The effect of self-orienting on speech perception in an acoustically complex audiovisual scene

Lubos Hladek, Bernhard U. Seeber

Audio Information Processing, Technical University of Munich, 80290 Munich, E-Mail:lubos.hladek@tum.de

Introduction

Listening to speech in a noisy environment is often a challenging task. Spatial release from masking (SRM) enables better speech understanding when the target speech is spatially separated from the interferer. However, the magnitude of SRM varies greatly with the head orientation [1], [2]. In daily conversations, people move around to improve understanding, but previous research brought mixed results in terms whether and how self-orienting affects speech understanding.

In a previous experiment by Brimijoin et al. [3] that studied unrestricted head movements, participants with unilateral hearing loss heard to sentences in noise while the signal-to-noise ratio was adapted. In most cases they oriented their heads 60 degrees away from the frontal target irrespective of the position of the interferer which led the authors to conclude that participants maximized the level of the target rather SNR. In another experiment by Grange and Culling [4], participants were asked to follow speech material in different spatial configurations and 'do whatever they would do normally in a social situation to understand the speech' while being seated on a chair. The authors did not observe an effect of SNR on head movement behavior but further experiments suggested that concrete instructions to exploit head movements may be beneficial for speech understanding [5]. In a recent study by Frissen et al. [6], the effect of self-motion during speech perception was investigated. The participants listened to spatially distributed sentences over headphones, the target was determined by a call sign, and the participants were asked to rotate the head during sound presentation. The sound was spatialized using head-related transfer functions (HRTFs) and the spatialization was either fixed (head-centered) or responded to the head movement to create a stable image (world-centered). Speech perception was negatively impacted in the world-centered condition, but the authors concluded that the effect is of negligible size.

In our pilot experiment [7], we studied speech perception during self-orienting movements when the participants were standing and freely orienting towards a target that originated in the front, or the back, or the side at ±90°. The interferer came always from the front at the start of each test sentence. The results suggest that speech perception improves slightly for the situation when the target was behind the participant but not much for the situation when the target was at the side, although we expected improvement since rotation towards the target (which was usually the case) brought them into an acoustically more favorable condition [8]. Hence, we suspected a negative effect of self-orienting or the dynamic change of speech cues on speech perception.

To address the question directly, we modified the pilot study 1) to create a more sensitive measure by adapting parameters of acoustical simulation 2) to create stimuli such that we can reconstruct what exactly people heard during self-motion. Thus, the aim of the current investigation is to assess speech perception during self-orienting and speech perception of a sound that was heard during self-motion, effectively isolating the effect of self-orienting from that of the movement-related cue change. This paper is still an interim report that shows mean performance of three participants and modelling of speech intelligibility for of one of the participants.

Methods

Three participants (1 female, mean age: 26.3 years) whose mother tongue was German were recruited for the experiment. Their hearing was checked by pure-tone audiometry at standard audiometric frequencies. All hearing thresholds were equal or below 20 dB HL at each frequency. One participant has not completed the screening but had good hearing by self-report. The participants provided written informed consent. The study was approved by the ethics committee of the Technical University of Munich, 65/18S.

The experiment was conducted in the Simulated Open Field Environment (SOFE v4, [9]), an audio-visual setup with loudspeakers and a four sided CAVE system inside an anechoic chamber [10]. The participant was standing in the middle of a square-shaped loudspeaker array composed of 36 loudspeakers (Dynaudio BM6A mkII, Dynaudio, Skanderborg, Denmark). The closest loudspeakers were at 2.1 m from the center of the square. The time, phase and frequency response of the loudspeakers was equalized in the range of 100 Hz to 18 kHz. Twelve low-latency motion-tracking cameras (OptiTrack Prime 17W cameras, NaturalPoint Inc. Corvallis, Oregon, USA) tracked the head of the participant. Four large screens were hung in front of the loudspeakers, four projectors (Barco F50 WQXGA, Barco, Kortrijk, Belgium) projected visual stimuli on the screens.

The target sound stimuli were OLSA [11] sentences presented at 61 dB SPL either from the front 0°, the back 180°, or the side ±90°. The interferer was speech-shaped noise created individually from and for each test sentence and presented at 70 dB SPL. The interferer had the same spectrum as the target sentence and a duration of 4.5 seconds. The target sentence was initiated always after 1 second of the interferer. The stimuli were placed in a virtual room (RT_{30} = 970 ms, 11 m x 13 m x 3 m, l x w x h) generated by rtSOFE [9], [10]. Room impulse responses for each loudspeaker channel were created using 17th-order Ambisonics with max_{RE} weighting [12] for reflections up to order 5, and nearest loudspeaker mapping for all remaining reflections up to order 100. In the audio-visual (AV) condition, a man-like virtual character appeared on the screen synchronously with the acoustic stimulus at the same

azimuth. The character remained on the same azimuth until the next stimulus was presented. In audio-only (A-only) condition, there was no visual component involved.

The task of the participants was to imagine that they were in the middle of a noisy situation and somebody talks to them from one of four possible azimuths and respond naturally with movement to this situation. They should listen to the target sentence and respond on a hand-held tablet that displayed the interface for our OLSA test implementation. The participants were further told not to leave the center of the array, and this was automatically checked at the beginning of each trial. On every trial, room simulation aligned with the actual rotation of the participant thus the stimuli were presented always relative to the participants' orientation. In the reference static condition (Static), participants remained still and looked forward. This was also automatically checked by the experimental script.

The experiment was organized in 6 blocks of 48 trials. One trial corresponds to one sentence in 4.5 s of noise stimulus. In each block, 12 sentences from 4 possible azimuths were presented. There were two blocks for each of the 3 conditions: AV, A-only, Static (Baseline). For each combination of the condition and azimuth, one OLSA list was randomly assigned for each participant. Each participant heard the last 24 sentences from each list. The order of blocks was randomly generated for each participant in a way that the first three blocks involved all three conditions. The experiment was conducted within 2 or 3 sessions.

Participants conducted 4 training blocks before the main experiment. The training protocol involved the same stimuli and procedures as the main experiment, except that only lists 32-40 (OLSA CD) were used for the training, these were then excluded for the main experiment. In the training, stimuli were presented at 64 dB SPL (first 2 blocks) and 62 dB SPL (second 2 blocks), i.e. at a slightly improved SNR. Participants remained still during training.

After the experiment, the self-motion trajectories in terms of yaw angle were used to generate stimuli that were heard by the participant during the motion. The stimuli were created by rotating the virtual room in 0.5° steps around the position of the virtual receiver at [4 m, 7 m, 1.8 m; relative to origin of the room] according to the trajectory of the actual rotation. To ensure synchrony of the motion tracking and sound presentation, the motion tracking device was synchronized with the sound card via the word-clock signal. Thus for each 123 samples of the soundcard (at 44100 Hz), one sample of the motion tracking signal was generated, which was produced with 2.8 ms latency. For latency assessment, the target sound (no interferer) was recorded by the head and torso simulator (HATS) (HMS II.3, Head Acoustics, Herzogenrath, Germany) when the HATS was manually rotated on a swivel chair. The motion trajectory was used to generate counter-rotating signal, which was recorded again with the HATS being static. To assess the latency between these two signals, ILDs [13] of the two recordings were cross-correlated. The analysis showed latency of about 2 ms which is close to the latency of motion tracking system.

The stimuli in which the sound moves according to the motion trajectories of one participant were analyzed by the Binaural Speech Intelligibility Model BSIM [2]. For this analysis, we choose the sluggish BSIM2010 version (with batch processing).

Results

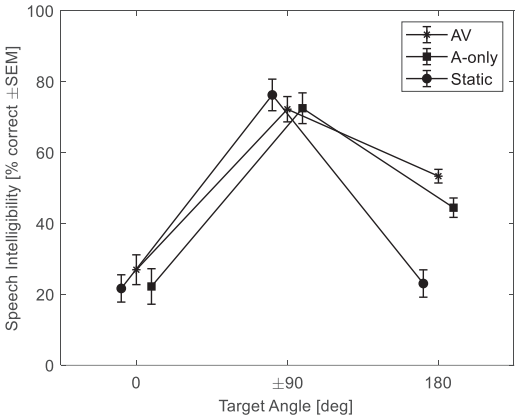

Figure 1. Preliminary results of three participants: Speech perception scores in per cent correct for each of the three conditions and all azimuths. The data for the targets at the left and right sides were pooled over.

Data in Figure 1 show mean performance in each of the experimental conditions. As expected, performance for the collocated condition (0° on x-axis) is the worst since there was no SRM, and there is no difference between the control static condition and the condition in which the participants could move.

Speech intelligibility for the targets at the side (90° on x-axis) improves drastically. Performance in AV (stars) and A-only (squares) condition is slightly below the Static Baseline condition. This replicates the result of the pilot study [7]. Targets behind the participant (180° on x-axis) were perceived with great difficulties in the static condition, as would be expected in the almost diotic condition. Participants improved considerably when they could move. They improved slightly more in the AV condition than in A-only condition. It can be argued that in the A-only condition, it was more difficult to find the sound behind them since this is easily confusable with the 0° condition in which they do not usually move. Thus, the visual cue helped people to orient in the scene, which improved speech intelligibility.

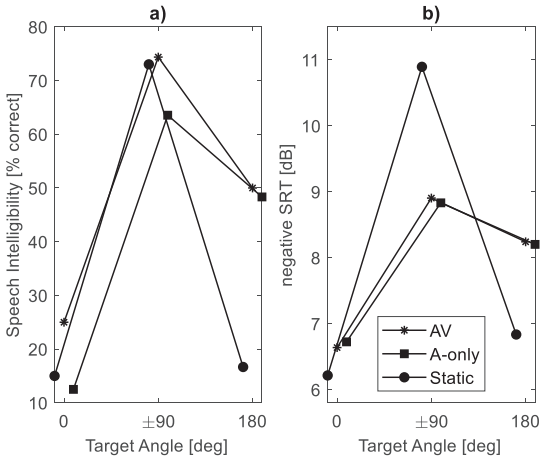

Figure 2. a) Speech intelligibility results of participant S03. b) Speech reception thresholds (SRT) modeled with the "sluggish BSIM2010" model. SRTs are displayed as negative values for easier comparison.

Data on Figure 2 show performance of one participant (a) and the predictions of the speech intelligibility model (b). For the static condition, the model shows 5 dB difference between the diotic condition and the target at the side. This value approximately corresponds to the predications of another speech model [8]. The mean SRT of the two conditions equals to -8.5 dB, and given the grand mean performance in the static condition 55% and the SNR of -9 dB, the model predicts performance well. The modeled speech intelligibility for the moving conditions (AV and A-only) for the target at 0 degrees is similar to the static condition, while performance for the two conditions for 90° targets is below the static condition. Performance for the 180° target is above the static condition. The model captures the trends in the data, but the A-only and AV conditions are underestimated for the 90° and 180° conditions when the participant was moving.

Discussion

This experiment shows that speech perception is improved when people naturally orient without explicit instructions in a complex acoustic scene. This holds mainly for the situation when the target is behind and the masker is at the front of the participant but not so much for the situation when the target is at the side. The data are in line with our pilot study [7] and the test shows much higher sensitivity. Results are generally in line with the previous studies in which the participants did not had the target behind (and interfere at the front) them and did not have to make big turns.

In addition, we analyzed the effect of self-motion cues by comparing performance of a single participant with the prediction of the speech intelligibility model. The predictions could capture the trends, but the model was not able to explain the data. Although this is only a case study, the analysis suggests that that speech perception might be influenced by self-orienting because we did not observe improvement of speech intelligibility (re. static condition) when the target is at the side and we expected higher improvement for the target behind the participant (at least to the value of 90° static condition), although a lack of motion of some participants might have smeared these differences. However, the current analysis is too preliminary to make conclusions.

In this interim report we showed that motion matters for speech intelligibility in a complex acoustic scene. We also showed a first attempt to model the data, but the model could capture only some aspects of the data. Subsequent studies will clarify whether the data could be explained by acoustic factors or whether non-acoustic factors related to self-motion influence speech intelligibility.

Acknowledgement

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] R. Beutelmann, T. Brand, and B. Kollmeier, 'Revision, extension, and evaluation of a binaural speech intelligibility model', J. Acoust. Soc. Am., vol. 127, no. 4, pp. 2479–2497, 2010.

[2] C. F. Hauth and T. Brand, 'Modeling sluggishness in binaural unmasking of speech for maskers with time-varying interaural phase differences', Trends Hear., vol. 22, pp. 1–10, 2018.

[3] 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.

[4] 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.

[5] J. A. Grange, J. F. Culling, B. Bardsley, L. I. Mackinney, S. E. Hughes, and S. S. Backhouse, 'Turn an Ear to Hear: How Hearing-Impaired Listeners Can Exploit Head Orientation to Enhance Their Speech Intelligibility in Noisy Social Settings', Trends Hear., vol. 22, pp. 1–13, 2018.

[6] I. Frissen, J. Scherzer, and H.-Y. Yao, 'The Impact of Speech-Irrelevant Head Movements on Speech Intelligibility in Multi-Talker Environments', Acta Acust. united with Acust., vol. 105, no. 6, pp. 1286–1290, Nov. 2019.

[7] 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.

[8] S. Jelfs, J. F. Culling, and M. Lavandier, 'Revision and validation of a binaural model for speech intelligibility in noise', Hear. Res., vol. 275, no. 1–2, pp. 96–104, 2011.

[9] 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.

[10] 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.

[11] K. C. Wagener, V. Kuhnel, B. Kollmeier, and T. Brand, ‘Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil II: Optimierung des Oldenburger Satztests Development and evaluation of a German sentence test Part II: Optimization of the Oldenburg sentence test.’, Z Audiol, vol. 38, no. 2, pp. 44–56, 1999.

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

[13] M. Dietz, S. D. Ewert, and V. Hohmann, ‘Auditory model based direction estimation of concurrent speakers from binaural signals’, Speech Commun., vol. 53, no. 5, pp. 592–605, May 2011.