PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Behavior and Speech Intelligibility in a Changing Multi-talker Environment

Ľuboš HLÁDEK and Bernhard U. SEEBER

Audio Information Processing, Technical University of Munich, Germany

ABSTRACT

At the auditory cocktail party, we often listen to speech coming from different directions. This project aims to investigate how behavior and speech perception interact when people have different goals when listening in an acoustically complex scene. In the experiment, participants stand in a simulated reverberant room and listen to target sentences of one talker from random directions in the horizontal plane. The interferers are presented in an ongoing fashion clustered in front of the participant but the participants can move during the presentation of the sentence. At the beginning of each trial, the scene is reset to be aligned with the current orientation of the participant. The experiment aims to test behavior without restricting the participant with instructions to move in a certain way and test the effect of visual cues. Stimuli are presented in free-field using the real-time Simulated Open Field Environment. The movement behavior is recorded using a video-based motion-tracking system, and the motion data are analyzed in relation to target position and visual conditions. The results are discussed in the context of developing realistic listening scenes for psychoacoustical and audiological research.

Keywords: Speech Perception, Head Movements

1. INTRODUCTION

Speech intelligibility testing is traditionally done with protocols in which the test subject is restricted in movement. However, people in real situations move and this may affect speech perception especially in situations with interfering sound sources. Movement within the acoustic scene is likely to affect speech intelligibility due to the spatial release from masking depending on head orientation (1).

Novel technologies for audio-visual virtual reality (2,3) allow the creation of complex acoustical scenes together with the immersive visual component of the scene that can be updated in real-time in order to respond to the behavior of the participant. This allows usage of experimental protocols that are less restrictive in terms of movement behavior and more realistic, which may elicit behavior that is similar to what can be observed in real situations.

Natural movement behavior is, however, quite subjective, since people vary in terms of their propensity to head movements (4). One example is an experiment in which subjects were instructed to follow two brief audio-visual speech targets at ±30 degrees from the midline. Some people had tendency to fully rotate their head towards the target, some only looked at the target but did not move their head (5). Brimijoin et al. (6) and Grange and Culling (1,7) investigated whether people adapt their orienting response to the acoustic scene. Brimijoin et al. (6) concluded that despite unilateral hearing aid users actively were using head orienting behavior while listening to speech in an adaptive procedure, they could not exploit the best SNR. Grange and Culling (1) also reported that normal-hearing, freely-orienting participants could not position themselves in the best SNR situation and even 44% of undirected trials were without movement. The effect of instruction was further evaluated by Grange and Culling (7) who found that explicit instructions to exploit the head orientation for better SNR lead to more pronounced head movements and better subjective SNR as measured for the frontal target when the distractor was coming from the side or from behind.

Visual cues are also likely to affect movement behavior (7,8). In general, people have the tendency to look at the sound even when the sound's position is not visually indicated. The visual stimulus, however, lessens the uncertainty of the sound source position. A picture of a person or a face may elicit cognitive bias of looking at the person, which is a natural response in social situations.

The ability to move freely in the above mentioned studies was often restricted either by the pose of the participant or the instructions. The aim of the present investigation was to test movement behavior and speech intelligibility during a conversation-like-situation when the source comes from different directions and the person is able to move freely without explicit instructions regarding the type of the head movement. In order to bridge a continuous type of presentation and trial-by-trial based type of presentation in the current experiment, the orientation of the acoustically simulated environment was always aligned with the current head orientation of the participant at the beginning of each trial without explicit notification of the participant. The primary questions we were interested in were: Will realistic orienting behavior interact with speech intelligibility when the interferer sound comes from the front at the beginning of the trial (worst acoustic configuration in terms of SNR) and when the targets emanate from different directions? How will visual cues interact with the head orienting behavior and speech intelligibility in such situation?

2. METHODS

2.1 Participants

Five participants (1 female, 4 male, 23±1.2 [mean±STD] years) took part in the experiment. All of them were native German speakers. Three participants had normal hearing as defined by the standard audiological pure tone threshold screening, two participants who did not undergo the screening did not report any hearing related problems. The participants provided written informed consent. The study was approved by the ethics committee of the Technical University of Munich, 65/18S.

2.2 Setup

The experiment was conducted in the anechoic chamber of the Audio Information Processing group (10 m x 6 m x 4 m; l x w x h) using the rtSOFE (2,3). The participant stood in the middle of the loudspeaker array (Dynaudio BM6A mkII, Dynaudio, Skanderborg, Denmark) with the square shape (4.8 m x 4.8 m) positioned at the height of 1.1 m from the wired net. The loudspeakers were coupled with digital-to-analog converters (RME 32DA, Audio AG, Haimhausen, Germany) and multi-channel sound card (RME HDSPe, Audio AG; Haimhausen, Germany) connected to the main experimental computer. The testing chamber was further equipped with 4 (Barco F50 WQXGA, Barco, Kortrijk, Belgium) high resolution projectors (32 dB(A) background noise), which were connected to another computer with two high performance video cards (NVidia P5000, NVidia, Santa Clara, California, USA). The projectors projected to 4 acoustically transparent screens, which were scrolled down in the front of the loudspeaker array for the duration of the experiment, creating a 360° immersive display. Additionally, 12 high-speed motion tracking cameras (OptiTrack Prime 17W cameras, NaturalPoint Inc. Corvallis, Oregon, USA) were used to monitor motion behavior. 8 cameras were positioned above the projection screens, 4 were positioned at ground level. Motion behavior was obtained by monitoring the position of the head. Each participant was equipped with the motion-tracking object consisting of three reflective spheres attached to the inside of a construction hard hat. The origin of the motion tracking object was manually adjusted to be approximately on the interaural axis of the participant.

The audio presentation involved acoustical simulation of a reverberant room (9 m x 15 m x 3 m, l x w x h) (RT20 = 900 ms) using rtSOFE (2,3), the receiver in the simulation was at 4 m, 7 m, 1.8 m relative to the origin of the x, y, z coordinates of the room. The room simulation produced 36 channel impulse responses (using the nearest speaker mapping method) which were convolved with the stimuli used in the experiment. Video presentation was controlled using Blender game engine (Blender v2.79, Blender Foundation, Amsterdam, Netherlands). The experiment was controlled using Matlab (Matlab v9.5, MathWorks, Natick, MA, USA) and Python 3.6.

.

2.3 Stimuli

The target stimulus was always a German OLSA (9) sentence which was presented from one of four possible directions (0°, +90°, 180°, -90°) at a virtual distance of 2.2 m relative to the origin. The

target sound was presented at 50 dB(A) at the center of the loudspeaker array. The sentence for each target location was selected from the one OLSA list spoken by a male speaker. Interferer sound consisted of 6 interleaved and randomly selected OLSA sentences spoken by a female talker always presented from 0° azimuth. The female OLSA stimuli were recorded and preprocessed in a previous project, the target male stimuli were from the OLSA test CD. Each of the six interferer streams was created by concatenating the sentences convolved with the room impulse response, such that the time between the offset of the previous sentence and the onset of the new sentence was randomly chosen between 0.5 s and 1.5 s. The interferer signal was presented at 57 dB(A) at the center of the loudspeaker array.

The azimuth of the target and interferer presentation was aligned with the current orientation of the head at the beginning of each trial. Therefore, the 0° azimuth always corresponded to the nearest loudspeaker relative to the head's horizontal orientation at the beginning of each trial. Each trial started with the realignment of the masker sound, only after 500 ms the target sound was presented.

The visual stimulus, used in one condition, consisted of a picture of a person which was presented at the position of the target sentence 500 ms before the onset of the target sound. The picture was generated by the MakeHuman software (10) and presented in dimensions of a regular person standing 2.2 m from the center of the loudspeaker array and looking to the middle of the array.

2.4 Procedures

Each participant underwent a training procedure before the main experiment to familiarize themselves with the OLSA corpus and the user interface used to provide responses. The procedure consisted of presentation of at least 90 OLSA sentences in noise from at least 3 lists. Upon arrival to the anechoic room, the experimenter equipped the participant with the head tracking device and explained safety procedures for the anechoic room. The participant together with the experimenter entered the chamber and the experimenter instructed the participant to imagine being in a noisy room where people are talking to the participant from different directions. They will be hearing the target sounds by a male talker coming from different directions and they should try to understand as much as possible since there will be also interferers with female voice. They should also move as if they were in such situation, but not to go away from the middle of the loudspeaker array (due to the acoustical sweet spot). Further, the experimenter explained that movement can affect performance but no further instructions regarding the movement behavior were provided. Following the instructions, the participant was provided with a hand held tablet computer which displayed all options of the OLSA test, current trial number, number of correct responses from the last run, and two buttons which controlled the beginning of the run and submission of responses (which lead to the presentation of the next trial).

The whole experiment consisted of three runs corresponding to three experimental conditions, each of 40 OLSA sentences, such that each target position was used 10 times. One OLSA sentence together with the response created a trial. The experiment employed: the audio-visual condition (AV), audio only condition (A-only), and the static condition. The first two conditions were identical with the exception that in the AV condition the target direction was indicated by the picture of the avatar. The static condition was identical to the A-only condition but the participant was instructed to stand still and not move the head. The experiment always started with the AV or the A condition, the order was counterbalanced across participants, and commenced with the Static condition.

3. RESULTS

Here we analyze preliminary data of the ongoing experiment in terms of speech intelligibility and movement behavior. The data are analyzed in terms of the observed trends since the number of participants (n=5) is not sufficient to make firm conclusions.

3.1 Speech Intelligibility

Figure 1 shows speech intelligibility in percent correct values where 0 corresponds to no intelligibility and 100 corresponds to full intelligibility shown as across-subject means with standard errors of the means. Symbols indicate different conditions, x-axis indicates target angle, the offset on x-axis is included to increase the visibility of the data. Data for the left and right lateral targets were collapsed since no laterality effect was expected.

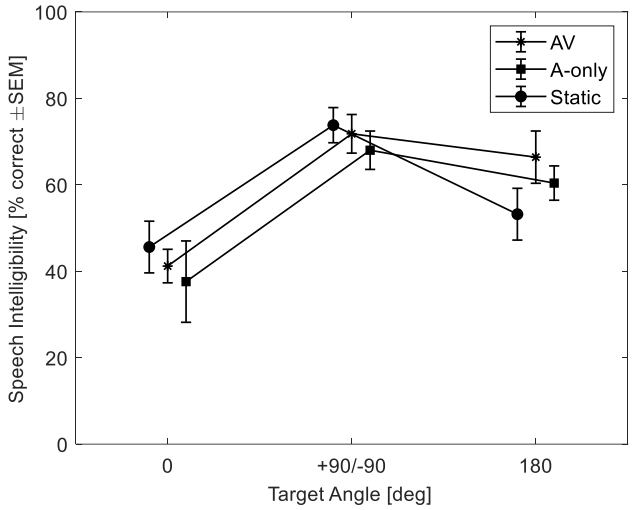

Figure 1 – Across-subject mean (±SEM) speech intelligibility as a function of target angle. Each symbol denotes data for each condition.

The data on the figure show that speech intelligibility was worst when the target was collocated with the masker (0°) since there was no benefit spatial separation of the target and the masker. Performance for the lateral target was better, but only due to the spatial separation of the target of the masker, since the visual inspection suggests no difference between the Static and the other two conditions. Therefore, head orienting behavior in the AV and A-only conditions had only a small effect on speech intelligibility when the target was at 90°. However, a difference between the conditions (based on visual inspection) can be seen for the target at 180°. Here, the performance in the AV condition improved relative to the Static condition. A-only condition is in between these two conditions. This suggests that natural orienting in this condition leads to head orientations enable them to use the head orientation benefit, even in moderately reverberant room.

3.2 Unrestricted Head Orienting Movements

Figure 2 shows raw movement patterns of one participant. The data are plotted with respect to the beginning of the trial labeled as 0. Data plotted separately for each trial are shown by separate curves. The vertical dashed line, within each panel, indicates the onset of the target. Horizontal lines indicate the target direction. Panels of the figure are organized according to the target positions (columns) and conditions (rows).

In the AV condition target at 0°, the person was moving less than in the condition A-only with target at 0°. This indicates that the person was actively involved in exploring possible benefits of speech orientation in the A-only condition. There is also a difference between the AV and the A-only condition for the target at 180°. In the AV condition, the movement is more precise, while in the A-only condition the movement is more random, reflecting the difficulty to localize the sound or the attempt the find a better head orientation. Also, it shows that the person sometimes did not know where the target was, so they did not move at all. Another visible pattern in the data is that orienting towards the lateral targets is fairly accurate, but orienting towards the 90° target in the A-only condition is more delayed and more complex than in the AV condition, reflecting the uncertainty of the sound location. In the AV condition, the movements are also slightly more precise, however, there is an evident tendency to undershoot the target. The participant was not moving in the Static condition, however, small head movements could not be avoided, which may have helped with speech perception.

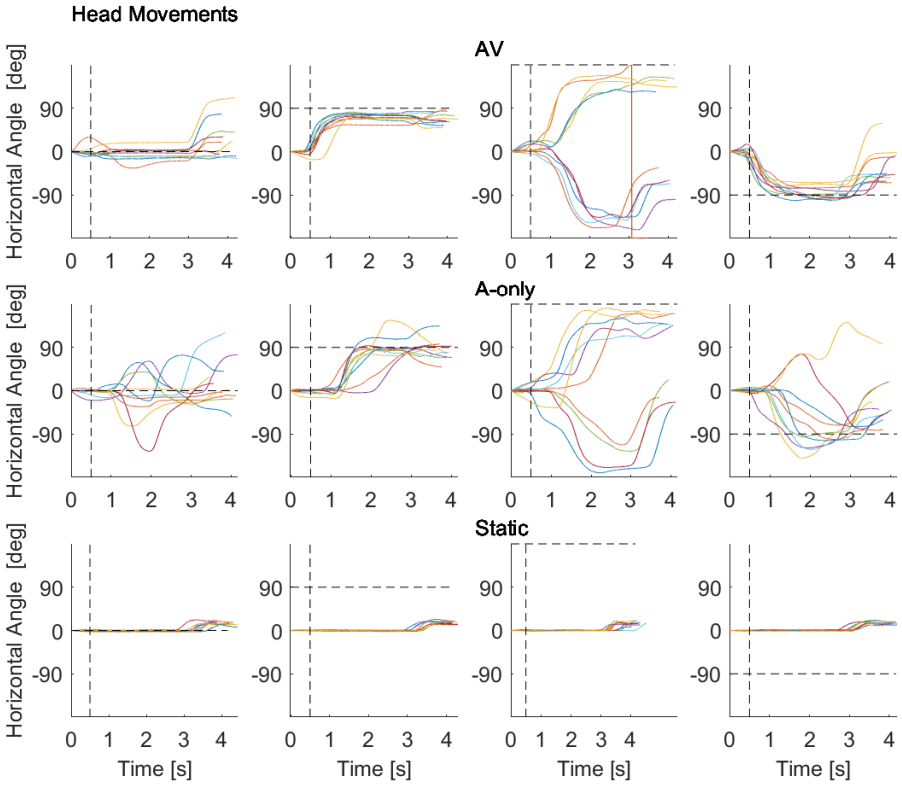

Figure 2 – Raw head orienting movements of a single participant. The data show yaw angle of the head during the presentation of the OLSA sentence. Each trace is one trial. Columns of the panels show data for different target positions, rows show data for different conditions indicated by the caption.

4. DISCUSSION

This study investigated natural orienting behavior and speech intelligibility in a moderately reverberant environment. The preliminary analysis indicates that speech intelligibility in the configuration with the masker at the front and target at 90° exhibits a spatial release from masking (7), and in this condition no additional movement related benefit could be observed. When the target was at 180°, the data suggest that the natural movements lead to better speech intelligibility. These patterns are facilitated by the procedure since the interferer sound always appeared in front of the participant at the beginning of each trial and then the participant naturally moved toward the target, which put them into more favorable SNR situation.

The data further show that participants tended to turn their head more complexly if they had problems localizing the sound in the A-only condition. In the AV condition, participants oriented their heads straight towards the targets. The head movement data of the participant in Figure 2 suggest that the orientation behavior was not straight towards the target (orienting was somewhat biased).

In agreement with previous studies (1,6), preliminary data do not suggest that the movement is driven by the best SNR for the best speech intelligibility but it cannot be completely excluded at this point. This may relate to a lack of sensitivity to the SNR changes, or movement patterns may be driven by localization rather than speech understanding processes.

This study also showed that utilizing interactive acoustic environments may elicit more natural patterns of behavior possibly avoiding the necessity to deliberately instruct participants, which interferes with the aim for naturalness.

ACKNOWLEDGEMENTS

This work was supported by DFG SFB 1330 project C5 and the rtSOFE by the Bernstein Center for Computational Neuroscience, BMBF 01 GQ 1004B.

REFERENCES

1. Grange JA, Culling JF. The benefit of head orientation to speech intelligibility in noise. J Acoust Soc Am [Internet]. 2016 Feb;139(2):703–12. Available from: http://asa.scitation.org/doi/10.1121/1.4941655
2. Seeber BU, Clapp SW. Interactive simulation and free-field auralization of acoustic space with the rtSOFE. J Acoust Soc Am [Internet]. 2017 May;141(5):3974–3974. Available from: http://asa.scitation.org/doi/10.1121/1.4989063
3. Seeber BU, Kerber S, Hafter ER. A system to simulate and reproduce audio–visual environments for spatial hearing research. Hear Res [Internet]. 2010 Feb;260(1–2):1–10. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0378595509002809
4. Fuller JH. Head movement propensity. Exp brain Res [Internet]. 1992;92(1):152–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1486950
5. Hládek L, Porr B, Brimijoin WO. The Effect of Eye-controlled Beamforming on Speech Intelligibility in a Dynamic 'Cocktail Party.' In: 41St Annual MidWinter Meeting of Association for Research in Otolaryngology. 2018. p. 730.
6. Brimijoin WO, McShefferty D, Akeroyd MA. Undirected head movements of listeners with asymmetrical hearing impairment during a speech-in-noise task. Hear Res [Internet]. 2012;283(1–2):162–8. Available from: http://dx.doi.org/10.1016/j.heares.2011.10.009
7. Grange JA, Culling JF, Bardsley B, Mackinney LI, Hughes SE, Backhouse SS. Turn an Ear to Hear: How Hearing-Impaired Listeners Can Exploit Head Orientation to Enhance Their Speech Intelligibility in Noisy Social Settings. Trends Hear. 2018;22:1–13.
8. Hendrikse MME, Llorach G, Grimm G, Hohmann V. Influence of visual cues on head and eye movements during listening tasks in multi-talker audiovisual environments with animated characters. Speech Commun [Internet]. 2018 Jul;101(June 2017):70–84. Available from: https://doi.org/10.1016/j.specom.2018.05.008
9. Wagener K, Kühnel V, Kollmeier B. Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test. Zeitschrift f{ü}r Audiol. 1999;38(1):4–15.
10. MakeHuman [Internet]. 2019. Available from: http://www.makehumancommunity.org