DOI: 10.61782/fa.2023.0127

forumacusticum 2023

EVALUATION OF BEHAVIOR-CONTROLLED HEARING DEVICES IN THE LAB USING INTERACTIVE TURN-TAKING CONVERSATIONS

Giso Grimm[1]* **Hendrik Kayser**[1]
Angelika Kothe[1] **Volker Hohmann**[1]
[1] Department of Medical Physics and Acoustics,
Carl von Ossietzky University of Oldenburg, Germany

ABSTRACT

In complex acoustic environments, spatial filtering offers a great potential for improving speech intelligibility with hearing devices. However, as the performance increases, knowledge of the user's personal listening preferences and identification of the attended and ignored sources becomes critical. In this approach, the hearing-device user's gaze and head movement behavior is set into the context of the current communication situation. Ideally, this would include knowledge of source positions, source types, but potentially also high-level features. Here, the context is provided by acoustic direction-of-arrival estimation. This way, the attended source can be identified from a mixture of sources in an audiovisual scene. Since the algorithm is driven by behavior, special care must be taken during its evaluation to ensure that user behavior is as ecologically valid as possible. This is achieved by establishing an interactive turn-taking conversation in virtual reality by representing remote interlocutors through their real-time animated avatars. The system provides access to isolated speech and noise signals, which allows for an instrumental evaluation, even in natural interactive turn-taking conversations. In addition, conversation success was analyzed. Results show that the proposed algorithm can provide a benefit in terms of SNR as well as conversational success.

Keywords: *hearing device, self motion, hearing device evaluation, virtual reality*

1. INTRODUCTION

In complex listening situations, the greatest hearing device benefit can typically be achieved with directional filters. The problem is that as the selectivity of the directional filters increases, the potential benefit increases. On the other hand, the dependence on a particular motion behavior also increases.

Behavior-dependent hearing devices are difficult to evaluate because they require subjects to behave as they would in natural conversational situations. However, movement behavior is often dictated by the experimental paradigm. The goal of this study is to improve ecological validity in the evaluation of such behavior-controlled hearing devices [1], while maintaining the sensitivity of the applied measures. The approach to increase ecological validity is to implement natural interactive communication between three speakers in a virtual reality (VR). This allows direct access to the clean speech signals and the noise signals independently so that signal-to-noise ratio (SNR) can be measured.

Ecological validity cannot usually be measured directly because it requires measurement in a natural situation without the influence of the measuring instrument and task. However, indicators of increased ecological validity can be used instead: Vertegaal et al. [2] were able to show that in a free conversation between four participants, the currently active speakers are looked at about 62% of the time, while the addressed participants are looked at about 40% of the time.

In this study, we compare SNR benefit from directional filtering as measured in interactive conversations in VR with the benefit in terms of speech reception thresholds (SRTs) measured using the Oldenburg sentence test (OLSA, [3]). Since the speech material of the sentence

**Corresponding author*: g.grimm@uol.de.

test is presented from the same positions and using the same room acoustic conditions, we hypothesize that the SRT benefit provided by directional filtering is comparable with SNR benefit measured during free interactive conversation. Furthermore, we analyze head movement and gaze behavior, for comparison with behavior described in literature [2]. We hypothesize that typical communication-related gaze behavior can be achieved also in this method of conversation via telepresence.

2. METHODS

2.1 General design and apparatus

The study involved a series of approximately 5-min-long triadic conversations between the participant and two confederates about casual topics. The participant was seated in a VR laboratory, the 'Gesture Lab' at the University of Oldenburg [4], surrounded by a loudspeaker array with a ring of 16 loudspeakers at ear level and 29 additional loudspeakers on a sphere. In front of the loudspeakers a cylindrical projection screen was mounted for a video projection with a field of view of 300 degrees.

The head movement of the participant was tracked by an optical tracking system (Qualisys Miqus M3). The position of a lightweight marker crown was tracked by six infrared cameras. Gaze direction relative to the head was estimated from electrooculography (EOG) data.

Two interlocutors were experimenters, who participated remotely in a separate room. They were represented in virtual reality by an avatars each. The experimenters were seated at distances and angles that corresponded to the virtual scene in order to elicit appropriate movements, and saw the participant via a transmitted camera image. The experimenters' audio was captured by an AKG headset. It was filtered and calibrated to produce a natural sounding audio signal in the lab. Head movements were captured with an inertial measurement unit (IMU) attached to the headset to animate the head movements of the avatars. The participant's audio was picked up with a microphone close to the mouth, for analysis and transmission to the experiments.

The 'ovbox' system was used to transmit the experimenters' audio and head movement to the virtual acoustic simulation [5]. This system allows low-delay exchange of audio signals, simulation of virtual acoustic environments, and low-delay exchange of arbitrary User Datagram Protocol (UDP)-based network data between multiple clients. With this system, delays between the microphone signal at one end and the loudspeaker signal at the other end have been achieved of about 20 ms, which is far below the delays achieved by typical videoconferencing systems. The delay between an actual head movement of the experimenter and the animated movement in the lab was 180 ms.

2.2 Virtual environment

The conversation was virtually set in a pub environment modeled after an existing location in Oldenburg [6]. Diffuse cafeteria background noise [7] was added with a sound level of 66 dB SPL C-weighted. The virtual acoustic environment was simulated using the Toolbox for Acoustic Scene Creation and Rendering (TASCAR) [8]. This software was also used for session management, data logging, and interfacing with all sensors and data streams.

The visual environment was rendered using the Blender game engine, version 2.79b [9]. Animation data was sent from TASCAR to Blender via OSC. The lips of the avatars were animated using a speech-based real-time lip simulation method [10].

2.3 Hearing device algorithms

All algorithms of this study were simulated in the playback system. For this purpose, the head orientation was measured and the direction-dependent gain was calculated in real time. Three different algorithms were realized. In the setting with the label *'none'*, no modifications to the playback signal were made. The *'dir1'* algorithm simulates the polar pattern of a cardioid microphone, with a null in the rear hemisphere, and 6 dB attenuation at +-90 degrees. In the third setting, labeled *'dir2'*, 6 dB attenuation was achieved at +-45 degrees. The attenuation of *'dir1'* and *'dir2'* was limited to -12 dB in the rear hemisphere. The steering direction of these virtual directional microphones was aligned with the head, as it would be the case for a fixed beamformer in a head-worn device.

2.4 Performance metrics

To assess the benefit of the different signal enhancement strategies, two measures were used in this study. The SRT was measured with the OLSA [3]. This matrix tests consists of syntactically correct nonsense-sentences of five words. We used the open version, i.e., the sentence was repeated by the participant and entered by one of the experimenters. Two lists of 20 sentences each were measured in a simultaneous interleaved measurement paradigm. OLSA sentences spoken by a male speaker were presented from the location of the male avatar, and OLSA sentences spoken by a female speaker were presented from the location of the female avatar. The speech level was adjusted to achieve 80% correct responses. The SRT benefit was calculated from the difference between the SRT in a signal enhancement condition and the SRT in the corresponding condition without signal enhancement.

To measure the SNR benefit in real time during conversation conditions, four additional virtual omnidirectional microphones were added to the virtual acoustic environment. Two of these microphones recorded only the target speech components in the scene, and the other two microphones recorded only the noise components. One of the two microphones in each signal set was subjected to the same virtual signal enhancement algorithm as the rendering system, controlled by the same head movement. In this way, these microphones recorded the sound signals that would have been present in a free field at the actual listening position, including the test participant's head movements, but without the head-shadow effect. The RMS levels of the microphone output signals were measured in 5.3 ms time windows. The SNR with and without processing was calculated by taking the difference between the logarithmic short-term levels of the microphone of the target signals and the microphone of the noise signals. To estimate the SNR benefit, only those time windows were taken in which at least one of the experimenters spoke and the subject did not speak. This selection was performed automatically based on the smoothed individual speech levels. The speech portions when the confederates spoke and the participant was listening were used for analysis of the SNR. This approach to determine the SNR benefit does not require a reference condition.

The direction of gaze relative to the head was extracted from the EOG signals by first removing the drift signal and then linearly mapping EOG voltage to gaze angle. Gaze direction in global coordinates was calculated by adding head rotation to gaze angle. A gaze toward the speaker was detected when the estimated gaze direction did not differ more than 15 degrees from the direction of the speaker.

2.5 Experimental conditions

Several experimental conditions were tested in the experiment. Each condition lasted approximately five minutes. The experiment always started with a training conversation condition, followed by a training SRT condition. All other conditions were randomized. The noise level L_n was varied so that it was either 40 dB SPL C-weighted or 66 dB SPL C-weighted. A list of all conditions can be found in Table 1. The total duration of the experiment was about 45 minutes.

Table 1. Overview of all test conditions.

	none	*dir1*	*dir2*
conv., L_n 40 dB	X		
conv., L_n 66 dB	X	X	X
SRT, L_n 66 dB	X	X	X

2.6 Participants

This study is a pilot test with five test participants. All of the participants were students from Oldenburg University, mean age 25.2 a (standard deviation 2.6 a), with self-reported normal hearing, and normal or corrected-to-normal vision.

3. RESULTS

To validate that the free conversation was roughly balanced across interlocutors, the speech contributions of the interlocutors in each condition were measured and related to the total duration of the respective condition. The data are shown in Figure 1. The median values range between 0.16 and 0.3, with a large overlap of the interquartile ranges in most conditions. There is a slight dominance of interlocutor 2, but the speech contributions are still reasonably balanced.

Absolute SRTs were at -4.4 dB (standard deviation 1.3 dB) without beamformer, at -9.9 dB (1.6 dB) with *dir1* and at -15.0 dB (1.1 dB) with *dir2*. The SNR during the listening phases in the interactive conversation was at 0.5 dB (0.7 dB) without beamformer, at -1.3 dB (2.5 dB) with *dir1*, and at -0.6 dB (0.7 dB) with *dir2*.

The algorithm benefit is shown in Figure 2. In the left panel, the benefit in SRT is shown. The median SRT benefit is about 5.5 dB for the *dir1* algorithm, and 10 dB for the *dir2* algorithm. In the right panel, the SNR benefit from the algorithm during free conversation is shown. Here, the median SNR benefit is 4.5 dB for *dir1* and 10 dB for *dir2*, which is comparable to the SRT benefit.

The benefit by the algorithm *dir1* is comparable to the benefit provided by real hearing devices. An artifact-free benefit of 10 dB is more than hearing devices can typically achieve.

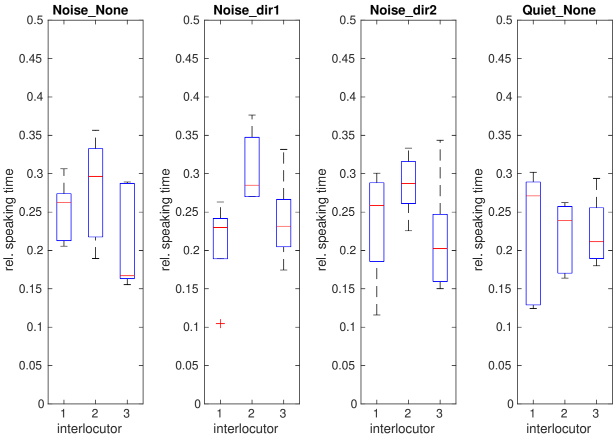

Figure 1. Talking time divided by the total duration of the conditions, in the four different conditions of free conversation (1 = test participant, 2,3 = experimenter).

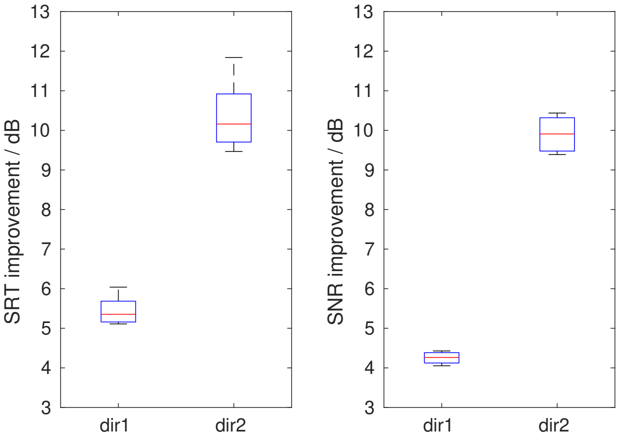

Figure 2. SRT benefit (left panel) and SNR benefit during free conversation (right panel). The benefit if *dir1* corresponds to benefit which can typically be achieved by real hearing devices.

Finally, the proportion of gaze direction to the active speaker is shown in Figure 3. According to [2] the proportion is about 62%. It can be noticed that in free conversation comparable values are reached in this experiment, but in the OLSA conditions some participants have much lower values, caused by a static forward gaze. This effect is visible by the large spread of the data in the OLSA conditions.

4. DISCUSSION

In this experiment, the SNR benefit during free interactive conversation was measured using a free-field microphone with the same directivity as that presented to the subject. The benefit is comparable whether measured by SRT or SNR. However, there is a slightly greater benefit of the algorithm in the SRT, which may be due to the simulated free-field microphone, which did not include effects such as head shadow, etc. Future studies will use simulated or measured HRTF to simulate the SNR benefit, which will likely provide more reliable results.

The SRTs measured with the OLSA speech matrix test were significantly lower than the SNRs during the active conversation, despite measuring the speech test at 80% word score convergence. Resulting in such low SNRs is a critical problem for evaluation of hearing devices, which often can not operate optimally in such unrealistic low SNRs.

The benefit of real hearing aid signal enhancement algorithms can also be evaluated by using hearing aid related transfer functions (HARTF) instead of HRTF. In this case, the two instances of the algorithm must be run simultaneously, which allows the SNR after processing to be estimated using the method of Hagerman and Olofson [11]. Again, the virtual receiver that generates the input signals to the signal enhancement algorithms would contain all the movement behavior applied by the test subjects. Thus, the proposed approach is not limited to the algorithms simulated in the rendering system.

5. CONCLUSIONS

By conducting interactive conversations in virtual reality using telepresence technology, we measured the benefit of signal enhancement algorithms in terms of SNR directly from ongoing triadic conversations. The benefit in terms of SNR was found to be similar to the benefit in terms of the SRT, while maintaining a more natural and communication-related head movement and gaze behavior. This suggests that the proposed method can lead to a more ecologically valid algorithm benefit estimate, yet under laboratory conditions and in a reasonable amount of time.

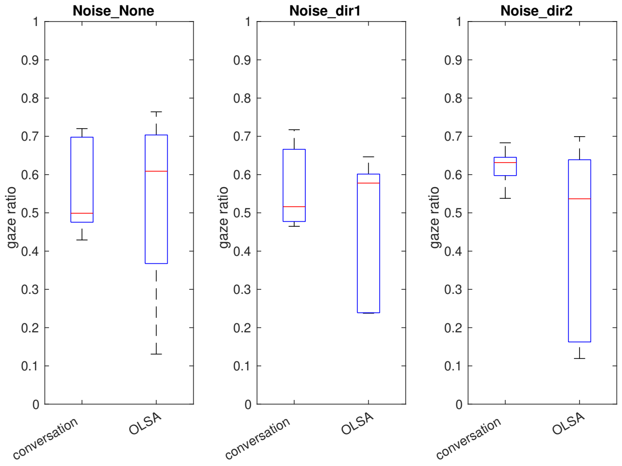

Figure 3. Amount of gaze to the active speaker while the participant was listening, in the different conditions.

6. ACKNOWLEDGMENTS

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 352015383 – SFB 1330 - B1.

7. REFERENCES

[1] G. Keidser, G. Naylor, D. S. Brungart, A. Caduff, J. Campos, S. Carlile, M. G. Carpenter, G. Grimm, V. Hohmann, I. Holube, S. Launer, T. Lunner, R. Mehra, F. Rapport, M. Slaney, and K. Smeds, "The quest for ecological validity in hearing science: What it is, why it matters, and how to advance it," *Ear & Hearing*, vol. 41, pp. 5S–19S, 11 2020.

[2] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt, "Eye gaze patterns in conversations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 301–308, ACM Press, Mar. 2001.

[3] K. C. Wagener and T. Brand, "Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: influence of measurement procedure and masking parameters," *International Journal of Audiology*, vol. 44, pp. 144–156, 2005.

[4] V. Hohmann, R. Paluch, M. Krueger, M. Meis, and G. Grimm, "The virtual reality lab: Realization and application of virtual sound environments," *Ear & Hearing*, vol. 41, pp. 31S–38S, 11 2020.

[5] G. Grimm, "ORLANDOviols Consort box (ovbox)." `https://ovbox.de/`, Apr. 2021.

[6] G. Grimm, M. Hendrikse, and V. Hohmann, "Pub environment." doi:10.5281/zenodo.5886987, Sept. 2021. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 352015383 – SFB 1330, Project B1.

[7] G. Grimm and V. Hohmann, "First order ambisonics field recordings for use in virtual acoustic environments in the context of audiology." doi:10.5281/ZENODO.3588303, 2019.

[8] G. Grimm, J. Luberadzka, and V. Hohmann, "A toolbox for rendering virtual acoustic environments in the context of audiology," *Acta Acustica united with Acustica*, vol. 105, pp. 566–578, 5 2019.

[9] T. Roosendaal and C. Wartmann, *The Official Blender Game Kit: Interactive 3d for Artists*. No Starch Press, 2003.

[10] G. Llorach, A. Evans, J. Blat, G. Grimm, and V. Hohmann, "Web-based live speech-driven lipsync," in *2016 8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*, (Barcelona, Spain), IEEE, 2016.

[11] B. Hagerman and Åke Olofsson, "A method to measure the effect of noise reduction algorithms using simultaneous speech and noise," *Acta Acustica united with Acustica*, vol. 90, pp. 356–361, 2004.