The Concurrent OLSA Test: A Method for Speech Recognition in Multi-talker Situations at Fixed SNR

Jan Heeren[1,2] (iD), Theresa Nuesse[2,3] (iD), Matthias Latzel[4],
Inga Holube[2,3], Volker Hohmann[1,2,5], Kirsten C. Wagener[1,2] (iD)
and Michael Schulte[1,2]

Abstract

A multi-talker paradigm is introduced that uses different attentional processes to adjust speech-recognition scores with the goal of conducting measurements at high signal-to-noise ratios (SNR). The basic idea is to simulate a group conversation with three talkers. Talkers alternately speak sentences of the German matrix test OLSA. Each time a sentence begins with the name "Kerstin" (call sign), the participant is addressed and instructed to repeat the last words of all sentences from that talker, until another talker begins a sentence with "Kerstin". The alternation of the talkers is implemented with an adjustable overlap time that causes an overlap between the call sign "Kerstin" and the target words to be repeated. Thus, the two tasks of detecting "Kerstin" and repeating target words are to be done at the same time. The paradigm was tested with 22 young normal-hearing participants (YNH) for three overlap times (0.6 s, 0.8 s, 1.0 s). Results for these overlap times show significant differences, with median target word recognition scores of 88%, 82%, and 77%, respectively (including call-sign and dual-task effects). A comparison of the dual task with the corresponding single tasks suggests that the observed effects reflect an increased cognitive load.

Keywords

speech recognition, speech-in-speech, cocktail party, attention, concurrent speech, OLSA

Received 30 September 2021; Revised received 7 May 2022; accepted 9 May 2022

Introduction

In clinical practice, the treatment of hearing impairment follows national policies prescribing diagnostics using single-talker speech recognition tests (e.g., American Speech-Language-Hearing Association, 1988; Gemeinsamer Bundesausschuss, 2021). Such tests typically measure speech recognition in quiet or in noise as a function of sound-pressure level or signal-to-noise ratio (SNR) and deliver reliable results for idealized listening conditions (e.g., Killion et al., 2004; Nilsson et al., 1994; Wagener et al., 1999). However, standard procedures discard several factors forming the complexity of real-life communication situations, such as competing talkers, turn-taking, and attention. Thus, patients may obtain good results in speech-recognition tests and yet have difficulties in real-life conversations (Bronkhorst, 2000; Desjardins, 2011; Humes et al., 2006; Singh et al., 2009).

A recent consensus paper defined that "in hearing science, ecological validity refers to the degree to which research findings reflect real-life hearing-related function, activity, or participation" (Keidser et al., 2020, p. 7S). Although the term "ecological validity" is contentious (see e.g., Beechey, 2022; Hammond, 1998; Holleman et al., 2020), it is used in this contribution in agreement with Keidser et al. (2020) and Wilhelm and Perrez (2013) in the sense of ecological

[1]Hörzentrum Oldenburg gGmbH, Oldenburg, Germany
[2]Cluster of Excellence Hearing4All, Oldenburg, Germany
[3]Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, Oldenburg, Germany
[4]Sonova AG, Stäfa, Switzerland
[5]Auditory Signal Processing, Department of Medical Physics and Acoustics, University of Oldenburg, Oldenburg, Germany

Corresponding Author:
Jan Heeren, Hörzentrum Oldenburg gGmbH, Marie-Curie-Straße 2, 26129 Oldenburg, Germany.
Email: heeren@hz-ol.de

generalizability, for which a representative design of the experiment is necessary (Dhami et al., 2004). To assess speech recognition that is more representative of everyday listening and incorporating multi-talker situations, approaches based on simultaneous speech (for a review see Bronkhorst, 2015) or alternating talkers (e.g., Best et al., 2016; Schneider et al., 2000; also see: Kitterick et al., 2010) have been proposed.

The approaches used for simultaneous speech range from comparisons of masker conditions (e.g., Arbogast et al., 2005; Brungart, 2001; Brungart et al., 2001; Francart et al., 2011; Holube, 2011; Kidd et al., 2016; Schneider et al., 2007) to the switching of target talkers (Best et al., 2008; Bolia et al., 2000; Lin & Carlile, 2015; Meister et al., 2018, 2020; Wood & Cowan, 1995), that includes turn-taking and other attentional processes and parameters. Interferer comparisons showed that speech-in-speech listening includes masking effects that cannot be explained energetically. Besides parameters like talker voices and SNR, it was found that meaningful speech maskers cause higher masking than meaningless speech maskers. Such masking-level differences that cannot be explained by energetic audibility are defined as informational masking (Arbogast et al., 2005; Bronkhorst, 2000; Brungart, 2001). For listening situations with one interfering talker, informational masking usually amounts to between 2 and 5 dB (Francart et al., 2011; Holube, 2011), but can be much higher under certain conditions (e.g., Brungart, 2001; Brungart et al., 2001; Kidd et al., 2016).

To examine the field of turn-taking and attention, so-called call-sign paradigms were designed that use simultaneous speech without a-priori knowledge of which talker is the target talker. Instead, test subjects have to detect a call sign that is spoken by one of the talkers. The coordinate-response measure corpus (CRM, Bolia et al., 2000; used by, e.g., Brungart, 2001; Brungart et al., 2001; Kidd et al., 2016) is a prominent example of such a call-sign test. It consists of sentences of the format "Ready (call sign), go to (color) (number) now", which are presented simultaneously by a specified number of talkers. The task is to recall the color-number combination (e.g., "white seven") said by the talker who spoke the call sign (e.g., "baron"). Shafiro and Gygi (2007) evaluated the call-sign-detection component in CRM and found that for configurations with 3–4 talkers (mixed sex), call-sign detection performance suffers more from the presence of the non-target talkers than color-number identification. This was interpreted as an effect of divided attention between the simultaneously presented call signs, while color-number identification is performed with selective auditory attention (Shafiro and Gygi, 2007). Selective attention is defined as the enhancement of the neural representation of an auditory object while the allocation of resources to competing objects is reduced (Shinn-Cunningham & Best, 2008). Humes et al. (2006) used a two-talker setting and tested elderly hearing-impaired

(EHI) and young normal-hearing participants (YNH). For the most ecologically valid listening conditions (different talker voices, binaural presentation), both EHI and YNH yielded color-number identification scores close to 100% if the target talker was announced in advance, reflecting negligible energetic masking effects. Keeping energetic factors fixed, both groups showed significant effects for call-sign based target-talker detection (approximately 10%-points decrease), and for a condition with a competing memory task (approximately 25%-points decrease in recognition score). Meister et al. (2018) investigated the influence of additional background noise in a similar memorize-and-recall task with two simultaneous sentences. At SNRs of +2 or +6 dB, word-recall scores of around 40% were measured for a certain word position, depending on the level of the background noise. This indicated a cognitive load induced by the background noise (Meister et al., 2018).

Brungart and Simpson (2007) investigated turn-taking factors in CRM by implementing target-talker changes between trials for 2–4 talkers and several spatial configurations. Target transitions either affected talker voices (type 1, fixed target position), target-talker positions (type 2, fixed target voice), or both (type 3). The main findings were that switching costs, which reflect decreases in correct responses due to target transitions, were minimal for two talkers and increased with increasing number of talkers, and that following the same voice with switching positions (transition type 2) is easier than the opposite (transition type 1). Meister et al. (2020) investigated this approach for German matrix sentences of the Oldenburg sentence test corpus (OLSA, Wagener et al., 1999), with results similar to those of Brungart and Simpson (2007). Studies on pauses and word rate in such paradigms showed that increasing gaps between speech tokens decreases switching costs, but when there are no switches, correct responses benefit from shorter gaps (Best et al., 2008). Best et al. (2008) concluded that "(…) when attention is sustained on one auditory object within a complex scene, attentional selectivity improves over time."

Studies of turn-taking using simultaneous speech have been criticized for a low ecological validity, because in real-life group conversations, talkers usually do not start simultaneously. The use of alternating talkers is a logical approach, but such designs have to deal with the fact that speech intelligibility is at ceiling if performed at realistic SNR (Bronkhorst, 2000; for realistic SNR see: Mansour et al., 2021; Smeds et al., 2015; Weisser & Buchholz, 2019; Wu et al., 2018). Kitterick et al. (2010) compared different starting times of phrases across talkers. For one-talker-at-a-time conditions, turn-taking effects broke down and speech recognition was as in single-talker conditions. To achieve sensitivity for these effects in such listening conditions, dual-task test designs with a primary and a secondary task have been applied (Gagné et al., 2017). By allocating cognitive resources to a parallel task, recognition scores can be

brought below ceiling such that turn-taking effects are revealed for alternating talkers (Schneider et al., 2007). This approach is based on limited cognitive resources and assumes that unattended speech is semantically processed (Aydelott et al., 2015; Bentin et al., 1995). An example of a one-talker-at-a-time test with high ecological validity is the National Acoustic Laboratories Dynamic Conversations Test (NAL-DCT, Best et al., 2016, 2018). It uses pseudo real-life conversations with 1–3 talkers combined with a written comprehension task. The test is sensitive to speech comprehension under realistic communication conditions and can be used for comparisons of turn-taking versus single-talker conditions, but these comparisons do not reveal switching costs (Best et al., 2016, 2018).

For the purposes of clinical diagnostics of speech recognition including turn-taking, existing tests either lack ecological validity (simultaneous speech tests) or specificity (alternating talkers plus side task). In alternating-talker tasks, non-listening side tasks are required to achieve sensitivity, and the assessment of turn-taking effects needs variation in the number of talkers. In simultaneous speech tests, the number of talkers determines the test sensitivity for the desired SNRs, resulting in significant changes of the paradigm. The aim of the current study was to develop a multi-talker paradigm that allows for comparisons of turn-taking versus single-target conditions, and selective versus divided attention, without changing the listening conditions. Furthermore, the ecological validity of an alternating-talker design was targeted, and non-listening side tasks avoided, as they potentially demand for off-topic skills from participants, or the interplay between the tasks may lead to

Figure 1. Sketch of the loudspeaker setup, with a participant seated in the center, and showing the positions of the talkers.

variances. Therefore, the proposed method, Concurrent OLSA (CCOLSA), combines the concept of call-sign tests with an alternating-talker design with three talkers, using the speed of triggering the OLSA sentences to achieve sensitivity for ecologically valid SNR. This approach leads to overlapping sentences across talkers and thus spans the space between simultaneous speech and one-talker-at-a-time approaches.

This study introduces the CCOLSA test and delivers a proof of concept based on data for young normal-hearing listeners for different overlap times, which is the essential parameter in the test design.

Method

Participants

Twenty two young, normal-hearing participants (eleven of each sex) were assessed. Their age ranged from 19 to 27 years with a mean of 22. All participants were inexperienced regarding the OLSA and showed pure-tone thresholds of ≤25 dB HL at all audiometric test frequencies between 0.125 and 8 kHz. The pure-tone-average of 0.5, 1, 2, and 4 kHz of the better ear (PTA4) was −1.5 dB HL (SD: 3.1 dB HL). All participants gave their informed consent prior to inclusion in the study, and received a reimbursement of 12 € per hour. The experiment was approved by the ethics committee ("Kommission für Forschungsfolgenabschätzung und Ethik") of the Carl von Ossietzky University in Oldenburg, Germany (Drs. 34/2017).

Stimuli

The targeted design demands for sentences with unpredictable last words (as target) and the possibility to define a first word as call sign that occurs in combination with different last words. Therefore, the new method CCOLSA was developed based on the German matrix test OLSA. Each OLSA sentence consists of five words, following the structure name-verb-numeral-adjective-object. For each of these word classes, the matrix contains ten alternatives (Wagener & Brand, 2005). OLSA speech recordings of three talkers were used in this study: the standard male OLSA talker (M1, Wagener et al., 1999), the standard female OLSA talker (F1, Wagener et al., 2014), and a second male talker (M2) from the multi-lingual OLSA (Hochmuth et al., 2014). Initially, speech levels for the three corpora were calibrated to 68 dB SPL. To achieve equal intelligibility of the three corpora, speech-recognition thresholds (SRT, i.e. SNR for a speech-recognition score of 50%) of six well-trained participants (normal-hearing lab members, age: 20–30 years) were measured for each corpus in cafeteria noise during a pre-test (method: word scoring based procedure A1, with adaptive speech level; Brand & Kollmeier, 2002). Differences in the mean SRTs for the three talkers occurred
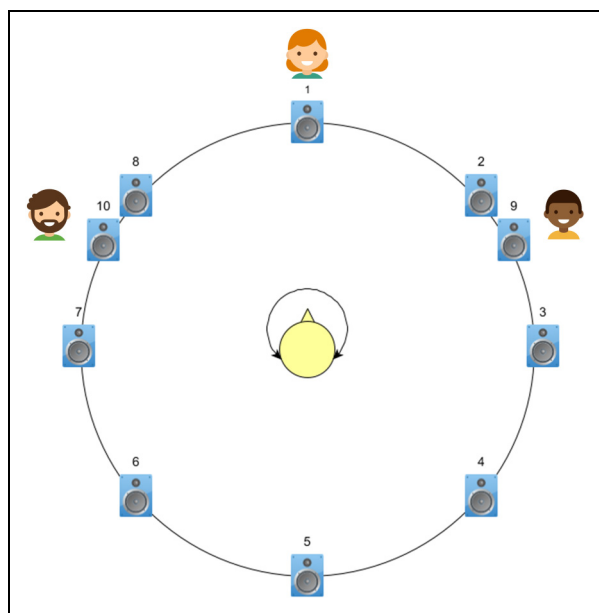
in the order of 1 dB, which is comparable with differences observed by Hochmuth et al. (2014). These were compensated for in the main experiment by calibrating the presentation level of talker M2 to 69.0 dB SPL and the level of talker F1 to 69.2 dB SPL (same level for all participants, no individual compensation). The speech was presented against a diffuse cafeteria noise (Grimm & Hohmann, 2019), which was also presented at 68 dB SPL. It was recorded in the cafeteria of the University of Oldenburg, campus Wechloy, in ambisonics B-format (first order). The character is highly diffuse and stationary with smaller, rather distant, fluctuating components.

Setup

Loudspeaker Array. Measurements were conducted in a lecture room with some acoustical optimizations (acoustic ceiling and curtains, moveable damping items) at Jade University of Applied Sciences in Oldenburg. The room had a volume of 262 m^3 with a T30 of 0.46 s. The loudspeaker setup is shown in Figure 1. Signals were presented via ten loudspeakers (Rokit 5, KRK Systems, Deerfield Beach, USA) positioned in a circle with a radius of 1.5 m. For the presentation of the diffuse background, eight loudspeakers were set up, equally spaced on the circle. The female speech was presented via the frontal speaker. For the presentation of the two male talkers, two loudspeakers were additionally placed at −60° and 60° azimuth. Participants were seated in the center of the loudspeaker setup on a rotatable office chair without armrests. The presentation of stimuli was controlled using the virtual acoustics toolbox TASCAR (Grimm et al., 2019). Speech was presented via single loudspeakers, and the cafeteria recording (recorded in 1st order ambisonics) was up-sampled and presented in 3rd order ambisonics. The CCOLSA test was implemented in MATLAB 2014 (MathWorks Inc., Natick, USA).

Head Tracking. The head orientation of the participants was monitored using a wireless head-tracking system based on a Bosch BNO055 integrated measurement unit (Bosch Sensortec GmbH, Reutlingen, Germany). The sensor was combined with a Teensy-LC microcontroller (PJRC, Sherwood, USA) and a HC-05 Bluetooth module. This set was boxed and attached to the participants' heads using a cap. The head-orientation data was buffered in the LabStreamingLayer (https://github.com/sccn/labstreaminglayer/) and recorded in TASCAR (Grimm et al., 2019) to store it with a time scale that matched that of the stimuli. In a pre-processing stage, the head-position data was smoothed (2-s-average). The analysis of the dual task (see below) used the head orientation at the time of a target-sentence presentation (2 s after the sentence onset). As participants do not necessarily turn their heads by the full amount of ±60° to face the loudspeaker of the male target talkers, the orientation was counted successful if the

2-s-average exceeded a threshold of ±25°. All head-tracker data was analyzed for calibration drift by comparing the cluster points in azimuth at the beginning, the middle, and the end of each dataset (30 s blocks). All deviations were <10°.

Paradigm

General Design. The CCOLSA paradigm follows the general idea of measuring speech recognition in multi-talker situations with talkers that call for the participant's attention. This was achieved by three talkers (M1 at −60°, F1 at 0°, and M2 at +60°) that speak OLSA sentences with a defined overlap between sentences from different talkers. One of the ten occurring names in the OLSA corpus, the name "Kerstin", was defined to be the call sign. In Figure 2, a sketch of a sample sequence of sentences is given.

Tasks. Three tasks were performed in this scenario, addressing three different listening modes:

Single task 1 (call-sign detection): "Name the number of the correct loudspeaker or point a finger in the direction if the word "Kerstin" is presented". In this single task, participants have to permanently monitor all talkers in order not to miss the first word of any sentence, because the call sign "Kerstin" might occur. The task was performed for 10 call signs, occurring every 10–13 s. The test was introduced by an initial phase of "non-Kerstin sentences" from random talkers with a duration of approximately 20 s.

Single task 2 (speech recognition for a fixed talker): "Repeat the last words of all sentences for a fixed talker." During this speech-recognition task, participants focus on the target talker, who does not change during the measurement. Sentences from the other two talkers can be ignored. In this task 2, OLSA test lists of 20 sentences were presented for the three talkers in separate measurements. The experimenter announced the target talker position before starting the measurement. During an initial phase of approximately 20 s, only the two non-target talkers presented sentences, so that the first sentence of the previously announced target talker was recognizable as the first target sentence without the need for a call sign. After completion of one test list, the new target talker was announced and the next speech recognition measurement started.

Dual task: "Repeat the last words of all sentences from the last talker who said "Kerstin". Always turn your body on a rotatable chair so that you face the talker you listen to." After a 20 s-initial phase of sentences from random talkers, the first sentence starting with the call sign "Kerstin" occurred (call-sign sentence). This indicated the first target talker for whom participants had to repeat the last words of all sentences (target sentences) until "Kerstin" occurred from another talker and the target talker changed. The dual task required simultaneous attention to the last word of the current target sentence (target word), and the first word of the following sentence (potentially a call sign), which
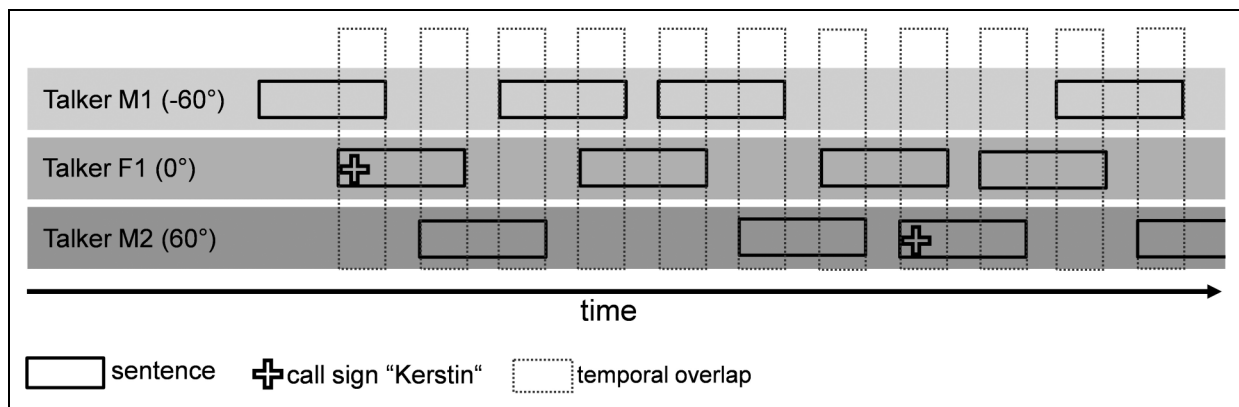
Figure 2. Sample sequence of sentences; sentences were timed with a defined temporal overlap and were presented randomly from the three talkers; some sentences started with the name "Kerstin", which was defined as the call sign.

appeared at the same time (see Figure 3). The overlap times tested were chosen so that the target word always completely overlapped the beginning of the following sentence from the next talker, and there were never more than two talkers active. A dual-task measurement contained 21 call-sign sentences (seven per talker), each of which was followed by 1–3 target sentences and one fake target sentence. In total, 60 target sentences were presented (20 per talker). To achieve a constantly high difficulty, all target-talker changes were introduced by a fake target sentence (S3 in Figure 3) from the previous target talker, meaning that a call-sign sentence was never presented next to a distractor sentence. Fake target sentences were not evaluated in the analysis of the results.

Responses to all tasks were given orally by the participants during the ongoing presentation, and were entered by the experimenter by pressing a button on a touch screen. The sentence presentations were continually controlled by the software, to adapt to the individual response times of the participants without presentation pauses. This was achieved by varying the number of distractor sentences (sentences from the non-target talkers) between two target sentences and did not affect the rate of presentations. Typically, two or three distractor sentences were presented during a response. Directly after a response, the next target sentence was presented. This design ensured that 1. the response for a past target word did not interfere with the perception of the next, which is mandatory for perception tests, and 2. that single and dual task listening conditions were equivalent during the most relevant stimulation periods (from responses to the following target words). The latter would not be the case if a fixed number of distractor sentences would be presented, because then quick responders would hear more presentations until target than slow responders. In the flow of targets and responses, target words occurred every 5 to 6 s and the experimenter was able to immediately judge if the participant could not respond. When this was the case, the experimenter was

instructed to quickly push the button for "target not recognized". The resulting button-press flow avoided indirect feedback through changes in the behavior of the experimenter, as would have occurred if misses had been judged by the software, e.g., by a countdown.

Outcome Variables. Outcomes of the two single tasks are the percentages of correct responses for call-sign detection and for speech recognition. In the dual task, the participants do not directly respond to the call signs (e.g., by specifying the number of the loudspeaker). To allow for an analysis of the call-sign detection during dual tasks, two criteria were considered: 1. correct response to target words between two call signs, and 2. the head orientation of the participant as measured with the head tracker. If one of these criteria was met, a call sign was counted as successfully detected. The impact of the two factors (responses and head tracking) on the results was analyzed separately.

The speech-recognition scores in the dual task were calculated as the percentage of correctly repeated target words and are given in the results as "speech recognition incl. missed-call-sign (CS) effects". To distinguish between the effects of missed call signs and failure in target-word recognition, a correction for misses caused by call-sign detection errors was applied, based on the two criteria given above. Therefore, the "net speech recognition" score was calculated as the percentage of correctly repeated target words while taking only target words for successfully detected call signs into account.

Conditions and Procedure

The CCOLSA paradigm was tested for three overlap times of the sentences: 0.6 s, 0.8 s, and 1.0 s (see Figures 2 and 3). Per overlap condition, the single tasks, i.e. call-sign detection and speech recognition for fixed-target talkers, were performed. Additionally, the dual task that included recognition of the last words for changing target talkers was conducted. Each
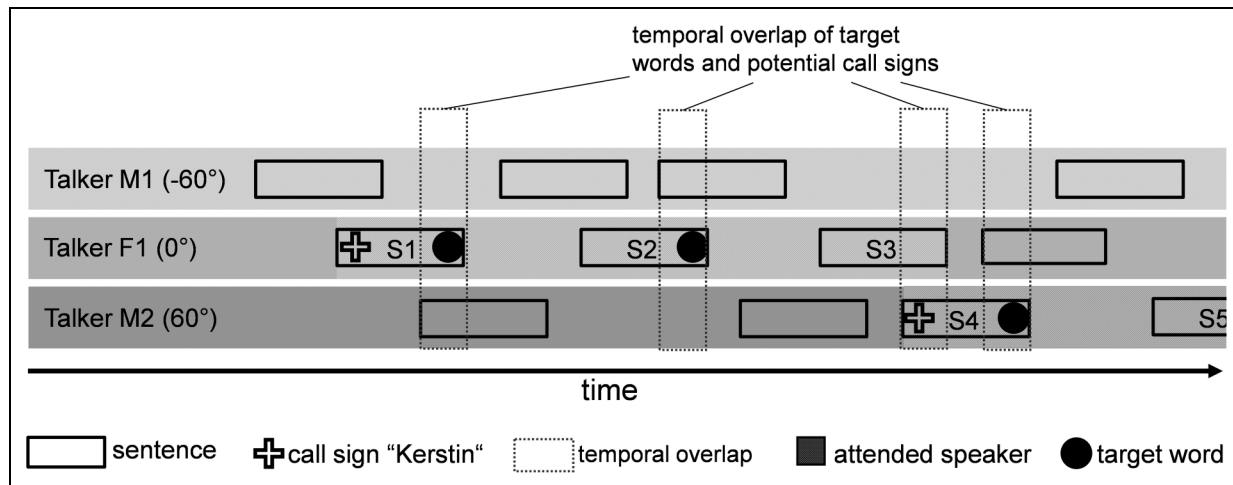
Figure 3. Attention pattern in the dual task for a sample sequence of sentences. Participants had to repeat the last words of all sentences for the last talker who spoke a sentence starting with the name "Kerstin" (call sign); target words and call signs occurred at the same time due to a defined overlap time between sentences across talkers.

task took approximately 5–6 min. In total, nine measurement tasks and two training tasks were conducted per participant. The procedure was split into two sessions of 2 h per participant, which were conducted within three weeks. In the first session, a general anamnesis, a pure-tone audiometry, and the single tasks (including a short training phase) were conducted. Call-sign detection was tested first and speech recognition (fixed talker) second. Additionally, two neurophysiological tests were carried out. The second session started after participants had performed another neuropsychological test sequence (approximately 20 min). Then, a longer training was conducted, including the two single- and the dual task at 0.6 s overlap, and the data collection for the dual task at the three overlaps, followed by another 20 min of neurophysiological tests. The overlap conditions were tested in randomized order for all tasks in both sessions. The analysis of the neurophysiological tests is beyond the scope of this contribution and is addressed in (Nuesse et al., in prep.).

## Results

Single and Dual Task

Figure 4 shows a boxplot of the results (medians, interquartile ranges, and whiskers) in the CCOLSA dual task and the two corresponding single tasks for the three overlap times of 0.6, 0.8, and 1.0 s. The left panel shows call-sign detection scores, the right panel shows target-word recognition scores. Target-word recognition scores for the dual task are given as a net value, i.e. considering only those target words with correct call-sign detection, and, additionally, as a total score including the missed-call-sign effects. Most of the data are not normally distributed (Shapiro-Wilk test,

$\alpha = 0.05$) and include ceiling effects. According to the recommendations of Šimkovic and Träuble (2019) for discrete data with ceiling effects, statistical analyzes were conducted using rank-based statistical tests. RAU transformation and model-based statistical analysis were considered, but this did not lead to any statistical or descriptive advantages; it was thus not applied.

Participants detected the call sign with a median score of $\geq$90% correct in both the single- and the dual-task condition (see Figure 4). No significant differences between the tasks were found. For the three overlaps in the single task, the median score in target-word recognition was 93–95%. In the dual task, this score dropped significantly to 81–90% (net scores), corresponding to a difference in speech recognition of 5–12%-points. Significance was tested using a Wilcoxon test (Bonferroni corrected; $p_{0.6s} = .015$, $r_{0.6s} = 0.60$, $T_{0.6s} = 213$; $p_{0.8s} < .001$, $r_{0.8s} = 0.86$, $T_{0.8s} = 250$, $p_{1.0s} < .001$, $r_{1.0s} = 0.81$, $T_{1.0s} = 244$). The median total target-word recognition scores, including the effect of missed call signs, ranged from 77–88% and were significantly lower than the net scores, corresponding to a difference in speech recognition in the single task of 7–16%-points (Wilcoxon test, Bonferroni corrected; $p_{0.6s} = .035$, $r_{0.6s} = 0.54$, $T_{0.6s} = 36$; $p_{0.8s} = .054$, $r_{0.8s} = 0.51$, $T_{0.8s} = 28$, $p_{1.0s} = .004$, $r_{1.0s} = 0.68$, $T_{1.0s} = 91$).

Overlap Dependency

The correct-response scores showed a dependency on the overlap time of the sentences, i.e. increasing the overlap time decreases the correct-response score (Figure 4). Significance was tested using Friedman tests (Bonferroni-corrected) and post-hoc Wilcoxon tests (Bonferroni-corrected) for the Friedman tests with results showing significant differences
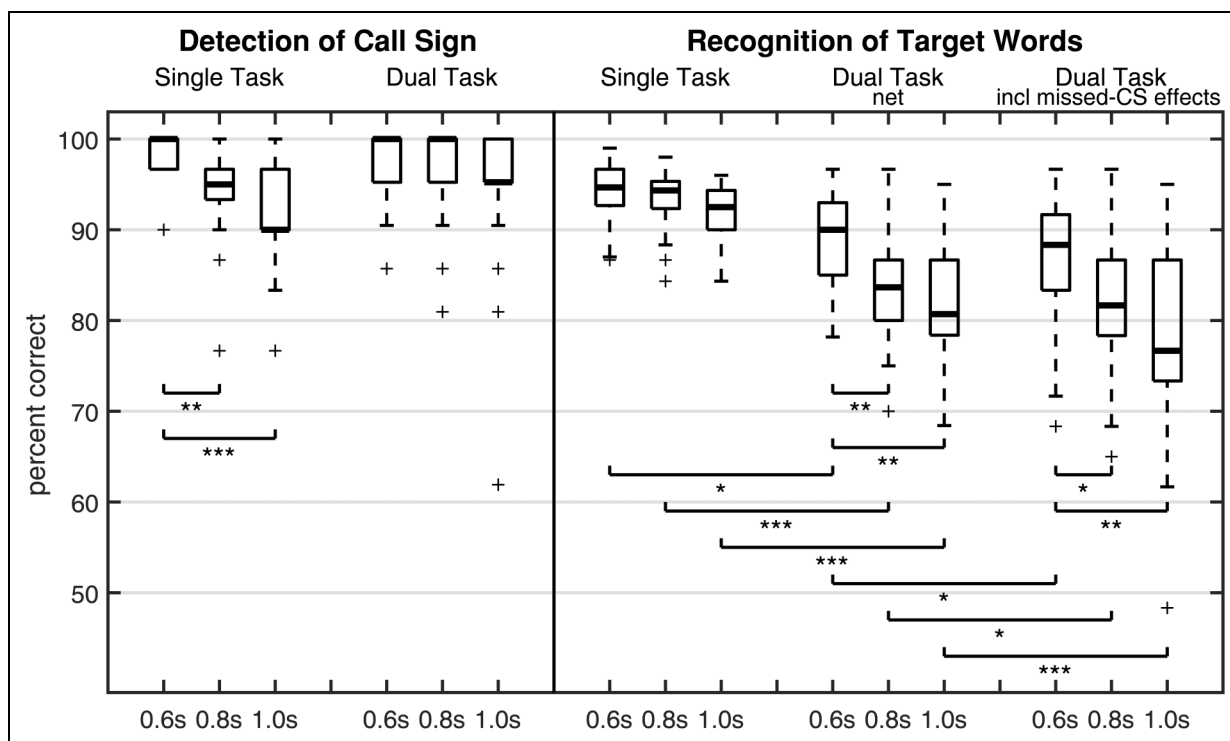
Figure 4. Boxplot of the call-sign detection scores and target-word recognition scores measured for the CCOLSA dual task and the corresponding single tasks for three overlap times. In the net values of the dual task target-word recognition, only those target words with correct call-sign detection were included. Statistical significance is indicated by asterisks (Wilcoxon test; *p < .05, **p < .01, ***p < .001).

(see Table 1 and Figure 4). This effect of overlap duration was significant for target-word recognition in the dual task (net and including missed call-sign effects), and for call-sign detection in the single task.

Head Orientation in Dual Task

Call-sign detection scores in the dual task were calculated either indirectly by reviewing the target-word responses after a call sign (if a correct response was given, the call must have been detected), or by additionally evaluating the participants' head orientation recorded using a head tracker. In the dual-task call-sign detection results, an influence of the head orientation was not detectable, because the scores were near to 100% in both cases.

Nevertheless, as one missed call sign caused 2–4 missed target words, there was a significant difference between speech-recognition scores including missed call-sign effects and net scores. Figure 5 shows the net target-word recognition scores as established considering or discarding head orientation of the head-tracker data. The condition including head-orientation data was identical to the net scores in Figure 4. The condition discarding head orientation incorporated only the first criterion (correct response of target words between two call signs) when calculating net scores. Discarding head orientation, the median values of percent

correct scores ranged from 86 to 91%, instead of 81 to 90% when head orientation was included. The effect of the overlap time was also statistically significant when head orientation was discarded (Friedman test, Bonferroni corrected, $\chi^2 = 9.2$, $p = .03$). In post-hoc tests, one significant paired comparison was found: 0.6 s versus 1.0 s (Wilcoxon test, Bonferroni corrected, $p = .013$, $r = 0.61$, $T = 39$). However, the difference caused by the head-tracker data is significant for all overlap times (Wilcoxon test, Bonferroni corrected; $p_{0.6s} = .043$, $r_{0.6s} = 0.43$, $T_{0.6s} = 0$; $p_{0.8s} = .001$, $r_{0.8s} = 0.68$, $T_{0.8s} = 0$, $p_{1.0s} = .003$, $r_{1.0s} = 0.63$, $T_{1.0s} = 0$).

Discussion

Single Tasks

Speech recognition in the single-task condition of CCOLSA was expected to be close to 100%, as the measurements were conducted at approximately 0 dB SNR, with two active talkers at a time (Bronkhorst, 2000; Humes et al., 2006). This expectation was confirmed for the shortest overlap time. Higher overlap times resulted in a decrease in target-word recognition scores, but this effect was not significant. Also, call-sign detection results in the single-task condition met the expectation that participants were able to perform it almost perfectly (Ericson et al., 2004; Hawley et al., 2004).

Table 1. Results of the Analysis of Performance Differences between Overlaps with the Friedman Test and Post-hoc Comparisons Using Wilcoxon Test; Bold Font Indicates Statistically Significant Values; p-values are Bonferroni-corrected to Account for Multiple Testing.

		Detection of call sign		Recognition of target words		
Friedman test		Single task	Dual Task	Single task	Dual task, net	Dual task, incl. missed-CS effects
χ^2		**25.77**	3.38	8.02	**18.70**	**13.00**
p		**<.001**	.553	.054	**<.001**	**.005**
Paired comparison between Overlaps / s	Wilcoxon test					
0.8 vs. 0.6	T	**4.5**			**36.0**	**34.0**
	p	**0.002**			**0.010**	**0.013**
	r	**0.738**			**0.627**	**0.606**
1.0 vs. 0.6	T	**0.0**			**20.0**	**27.5**
	p	**<0.001**			**0.002**	**0.004**
	r	**0.800**			**0.737**	**0.686**
1.0 vs. 0.8	T	28.0			73.5	61.0
	p	0.198			0.719	0.299
	r	0.392			0.251	0.351
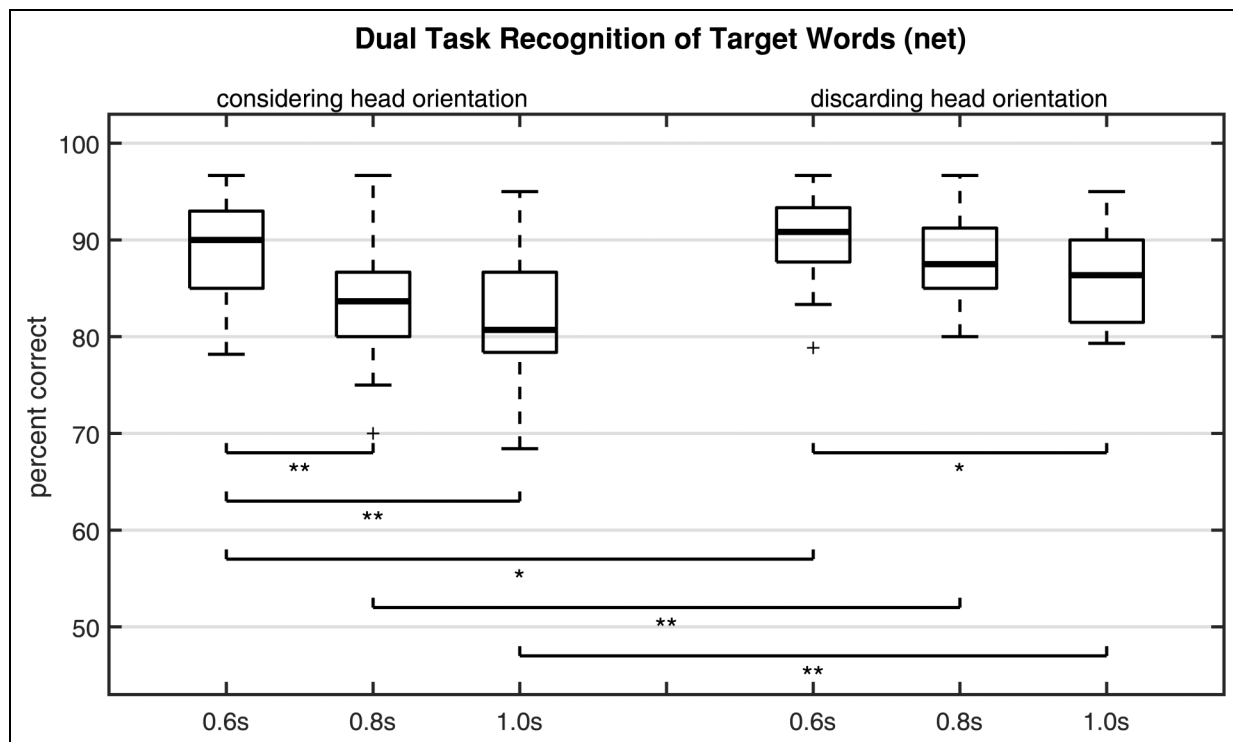
Figure 5. Influence of head-tracker data on the compensation of target-word recognition scores for call-sign detection errors; boxes 1–3 are identical with boxes 10–12 in Figure 4; statistical significance is indicated by asterisks (Wilcoxon test; *p < .05, **p < .01).

Nevertheless, this condition showed a significant effect of the overlap time. This is consistent with Best et al. (2008), who found that switching costs increase with increasing follow-up speed of sentences, due to interruptions in the continuity of auditory objects. Moreover, more speech per time increases cognitive load (Müller et al., 2019), even if only unattended speech is added (Aydelott et al., 2015; Bentin et al., 1995).

This may influence recognition of attended speech by blocking resources (Schneider et al., 2007).

Dual Task

In the dual task, target-word recognition scores were significantly lower than for the single task. This means that the

concurrence of target words and call signs leads to a decrease in speech-recognition scores. The net target-word recognition (corrected for call-sign detection errors) is 5–12%-points lower than for the single-task condition. As this decrease reflects the dual-task costs compared to selective attention listening, it is comparable to the decrease by 10%-points in CRM results caused by call-sign-based target announcement (Humes et al., 2006). It may be interpreted as switching costs that arise from divided attention and cognitive load. The overall recognition score in the dual task (recognition of target words including missed-CS effects) was additionally decreased by 2–4%-points, and indicates that the detection of call signs also slightly suffers in the dual task, although the direct comparison of single- and dual task for call-sign detection showed no significant difference. This indicates that missing a single call sign led to 2–4 missed target words. The observed differences between the overlap conditions prove that the speech-recognition rates can be adjusted by the overlap time.

While the head-turning component of the dual task was included to record call-sign responses, it may have resulted in small changes in the spatial release from masking afforded by the location of the competing talkers relative to the head. Given that the talkers were always spatially separated, and could also be distinguished based on differences in voice, it is unlikely that these small changes in configuration would have greatly impacted talker segregation (and thus response scores; compare: Ericson et al., 2004). However, if the paradigm were to be applied with more similar voices, changes in spatial release from masking might need to be considered. In any case, participants might be able to find orientation strategies to increase their performance if they were encouraged to, and not guided by the task.

The analysis of the net target word recognition with/without head-tracker data shows that call-sign detection was underestimated if the head-tracker data was discarded and only correct responses considered. Implicitly, head-tracking is necessary if analysis of net target word recognition is desired. It might be obsolete for conditions in which call sign detection is at ceiling, as for the overlap condition of 0.6 s in this study.

Advantages of the CCOLSA

Compared to multi-talker tests with simultaneous speech, CCOLSA represents a step towards higher ecological validity by a simulation of turn-taking through direct approaches to the participant's attention in an ongoing group conversation. It allows for analyzes of switching costs and divided attention, based on an alternating-talkers approach. In the past, these outcome measures were only discernable through simultaneous speech approaches (e.g., Bolia et al., 2000; Meister et al., 2018, 2020). Furthermore, the test sensitivity can be adjusted to a targeted SNR using the overlap parameter without a need for a side task. The number of

talkers is fixed, both in the moment of call-sign/ target-word presentation (two talkers) and in total (three talkers). It is not necessary to vary the number of talkers for analyzes of turn-taking effects, as it is in NAL-DCT (Best et al., 2016). As CCOLSA is calibrated for equal intelligibility of the three talkers, biases due to variances in speech recognition across talkers are minimized. In future applications, test difficulty can potentially be adjusted to the target group, e.g., elderly people or hearing-aid users. Presumably, measurements at lower background-noise levels (corresponding to a higher SNR) are possible for non-normal-hearing subjects. Additionally, the CCOLSA test allows for testing speech recognition combined with directional effects. This combination might be explicitly useful for testing spatial noise reduction algorithms in hearing devices, where trade-offs between sound quality and spatial-cue preservation are a common issue.

Technical Issues and Limitations

For matrix tests such as the OLSA, training effects are known and were extensively examined for the original male and a synthetic female talker (Nuesse et al., 2019; Schlueter et al., 2016). In this study, an explicit training phase with speech-recognition measurements of 60 training sentences, as suggested in the literature, was not included. Instead, shorter training phases were conducted for every specific task. The measurements always started with the detection of call signs in single task, for which no data about training effects is yet available. After the single-task detection of call signs, speech recognition in single task was assessed. Because of the familiarization with the test material during the detection task, and the high SNR leading to speech-recognition scores near 100% (above 90% in median for all overlaps), the impact of training effects is estimated to be low in the CCOLSA. Nevertheless, it cannot be ruled out that participants went through various training effects in the different tasks, including the effects due to speech material and familiarization with the different talkers, as well as with the spatial situation.

In future applications, the ceiling effects observed in this study might lead to problems in interpretation of the results or of the statistical analyzes, and should be carefully considered and avoided, if necessary in the individual context. Especially the single task conditions have a sensitivity range close to single talker speech tests (compare Kitterick et al., 2010) and limit the range of applicable SNR, e.g., if needed to compute dual-task costs.

Furthermore, measurements were conducted in an acoustically treated, yet reverberant, lecture room. The influence of room acoustics is expected to be small, but could not be excluded. To evaluate the CCOLSA setup in other settings, comparative measurements in a multi-center study might help to estimate the influence of room acoustics.

Conclusion

The CCOLSA paradigm was introduced to measure speech recognition in multi-talker situations based on different types of attentional processes. The OLSA corpus was combined with the concept of call-sign tests for three talkers and a novel timing control, so that call signs and target words occurred with a defined and overlapping timing. The difficulty of the test can be adjusted by the overlap time between call sign and target word, which allows for its use at high SNRs. For the same listening conditions, a call-sign detection test and a speech-recognition test can be performed as single tasks and, concurrently, as a dual task. The paradigm allows for analyzes of turn-taking and attentional factors and, thus, extends the analytical spectrum of the OLSA towards complex multi-talker situations. As the talkers are spatially separated, the test can, for example, be applied to the evaluation of beamformers in hearing devices or any other algorithms influencing the binaural listening conditions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID iDs

Jan Heeren 🆔 https://orcid.org/0000-0002-8489-8705
Theresa Nuesse 🆔 https://orcid.org/0000-0002-6202-8952
Kirsten C. Wagener 🆔 https://orcid.org/0000-0003-2858-2789

References

American Speech, Language and Hearing Association (1988). Determining Threshold Level for Speech [Guidelines]. www.asha.org/policy

Arbogast T. L., Mason C. R., & Kidd G.Jr. (2005). The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *117*(4), 2169–2180. https://doi.org/10.1121/1.1861598

Aydelott J., Jamaluddin Z., & Nixon Pearce S. (2015). Semantic processing of unattended speech in dichotic listening. *The Journal of the Acoustical Society of America*, *138*(2), 964–975. https://doi.org/10.1121/1.4927410

Beechey T. (2022). Ecological validity, external validity, and mundane realism in hearing science. *Ear and Hearing*. https://doi.org/10.1097/AUD.0000000000001202

Bentin S., Kutas M., & Hillyard S. A. (1995). Semantic processing and memory for attended and unattended words in dichotic listening: Behavioral and electrophysiological evidence. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 54. https://doi.org/10.1037/0096-1523.21.1.54

Best V., Keidser G., Freeston K., & Buchholz J. M. (2016). A dynamic speech comprehension test for assessing real-world listening ability. *Journal of the American Academy of Audiology*, *27*(7), 515–526. https://doi.org/10.3766/jaaa.15089

Best V., Keidser G., Freeston K., & Buchholz J. M. (2018). Evaluation of the NAL dynamic conversations test in older listeners with hearing loss. *International Journal of Audiology*, *57*(3), 221–229. https://doi.org/10.1080/14992027.2017.1365275

Best V., Ozmeral E. J., Kopčo N., & Shinn-Cunningham B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences*, *105*(35), 13174–13178. https://doi.org/10.1073/pnas.0803718105

Bolia R. S., Nelson W. T., Ericson M. A., & Simpson B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, *107*(2), 1065–1066. https://doi.org/10.1121/1.428288

Brand T., & Kollmeier B. (2002). 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), 2801–2810. https://doi.org/10.1121/1.1479152

Bronkhorst A. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica United with Acustica*, *86*(1), 117–128.

Bronkhorst A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, *77*(5), 1465–1487. https://doi.org/10.3758/s13414-015-0882-9

Brungart D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109. https://doi.org/10.1121/1.1345696

Brungart D. S., & Simpson B. D. (2007). Cocktail party listening in a dynamic multitalker environment. *Perception & Psychophysics*, *69*(1), 79–91. https://doi.org/10.3758/BF03194455

Brungart D. S., Simpson B. D., Ericson M. A., & Scott K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *110*(5), 2527–2538. https://doi.org/10.1121/1.1408946

Bundesausschuss G. (2021). Hilfsmittel-Richtlinie.

Desjardins J. L. (2011). Age-Related Changes in Listening Effort for Various Types of Masker Noises. Communication Sciences and Disorders—Dissertations 14. https://surface.syr.edu/csd_etd/14

Dhami M. K., Hertwig R., & Hoffrage U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, *130*(6), 959–988. https://doi.org/10.1037/0033-2909.130.6.959

Ericson A., Brungart D. S., & Simpson B. D. (2004). Factors that influence intelligibility in multitalker speech displays. *Mark International Journal of Aviation Psychology*, *14*(3), 313–334. https://doi.org/10.1207/s15327108ijap1403_6

Francart T., Van Wieringen A., & Wouters J. (2011). Comparison of fluctuating maskers for speech recognition tests. *International Journal of Audiology*, *50*(1), 2–13. https://doi.org/10.3109/14992027.2010.505582

Gagné J.-P., Besser J., & Lemke U. (2017). Behavioral assessment of listening effort using a dual-task paradigm. *Trends in Hearing*, *21*, 1–25. https://doi.org/10.1177/2331216516687287

Grimm G., & Hohmann V. (2019). First order ambisonics field recordings for use in virtual acoustic environments in the context of audiology. *Zenodo*. https://doi.org/10.5281/zenodo.3588303

Grimm G., Luberadzka J., & Hohmann V. (2019). A toolbox for rendering virtual acoustic environments in the context of audiology. *Acta Acustica United with Acustica*, *105*(3), 566–578. https://doi.org/10.3813/AAA.919337

Hammond K. R. (1998). Ecological validity: Then and now. http://www.brunswik.org/notes/essay2.html (05.08.2019)

Hawley M. L., Litovsky R., & Culling J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, *115*(2), 833–843. https://doi.org/10.1121/1.1639908

Hochmuth S., Jürgens T., Brand T., & Kollmeier B. (2014). Multilinguale Cocktailparty—Einfluss von sprecher- und sprachspezifischen Faktoren auf die Sprachverständlichkeit im Störschall. In Proceedings of the 17. Jahrestagung für Audiologie, Oldenburg.

Holleman G. A., Hooge I. T., Kemner C., & Hessels R. S. (2020). The "real-world approach" and its problems: A critique of the term ecological validity. *Frontiers in Psychology*, *11*, 721. https://doi.org/10.3389/fpsyg.2020.00721

Holube I. (2011). Speech intelligibility in fluctuating maskers. In Proceedings of the international symposium on auditory and audiological research (Vol. *3*, pp. 57–64).

Humes L. E., Lee J. H., & Coughlin M. P. (2006). Auditory measures of selective and divided attention in young and older adults using single-talker competition. *The Journal of the Acoustical Society of America*, *120*(5), 2926–2937. https://doi.org/10.1121/1.2354070

Keidser G., Naylor G., & Brungart D. Caduff A S.., Campos J., Carlile S., Carpenter M. G., Grimm G., Hohmann V., Holube I., Launer S., Lunner T., Mehra R., Rapport F., Slaney M., & Smeds K. (2020). The quest for ecological validity in hearing science: What it is, why it matters, and how to advance it. *Ear and Hearing*, *41*(Suppl 1), 5S. https://doi.org/10.1097/AUD.0000000000000944

Kidd G.Jr., Mason C. R., Swaminathan J., Roverud E., Clayton K. K., & Best V. (2016). Determining the energetic and informational components of speech-on-speech masking. *The Journal of the Acoustical Society of America*, *140*(1), 132–144. https://doi.org/10.1121/1.4954748

Killion M. C., Niquette P. A., Gudmundsen G. I., Revit L. J., & Banerjee S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *116*(4), 2395–2405. https://doi.org/10.1121/1.1784440

Kitterick P. T., Bailey P. J., & Summerfield A. Q. (2010). Benefits of knowing who, where, and when in multi-talker listening. *The Journal of the Acoustical Society of America*, *127*(4), 2498–2508. https://doi.org/10.1121/1.3327507

Lin G., & Carlile S. (2015). Costs of switching auditory spatial attention in following conversational turn-taking. *Frontiers in Neuroscience*. https://doi.org/10.3389/fnins.2015.00124

Mansour N., Marschall M., May T., Westermann A., & Dau T. (2021). A method for realistic, conversational signal-to-noise ratio estimation. *The Journal of the Acoustical Society of America*, *149*, 1559–1566. https://doi.org/10.1121/10.0003626

Meister H., Rählmann S., & Walger M. (2018). Low background noise increases cognitive load in older adults listening to competing speech. *The Journal of the Acoustical Society of America*, *144*(5), EL417–EL422. https://doi.org/10.1121/1.5078953

Meister H., Wenzel F., Gehlen A. K., Kessler J., & Walger M. (2020). Static and dynamic cocktail party listening in younger and older adults. *Hearing Research*, *395*, 108020. https://doi.org/10.1016/j.heares.2020.108020

Müller J. A., Wendt D., Kollmeier B., Debener S., & Brand T. (2019). Effect of speech rate on neural tracking of speech. *Frontiers in Psychology*, *10*, 449. https://doi.org/10.3389/fpsyg.2019.00449

Nilsson M., Soli S. D., & Sullivan J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, *95*(2), 1085–1099. https://doi.org/10.1121/1.408469

Nuesse T., Heeren J., Latzel M., Schulte M., & Holube I. (in prep.). Cognitive abilities affect speech recognition in multi-talker situations of young and elderly normal-hearing listeners using the Concurrent OLSA test.

Nuesse T., Wiercinski B., Brand T., & Holube I. (2019). Measuring speech recognition with a matrix test using synthetic speech. *Trends in Hearing*, *23*, 2331216519862982. https://doi.org/10.1177/2331216519862982

Schlueter A., Lemke U., Kollmeier B., & Holube I. (2016). Normal and time-compressed speech: how does learning affect speech recognition thresholds in noise? *Trends in Hearing*, *20*, 1–13. https://doi.org/10.1177/2331216516669889

Schneider B. A., Daneman M., Murphy D. R., & See S. K. (2000). Listening to discourse in distracting settings: The effects of aging. *Psychology and Aging*, *15*(1), 110. https://doi.org/10.1037/0882-7974.15.1.110

Schneider B. A., Li L., & Daneman M. (2007). How competing speech interferes with speech comprehension in everyday listening situations. *Journal of the American Academy of Audiology*, *18*(7), 559–572. https://doi.org/10.3766/jaaa.18.7.4

Shafiro V., & Gygi B. (2007). Perceiving the speech of multiple concurrent talkers in a combined divided and selective attention task. *The Journal of the Acoustical Society of America*, *122*(6), EL229–EL235. https://doi.org/10.1121/1.2806174

Shinn-Cunningham B. G., & Best V. (2008). Selective attention in normal and impaired hearing. *Trends in Amplification*, *12*(4), 283–299. https://doi.org/10.1177/1084713808325306

Šimkovic M., & Träuble B. (2019). Robustness of statistical methods when measure is affected by ceiling and/or floor effect. *PloS one*, *14*(8), e0220889. https://doi.org/10.1371/journal.pone.0220889

Singh G., Pichora-Fuller M. K., Behrens T., & Neher T. (2009). The effect of hearing loss on auditory spatial attention. In Proceedings of ISAAR 2009.

Smeds K., Wolters F., & Rung M. (2015). Estimation of signal-to-noise ratios in realistic sound scenarios. *Journal of*

the American Academy of Audiology, 26(2), 183–196. https://doi.org/10.3766/jaaa.26.2.7

Wagener K. C., & Brand T. (2005). Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters. *International Journal of Audiology*, 44(3), 144–157. https://doi.org/10.1080/14992020500057517

Wagener K. C., Brand T., & Kollmeier B. (1999). Entwicklung und evaluation eines Satztests in deutscher Sprache III: Evaluation des Oldenburger Satztests. *Z Audiology*, 38(3), 86–95.

Wagener K. C., Hochmuth S., Ahrlich M., Zokoll M. A., & Kollmeier B. (2014). Der weibliche Oldenburger Satztest. In 17th annual conference of the DGA Oldenburg, CD-Rom.

Weisser A., & Buchholz J. M. (2019). Conversational speech levels and signal-to-noise ratios in realistic acoustic conditions. *The Journal of the Acoustical Society of America*, 145, 349–360. https://doi.org/10.1121/1.5087567

Wilhelm P., & Perrez M. (2013). A history of research conducted in daily life. Scientific report Nr. 170, Department of Psychology, University of Friborg, Switzerland. http://www.unifr.ch/psycho/site/assets/files/klinisch/Wilhelm%20&%20Perrez%20(2013).%20A%20history%20of%20research%20coducted%20in%20daily%20life.pdf (03.08.2019)

Wood N., & Cowan N. (1995). The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1), 255. https://doi.org/10.1037/0278-7393.21.1.255

Wu Y. H., Stangl E., Chipara O., Hasan S. S., Welhaven A., & Oleson J. (2018). Characteristics of real-world signal-to-noise ratios and speech listening situations of older adults with mild-to-moderate hearing loss. *Ear and Hearing*, 39(2), 293. https://doi.org/10.1097/AUD.0000000000000486