



The effect of self-motion cues on speech perception in an acoustically complex scene

Lubos Hladek, Bernhard Seeber

► To cite this version:

Lubos Hladek, Bernhard Seeber. The effect of self-motion cues on speech perception in an acoustically complex scene. Forum Acusticum, Dec 2020, Lyon, France. pp.1283-1285, 10.48465/fa.2020.0364 . hal-03234211

HAL Id: hal-03234211

<https://hal.science/hal-03234211>

Submitted on 26 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THE EFFECT OF SELF-MOTION CUES ON SPEECH PERCEPTION IN AN ACOUSTICALLY COMPLEX SCENE

Ľuboš Hládek

Audio Information Processing, Technical University of Munich, Theresienstr. 90, 80 333 Munich, Germany

Bernhard U Seeber

Correspondence lubos.hladek@tum.de

ABSTRACT

In a cocktail party, people talk to each other from different directions and the listener often turns toward the speaker. Previously, we observed a decrease in speech intelligibility during self-orienting movement but we could not isolate the contribution of self-motion cues. In the present study, we isolate self-motion cues by comparing speech perception during self-orienting movement and a situation when the participant stands still and hears sound rotating around them in the same way, as they were moving. The participant thus receives the same auditory cues as during self-motion. The task of the participants is to listen to an OLSA sentence, which may emanate from any of four directions (front, back, sides) and orient naturally. A speech-shaped noise interferer is presented always from the front of the participant, relative to the beginning of the trial. A virtual room acoustic environment is auralized via the loudspeakers of the Simulated Open Field Environment in the anechoic chamber. The stimuli are created by convolving the sounds with impulse responses that were obtained by rotations of the simulated room according to actual self-motion. The results suggest that artificial rotation of the scene leads to realistic speech intelligibility.

1. INTRODUCTION

People are exceptionally good at listening to speech in noise, so called the ‘cocktail party problem’. One of the parameters is the position of the talker and the listener which leads to spatial release from masking (SRM) when the sound sources are displaced from interfering sounds [1]. The SRM further depends on head orientation relative to the talker and the masker, and people could obtain up to 16 dB head orientation benefit in an anechoic environment [2].

In our previous experiment [3], [4], people were listening to OLSA sentences in a reverberant room, which could emanate from one of four possible directions (0° , $\pm 90^\circ$, 180°) equidistant from the listener, in speech-shaped-noise coming from direction 0° . The standing listener was asked to naturally orient towards the target speech as if somebody was approaching them from that direction. Either a static visual character that indicated the direction (AV condition) accompanied the target speech or there was no visual character (A-only condition). For the lateral target, performance was slightly worse or equal than

the baseline condition during which the participants were standing (Static condition), despite we expected that they should have improved according to a prediction of SRM for the corresponding directions of rotation [5]. For the rear target, we observed only slight improvement with respect to the Static condition, despite a substantial benefit people obtain during the rotation, but less as we would expect based only on the spatial unmasking.

The aim of this pilot study was to isolate self-rotation cues from the sound-rotation cues. We asked whether speech perception is influenced by non-auditory rotation cues or whether speech perception is determined only by the changing acoustical cues due to sound rotation around the listener. In the current experiment, we thus compare speech perception from the previous experiment during active self-rotation with speech perception of the same material and same listeners during passive sound rotation.

2. METHODS

Sound rotation was created using sound virtualization techniques using corresponding rotation trajectories from the previous experiment. Thus, participants in this experiment were standing still but heard the same sounds as in the previous experiment when they were actively rotating. For the details of the previous study please see the references [3], [4].

The target stimuli were OLSA sentences presented at 60 dB SPL (only direct sound was measured) together with a stationary speech-shaped-noise presented at 70 dB SPL. The noise had the same spectrum as each target sentence. The masker sound had a duration of 4.5 seconds and the onset of the target sound was delayed by 1 second from the onset of the masker. The target sound could come from one of four possible directions (0° , $\pm 90^\circ$, 180°), the masker sound came always from the front of the listener (0°). The moving sound (M-S) stimuli were created by counter rotation of the acoustic scene according to the rotation of the participant in the previous experiment during the same target sentence. The stimuli were created by convolution of the anechoic target sentence with impulse responses belonging to the correspondingly rotated room acoustical condition. The rotation was sampled at 0.5° steps with the minimum length of the convolution window of 50 samples of the input sound and maximum step of 0.5° (max rotation speed that could be reconstructed was $442^\circ/\text{s}$ but on average it was much slower). The input signal was linearly

cross-faded between two successive windows before the convolution with the full impulse response. Each window was convolved with the impulse response that corresponded to the head orientation in that moment. The resulting stimuli reconstructed the scene with high spatial and temporal fidelity assuring that participants in this experiment heard stimuli that were very close to what they experienced in the previous experiment.

The stimuli were spatialized in a virtual reverberant room ($RT_{30} = 1.09$ s) which was created using an image-source method [6] implemented in the Simulated Open Field Environment (SOFE v4) [7]. The sounds were auralized over the loudspeakers using the Ambisonics technique with Max-rE coding [8] (reflections up to order 5) and the nearest loudspeaker mapping (reflections up to order 100).

The task of the participant was to stand still, look forward, and report the perceived sentence into a custom GUI on a hand-held tablet. The GUI displayed all possibilities of the matrix sentence test and provided feedback on the performance of the last sentence. The participant's position was tracked and checked by the experimenter at the beginning of the experiment.

During the experiment, participants heard the same sentences of the AV and the A-only condition as in the previous experiment, with a difference that there was no visual component in the current experiment and the order of all sentences was shuffled and irrespective of condition. Therefore, the experiment involved 192 trials, which were split into 4 blocks of 48 sentences to allow short breaks for the participants.

The current experiment was conducted in the same environment as the previous experiment. For the description of the acoustic system in the anechoic chamber see the methods of the previous studies [3], [4]. Four huge white acoustically transparent screens were placed in front of the loudspeakers. The screens showed static pictures of virtual characters. The chamber was further equipped with a motion capture system (OptiTrack Prime 17W, NaturalPoint Inc. Corvallis, Oregon, USA) running at 359 Hz which was synchronized (eSync 2, NaturalPoint Inc. Corvallis, Oregon, USA) with the sound card to allow very accurate reconstruction of the trajectories with respect to the sound. The synchrony was also assessed by a procedure in which an experimenter manually rotated a head and torso simulator (HATS) (HMS II.3, Head Acoustics, Herzogenrath, Germany) on a swivel chair while he recorded the in-ear signals on the HATS. Successively, the reconstructed signals were compared with the recordings in terms of the interaural cues, which showed that the motion capture and the sound card were in synchrony. The background noise from the projection was 32 dBA at the place of the participant (c.f. masker noise 70 dB SPL). The experimental scripts were written and run in MATLAB (v9.5.0, Mathworks, Natick, MA, USA) and Python (v3.6).

For this pilot study, we recruited three young participants without hearing problems; their hearing was assessed by a standard pure-tone audiometry procedure using a calibrated audiometer (MADSEN Astera², type 1066, Natus Medical Denmark Ap, Denmark). All thresholds were equal or below 20 dB HL. All participants took part in the previous study and provided written informed consent. The ethical committee of the Technical University of Munich approved the procedures (65/18S).

3. RESULTS

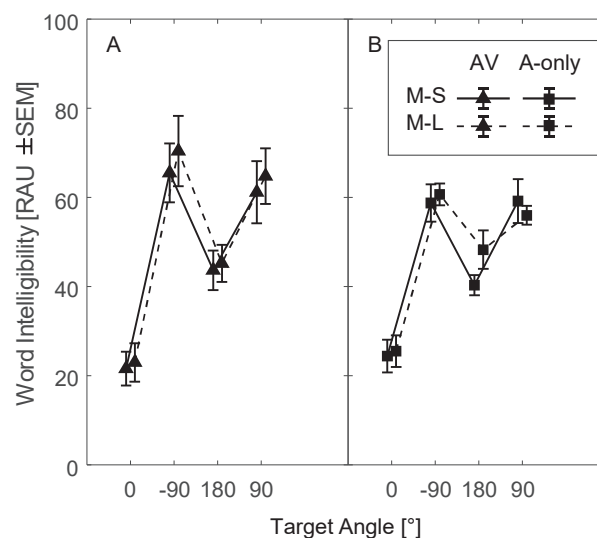


Figure 1. Speech intelligibility of the moving sound condition (M-S) of the current experiment and performance in the previous experiment – moving listener condition (M-L). Panel A show data of the AV condition, panel B show data of the A-only condition.

Figure 1 shows word intelligibility averaged across whole sentence for the M-S (current exp.) and the M-L (previous exp.) conditions. The x-axis shows the relative lateral angle of the target with respect to the masker. The two panels show performance in in the AV and A-only conditions for each of the target positions. The word intelligibility was computed as a mean intelligibility of each individual word of the target sentence. There is no systematic difference between the two conditions. The difference of the RAU values of two conditions was submitted to two-way repeated measures ANOVA with factors of target angle and visual condition. Neither factor nor interaction was significant. This indicates that non-auditory self-motion cues did not affect performance in this test.

4. DISCUSSION

The experiment tested speech intelligibility in which the participants were standing still and received the same auditory inputs as they received during active self-rotation

in the previous experiment. The data of this pilot study suggest that non-auditory self-rotation cues did not positively or negatively influence speech intelligibility. However, due to a small sample of participants the result needs to be confirmed with a larger set of participants and the conclusions cannot be drawn from this pilot experiment.

This experiment provides a methodological step for speech perception research since it has not been clear how active self-rotation contributes to speech intelligibility. In a previous study, we saw that speech intelligibility was both improved and degraded, depending on the target angle, in the moving condition with respect to a static baseline. The present experiment shows that the acoustic cues from sound motion are sufficient to explain speech intelligibility during self-rotation. Therefore, artificially created sound rotation trajectories, as for instance has been previously done [9], may represent the acoustic situation as during active self-movement.

In addition, the results indicate that the difference between the AV and A-only condition arise only from the sound motion, not due to the presence of the visual cue. In the previous experiment, the visual cue was synchronously presented with the target which could provide some benefit for the later parts of the sentence since previous research suggested that the prior knowledge about target position improves intelligibility [10].

Taken together, the data for the first three participants suggest that speech intelligibility during self-rotation in the previous experiment was likely related to sound rotation and was not influenced by non-acoustic self-rotation cues. The data also suggest that the presence of visual cues changed the behavior, which led to a change in intelligibility, but the mere knowledge about sound location did not strongly affect intelligibility. However, the visual cue in the previous experiment was synchronous and the prior information about sound position was limited. To confirm these findings we need to collect more data.

5. ACKNOWLEDGEMENTS

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

6. REFERENCES

- [1] J. F. Culling, M. L. Hawley, and R. Y. Litovsky, “The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources,” *J Acoust Soc Am*, vol. 116, no. 2, pp. 1057–65, Aug. 2004.
- [2] 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.
- [3] L. 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.
- [4] L. Hládek and B. U. Seeber, “The effect of self-orienting on speech perception in an acoustically complex audiovisual scene,” in *Fortschritte der Akustik -- DAGA '20*, 2020, pp. 91–94.
- [5] 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, doi: 10.1016/j.heares.2010.12.005.
- [6] J. Borish, “Extension of the image model to arbitrary polyhedra,” *J. Acoust. Soc. Am.*, vol. 75, no. 6, pp. 1827–1836, Jun. 1984, doi: 10.1121/1.390983.
- [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] P. Stitt, S. Bertet, and M. van Walstijn, “Extended Energy Vector Prediction of Ambisonically Reproduced Image Direction at Off-Center Listening Positions,” *J. Audio Eng. Soc.*, vol. 64, no. 5, pp. 299–310, May 2016, doi: 10.17743/jaes.2016.0008.
- [9] M. M. E. Hendrikse, G. Grimm, and V. Hohmann, “Evaluation of the Influence of Head Movement on Hearing Aid Algorithm Performance Using Acoustic Simulations,” *Trends Hear.*, vol. 24, p. 233121652091668, Jan. 2020, doi: 10.1177/2331216520916682.
- [10] G. Kidd, T. L. Arbogast, C. R. Mason, and F. J. Gallun, “The advantage of knowing where to listen,” *J. Acoust. Soc. Am.*, vol. 118, no. 6, pp. 3804–3815, Dec. 2005, doi: 10.1121/1.2109187.