Speech intelligibility in reverberation is reduced during self-rotation

Ľuboš Hládek and Bernhard U. Seeber

Audio Information Processing, Technical University of Munich, Arcisstraße 21,

80333 Munich, Germany

Corresponding author:

Ľuboš Hládek

Address:

Audio Information Processing

Technical University of Munich

Arcisstr. 21

80333 Munich, Germany

Email: lubos.hladek@tum.de

Phone: +49 (0)89 289-28564

Abstract

Spatial unmasking helps speech intelligibility in a cocktail party but its effects have been studied mainly for stationary participants. Here, we investigate behavior and speech intelligibility during active self-rotation of standing participants and we assess the impact of motion and visually presented location cues. We employed a spatialized speech test in a controlled reverberant space with target sentences randomly appearing on each trial at one of four possible locations (0°, ±90°, 180°), while speech-shaped noise was presented from 0° respective to the participant's orientation. Participants responded naturally with motion as in a social situation. Target sentences were presented either without (A-only) or with a picture of an avatar (AV). In a baseline (Static) condition, people were standing still without visual location cues. Participants undershot the targets, often in the acoustically optimal way, but they also oriented away from the frontal target where there was no acoustic benefit. They performed equally in AV and A-only. They performed better in the A-only than in the Static condition for the rear target, but worse for the lateral target. While the first can be partly explained by spatial unmasking, the latter cannot. The speech intelligibility model by Jelfs et al., (2011), extended to consider self-rotation, overestimated participant performance during motion. The experimental and modeling results suggest that listeners have a limited access to the spatial unmasking cues during self-rotation. The results are discussed in context of binaural sluggishness and non-acoustic factors.

Keywords: Speech understanding, head rotation, spatial unmasking, speech intelligibility model

Introduction

In cocktail parties, people are often moving. In conversations, people come closer to each other and increase their voices when there is a high level of background noise (Beechey et al., 2018a; Cheyne et al., 2009; Hadley et al., 2019; Latif et al., 2014). When they turn their head toward the speaker, they could benefit from improved acoustic cues such as an increased signal-to-noise ratio in one of the ears (Grange & Culling, 2016b). On the contrary, other research reported that dynamic changes of the sound source could negatively affect speech intelligibility (Viveros Muñoz et al., 2019). Research on hearing aids has suspected movement to be a limiting factor for hearing aid benefit in real situations (Bentler, 2005; Ching et al., 2009; Cord et al., 2004). Therefore, numerous studies aimed to create ecologically valid and individualized assessment of hearing abilities to provide more accurate predictions and better treatment of hearing loss (Beechey et al., 2018b; Best et al., 2007; Brungart et al., 2020; Gonzalez-Franco et al., 2017; Hendrikse et al., 2018, 2019; Kerber & Seeber, 2011, 2013; Kishline et al., 2020; Stecker, 2019; Viveros Muñoz et al., 2019). Here, we aim to create an ecologically relevant testing scenario by directly studying the effects of natural self-rotation with or without location cues.

The research traditionally finds that normal hearing people are exceptionally good in understanding speech in ‘cocktail parties’. One of the prominent factors is spatial separation of the target from the interferer – spatial unmasking or spatial release from masking (Bronkhorst, 2000, 2015; Culling et al., 2004). Here, we will refer to spatial unmasking as an acoustic and auditory phenomenon. The head creates an acoustic shadow, which leads to a higher signal-to-noise ratio at one of the ears, hence one component of spatial unmasking is ‘better ear listening’, while the other component, ‘binaural unmasking’, is considered as a contribution of the binaural system to improve detection of the target in the interferer. Spatial unmasking depends on a relative orientation of the listener to the sound sources and the

maximal benefits are usually found in the range of 10 – 14 dB (Grange & Culling, 2016a; Jelfs et al., 2011; Marrone et al., 2008). Although the values vary with reverberation time, age and other factors, the phenomenon is strong in rooms with standard reverberation profile and across different age groups (Bronkhorst, 2000, 2015; Marrone et al., 2008). These tests are traditionally conducted for static sounds and static participants; however, less is known about how these benefits translate to moving people or sources.

A study by Brimijoin et al. (2012) was one of the first to investigate the effect of self-rotation on a speech intelligibility. The experimenters expected that people would adapt head rotations for better understanding but instead they observed that people always maintained an off-center rotation relative to the target. Therefore, participants effectively ignored the position of the interferer, suggesting that movement induced changes of intelligibility did not play a critical role in behavior. One possible reason is that the stimuli had a fixed target-interferer configuration within a block, and people may have adapted to it, but in real situations, we often turn towards unexpected directions. In another study, Grange & Culling (2016a) observed a high variability in undirected head orienting movements suggesting a weak contribution of motion induced intelligibility changes to behavior but the participants were seated on a chair which limited the movement only to the head rotations. Shen et al. (2017) also observed a difference in the propensity to head movements in an experiment where both speech and masker came from behind. Head movement did not affect speech perception. High variance in propensity to head movements has been also observed in an experiment in which the participants were instructed to follow speech signals originating at two locations in pseudo-random intervals (Hládek et al., 2019) suggesting that people may use different strategies during speech perception.

Grange et al. (2018) observed that participants could maximize speech intelligibility by movement, if they were explicitly instructed to do so and Hendrikse et al. (2018) showed that

when participants were instructed to follow the active speaker in a pre-recorded spatialized conversation, head movement patterns influenced signal-to-noise ratio. The latter study also showed that the presence of visual cues (lip movements and gaze of other avatars) influences the pattern of head rotations, gaze fixations and performance in terms of a combined localization and speech intelligibility. However, the study did not identify a consistent strategy to optimize speech intelligibility by means of head movements. These results suggest that people may adapt the head movements for better intelligibility but it depends on whether they are sitting or standing (Hendrikse et al., 2019), whether the SNR is changing (Brimijoin et al., 2012), instructions (Grange et al., 2018), content, visual cues (Hendrikse et al., 2018) and possibly other attention and experience-related factors (Best et al., 2008; Kidd et al., 2005). The effects of self-motion in the context of hearing have been recently briefly reviewed by Grimm et al. (2020).

When people rotate during speech perception, acoustic and binaural cues vary over time, which may influence speech intelligibility. In experiment of Frissen et al. (2019), the participants were listening over headphones to target speech among spatially distributed masker sentences and they were actively turning their head; the movement was unrelated to speech. Using virtual acoustics and motion tracking, the speech material was made world-centered (the stimulus was perceived as static), or head-centered (i.e., no compensation for head rotation) to isolate the effect of motion. However, the study reports only a negligible effect of head-rotation on speech intelligibility.

While the abovementioned studies mainly considered head rotations of participants and stationary sound sources, Pastore & Yost (2017) investigated the effect of a moving target sound source, Viveros Muñoz et al. (2019) the effect of a moving interfering sound source, and Davis et al. (2016) considered both, a moving target and an interfering sound source. Davis et al. (2016) and Pastore & Yost (2017) reported relatively small effects of motion,

which could be potentially accounted to changes in binaural and acoustic cues that correspond to the change in position. Viveros Muñoz et al. (2019) reported a negative effect of motion on speech intelligibility in a group of older listeners in reverberant conditions but in other conditions, the effect was small or negligible. A possible reason why these studies had difficulties to isolate the effect of motion on speech perception could relate to either cognitive factors (since the older participants might be more susceptible to attentional effects) or possibly slow rotation speeds. Viveros Muñoz et al. (2019) used speed of 32.73 °/s (circular moving masker), Pastore & Yost (2017) used 53 °/s, but people naturally rotate 150 °/s or even faster (Brimijoin et al., 2010).

The aim of the present study is to investigate the effect of natural self-rotation on speech intelligibility of standing participants and to assess the effect of presence of a visual cue indicating target location on self-rotating behavior and speech intelligibility. In order to study the dynamics of speech intelligibility, we use a structured sentence of five words as a target sound and measured intelligibility separately for each word. First, we hypothesize (H1) that people naturally perform self-rotations, which help speech intelligibility. Therefore, we expect that people would employ a strategy that would maximize the acoustic and binaural benefits even if the target comes from an unexpected direction. For instance, if an interferer was at the front and a target on the side or behind, a close-to-optimal strategy would be to rotate approximately half-way between the interferer and the target. In this position, the acoustic shadow of the head maximizes signal-to-noise ratio at the ear closer to the target, while it also gives a substantial benefit in terms of the binaural cues. Second, we hypothesize (H2) that a visual indication of target location would help participant to rotate even more optimally and obtain better speech intelligibility. Third, we hypothesize (H3) that the dynamic effect of self-rotation could also influence speech intelligibility in a negative way, similar to what is known from binaural sluggishness (Grantham & Wightman, 1978, 1979).

To address these hypotheses, we created a testing environment that stipulated realistic behavioral responses but also provided different degrees of dynamics in terms of the change of target-interferer configuration with different opportunities for motion-induced speech intelligibility benefit. During the experiment, we encouraged participants to behave naturally with as little limitations as possible in terms of self-rotation. To make a more direct link between the spatial unmasking benefits during motion and speech intelligibility, we employ a speech perception model (Jelfs et al., 2011) to predict intelligibility for different head orientations under an assumptions that the sound and the person are stationary.

Methods

Participants

Young volunteers (n = 9, age: 26.6±6 (median±iqr), 1 female), native German speakers, took part in the study. Their hearing thresholds were checked with a calibrated audiometer (MADSEN Astera², type 1066, Natus Medical Denmark Ap, Denmark). All pure-tone thresholds at standard audiological frequencies (250 Hz - 8 kHz) were below or equal to 20 dB HL. One additional participant did not finish the study because of problems with hearing the target sounds in the majority of conditions although the participant did not report problems with hearing. All participants provided written informed consent. Methodology and procedures were approved by the ethics committee of the Technical University of Munich (65/18S).

Environment

The study was conducted in the Simulated Open Field Environment (SOFE v4) (Seeber et al., 2010; Seeber & Clapp, 2017). The SOFE v4 setup consists of a high-fidelity sound reproduction system with a four-sided video projection inside an anechoic chamber (10 m x 6 m x 4 m; l x w x h). Thirty-six equally spaced active loudspeakers (Dynaudio BM6A mkII,

Dynaudio, Skanderborg, Denmark) are positioned on a square-shaped construction (4.39 m x 4.39 m) at the height of 1.4 m and they point to the center of the square (10° separation between the loudspeakers). The audio signals are played via a multi-channel sound card (RME HDSPe, Audio AG, Haimhausen, Germany) and digital-analog converters (RME 32DA, Audio AG, Haimhausen, Germany). The audio presentation system was calibrated and loudspeakers equalized in frequency response, time of arrival and phase for frequencies between 100 Hz and 18 kHz by a set of finite impulse response filters of 512 taps length at 44.1 kHz sampling frequency. The visual presentation system consisted of four high-resolution projectors (Barco F50 WQXGA, Barco, Kortrijk, Belgium) with low background noise (total of 32 dB(A) in the middle of the loudspeaker array) that project to four large acoustically transparent screens with projection area of 4.3 m x 2.7 m at distance of 2.15 m positioned right in front of the loudspeakers. The SOFE is further equipped with twelve high-speed optical motion-tracking cameras (OptiTrack Prime 17W, NaturalPoint Inc. Corvallis, Oregon, USA) which run in synchrony (eSync 2, NaturalPoint Inc. Corvallis, Oregon, USA) with the sound card via a word-clock signal. With the sound presented at 44.1 kHz, the motion tracking ran on 358.6 Hz such that each sample corresponded to 123 samples on the sound card. The experiment was controlled with three PCs using custom scripts written in MATLAB (v9.8.0 and v9.9.0, Mathworks, Natick, MA, USA) and Python (v3.6). The synchrony of the motion capture system and the sound presentation system was assessed by recoding the in-ear signals of an artificial head (HMS II.3-33, Head Acoustics, Herzogenrath, Germany) which was rotated in the place of the participant. The motion trajectory was then used to re-create a ‘moving’ stimulus with a 0.5° resolution, which was recorded again by the static artificial head. We observed that interaural level differences of the two recordings were aligned.

The participants held a tablet touch-screen displaying all ten possibilities for each word of the OLSA (Wagener, Kühnel, et al., 1999) sentences in matrix format, which could be tapped on.

The GUI also displayed information whether the participants can move or whether they should stand still. The GUI gave feedback on performance from the previous trial (i.e., number of correct words out of five). Participants wore a motion-tracking crown, which was used to determine the position and rotation of the head. The position of the crown on the head was calibrated at the beginning of each experimental block (6 times during the experiment) to ensure precise measurement of head rotations. The experimental program checked, at the beginning of each trial, if the participant was standing within 20 cm of the center of the loudspeaker array and in the Static condition, it checked whether they were facing the frontal loudspeaker with a tolerance of 3 degrees.

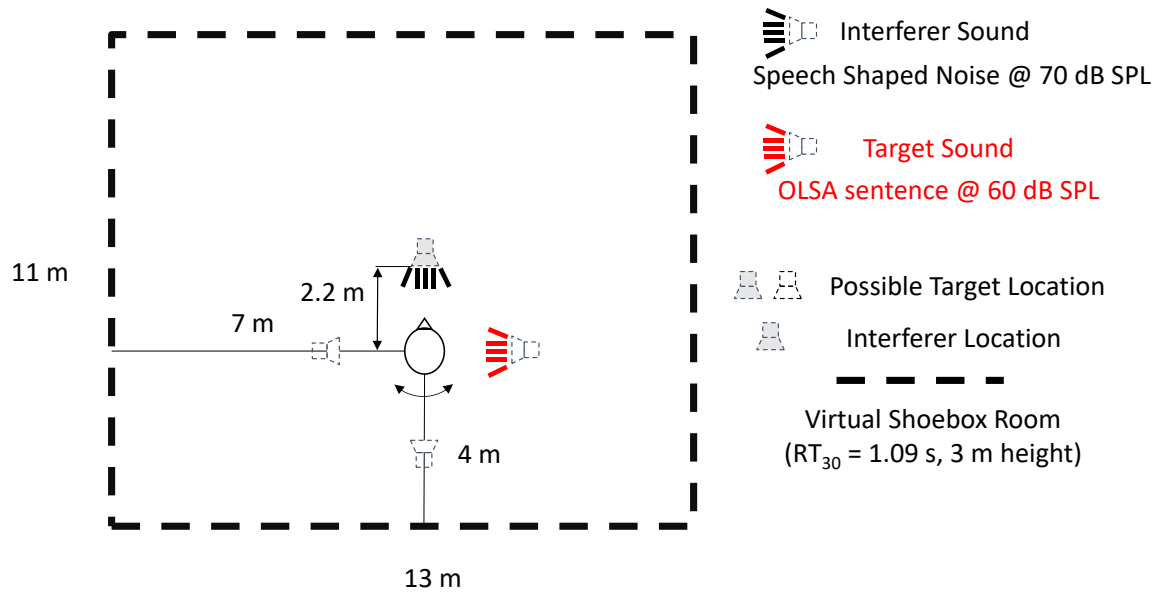


Figure 1 – Position of the subject and of target and interferer sounds in the virtual room.

Stimuli

The target sound stimuli consisted of twelve unique sentence lists from the OLSA matrix test (Wagener, Kühnel, et al., 1999) presented at 60 dB SPL, which were randomly assigned to each participant from a total set of 32 lists. Every single list was fixed to one of four possible

locations (0°, ±90°, 180°) and a condition. OLSA lists consist of sentences with fixed structure (e.g., ‘Britta gibt vier alte Bilder.’) such that each word is taken from a closed set of ten options. From each list, we used sentences 6-30.

The interferer sound was 4.5-seconds-long and it always started one second before the target. The noise was stationary speech-shaped noise presented at 70 dB SPL with the same spectrum as the target sentence, which was computed for each sentence by taking the Fourier transform of the speech signal and randomizing the phase. Each token was ramped at the onset and the offset with a 50 ms Gaussian slope. The sound level of stimuli was defined as the level of the direct sound (anechoic part without reflections) in the middle of the loudspeaker array. The level was verified by a calibrated hand-held sound level meter (XL2, NTi Audio, Schaan, Liechtenstein) by measuring the level of speech-shaped noise played from one of the equalized loudspeakers.

All stimuli, targets and interferers, were spatialized in a virtual reverberant room shown on Figure 1. Dimensions of the virtual room were 11 m x 13 m x 3 m (l x w x h) and the virtual listener was placed off-center (4 m, 7 m, 1.8 m; l x w x h). The sound sources were positioned at 2.2 m distance from the listener. The acoustics of virtual room was modeled using the image source method (Borish, 1984) implemented in the real-time SOFE (Seeber et al., 2010; Seeber & Clapp, 2017), for reflections of order up to one hundred. Individual reflections were rendered via loudspeakers using the Ambisonics technique with max-rE weighting (Stitt et al., 2016) for the reflections up to fifth order. Higher-order reflections were mapped to the nearest loudspeaker. Reverberation time of the simulated impulse response (T_{30} = 1.16 s @ 250 Hz, 1.34 s @ 500 Hz, 1.15 s @ 1 kHz, 1.02 s @ 2 kHz, 0.85 s @ 4 kHz) was determined by ita_room_acoustics function of ITA toolbox (Berzborn et al., 2017).

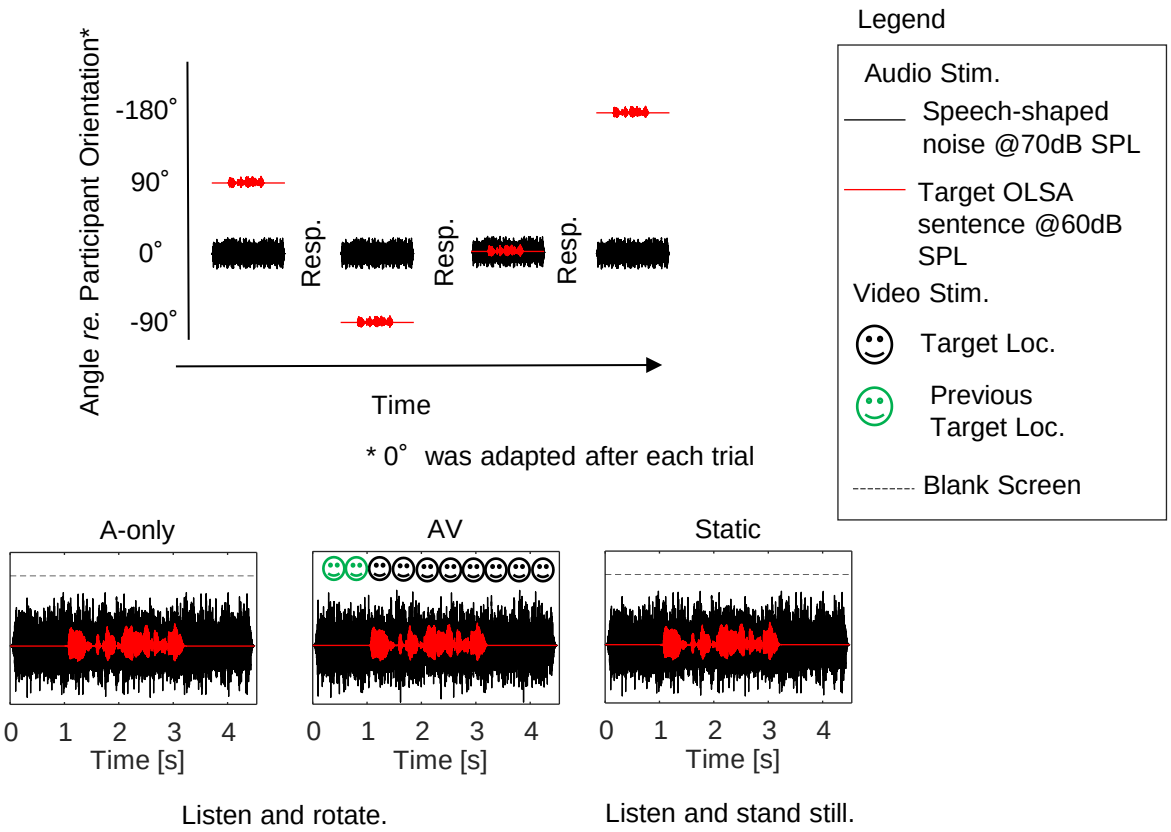


Figure 2 – Experimental procedures and conditions. The top graph shows an example of four experimental trials with time on the abscissa and target angle relative to the participant on the ordinate. Each trial (one sentence and noise) was followed by a response. The bottom part of the figure shows three experimental conditions and the corresponding instructions.

The visual stimuli consisted of a human-sized virtual character (*MakeHuman*, 2019) video-projected on the four screens surrounding the participant. The character appeared at the target azimuth synchronously with the onset of the target sentence and was visible until the start of the new sentence in the next trial. The synchrony of audio-visual experimental stimuli was assessed using a photosensitive LED and a pre-amplified measurement microphone connected to a storage oscilloscope (HMO724, Rhode & Schwarz). The analysis of 10 repetitions showed an offset of 80±13 ms (mean±std) between two stimuli which was accounted for in the experimental code.

Conditions and procedures

The experiment involved three conditions: audio-visual (AV), audio-only (A-only), and a static baseline (Static), with the condition fixed within a block (Figure 2). The conditions

differed in type of stimuli and instructions. In the Static condition, the participants were standing still and there was no visual stimulus, only sounds. In the A-only condition, the participants heard auditory stimuli without visual stimuli (same as in Static) and they performed self-rotations. The AV condition was identical to A-only, but the target sound was accompanied by the visual virtual character.

The participants were instructed to imagine a social situation in which somebody was talking to them. The participants were asked to listen to the target sentence, as if the approaching person was saying it, and behave naturally. For instance, they could rotate toward the person, if it is something they usually do. Participants were told that in some blocks a virtual character would appear at the target direction while in other blocks there will be no virtual character, only sound. The participants were also told not to walk away from their position (participant could only rotate on place) and that they did not have to return to the initial orientation to initiate the next trial. In the Static condition, participants were instructed to stand still and listen to the target sentence. After each trial, participants reported the five words of each sentence using a hand-held touchscreen tablet.

The experiment was organized in six blocks with the condition fixed for each block. Each condition was presented in two blocks such that the first block involved sentences from the first part of the sentence list, and the second block involved sentences from second part of the sentence list. The order of blocks was random and unique for each participant, with the limitation that the first three blocks involved all three conditions. Each block consisted of forty-eight trials; one trial consisted of one sentence presented together with interfering noise, followed by a response. The target angle varied pseudo-randomly on a trial-by-trial basis, and it could be one of four possible target locations (0°, ±90°, 180°), but the target angle changed after each trial. The noise stimulus was presented always from the front of the participant (0°) (Figure 2). On each trial, the virtual acoustic presentation was aligned with the head

orientation of the person (by using motion-tracking), thus the physical target angles were changing according to the movement of the participant to be always correct with respect to the momentary orientation at the beginning of each trial. This was done to suppress any potential influence of a fixed, room centered spatial frame of reference. Over the whole block, participants heard twelve sentences from each of the four possible positions.

Before the start of the main experiment, all participants underwent four training blocks. The training consisted of the blocks of the Static condition with the same procedures as described for the main experiment. However, only lists 33-40 from the OLSA CD were used during the training (these lists were not used in the main experiment), and the sound level was set to 64 dB SPL for the first two blocks, and 62 dB SPL for the second two blocks. For some participants this was increased to 67 dB SPL and 63 dB SPL.

Analysis

We analyzed self-rotations by analyzing the horizontal rotations of the head. The data were obtained from the output of the motion tracking system, which provided rotation values of the tracking object in quaternions. Before the analysis, data were rotated to align the motion tracking reference frame and the experimental reference frame. Then the data were transformed to Euler angles that provide the yaw angles.

To analyze yaw rotations, first, we took trajectories of individual trials and computed initial position at the beginning of each trial, which was marked as angle 0°. The data were unwrapped to avoid any discontinuities and smoothed with a 27 ms Kaiser window (twice, zero-phase, using Matlab function filtfilt). In the next step, each individual trajectory was split into five sub-parts according to the duration of words in each sentence. The durations were determined by listening to all OLSA sentences (that were used in the experiment) and we manually labelled the beginnings of all words in each sentence using a visual interface. In this way, we obtained precise temporal positions for each word in the experiment with respect to

the rotation trajectory of each participant. From the pre-processed trajectories, we computed the median of the rotation angle for each individual word of target sentence. These values were then used to compute the absolute angular distance to the target and angular rotations.

To analyze the time course of speech intelligibility, we computed the number of correct words across all 24 test sentences in each condition for each of the five words of the target sentences. For the purpose of plotting, per cent correct values were transformed into RAU scores (Studebaker, 1985). The guessing rate for each word is 10%.

We used a speech intelligibly model (Jelfs et al., 2011) from the Auditory Modeling Toolbox (Søndergaard & Majdak, 2013) to create predictions of speech intelligibility for different head orientations during self-rotation, but the model assumes that the acoustic scene is static for a particular head orientation; it does not include short-time processing or binaural sluggishness. This model characterizes the spatial unmasking in reverberant environments in terms of the effective target-to-interferer ratio in dB. To compute the predictions, we used the individual self-rotation trajectories and the spatialized room impulse responses from the experiment (for SOFE-loudspeakers placed every 10° in horizontal plane). The predictions of the model for different head orientations (using 1° resolution) were interpolated along each individual horizontal rotation trajectory. Subsequently, for each trajectory, we extracted five values in dB by taking the median value for each word. In order to map the dB values to per cent correct performance, we fit the participant performance in the Static condition to the predicted dB values. The data of the A-only and the AV conditions were obtained from this model using the respective self-rotation trajectories.

The performance, the model output and behavior data were statistically analyzed in MATLAB using Generalized Linear Mixed-Effect Models (GLM). In the statistical analysis, we considered the experimental factors (target angle, visual cue, word position and their interactions) as predictors and these are specified further in text. We fit the responses of the

participants (correct word / incorrect word) or the speech model predictions using a model with the ‘binomial’ distribution and ‘logit’ link function. For the analysis of angular self-rotations, we used a similar approach, but the statistical model had the ‘Gaussian’ distribution and ‘identity’ link function. We used the MATLAB function ‘fitglme’ with parameter ‘DummyVarCoding’ set to ‘effects’, parameter ‘CheckHessian’ set to ‘true’, ‘CovariancePattern’,’CompSymm’ and the fitting method was set to ‘ApproximateLaplace’. The fixed-effects design matrix included all main effects and their interactions. The random-effects design matrix had the same structure as the fixed-effects design matrix, and participants were the grouping variable.

Results

Self-orienting behavior

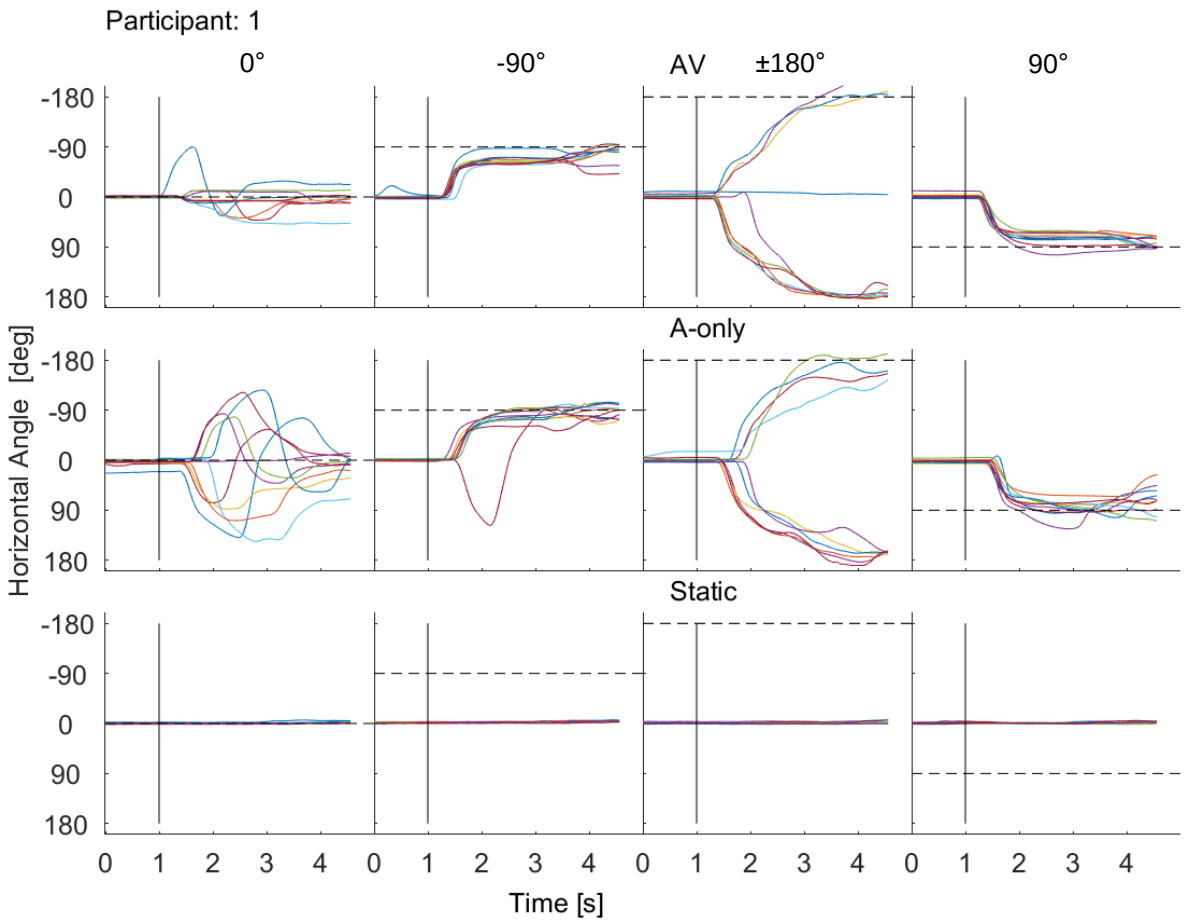


Figure 3 – An example of raw head orientation trajectories for one participant. The first row shows data for the AV, the second row for the A-only and the third row for the Static condition. The columns show data for different target angles which are indicated by dashed lines. The vertical line at the time of 1 second indicates the onset of the target sound. The data of the other participants are in the Supplementary material.

Figure 3 shows raw self-rotation trajectories of Participant 1. Visual inspection suggests that the participant was usually turning towards the target and was standing still in the Static condition. A notable difference between the AV and A-only condition is could be seen for the frontal target (Figure 3; first column) where we see much more variance and more complexity in the A-only condition.

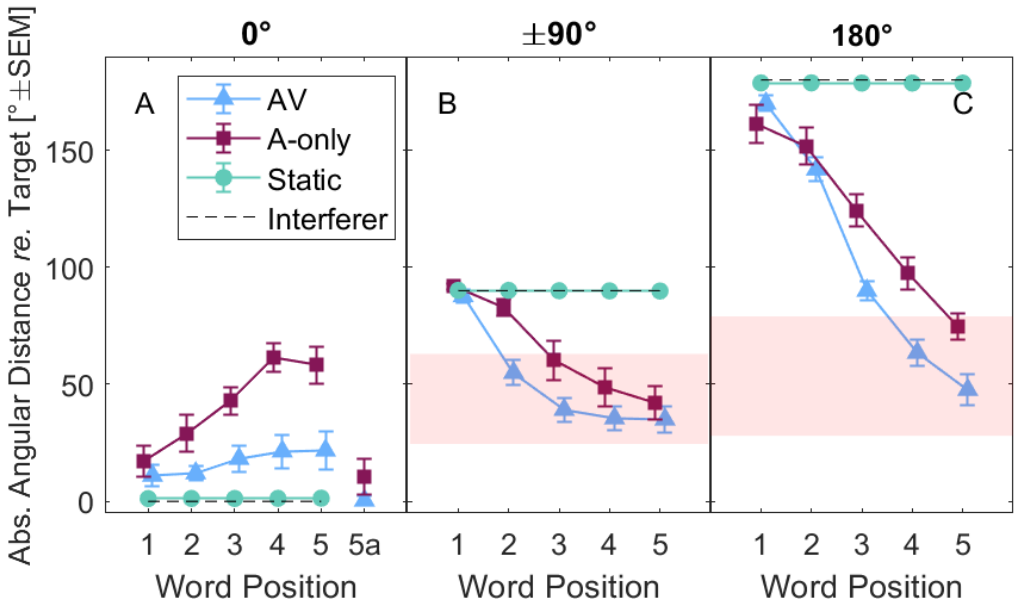


Figure 4 – Mean absolute angular distance to the target as a function of word position. Panels A-C show different target angles. Panel A also shows the mean angular orientation at the final word (5a on x-axis) – positive means to the left. The middle panel (B) pools the data for the left and the right target angles. The red-shaded area is an acoustically optimal region determined from Jelfs et al. (2011) model, the maximum output with a 1 dB margin. Data show across-subject means.

To see the trends across subject, Figure 4 analyzes participants' self-orienting behavior in terms of absolute angular distance to target during the target sentence as a function of word position. We test whether the visual cue condition AV (▲) and A-only (■) had an effect on the orientation. We also test whether the trajectory falls into the region with optimal speech intelligibility benefit, expressed as a deviation of less than 1 dB from the best possible unmasking according to Jelfs et al. (2011). For the frontal target, the trajectory endpoint is shown with the '5a' label because the mean trajectory differs from the mean of the absolute angular distances to the target. For the lateral and the rear targets, the mean trajectories are closely represented by the absolute angular distance shown on the figures. The optimal region is shown by the red-shaded area in Figure 4 B and C. The Static (●) data are shown as reference values when the participants were instructed to stand still.

Overall, the data in Figure 4 show that the angular distance was closer to the target in the AV condition than in the A-only condition ($F(1,8610)=40.62$, $p<0.001$), and this was the case for

all target directions (the interaction of target and condition was insignificant). However, due to different profiles of movement trajectories, the data show a significant three-way interaction ($F(8,8610)=3.12$, $p=0.017$). This relates to different trajectory endpoints for the lateral and rear targets. The statistical tests were conducted using a three-way ANOVA on GLM (conditions AV and A-only), we report only the significant effects which involve factor of condition.

For the frontal target (Figure 4A), the absolute distance to the target for the final word was at 24° in the AV condition at the trajectory endpoint, while it was 58° in the A-only condition. The mean trajectory endpoints (‘5a’) were at zero (t-test for AV and A-only: $p>0.05$) indicating that people were rotating to the left or right around the target, but on average they turned to the target. To analyze further whether people preferred to orient towards the target or away from the target at the trajectory endpoints, we counted the percentage of endpoints that were within 10° of angular distance from the target and the percentage of endpoints that were between 10° and 90°. People were pointing towards the target in the 42% of trials in the AV condition and 9% of trials in the A-only condition, while they turned away from target in 56% trials in the AV condition and in 69% trials in the A-only. The percentages are computed across all trials and participants.

For the lateral targets (Figure 4 B), the angular distance endpoint was at 31° in the AV condition and 44° in the A-only condition and both these values fall well within the range of optimal orientation. In the AV condition, four out of five target words fall within the optimal region, in the A-only condition, three words fall into the region. For the rear targets (Figure 4 C), the trajectory endpoint was at 44° in the AV condition and 74° in the A-only condition. Both these values fall within the range of optimal orientation. In the AV condition, two words fall into the optimal region, while for the A-only condition, only one word just falls into the optimal region.

The analysis shows that people oriented closer towards the target in the AV condition than in the A-only condition. However, in most cases, they maintain a strong off-target bias, this was the case also for the frontal target, even though there was no acoustical benefit of self-rotation (no acoustically optimal region). This can be explained by a ‘polite’ communication strategy, since people prefer to maintain an off-center self-orientation towards other people. Although, in the A-only condition, the data may indicate a ‘search strategy’ (higher variability and less likely to end at the target direction). For the lateral and rear targets we observed that self-orientation was often in the acoustically optimal region, which was more quickly reached with visual cues, but orientation did not approach the visual target. Therefore, it seems likely that people employ various strategies when positioning themselves in an acoustic scene.

Speech intelligibility during self-rotation

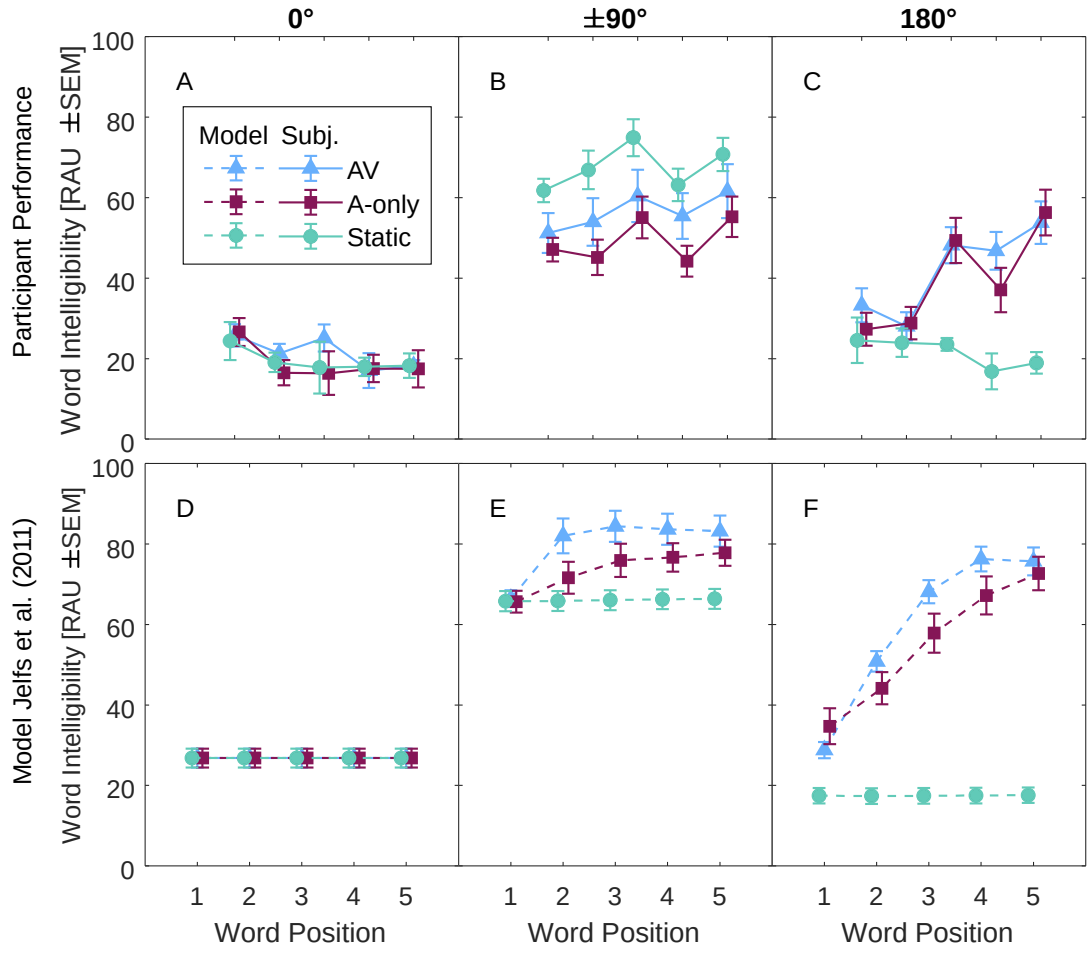


Figure 5 – Analysis of intelligibility of each word in the sentence in terms of participant performance and model predictions (Jelfs et al., 2011) for conditions with self-rotation (Blue triangles – AV, burgundy squares – A-only) and for the static baseline (green circles – Static). The abscissa shows word position in the target sentence. A,D - data for the frontal target, B,E - combined data of the lateral targets, C,F - data of the rear target.

Figure 5 shows participant's speech intelligibility (top row, solid lines) and model predictions (bottom row, dashed lines). The symbols code conditions; different subpanels show different target angles.

Static Baseline

The speech intelligibility in the Static condition (Figure 5, green circles) was worst for the frontal and the rear target which is expected for a (binaurally) co-located target and masker. For the lateral target, intelligibility improved. Word position only slightly influenced performance. A two-way ANOVA on GLM with factors target angle (0°, ±90°, 180°), and word position (1-5) confirmed these trends showing that the factor target angle

(F(1,4305)=221.12; p<0.001) was significant while the factor word position and the interaction were not significant. The effect of target location can be explained by spatial unmasking in the different target configurations. The frontal co-located target and rear target provide no spatial unmasking, while there is a substantial spatial unmasking for the lateral target. The Static condition data were used to obtain the parameters for the prediction model.

Effect of self-rotation

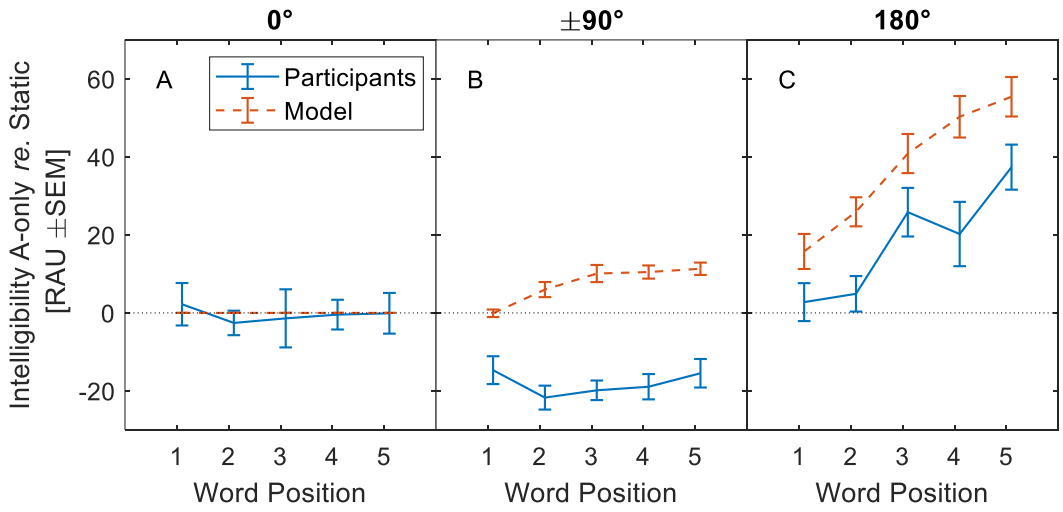


Figure 6 – Effect of self-rotation on speech intelligibility. Data show change of intelligibility scores in RAU due to movement (A-only vs. Static) as a function of word position. Solid blue lines – participants' data, dashed lines – model predictions. (A-C) Data are split according to target location.

In Figure 5, we could see the effect of self-rotation by comparing the Static and the A-only conditions (green and burgundy lines). For the frontal target (Figure 5A), the A-only data overlap with the Static, therefore we observe no effect. For the lateral targets (Figure 5B), the overall intelligibility improved relative to the frontal target, due to spatial unmasking, but the speech intelligibility in the A-only did not improve as much as the Static, which indicates a negative effect of self-rotation. Speech intelligibility for the rear target (Figure 5C), was as low as for the frontal target in the Static condition but continuously improved during the movement in the A-only condition.

Figure 6 further analyses the contrasts of A-only and the Static conditions in terms of participant performance (solid lines) and model predictions (dashed lines). The participant data show no effect for the frontal target, a negative effect for the lateral target and a positive effect with an increasing trend for the rear target. A statistical analysis was conducted using a three-way ANOVA on a GLM with factors of target angle (0°, ±90°, 180°), condition (Static, A-only), and word position (1-5) and we analyzed the interactions with the factor of condition (to see the effect of self-rotation). The analysis showed an interaction of target angle and condition ($F(2,8610)= 34.705$, $p<0.001$), and the three-way interaction of target angle, word position and condition ($F(8,8610)= 2.06$, $p=0.035$). Further, we conducted partial post-hoc analysis by fitting new GLMs for each target angle. The analysis showed a main effect of condition ($F(1,4310)= 15.993$, $p<0.001$) for the lateral angles. Further it showed a main effect of condition ($F(1,2150)= 30.819$, $p<0.001$) and significant interaction of condition and word position ($F(4,2150)= 5.7109$, $p<0.001$) for the rear angle (Bonferroni-Holm corrected α for 12 additional tests). To confirm the trends of the performance in the A-only condition vs. Static condition, we performed four additional paired one-tailed t-tests (Bonferroni-Holm corrected) for the lateral and the rear targets (data averaged across word positions), which confirmed the negative trend (worse performance in A-only than the Static, $t(8)=-13.31$, $p<0.001$) for the lateral targets and the positive trend for the rear targets ($t(8) =5.77$, $p<0.001$).

The model data on Figure 5D-F represent predictions of a speech intelligibility model that considers each participant's self-orientation trajectories and they generally follow the trends of the participant data, but overall they overestimate the participant performance. Dashed lines on Figure 6 analyze the contrast between the A-only and Static. The predictions are trivial for the frontal target since the model predicts no benefit for the co-located target and interferer. However, the model shows performance better than the Static condition and it overestimates the participant's performance for the lateral and rear target. The magnitude of overestimation slightly varies with word position. We conducted a statistical evaluation of the

model predictions (per cent correct for each trial are obtained from a GLM model, same as on Figure 5D-F) and participant data (per cent correct for each trial are obtained from a GLM model with factors of target angle (0°, ±90°, 180°), condition (AV, A-only, Static), word position (1-5) that was fit on all participant data). The per cent correct values were RAU transformed and the participant data were subtracted from the model predictions. The difference data were further fit to a GLM with factors of target angle (0°, ±90°, 180°), condition (Static, A-only), and word position (1-5). An ANOVA showed a significant intercept ($F(1,8610)= 204.96$,$p<0.001$), indicating a significant difference between the model and the participants data across all conditions. Further it showed a significant interaction of target and condition ($F(2, 8610)= 118.94$, $p<0.001$), and word position and condition ($F(4, 8610)= 11.29$, $p<0.001$), and a three-way interaction of target, word position and condition ($F(8, 8610)= 6.3859$,$p<0.001$). Partial GLM models for each target showed a significant main effect of condition (lateral targets: $F(1, 4310)=54.31$, $p<0.001$; rear target: $F(1,2150)= 64.693$, $p<0.001$) and the interaction of the condition and word position (lateral targets: $F(4, 4310)= 5.1682$, $p<0.001$; rear target: $F(4, 2150)= 6.2582$, $p<0.001$). All these effects are significant after Bonferroni-Holm correction. These results show strongly significant differences between participant performance and model predictions for lateral and rear target angles.

The analysis indicates that the movement had a significant effect on speech intelligibility. The increase of speech intelligibility when the people were moving towards the rear target could be related the change in spatial unmasking. On the other hand, the decrease of speech intelligibility for the lateral targets could not be accounted to it. The predictions of the model indicate that changes in spatial unmasking due to self-rotation could not be fully translated to speech intelligibility benefits.

Effect of visual cues that indicate target location

We assess the effect of the visual cue indicating the target location by comparing the participant performance in the AV and A-only conditions. In Figure 5, the difference can be seen by comparing the burgundy (AV) and the blue (A-only) conditions. Although the visual inspection may indicate an effect for the lateral (Figure 5B) and the rear targets (Figure 5C), the statistical analysis did not show any significant interactions with the factor condition ($p>0.05$). The statistical test was conducted on the participant responses using the GLM model with factors of target angle (0°, ±90°, 180°), condition (AV, A-only), and word position (1-5). The partial tests for individual target directions did not show any significant results either. Post-hoc Bonferroni-Holm correction was applied.

To get further insight into the effect of the visual cue on behavior, we question how the predictions of the model (Figure 5D-F; AV vs A-only) were influenced by the visual cue condition through the altered self-rotating behavior. We analyzed the predictions of the model using a GLM model with factors of target angle (0°, ±90°, 180°), condition (AV, A-only), and word position (1-5) and it showed a significant interaction of condition and word position ($F(4,8610)= 2.5812$, $p=0.035$). We conducted a post-hoc analysis with Bonferroni- Holm correction using ANOVA on the GLMs for all target locations, but it did not show significant interactions that involved the factor of condition ($p>0.05$).

These results indicate that speech intelligibility was not statistically significantly different in the A-only and AV conditions. We found a significant difference between the conditions in terms of the model outputs, but the magnitude of the effect is rather small. This however confirms that people had a tendency to move in a way that would be beneficial for speech intelligibility but the effect on speech intelligibility was limited in the current situation.

Discussion

Self-rotation behavior

Participants preferred to maintain an off-target self-orientation for all target directions, even for the co-located target and interferer, a situation in which the participants could not obtain a speech intelligibility benefit by self-rotation. We speculate that this can reflect a polite strategy or a social bias. For instance, such undershooting has been previously observed in studies with participants who had live conversations (Lu et al., 2021; Vertegaal et al., 2001) but has also been observed in sound localization studies (Lewald et al., 2000) when people were asked to turn their head towards a sound source or a visual target. Such underestimation in the previous studies could have been related to the sitting position, while in the current study the participants were standing and freely rotating. On the other hand, the behavior was acoustically optimal for the lateral and rear targets. When people were orienting towards lateral or rear targets, their endpoint usually fell within the acoustically optimal region, despite them not being explicitly instructed to find the best possible head orientation, a prerequisite seen in other studies (Grange et al., 2018). In laboratory studies, participant's behavior might be affected by the novelty of the test environment, which may hinder a natural behavioral response. When participants have no, or very little, instructions, the situation might be awkward for them. On the other hand, giving explicit instructions may lead to exaggerated movements not typical for social situations. In our experiment, we asked participants to mimic natural behavior, and we cued them indirectly by indicating the target location. Although the virtual character did not provide speech intelligibility cues, the visual indication of location decreased the angular distance from the target, which had a slight effect on model predictions. We cannot completely tease apart different strategies, but it is possible that people combine a learned social strategy for communication with a strategy, in which they seek an acoustical benefit to improve speech intelligibility. It is also possible, that they employ a search strategy in case they have problems localizing the sound.

Speech intelligibility

Baseline

When participants were standing still, speech intelligibility was poor for the frontal and rear targets, and it was better for the lateral targets due to spatial unmasking. In similar experimental conditions as here in the Static condition, Best et al. (2018) observed an increase of speech intelligibility in a series of digit words, which they called a build-up of speech intelligibility. Although our statistical test did not indicate a significant interaction of target and word position, this could have been due to splitting the sentence into five words while the speech material was calibrated per word groups (Wagener, Brand, et al., 1999). This could have led to a speech intelligibility imbalance on the fourth word (Llorach & Hohmann, 2019). Furthermore, the study by Best et al. (2018) included only target-interferer configurations which were spatially separated and therefore people could experience spatial unmasking in all trials. Our further statistical testing by the individual target angles did not show significant results, therefore we cannot the support the existence of the effect (despite the data of Figure 5B show a trend in the expected direction) but also we cannot exclude it.

The effect of motion

The comparison of the Static and A-only conditions showed that speech intelligibility improved during self-rotation toward the rear target, but it slightly decreased during self-rotation toward the lateral targets and it did not change for the frontal target. The improvement should be related to the change in spatial unmasking during the rotation. However, the decrease could not be related to the change in self-orientation and spatial unmasking. Especially, the decrease seems to be a constant offset even at the beginning, when the movement did not have a big effect. If there was no spatial unmasking possible, intelligibility did not decrease nor increase. These results suggest that movement has an effect

on speech intelligibility but only if the target and interferer are spatially separated to evoke spatial unmasking.

The contribution of spatial unmasking was investigated with a model predicting speech intelligibility across the head turn. Predictions follow trends in the participant data. For instance, the predicted intelligibility for the lateral target is higher than for the frontal target, and the predicted speech intelligibility for the rear target increases during the sentence due to the head rotation. An intriguing observation is that the model predicts an increase in intelligibility for the lateral target with respect to the Static condition, and a steep increase in speech intelligibility for movement toward the rear target. The trend is consistent with the performance for the rear target although the model overestimates participant's performance. However, for the lateral target, the model goes in the opposite direction than the behavioral data. This indicates that the reduction in speech intelligibility relates to the perceptual organization of the acoustic scene, which is not captured by the model.

A possible mechanism could be that the self-rotation limits the benefits of spatial unmasking due to binaural sluggishness (Grantham & Wightman, 1978, 1979) which could be modeled by applying an integration window into or after the cue extraction stage (Hauth & Brand, 2018; Kolotzek et al., 2021). In this context, the speech model used in our study could be seen as a model without temporal limits, which extracts all cues optimally, a possible reason for it outperforming the participants.

The non-acoustic factors could also interact with the perceptual organization, for instance on attentional level such that people head to realign the visual references after rotation with the acoustic representation. In terms of the self-motion cues, Kondo et al. (2012) showed that they did not play a role in the 'resetting' of the perceptual organization of the acoustic scene, however, they used a classical streaming paradigm. Therefore it is not clear how that would translate to the speech perception task.

Frissen et al. (2019) studied the effect of speech-irrelevant head movements on speech intelligibility and observed a small, almost negligible negative effect of head movements on speech intelligibility. A reason that the effect was not larger might be that the possible dynamic spatial unmasking is limited in complex scenes with multiple maskers, as used in their study. Shen et al. (2017) measured speech intelligibility during head rotation with seated participants. Although the data suggest a slight benefit for non-head turners, the difference was not significant. This could relate to the small spatial separation of the target and interferer sound, as well as the type of head movements in their experimental conditions, which were limited to a small range.

Effect of visual cues

The comparison of the AV and A-only conditions showed only a modest change in speech intelligibility for the lateral target (but not statistically significant), and no change of intelligibility due to the visual cue for the rear target. Although the self-rotating behavior changed in the AV condition with respect to the A-only condition, the change in trajectories likely did not bring sufficient change to speech intelligibility. Possibly, the room for improvement was not big enough. The analysis of model predictions showed similar trends, but the small significant effect suggests that the participants were likely to use the visual cue for their benefit.

Hendrikse et al. (2018) observed that target location cues helped participants to identify the speech target out of multiple options, therefore they suggested prior knowledge of the target position could help participants orient in the scene. In our experiment, people were visually cued at the onset of the target while in Hendrikse et al. (2018) the target phrase started with a keyword and the target words were presented only two seconds later, which may have helped with attentional focus. In addition, they used multiple competing talkers as distractors, which

might have required more attentional resources to focus on the target than with the continuous noise masker used in our study.

Limiting factors

The speech intelligibility model employed in this study does not take into account speech masking (Rennies et al., 2011), but our speech target included reverberation. This may have affected the outcomes of the speech model; however, in our evaluations we use relative comparisons, which would minimize this type of effect. Additionally, the model uses head-related transfer functions (HRTFs) recorded from an artificial head in situ in the experimental setup. Individualized HRTFs capture individual signal-to-noise ratio differences, interaural cross-correlation differences and binaural cues with higher fidelity; however, the general trends should also be reproduced with non-individual HRTFs. In the analysis using HRTFs, we considered only horizontal rotations of the whole upper body (manikin), but shoulder reflections for different head orientations can alter interaural cues at higher frequencies substantially (Kolotzek, 2017). Nevertheless, we do not assume that these acoustic effects would substantially change the outcomes of the study.

In the experimental design, we aimed to align the orientation of the reference target angle (0°) with respect to the listener always at the beginning of each trial. However, the participant's orientation was recorded about a second or two before the onset the acoustic stimulus (due to a lag in the computer program), thus in some trials people may have moved and were offset form the 0° during the masker onset (see Supplemental material). The angles are correctly taken into account by the acoustic model, but in addition to that we also reanalyzed the behavioral data without three participants where this problem was the most prominent and this did not affect the results.

Conclusions

H1: The study showed that people used an acoustically optimal strategy of self-rotation in situations in which they could get a spatial benefit for speech intelligibility. However, they exhibited similar movement patterns in situations, in which they could not obtain any benefit. Thus, people either use a mixture of strategies (learned social strategy, search strategy and acoustically optimal strategy) or a learned social strategy, which might be optimal in many cases.

H2: We did not find evidence supporting an effect of visual location cueing on speech intelligibility; the movement patterns in the AV were only slightly more accurate and closer to be acoustically optimal. The speech intelligibility model outputs (obtained from head rotations) indicated a small benefit of the visual cues, suggesting that the more direct movement may provide only a limited benefit in terms of speech intelligibility.

H3: The effect of movement on speech intelligibility can be partly, but not fully, accounted to the self-rotation-induced changes of spatial unmasking, indicating that speech intelligibility could be reduced during self-rotation. This may relate to a mechanism that limits the access to cues of spatial unmasking during self-rotation.

Acknowledgement

We thank all participants and all who helped during the piloting phase of the experiment.

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

Declaration of conflicting interests

The Authors declare that there is no conflict of interest.

References

Beechey, T., Buchholz, J. M., & Keidser, G. (2018a). Measuring communication difficulty through effortful speech production during conversation. *Speech Communication*, *100*(June 2017), 18–29. https://doi.org/10.1016/j.specom.2018.04.007

Beechey, T., Buchholz, J. M., & Keidser, G. (2018b). Measuring communication difficulty through effortful speech production during conversation. *Speech Communication*, *100*, 18–29. https://doi.org/10.1016/j.specom.2018.04.007

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), 473–484. https://doi.org/10.3766/jaaa.16.7.7

Berzborn, M., Bomhardt, R., Klein, J., Richter, J.-G., & Vorländer, M. (2017). The ITA-Toolbox: An Open Source MATLAB Toolbox for Acoustic Measurements and Signal Processing. *43th Annual German Congress on Acoustics, Kiel, Germany*, *June*, 222–225.

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

Best, V., Ozmeral, E. J., & Shinn-Cunningham, B. G. (2007). Visually-guided Attention Enhances Target Identification in a Complex Auditory Scene. *Journal of the Association for Research in Otolaryngology*, *8*(2), 294–304. https://doi.org/10.1007/s10162-007-0073-z

Best, V., Swaminathan, J., Kopčo, N., Roverud, E., & Shinn-Cunningham, B. (2018). A "Buildup" of Speech Intelligibility in Listeners With Normal Hearing and Hearing Loss. *Trends in Hearing*, *22*, 233121651880751. https://doi.org/10.1177/2331216518807519

Borish, J. (1984). Extension of the image model to arbitrary polyhedra. *The Journal of the Acoustical Society of America*, *75*(6), 1827–1836. https://doi.org/10.1121/1.390983

Brimijoin, W. O., McShefferty, D., & Akeroyd, M. A. (2010). Auditory and visual orienting responses in listeners with and without hearing-impairment. *The Journal of the Acoustical Society of America*, *127*(6), 3678–3688. https://doi.org/10.1121/1.3409488

Brimijoin, W. O., McShefferty, D., & Akeroyd, M. A. (2012). Undirected head movements of listeners with

asymmetrical hearing impairment during a speech-in-noise task. *Hearing Research*, *283*(1–2), 162–168. https://doi.org/10.1016/j.heares.2011.10.009

Bronkhorst, A. W. (2000). The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions. In *Acta Acustica united with Acustica* (Vol. 86, Issue 1, pp. 117–128). https://doi.org/10.1306/74D710F5-2B21-11D7-8648000102C1865D

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., Barrett, M. E., Cohen, J. I., Fodor, C., Yancey, C. M., & Gordon-Salant, S. (2020). Objective Assessment of Speech Intelligibility in Crowded Public Spaces. *Ear & Hearing*, *41*(Supplement 1), 68S-78S. https://doi.org/10.1097/AUD.0000000000000943

Cheyne, H. A., Kalgaonkar, K., Clements, M., & Zurek, P. (2009). Talker-to-listener distance effects on speech production and perception. *The Journal of the Acoustical Society of America*, *126*(4), 2052. https://doi.org/10.1121/1.3205400

Ching, T. Y. C., O'Brien, A., Dillon, H., Chalupper, J., Hartley, L., Hartley, D., Raicevich, G., & Hain, J. (2009). Directional Effects on Infants and Young Children in Real Life: Implications for Amplification. *Journal of Speech Language and Hearing Research*, *52*(5), 1241–1254. https://doi.org/10.1044/1092-4388(2009/08-0261)

Cord, M. T., Surr, R. K., Walden, B. E., & 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), 353–364. https://doi.org/10.3766/jaaa.15.5.3

Culling, J. F., Hawley, M. L., & Litovsky, R. Y. (2004). The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. *The Journal of the Acoustical Society of America*, *116*(2), 1057–1065. https://doi.org/10.1121/1.1772396

Davis, T. J., Grantham, D. W., & Gifford, R. H. (2016). Effect of motion on speech recognition. *Hearing Research*, *337*, 80–88. https://doi.org/10.1016/j.heares.2016.05.011

Frissen, I., Scherzer, J., & Yao, H.-Y. (2019). The Impact of Speech-Irrelevant Head Movements on Speech Intelligibility in Multi-Talker Environments. *Acta Acustica United with Acustica*, *105*(6), 1286–1290.

https://doi.org/10.3813/AAA.919408

Gonzalez-Franco, M., Maselli, A., Florencio, D., Smolyanskiy, N., & Zhang, Z. (2017). Concurrent talking in immersive virtual reality: on the dominance of visual speech cues. *Scientific Reports*, *7*(1), 3817. https://doi.org/10.1038/s41598-017-04201-x

Grange, J. A., & Culling, J. F. (2016a). The benefit of head orientation to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *139*(2), 703–712. https://doi.org/10.1121/1.4941655

Grange, J. A., & Culling, J. F. (2016b). Head orientation benefit to speech intelligibility in noise for cochlear implant users and in realistic listening conditions. *The Journal of the Acoustical Society of America*, *140*(6), 4061–4072. https://doi.org/10.1121/1.4968515

Grange, J. A., Culling, J. F., Bardsley, B., Mackinney, L. I., Hughes, S. E., & Backhouse, S. S. (2018). Turn an Ear to Hear: How Hearing-Impaired Listeners Can Exploit Head Orientation to Enhance Their Speech Intelligibility in Noisy Social Settings. *Trends in Hearing*, *22*, 1–13. https://doi.org/10.1177/2331216518802701

Grantham, D. W., & Wightman, F. L. (1978). Detectability of varying interaural temporal differences. *The Journal of the Acoustical Society of America*, *63*(2), 511–523. https://doi.org/10.1121/1.381751

Grantham, D. W., & Wightman, F. L. (1979). Detectability of a pulsed tone in the presence of a masker with time-varying interaural correlation. *The Journal of the Acoustical Society of America*, *65*(6), 1509–1517. https://doi.org/10.1121/1.382915

Grimm, G., Hendrikse, M. M. E., & Hohmann, V. (2020). Review of Self-Motion in the Context of Hearing and Hearing Device Research. *Ear & Hearing*, *41*(Supplement 1), 48S-55S. https://doi.org/10.1097/AUD.0000000000000940

Hadley, L. V., Brimijoin, W. O., & Whitmer, W. M. (2019). Speech, movement, and gaze behaviours during dyadic conversation in noise. *Scientific Reports*, *9*(1), 10451. https://doi.org/10.1038/s41598-019-46416-0

Hauth, C. F., & Brand, T. (2018). Modeling sluggishness in binaural unmasking of speech for maskers with time-varying interaural phase differences. *Trends in Hearing*, *22*, 1–10. https://doi.org/10.1177/2331216517753547

Hendrikse, M. M. E., Llorach, G., Grimm, G., & Hohmann, V. (2018). Influence of visual cues on head and eye movements during listening tasks in multi-talker audiovisual environments with animated characters.

Speech Communication, *101*(June 2017), 70–84. https://doi.org/10.1016/j.specom.2018.05.008

Hendrikse, M. M. E., Llorach, G., Hohmann, V., & Grimm, G. (2019). Movement and Gaze Behavior in Virtual Audiovisual Listening Environments Resembling Everyday Life. *Trends in Hearing*, *23*, 233121651987236. https://doi.org/10.1177/2331216519872362

Hládek, Ľ., Porr, B., Naylor, G., Lunner, T., & Owen Brimijoin, W. (2019). On the Interaction of Head and Gaze Control With Acoustic Beam Width of a Simulated Beamformer in a Two-Talker Scenario. *Trends in Hearing*, *23*, 233121651987679. https://doi.org/10.1177/2331216519876795

Jelfs, S., Culling, J. F., & Lavandier, M. (2011). Revision and validation of a binaural model for speech intelligibility in noise. *Hearing Research*, *275*(1–2), 96–104. https://doi.org/10.1016/j.heares.2010.12.005

Kerber, S., & Seeber, B. U. (2011). Towards quantifying cochlear implant localization performance in complex acoustic environments. *Cochlear Implants International*, *12*(sup2), S27–S29. https://doi.org/10.1179/146701011X13074645127351

Kerber, S., & Seeber, B. U. (2013). Localization in Reverberation with Cochlear Implants. *Journal of the Association for Research in Otolaryngology*, *14*(3), 379–392. https://doi.org/10.1007/s10162-013-0378-z

Kidd, G., Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005). The advantage of knowing where to listen. *The Journal of the Acoustical Society of America*, *118*(6), 3804–3815. https://doi.org/10.1121/1.2109187

Kishline, L. R., Colburn, S. W., & Robinson, P. W. (2020). A multimedia speech corpus for audio visual research in virtual reality (L). *The Journal of the Acoustical Society of America*, *148*(2), 492–495. https://doi.org/10.1121/10.0001670

Kolotzek, N. F. (2017). *Efect of head turning on localization in the horizontal plane with hearing aid satellites* [Technical University of Munich, Germany]. (Unpublished master's dissertation)

Kolotzek, N. F., Aublin, P. G., & Seeber, B. U. (2021). Fast processing explains the effect of sound reflections on binaural unmasking. *Submitted*.

Kondo, H. M., Pressnitzer, D., Toshima, I., & Kashino, M. (2012). Effects of self-motion on auditory scene analysis. *Proceedings of the National Academy of Sciences*, *109*(17), 6775–6780. https://doi.org/10.1073/pnas.1112852109

Latif, N., Barbosa, A. V., Vatiokiotis-Bateson, E., Castelhano, M. S., & Munhall, K. G. (2014). Movement

coordination during conversation. *PLoS ONE*, *9*(8), 1–10. https://doi.org/10.1371/journal.pone.0105036

Lewald, J., Dörrscheidt, G. J., & Ehrenstein, W. H. (2000). Sound localization with eccentric head position. *Behavioural Brain Research*, *108*(2), 105–125. https://doi.org/10.1016/S0166-4328(99)00141-2

Llorach, G., & Hohmann, V. (2019). Word error and confusion patterns in an audiovisual German matrix sentence test (OLSA). *Proceedings of the 23rd International Congress on Acoustics, September*, 5749–5751.

Lu, H., McKinney, M. F., Zhang, T., & Oxenham, A. J. (2021). Investigating age, hearing loss, and background noise effects on speaker-targeted head and eye movements in three-way conversations. *The Journal of the Acoustical Society of America*, *149*(3), 1889–1900. https://doi.org/10.1121/10.0003707

MakeHuman (1.2.0 alpha3). (2019). http://www.makehumancommunity.org

Marrone, N., Mason, C. R., & Kidd, G. (2008). *The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms*. https://doi.org/10.1121/1.2980441

Pastore, M. T., & Yost, W. A. (2017). Spatial Release from Masking with a Moving Target. *Frontiers in Psychology*, *8*(DEC), 1–8. https://doi.org/10.3389/fpsyg.2017.02238

Rennies, J., Brand, T., & Kollmeier, B. (2011). Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet. *The Journal of the Acoustical Society of America*, *130*(5), 2999–3012. https://doi.org/10.1121/1.3641368

Seeber, B. U., & Clapp, S. W. (2017). Interactive simulation and free-field auralization of acoustic space with the rtSOFE. *The Journal of the Acoustical Society of America*, *141*(5), 3974–3974. https://doi.org/10.1121/1.4989063

Seeber, B. U., Kerber, S., & Hafter, E. R. (2010). A system to simulate and reproduce audio–visual environments for spatial hearing research. *Hearing Research*, *260*(1–2), 1–10. https://doi.org/10.1016/j.heares.2009.11.004

Shen, Y., Folkerts, M. L., & Richards, V. M. (2017). Head movements while recognizing speech arriving from behind. *The Journal of the Acoustical Society of America*, *141*(2), EL108–EL114. https://doi.org/10.1121/1.4976111

Søndergaard, P., & Majdak, P. (2013). The Auditory Modeling Toolbox. In J. Blauert (Ed.), *The Technology of*

Binaural Listening (pp. 33–56). Springer.

Stecker, G. C. (2019). Using Virtual Reality to Assess Auditory Performance. *The Hearing Journal*, *72*(6), 20. https://doi.org/10.1097/01.HJ.0000558464.75151.52

Stitt, P., Bertet, S., & van Walstijn, M. (2016). Extended Energy Vector Prediction of Ambisonically Reproduced Image Direction at Off-Center Listening Positions. *Journal of the Audio Engineering Society*, *64*(5), 299–310. https://doi.org/10.17743/jaes.2016.0008

Studebaker, G. A. (1985). A "rationalized" arcsine transform. *Journal of Speech and Hearing Research*, *28*(3), 455–462.

Vertegaal, R., Slagter, R., van der Veer, G., & Nijholt, A. (2001). Eye gaze patterns in conversations. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '01*, 301–308. https://doi.org/10.1145/365024.365119

Viveros Muñoz, R., Aspöck, L., & Fels, J. (2019). Spatial Release From Masking Under Different Reverberant Conditions in Young and Elderly Subjects: Effect of Moving or Stationary Maskers at Circular and Radial Conditions. *Journal of Speech, Language, and Hearing Research*, *62*(9), 3582–3595. https://doi.org/10.1044/2019_JSLHR-H-19-0092

Wagener, K. C., Brand, T., & Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests in deutscher Sprache Teil II: Optimierung des Oldenburger Satztests. *ZEITSCHRIFT FUR AUDIOLOGIE*, *38*(2), 44–56.

Wagener, K. C., Kühnel, V., & Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. *ZEITSCHRIFT FUR AUDIOLOGIE*, *38*(1), 4–15.

Figure 1 – Position of the subject and of target and interferer sounds in the virtual room.

Figure 2 – Experimental procedures and conditions. The top graph shows an example of four experimental trials with time on the abscissa and target angle relative to the participant on the ordinate. Each trial (one sentence and noise) was followed by a response. The bottom part of the figure shows three experimental conditions and the corresponding instructions.

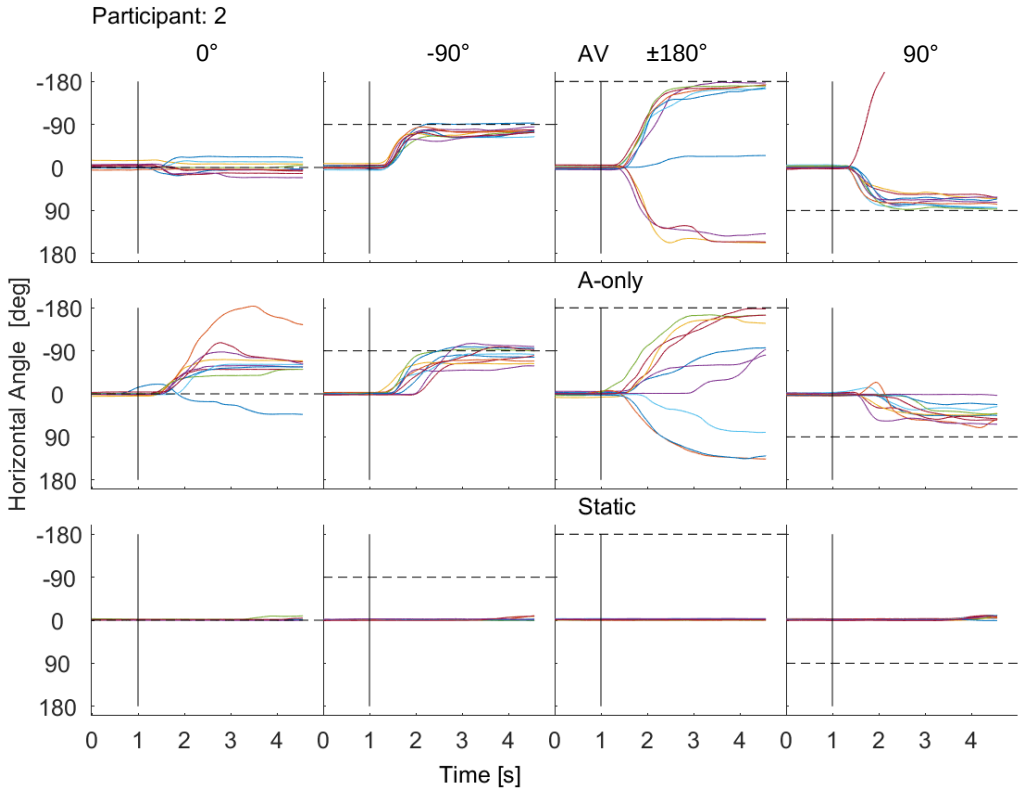
Figure 3 – An example of raw head orientation trajectories for one participant. The first row shows data for the AV, the second row for the A-only and the third row for the Static condition. The columns show data for different target angles which are indicated by dashed lines. The vertical line at the time of 1 second indicates the onset of the target sound. The data of the other participants are in the Supplementary material.

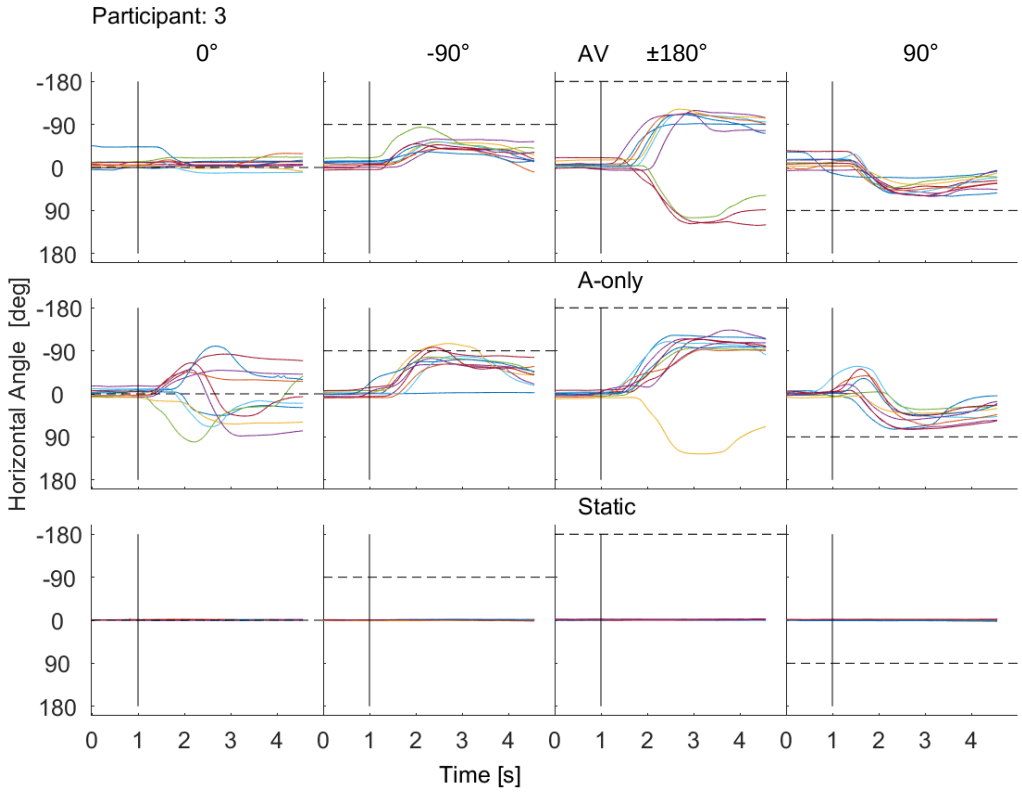
Figure 4 – Mean absolute angular distance to the target as a function of word position. Panels A-C show different target angles. Panel A also shows the mean angular orientation at the final word (5a on x-axis) – positive means to the left. The middle panel (B) pools the data for the left and the right target angles. The red-shaded area is an acoustically optimal region determined from Jelfs et al. (2011) model, the maximum output with a 1 dB margin. Data show across-subject means.

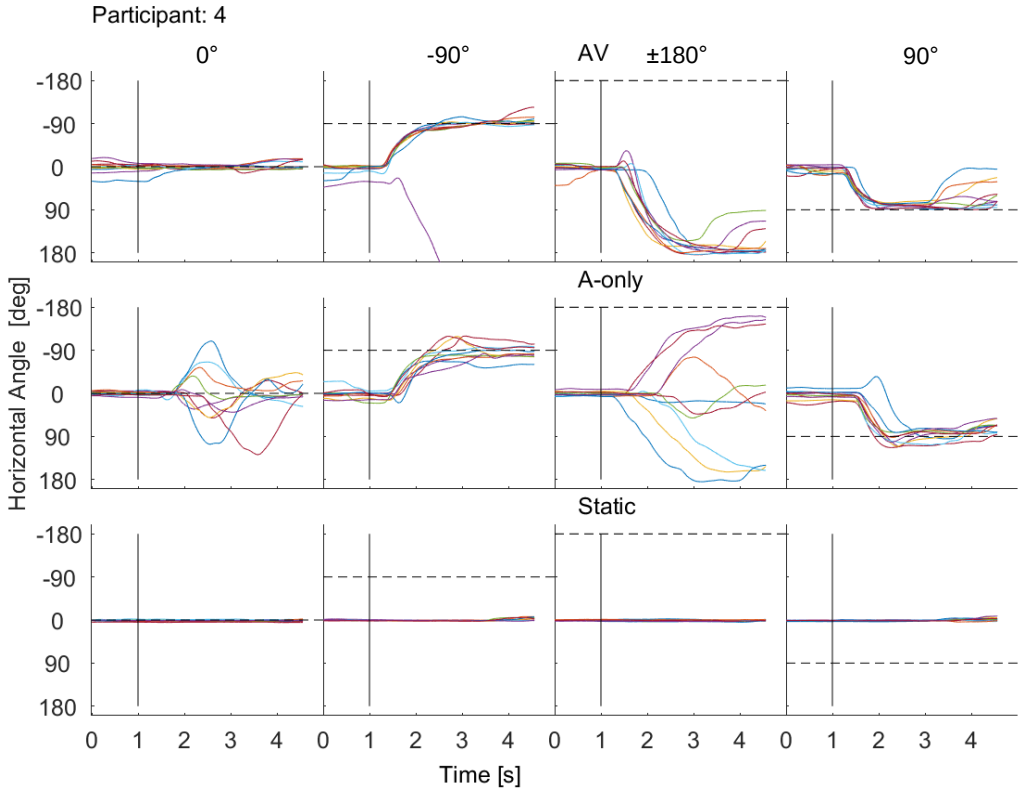
Figure 5 – Analysis of intelligibility of each word in the sentence in terms of participant performance and model predictions (Jelfs et al., 2011) for conditions with self-rotation (Blue triangles – AV, burgundy squares – A-only) and for the static baseline (green circles – Static). The abscissa shows word position in the target sentence. A,D - data for the frontal target, B,E - combined data of the lateral targets, C,F - data of the rear target.

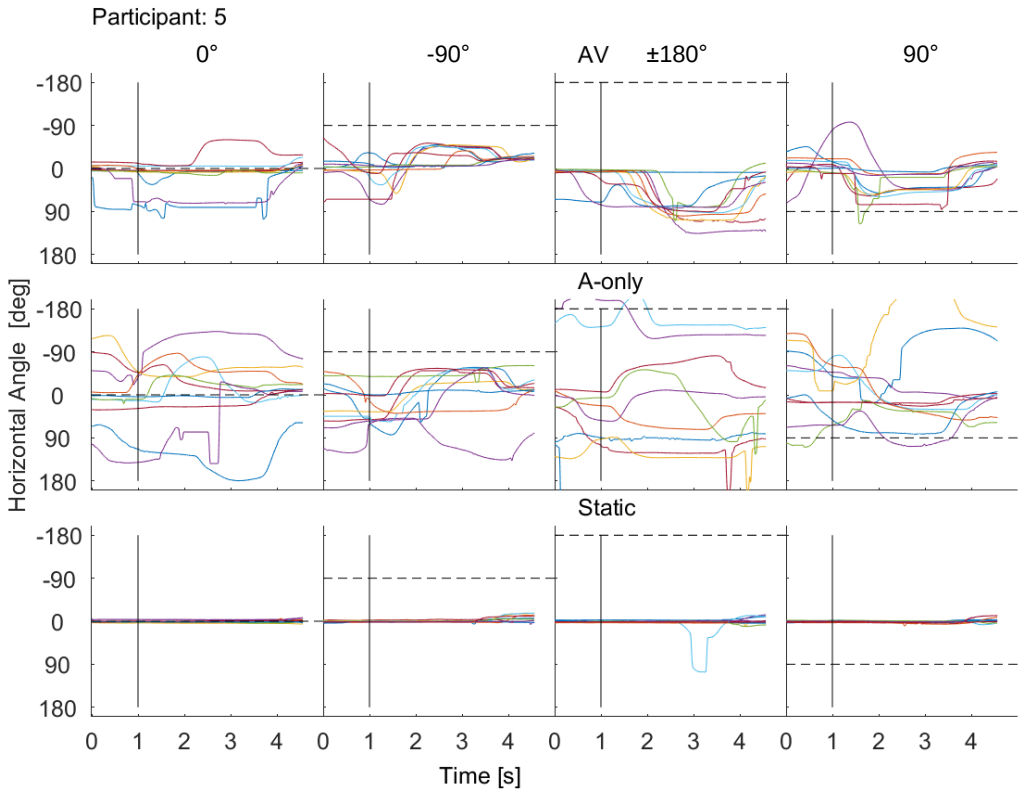
Figure 6 – Effect of self-rotation on speech intelligibility. Data show change of intelligibility scores in RAU due to movement (A-only vs. Static) as a function of word position. Solid blue lines – participants' data, dashed lines – model predictions. (A-C) Data are split according to target location.

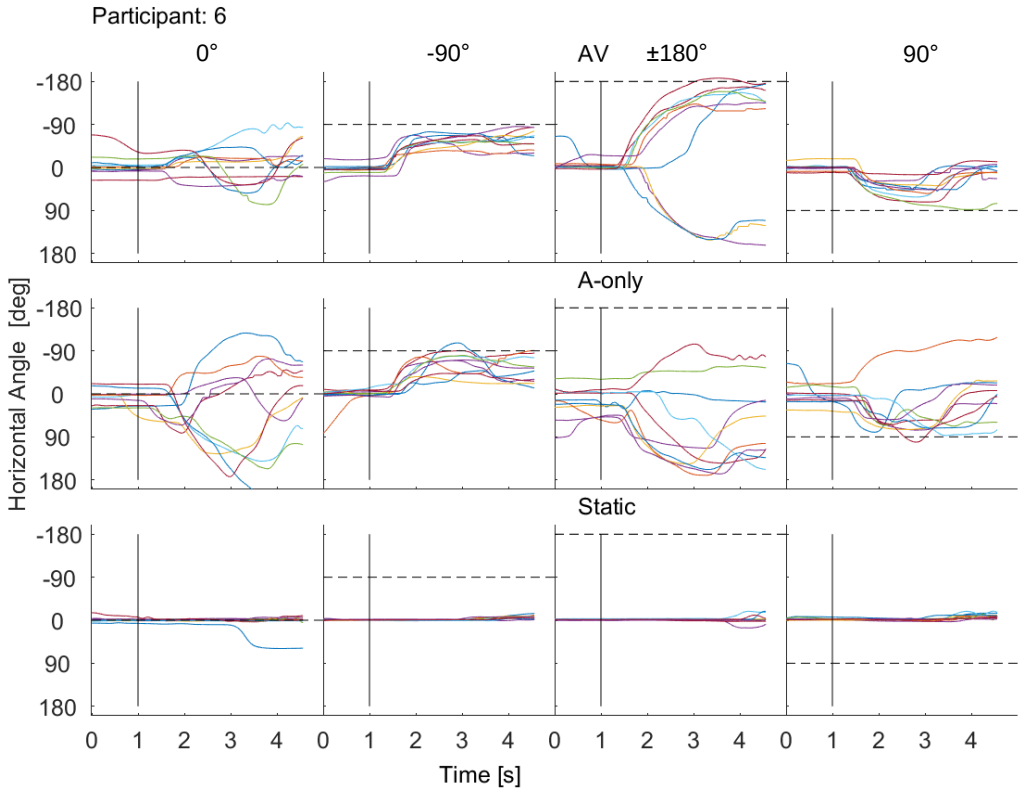
Supplemental material

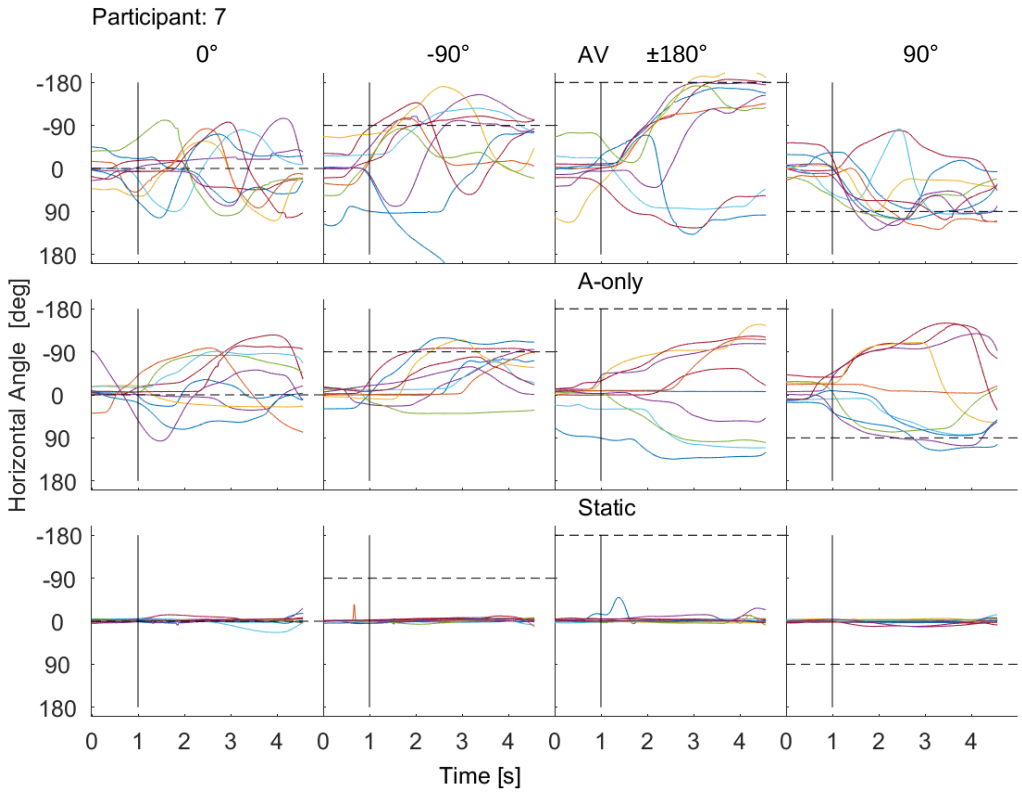

Participant: 2
0°
-90°
AV ±180°
90°
A-only
Static
Horizontal Angle [deg]
-180
-90
0
90
180
0 1 2 3 4
Time [s]

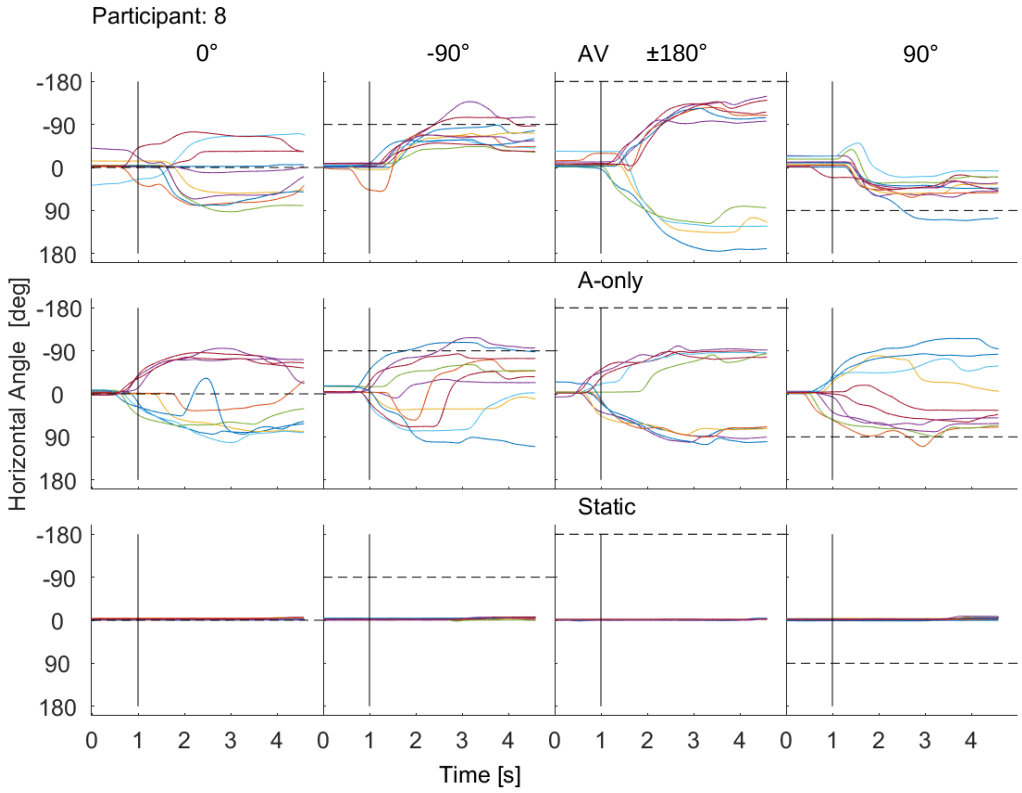

Participant: 3
0°
-90°
AV ±180°
90°
A-only
Static
Horizontal Angle [deg]
-180
-90
0
90
180
0 1 2 3 4
Time [s]

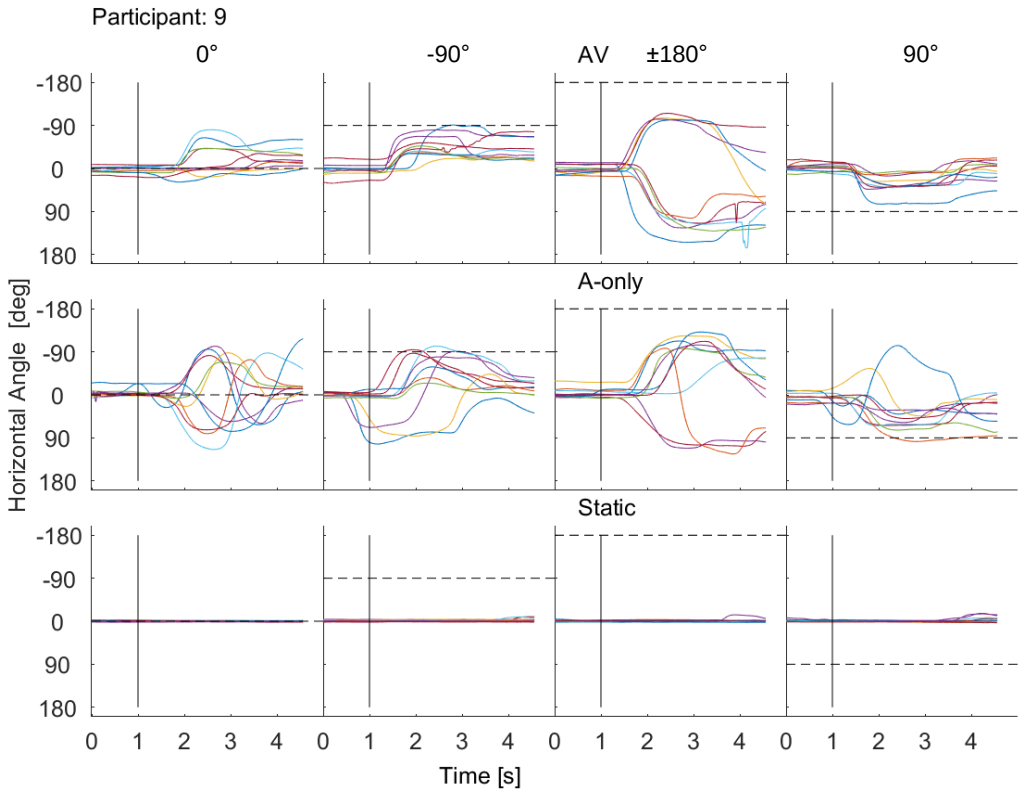

Participant: 4
0°
-90°
AV
±180°
90°
A-only
Static
Horizontal Angle [deg]
-180
-90
0
90
180
0 1 2 3 4
Time [s]


Participant: 5
0°
-90°
AV
±180°
90°
A-only
Static
Horizontal Angle [deg]
-180
-90
0
90
180
0 1 2 3 4
Time [s]


Participant: 6
0°
-90°
AV
±180°
90°
A-only
Static
Horizontal Angle [deg]
-180
-90
0
90
180
0 1 2 3 4
Time [s]


Participant: 7
0°
-90°
AV
±180°
90°
A-only
Static
Horizontal Angle [deg]
-180
-90
0
90
180
0 1 2 3 4
Time [s]


Participant: 8
0°
-90°
AV
±180°
90°
A-only
Static
Horizontal Angle [deg]
-180
-90
0
90
180
0 1 2 3 4
Time [s]


Participant: 9
0°
-90°
AV
±180°
90°
A-only
Static
Horizontal Angle [deg]
-180
-90
0
90
180
0 1 2 3 4
Time [s]