results of fitting the distributions to the data from the two sessions are displayed in Fig. 3 together with the average probability to find the target by distance. The confidence intervals (CIs) are obtained by randomly sampling the dataset 10 000 times with replacements 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. Additionally, the p-values of the difference in mean thresholds between the conditions are obtained by comparing the bootstrapped pseudosamples' mean difference to a null sample (no difference in means) obtained by subtracting the mean difference in each condition. In Fig. 3, the red circle data points stem from the study at hand and the blue triangle data points are from Rummukainen *et al.* (2018).

Figure 3 shows a significant discrepancy in probabilities to differentiate the target sound event between the self-translation and source-translation sessions (bootstrap $p = 0$). A threshold for 79.4% correct response level (Levitt, 1971) in the self-translation session was found to be 4.25 m (95% CI 3.85–4.84 m). This value with a 0.25 m lateral movement range corresponds to the minimum audible angle of 3.4° (95% CI 3.0°–3.7°).

The audio-visual source-translation session results are 8.47 m (95% CI 7.78–9.31 m). This value with a 0.25 m lateral movement range corresponds to the minimum audible angle of 1.7° (95% CI 1.5°–1.8°).

Audio only ST-MAA: For readability, the results from Rummukainen *et al.* (2018) are summarized here. A threshold for 79.4% correct response level in the audio only self-translation session was found to be 4.33 m (95% CI 3.99–5.19 m). This value with a 0.25 m lateral movement range corresponds to the minimum audible angle of 3.3° (95% CI 2.8°–3.6°). The audio only source-translation session results are 13.12 m (95% CI 11.85–14.86 m). This value with a 0.25 m lateral movement range corresponds to the minimum audible angle of 1.1° (95% CI 1.0°–1.2°). The difference between these thresholds is significant (bootstrap $p = 0$). The thresholds are collected in Fig. 4.

3. Discussion

The audio-visual ST-MAA was found to be significantly larger than the audio-visual source-translation MAA. This finding confirms our first hypothesis and is in line with the audio-only ST-MAA thresholds reported in Rummukainen *et al.* (2018). The audio-only ST-MAA of 3.3° found there does not differ from the audio-visual ST-MAA of 3.4° found in this study (bootstrap $p = 0.485$). However, in previous studies, visual motion is shown to affect the direction of

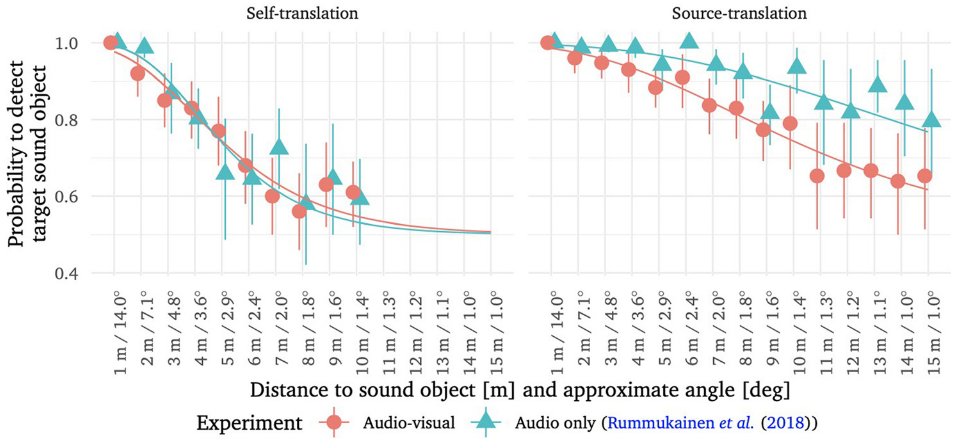

Fig. 3. (Color online) Psychometric functions for source-translation and self-translation sessions modeled according to the Weibull-distribution. The data points are the average of each participants' average of four trials at each distance for the ±0.25 m lateral movement range. The whiskers denote the bootstrapped 95% confidence interval of the mean.

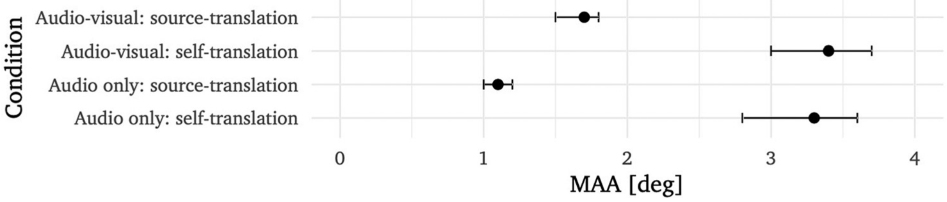

Fig. 4. MAAs in audio only (Rummukainen *et al.*, 2018) and audio-visual dynamic scenarios. The whiskers denote the bootstrapped 95% confidence interval of the mean.

auditory motion (Mateeff *et al.*, 1985) and to increase the MAA (Stawicki *et al.*, 2019). Similarly, in this study, the source-translation MAA was found to be larger compared to the audio-only condition reported in Rummukainen *et al.* (2018) (1.7° versus 1.1°, bootstrap $p = 0.001$). Therefore, we reject our second hypothesis and conclude that ST-MAA is not affected by visual capture in contrast to the source-translation case, where the visual influence is significant.

Previous literature on dynamic visual capture also suggests that the ST-MAA should be increased due to the visual cue. In the self-translation case, the listener has vestibular and proprioception cues in addition to visual cues available to determine the position and orientation of the head in space and its relative location with regards to the dynamic auditory cues. The findings of this study suggest that the self-motion cues are potentially able to counteract the visual capture of auditory perception. Either vestibular stimulation or visual stimulation alone has been shown to suffice to resolve the direction of sound when head rotation is simulated by rotating either the listener or their visual field (Wallach, 1940). Motion speed was not controlled in this study for the self-translation session. In previous work the angular velocity has been found to have an effect on the minimum audible movement angle (Carlile and Leung, 2016). Future work could examine the effect of motion speed and movement strategies on the ST-MAA.

4. Conclusions

This study investigated the ST-MAA under the influence of a visual cue. No effect of visual capture on the ST-MAA was found, which is in contrast to previous studies where the MAA has been shown to increase under corresponding conditions. In line with previous studies, the effect of dynamic visual capture was evident on the source-translation condition, where the MAA increased. In conclusion, based on empirical evidence presented here, the ST-MAA appears to be robust against visual capture.

References and links

Alais, D., and Burr, D. (**2004**). "The ventriloquist effect results from near-optimal bimodal integration," Curr. Biol. **14**(3), 257–262.

Algazi, V. R., Avendano, C., and Duda, R. O. (**2001**). "Estimation of a spherical-head model from anthropometry," J. Audio Eng. Soc. **49**(6), 472–479.

Carlile, S., and Leung, J. (**2016**). "The Perception of auditory motion," Trends Hear. **20**, 1–19.

Duda, R. O., and Martens, W. L. (**1998**). "Range dependence of the response of a spherical head model," J. Acoust. Soc. Am. **104**(5), 3048–3058.

Kitajima, N., and Yamashita, Y. (**1999**). "Dynamic capture of sound motion by light stimuli moving in three-dimensional space," Percept. Motor Skills **89**(3), 1139–1158.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**(2), 467–477.

Mateeff, S., Hohnsbein, J., and Noack, T. (**1985**). "Dynamic visual capture: Apparent auditory motion induced by a moving visual target," Perception **14**(6), 721–727.

Niehorster, D. C., Li, L., and Lappe, M. (**2017**). "The accuracy and precision of position and orientation tracking in the HTC Vive virtual reality system for scientific research," i-Perception **8**(3), 1–23.

Perrott, D. R., and Saberi, K. (**1990**). "Minimum audible angle thresholds for sources varying in both elevation and azimuth," J. Acoust. Soc. Am. **87**(4), 1728–1731.

Rummukainen, O. S., Schlecht, S. J., and Habets, E. A. P. (**2018**). "Self-translation induced minimum audible angle," J. Acoust. Soc. Am. **144**(4), EL340–EL345.

Soto-Faraco, S., Lyons, J., Gazzaniga, M., Spence, C., and Kingstone, A. (**2002**). "The ventriloquist in motion: Illusory capture of dynamic information across sensory modalities," Cogn. Brain Res. **14**, 139–146.

Stawicki, M., Majdak, P., and Baskent, D. (**2019**). "Ventriloquist illusion produced with virtual acoustic spatial cues and asynchronous audiovisual stimuli in both young and older individuals," Multisens. Res. **32**(8), 745–770.

Strybel, T. Z., and Vatakis, A. (**2004**). "A comparison of auditory and visual apparent motion presented individually and with crossmodal moving distractors," Perception **33**(9), 1033–1048.

Wallach, H. (**1940**). "The role of head movements and vestibular and visual cues in sound localization," J. Exp. Psychol. **27**(4), 339–368.