A pilot study of realistic communication in a simulated restaurant acoustic environment

Julia Schütze[1,a], Manuj Yadav[2,5], Maartje Hendrikse[1,4], Birger Kollmeier[1,2,3,4], Kirsten C. Wagener[2,3,4]

[1] *CvO Universität Oldenburg, Deptm. f. med. Physik u. Akustik*
[2] *HörTech gGmbH, 26129 Oldenburg*
[3] *Hörzentrum GmbH Oldenburg, 26129 Oldenburg*
[4] *Exzellenzcluster 'Hearing4all', 26111 Oldenburg*
[5] *Institut für Technische Akustik, RWTH Aachen Universität, 52074, Aachen*
[a] *julia.schuetze@uni-oldenburg.de (corresponding author)*

Introduction

When the benefit of hearing aids is measured in the field in everyday life, the outcomes are supposed to be ecologically much more valid than in the laboratory. However, they are difficult to reproduce reliably, as it is hardly possible to determine all variables systematically. In the laboratory, on the other hand, the everyday communication situations of hearing-impaired people are only marginally represented, e.g., by measuring speech recognition using standardized speech material in quiet and in noise. The patient is asked to repeat the presented word or sentence, which gives reliable information about her or his speech recognition [1]. But these traditional measurements in the lab do not represent real-life situations [2, 3]. Therefore, not only speech recognition but also the benefit of hearing aids determined in the laboratory differ greatly from everyday communication, which also requires understanding of the content. For this reason, an ecologically valid communication task (details in methods) is evaluated here to determine the ability to communicate in comparison with speech recognition measurements.

Methods

Communication task

The interactive Diapix communication task [4], translated into German from the original English material, was chosen in this study to facilitate a highly natural conversation between a pair of participants in controlled laboratory conditions: Each participant is given an image on an A4 sheet of paper, where the images are the same except for twelve differences (Figure 1) between them; each quadrant having three differences. The participants are then asked to verbally communicate to 'spot the differences' between the pair of images. The ensuing conversation closely follows a question-answer form, where each participant tries to find out whether the objects in their image match with those in the other participant's image. This continues until all the differences were found or ends if the trial duration is exceeded. To distribute the talking approximately equally between participants, they were instructed to take turns leading the conversation in each quadrant. Hereby, one participant would first describe the objects in the quadrant, typically involving more of the talking, and the other participant would respond accordingly with a typically shorter duration of speech. Overall, this task allows elicitin spontaneous speech from the participants in an ecologically valid way. This method has been used previously in studies with normal and hearing-impaired participants [5, 6]. A similar method that uses a different task was proposed by [7].

Participants

Ten normal-hearing volunteers (eight females, two males; four younger than 30 years grouped in 2 pairs, six older than 60 years grouped into 3 pairs, with mean ages of 23,3 years and 64,7 years, respectively; hourly payment) participated in the measurements. They had hearing thresholds for pure tones better than 25 dB HL at the frequencies 125, 250, 500, 750, 1000, 1500, 2000, 3000, 4000, and 6000 Hz for the better ear. Two participants of the younger group took part in the questionnaire, and two participants of the older group only took part in the Test questionnaire.

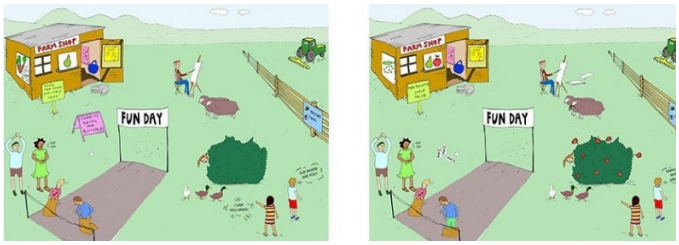

Figure 1: DiapixUK communication task image material Farm scene 1 (version A on left; version B on right). Twelve differences can be found between both images [4].

Experimental design

The virtual acoustic environment that was simulated for the participants represents a restaurant scene based on an actual restaurant (OLs Brauhaus, Oldenburg). The acoustic scene consisted of noise from several conversations in German (mostly babble-like) at varying distances from the simulated listening position, as well as music, clinging glasses, noise of chairs sliding on the floor and noise from the bar and kitchen. The acoustic scene was modelled using Toolbox for Acoustic Scene Creation And Rendering (TASCAR) [8], which uses first-order image source model per virtual source in the model to simulate the early reflections. Following these early reflections, the late reverberation was based on the room impulse responses (truncated to remove early reflections) that were recorded in the OLs Brauhaus in Oldenburg. The measurement apparatus included a CoreSound Tetramic microphone positioned at the listener

position, a omnidirectional loudspeaker positioned at several locations in the room, a computer and a sound interface. A logarithmic frequency sweep [9] was used as a measurement signal, the average over 10 repetitions was taken according to ISO 3382-2 [10]. The measured reverberation time was $T_{60} = 753\,ms$.

The acoustic scene was presented using a horizontal array of 16 loudspeakers (Genelec 8020B) arranged in a 3 m circle in a sound-treated testing room. The reproduction method used was 7th order Ambisonics panning with max-rE decoding for the virtual sources and first-order Ambisonics for the diffuse sources and late reverberation (all within TASCAR).

For the experiments, pure-tone audiometry and speech recognition threshold (SRT) measurements using the Oldenburg sentence test (OLSA) [11] were performed for each participant separately. The OLSA was used both in an adaptive procedure [12] and with a fixed speech presentation level. The fixed speech level represents a talker who cannot speak louder than a certain level, therefore understanding a conversation in a noisy environment becomes difficult. A talker who is able to increase the level of speech is represented by measuring the OLSA adaptively.

This was followed by the communication task for a pair of two participants. The participants were seated at 1.5 m distance from each other and 0.75 m each from the center of the loudspeaker circle. The session started with a short orientation and a practice trial (data not used) where the participants solved a communication task in quiet (no virtual scene reproduction) and filled in a subjective questionnaire. This was followed by five experimental trials, each with a different sound reproduction level, each using a different pair of randomly selected Diapix image. Both the practice and the experimental trials lasted three minutes each, followed by filling in the subjective questionnaire in silence. The subjective questionnaire included the following items:

Effort of speaking, Effort of understanding the other participant and Estimating the other participant's speaking effort. Each of these items were rated on separate ordinal scales from 0-12, with descriptors on even numbers: 0: "no effort", 2: "very little effort", 4: "little effort", 6: "moderate effort", 8: "considerable effort", 10: "much effort" and 12: "extreme effort" (the questionnaire was in German).

The virtual restaurant scene was calibrated to be reproduced at 50, 60, 70 and 80 dB A-weighted equivalent sound pressure levels (SPLs) along with the silence condition (~25 dB(A)) at the sitting positions for each participant for both the communication task and the adaptively measured OLSA. For the OLSA measured with a fixed speech presentation level the restaurant scene was presented at 50, 60, 65, 70 dB(A). The order of different sound presentation levels was randomized in each experiment. Presentation levels were chosen based on two studies where typical background levels in occupied restaurants were investigated [13, 14]. The speech per participant was recorded using a headset microphone each (DPA 4288), using the software Adobe Audition running on a desktop computer with a RME Babyface Pro audio interface at a sampling rate of 44100 Hz and 24-bit resolution.

Speech intelligibly and communication task were measured as Test and ReTest on two different days each.

Analysis

In this study, both the recorded speech and the subjective questionnaire ratings were analyzed from the communication task sessions as well as the speech intelligibility data. All data were analyzed separately for the younger and older groups. The variable performance, i.e., the number of differences found, and the variables number of words and number of conversation swaps were measured from the recorded speech for a content-based analysis. For the former, all recordings were annotated using the computer software Praat [15], and processed further in Matlab® (Mathworks, USA).

Results

In Figure 2 the mean speech recognition rate measured with the OLSA and fixed speech presentation level is shown. Figure 3 shows the mean SRT results of the adaptively measured OLSA. Both, Test and ReTest data are shown.

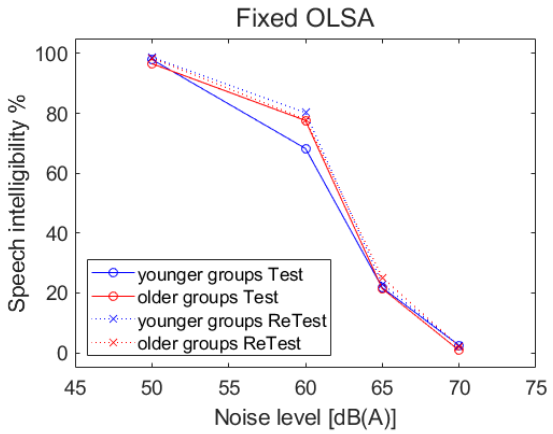

Figure 2: Speech recognition rate (OLSA, fixed speech presentation level) in Test (solid lines) and ReTest (dotted lines) was measured at noise presentation levels of 50, 60, 65 and 70 dB(A), respectively. Mean across the 2 younger groups are shown in blue, mean across the 3 older groups in red.

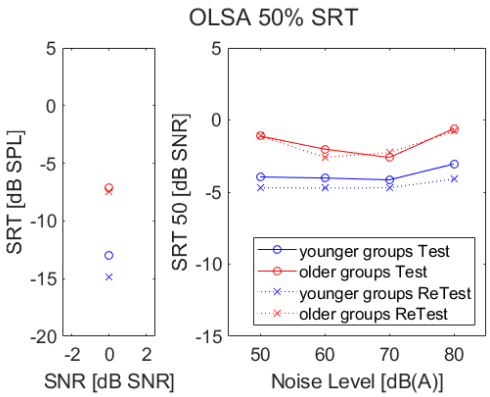

Figure 3: Mean Test and ReTest SRT results of the adaptively measured OLSA. Same representations as in Figure 2.

For the communication task, different solution strategies could be observed between the different trials and groups:

Most groups started to describe the image in the upper left corner proceeding clockwise, while other approaches include starting in the middle or at an especially striking object.

Figure 4 shows how many differences the pair of participants were able to find in the given time, where the younger pairs performed better than the elderly pairs. No clear trend can be seen in the comparison of Test and ReTest sessions. Furthermore, performance seems to remain stable across the different noise presentation levels as well as in quiet.

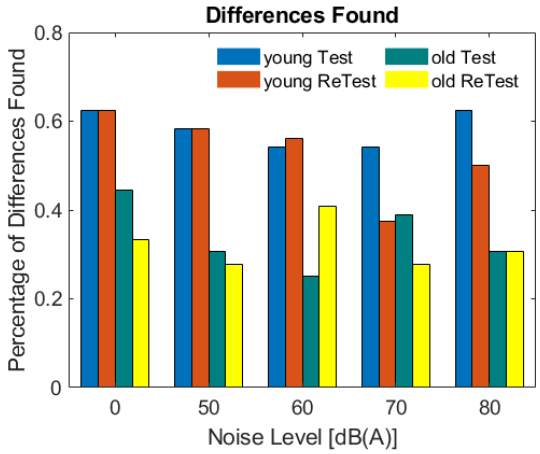

Figure 4: Percentage of differences found in each condition.

The younger pairs also used more words overall compared to the elderly pairs (Figure 5), with no clear trend over the various noise levels and relatively stable results for Test and ReTest sessions.

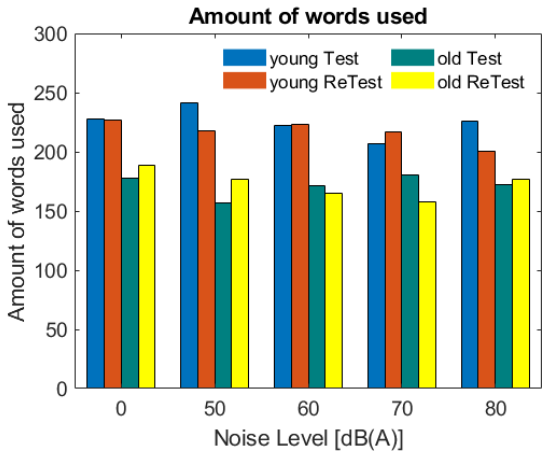

Figure 5: Amount of words used in each condition.

A third parameter which shows a similar lack of trend is the number of conversation swaps, as shown in Figure 6. Here, the term conversation swap means that the role of describing the image changes to the second participant. The first participant is then in the position to draw attention to a difference as soon as she or he notices a difference between his picture and the description by her or his partner.

In Figure 7, a trend for effort rising in all three items with higher presentation levels can be seen. The rating of effort does not reach the end of the scale, even for the higher levels.

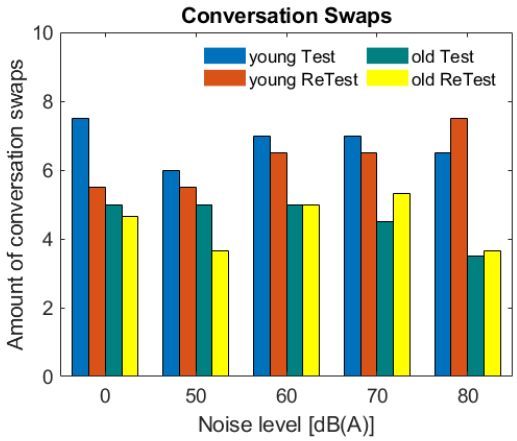

Figure 6: Amount of conversation swaps.

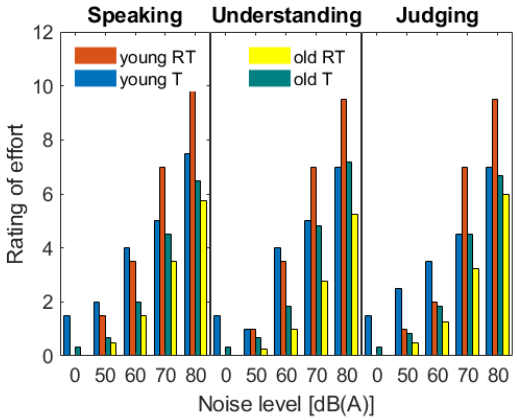

Figure 7: Questionnaire ratings on the effort of speaking and understanding as well as estimation of the partners effort.

Discussion

As the performance differs between the younger and older groups (see Figure 4), for these groups the difference appears to be consistent across all sound presentation levels. The trend in effort rises with increasing noise presentation level (see Figure 7). The differences between younger and older groups will be further investigated in the future by setting the performance in relation to cognitive measures.

While the content-based measures seem mostly unchanged with varying sound presentation levels, the subjective effort ratings in the questionnaire clearly increase with increasing noise level. The effort for speaking and understanding appears to be larger with higher levels. Also, the estimated effort of the partner seems to be aligned with the own effort in speaking and understanding. Since it appears that normal hearing participants can compensate higher noise presentation levels through effort, a hypothesis consistent with the current results is that the compensation effort will be different in hearing impaired participants. Therefore, the experiment will be repeated with hearing impaired participants. Here, it will be investigated, if the content-based measures in the communication task will break down with higher sound presentation levels, or if an increase in effort is sufficient to achieve the same results.

In the communication task, a trial length of three minutes was chosen to avoid fatigue and to avoid the participants

finding all the differences before the three minutes are up. The mean length of recordings in [4] was 7.7 min for each trial, which referred to 2.6 min of talking of each participant, when pauses and breaks are excluded. The mean amount of words elicited per participant was 613 words. In the three minutes trial in our communication task, the mean number of words was 191, which would correspond to 490 in 7.7min. The mean age of participants in [4] is 22,6 years, which might explain the difference to our mean amount of words, as the older groups tend to use less words. When comparing the mean amount of words only for the younger groups, i.e., 570 words, to the results of [4], the difference is lower. An additional explanation for the difference is provided by the average word length, which in German is 1.7 syllables per word, whereas in English it is 1.4 syllables [16].

When comparing OLSA and content-based measures, one would expect that when participants cannot talk louder, the resulting content-based measures should exhibit a decline. In fact, there is a small gap in the content-based measure between the younger and older groups (Figure 4), which is consistent with the adaptively measured OLSA (Figure 3).

Overall, this study shows the feasibility of using communication tasks that are more representative of natural talking between a pair of people in varying sound presentation levels. However, more work is needed to determine whether the trends seen here are valid for more fine-grained acoustic-phonetic features that can be extracted from the speech recordings (in progress), and for different groups including those with hearing impairment.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 352015383 – SFB 1330 C4.
Supported by “Innovation network for integrated, binaural hearing system technology (VIBHear)” and SFB1330-B1.

References

[1] Wolters, Florian, et al. "Common sound scenarios: A context-driven categorization of everyday sound environments for application in hearing-device research." *Journal of the American Academy of Audiology* 27.7 (2016): 527-540.

[2] Bentler, R.A., 2005. Effectiveness of directional microphones and noise reduction schemes in hearing aids: A systematic review of the evidence. Journal of the American Academy of Audiology, 16(7), pp.473-484.

[3] Cord, M.T., Surr, R.K., Walden, B.E. and Dyrlund, O., 2004. Relationship between laboratory measures of directional advantage and everyday success with directional microphone hearing aids. Journal of the American Academy of Audiology, 15(5), pp.353-364.

[4] Baker, R., Hazan, V. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behav Res* **43,** 761–770 (2011). https://doi.org/10.3758/s13428-011-0075-y

[5] Hazan, Valerie, et al. "Clear speech adaptations in spontaneous speech produced by young and older adults." *The Journal of the Acoustical Society of America* 144.3 (2012): 1331-1346.

[6] Knoll, Monja Angelika, Melissa Johnstone, and Charlene Blakely. "Can you hear me? Acoustic modifications in speech directed to foreigners and hearing-impaired people." Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[7] Beechey, Timothy, Jörg M. Buchholz, and Gitte Keidser. "Eliciting naturalistic conversations: a method for assessing communication ability, subjective experience, and the impacts of noise and hearing impairment." Journal of Speech, Language, and Hearing Research 62.2 (2019): 470-484.

[8] Grimm, Giso; Luberadzka, Joanna; Hohmann, Volker. A Toolbox for Rendering Virtual Acoustic Environments in the Context of Audiology. Acta Acustica united with Acustica, Volume 105, Number 3, May/June 2019, pp. 566-578(13), https://doi.org/10.3813/AAA.919337

[9] Farina, A., Bellini, A., & Armelloni, E. (2001). Non-linear convolution: a new approach for the auralization of distorting systems. *110th Convention of the Audio Engineering Society*. Amsterdam, The Netherlands.

[10] DIN, ENISO. "3382-2: Akustik–Messung von Parametern der Raumakustik–Teil 2: Nachhallzeit in gewöhnlichen Räumen (ISO 3382-2: 2008)." *Deutsche Fassung EN ISO* (2008): 3382-2.

[11] Wagener K, Brand T, Kollmeier B (1999) Entwicklung und Evaluation eines Satztests für die deutsche Sprache I-III: Design, Optimierung und Evaluation des Oldenburger Satztests. ZfA 38 (1-3), 4-15, 44-56, 86-95.

[12] Brand, Thomas, and Birger Kollmeier. "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests." *The Journal of the Acoustical Society of America* 111.6 (2002): 2801-2810.

[13] Hodgson, Murray, Gavin Steininger, and Zohreh Razavi. "Measurement and prediction of speech and noise levels and the Lombard effect in eating establishments." *The Journal of the Acoustical Society of America* 121.4 (2007): 2023-2033.

[14] Lebo, Charles P., et al. "Restaurant noise, hearing loss, and hearing aids." *Western Journal of Medicine* 161.1 (1994): 45.

[15] Boersma, Paul & Weenink, David (2020). Praat: doing phonetics by computer [Computer program], URL: http://www.praat.org/

[16] Fucks, Wilhelm. Nach allen Regeln der Kunst. Deutsche Verlags-Anstalt, 1968.