Self-rotation behavior during a spatialized speech test in reverberation

Ľuboš Hládek, Bernhard U. Seeber

Audio Information Processing, Technical University of Munich, 80333 Munich, E-Mail:lubos.hladek@tum.de

Introduction

Speech perception in noise has been traditionally tested in static environments [1]. A typical scenario has a static target speech sound source presented via a loudspeaker in the front and a static interfering sound source presented via a loudspeaker at the same or a different direction, e.g. at 90°. Such testing can capture and simulate many important aspects of the cocktail party situations, for instance the effect of spatial unmasking [2]. Spatial unmasking describes the ability of the auditory system to improve speech perception when the speech and the interfering noise emanate from different places. However, in real cocktail parties people move, for instance they make head rotations, or whole body rotations, which drastically changes in-ear signals. This leads to a change of configuration between the listener and the sound sources, thus it affects the spatial unmasking [3]. If we think about a person who is standing at one place and self-rotating, certain self-rotations will lead to acoustically more favorable situations in terms of spatial unmasking. This opens the question whether people develop a movement strategy, whether they adapt their position and behave in an 'acoustically optimal' way to improve speech perception.

In a study by Brimijoin et al. [4], participants with unilateral hearing loss were listening to speech stimuli in noise, such that the speech came from a fixed position and the interferer had a varying position. Participants usually maintained a constant off-target self-orientation at about 50° during listening and they, overall, ignored the position of the interferer. A study by Grange and Culling [3] indicated high across-subject variance in terms of propensity to head movements, suggesting that people do not employ a unified and consistent strategy. However, in the previous studies, participants were sitting on chairs and that may have biased their behavior. The design of these experiments involved only static targets at a priori known positions, while in many real situations the position of the speech source is not immediately known.

In this study, we conduct a preliminary analysis of behavioral data of a spatialized speech test experiment[5]. In this study, the participants were standing and performing whole body self-rotations, while the target position was varied randomly on a trial-by-trial basis from all around the listener.

In the current work, we address the question whether the presence of the target location cue would affect the complexity of the movements by analyzing the variability of self-rotation trajectories during the experiment.

Methods

Participants

A preliminary data set of two young people, native German speakers, with pure-tone thresholds at standard audiometric frequencies (from 250 Hz – 8 kHz) below 20 dB HL were analyzed. The participants provided written informed consent and the study was approved by the ethics committee of the Technical University of Munich, 65/18S.

Environment

The experiment was conducted in the Simulated Open Field Environment (SOFE v4) [6], [7]. The environment involved a system for delivery of auditory-visual stimuli used in the experiment. The audio reproduction system consisted of an array of 36 horizontally and uniformly spaced loudspeakers (BM6A MKII, Dynaudio) around the listener at height of 1.6 m from the floor inside of an anechoic chamber. The loudspeaker array was arranged in a square such that the closest loudspeaker was at 2.1 m from the listener. The active loudspeakers were controlled over DA converters (M-32 DA, RME) coupled with a multi-channel sound card (HDSPe MADI FX, RME). The visual stimuli were projected from four projectors (32 dB SPL(A) background noise), controlled over the Blender game engine, onto four acoustically-transparent screens in front of the loudspeakers, which surround the listener from all horizontal directions. The synchronization of the audio-visual stimuli was independently assessed using a storage oscilloscope. The room was further equipped with a set of 12 motion-tracking cameras, which were synchronized with the audio presentation system. The participants wore a crown with small reflective spheres for the motion tracking. The participants were standing in the middle of the loudspeaker array on the mesh floor of the anechoic chamber. Participants held a wireless tablet that had a graphical interface to provide responses for the speech test.

Stimuli and procedures

Participants performed a spatialized version of the OLSA test, with a close set of responses [8]. The stimuli were spatialized in a virtual shoebox room (11 m x 13 m x 3 m, l x w x h) and the material was presented over the loudspeakers. The auditory stimuli consisted of the OLSA sentence lists. The interferer was a stationary speech-shaped noise. The noise stimuli were created such that each token had a frequency spectrum identical to the spectrum of the target sentence but started 1 second before the onset of the target sentence and it ended after 4.5 seconds, always after the sentence. A single interferer sound source was positioned always at the front with respect to the listener (0°), the target sound could be spatialized at one of four azimuthal positions with respect to the listener: 0° – front, ±90° – right/ left, 180° – rear. All sounds were presented at a fixed virtual egocentric distance 2.1 m with respect to the listener position. The loudspeaker stimuli were created by convolving the target and the interferer sounds with respective multi-channel impulse responses, which were created by the image source method and mapped to the loudspeakers using 17th order Ambisonics using the sampling decoder with max_re weighting [9]. The

room acoustic model was implemented in the program real-time SOFE developed at the institute [6]. The visual stimuli consisted of a static picture of an avatar in human size that was presented synchronously with the onset of the target sentence at the direction of the target sentence. After the sound stimulus, the avatar stayed on the screen and switched its position only with the next trial at the onset of a new sentence.

The participants could freely perform self-rotations during the test. In two experimental conditions, A-only and AV, the participants were instructed to imagine a cocktail party situation in which somebody approached them from a random direction and react with their own rotation as if they were in such a situation. After presentation of the sentence, participants had to provide the response. They had to enter five words of the OLSA sentence on each trial. The GUI displayed 10 options for each word. Participants had to guess the word if they did not hear that particular word.

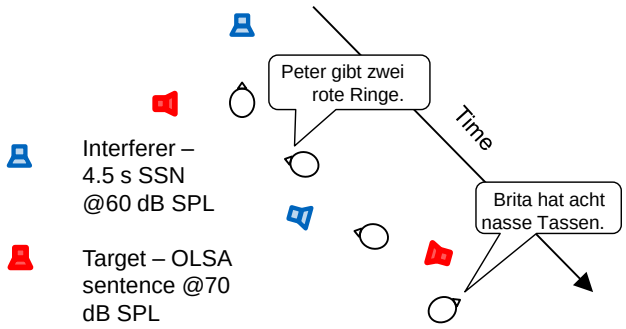

Figure 1: Two example trials with the target at -90° and 180° with exemplary self-rotations. Target angle could be 0°, ±90° and 180° chosen at random relative to the current end orientation of the person. Interferer was speech-shaped noise at 0°, the target was a sentence from the OLSA list. The task was to imagine a noisy situation when somebody is approaching the listener from a random direction, listen to the target and respond with self-rotation as if one was in such situation. All stimuli were delivered in free field. Target angle could be visually indicated (AV) or not (A-only).

Figure 1 shows two example trials from the experiment. It shows that at the end of each trial (the time after they give response) participants usually had a different orientation than at the trial's beginning. The figure shows that the virtual room always aligned with the orientation of the participant at trial start, using the motion tracking data, so the angle of 0° was set to the current orientation, and the participants did not have to return to their original orientation but could just continue the experiment from their current orientation.

The A-only and AV conditions were identical, except that in the AV condition the avatar appeared at the target location, otherwise the screens were blank. In another condition, Static, the stimuli were identical to the A-only condition but the participants were asked to stand still and perform the speech test without any motion, looking straight ahead. The condition was held constant over a block of trials.

The experiment was organized in 6 blocks (2 repeats x 3 conditions) of 48 trials (12 sentences x 4 target positions). For a combination of the three conditions and the four target angles, the target sentences were selected from one list. Each participant was randomly assigned 12 lists, 24 sentences were used from each list during the experiment. One trial consisted of the presentation of one target sentence and the response. Each trial started with the presentation of the interferer sound in front of the listener and a target sentence at a pseudo-randomly chosen target angle from one of four possible target angles. The target angle changed after each trial. Further details of the randomization procedures and the training procedures can be found elsewhere [5].

Analysis

In this preliminary analysis, the complexity of movement was assessed in terms of trial-to-trial variability of horizontal self-orientation angles. The variability was measured as the interquartile range (IQR) of the median orientation during each word of the target sentence. The IQR was computed from the data pooled for a given combination of condition and target angle and averaged across word positions. The analysis was performed in Matlab (v9.9, Mathworks).

Results

The analysis aims to identify whether the information about the target location influenced the complexity of movements. Higher trial-to-trial variability would indicate higher complexity.

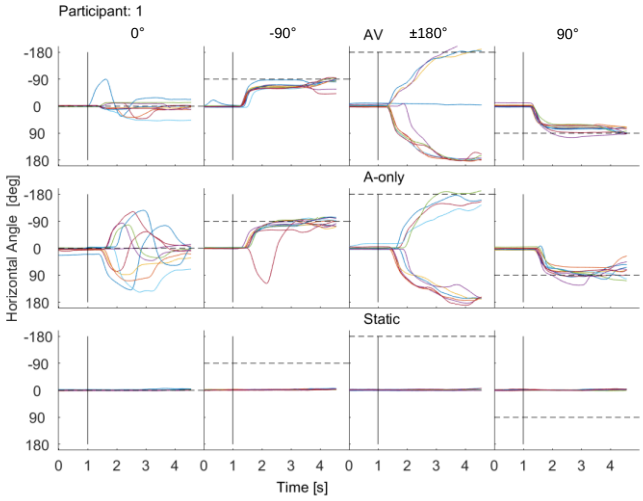

Figure 2: Raw data of self-rotation trajectories for each trial of the experiment for one participant. The panels are organized according the target angle (columns: 0°, -90°, 180°, 90°) and conditions (rows: AV, A-only, Static). Data are re-plotted from [10].

Figure 2 shows raw self-rotation trajectories for all experimental trials for one participant. By visual inspection of the data, it is obvious that the trajectories in the A-only condition (middle row) have higher trial-to-trial variability than the trajectories in the AV condition. This is further summarized in Figure 3 for both participants. The data also

show that the participant tended to undershoot the target, when the target was at the side or behind the participant.

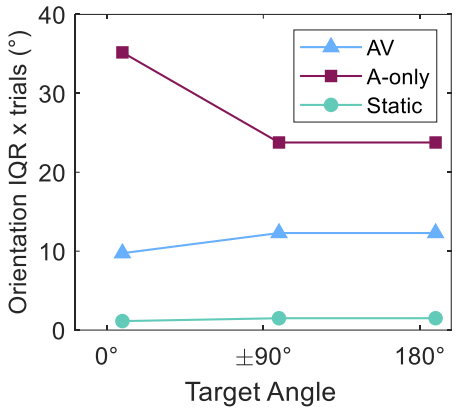

Figure 3: Trial-to-trial variability of orienting angles as a function of target angle. Lines indicate conditions.

Figure 3 shows the across-subject median of the inter-trial interquartile range as a function of target angle for all three conditions (shown with different symbols and lines). The data show that the trial-to-trial variability was much higher in the A-only condition than in the AV condition for all target angles; the difference was highest for the frontal angle, which could relate to different movement profiles between the target angles. The analysis shows a trend in the expected direction, but the firm conclusions cannot be drawn in this preliminary analysis with the data from two participants. The data in the Static condition were close to zero as expected. The small deviation from zero could be related to the sway when standing still.

Discussion

This preliminary analysis suggests that the visual target location cue reduced the trial-to-trial variability of self-rotation trajectories in an auditory speech intelligibility task. Therefore, the trajectories in the AV condition were more consistent across the experiment than in the A-only condition.

This indicates that when the participants knew the target location, the trajectories were less complex and more directed. In turn, when participants did not know the target location, they made more complex and pronounced movements which might indicate an active search behavior. The role of head movements in sound localization has been extensively studied even in the early literature [11]. The current results suggest that the participants localized the target in this speech intelligibility task even if they were not explicitly instructed to do so. These results suggest that people actively use a search behavior in cocktail party situations when the location cues are absent or less salient. On the other hand, if the location cues are present people tend to perform a single rotational movement towards the target such that they undershoot the lateral or rear target, which is often the acoustically optimal orientation [3].

In conclusion, people seem to actively use whole body rotations, and not only head movements, in the cocktail party situation when they are instructed to behave naturally. The preliminary data indicate that the location cue made the self-rotation more consistent, while the absence of the location cue led to more diverse and complex trajectories.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 352015383 – SFB 1330, Project C5. rtSOFE development is supported by the Bernstein Center for Computational Neuroscience, BMBF 01 GQ 1004B.

References

[1] A. W. Bronkhorst, ‘The cocktail-party problem revisited: early processing and selection of multi-talker speech’, *Attention, Perception, Psychophys.*, vol. 77, no. 5, pp. 1465–1487, Jul. 2015, doi: 10.3758/s13414-015-0882-9.

[2] R. L. Freyman, K. S. Helfer, D. D. McCall, and R. K. Clifton, ‘The role of perceived spatial separation in the unmasking of speech’, *J. Acoust. Soc. Am.*, vol. 106, no. 6, pp. 3578–3588, 1999.

[3] J. A. Grange and J. F. Culling, ‘The benefit of head orientation to speech intelligibility in noise’, *J. Acoust. Soc. Am.*, vol. 139, no. 2, pp. 703–712, Feb. 2016, doi: 10.1121/1.4941655.

[4] W. O. Brimijoin, D. McShefferty, and M. A. Akeroyd, ‘Undirected head movements of listeners with asymmetrical hearing impairment during a speech-in-noise task’, *Hear. Res.*, vol. 283, no. 1–2, pp. 162–168, 2012, doi: 10.1016/j.heares.2011.10.009.

[5] Ľ. Hládek and B. U. Seeber, ‘Behavior and Speech Intelligibility in a Changing Multi-talker Environment’, in *Proc. of the 23rd International Congress on Acoustics 9 to 13 September 2019 in Aachen, Germany*, 2019, pp. 1–6.

[6] B. U. Seeber and S. W. Clapp, ‘Interactive simulation and free-field auralization of acoustic space with the rtSOFE’, *J. Acoust. Soc. Am.*, vol. 141, no. 5, pp. 3974–3974, May 2017, doi: 10.1121/1.4989063.

[7] B. U. Seeber, S. Kerber, and E. R. Hafter, ‘A system to simulate and reproduce audio–visual environments for spatial hearing research’, *Hear. Res.*, vol. 260, no. 1–2, pp. 1–10, Feb. 2010, doi: 10.1016/j.heares.2009.11.004.

[8] K. C. Wagener, V. Kühnel, and B. Kollmeier, ‘Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests’, *ZEITSCHRIFT FUR Audiol.*, vol. 38, no. 1, pp. 4–15., 1999.

[9] F. Zotter and M. Frank, *Ambisonics*, vol. 19. Cham: Springer International Publishing, 2019.

[10] Ľ. Hládek and B. U. Seeber, ‘Speech intelligibility in reverberation is reduced during self-rotation.’, *Zenodo*, 2021, doi: 10.5281/zenodo.5069533.

[11] H. Wallach, ‘The role of head movements and vestibular and visual cues in sound localization’, *J. Exp. Psychol.*, vol. 27, pp. 339–368, 1940.