NOVEMBER 14 2025

Exploring auditory selective attention shifts in virtual reality: An approach with matrix sentences

Carolin Breuer ; Janina Fels

J. Acoust. Soc. Am. 158, 3805–3813 (2025)
https://doi.org/10.1121/10.0039864

JASA ARTICLE

Exploring auditory selective attention shifts in virtual reality: An approach with matrix sentences

Carolin Breuer[a] and Janina Fels
Institute for Hearing Technology and Acoustics, RWTH Aachen University, Aachen 52074, Germany

ABSTRACT:
This study explores the voluntary switching of auditory selective attention using more natural stimuli and complex acoustic conditions. Building on previous categorization tasks with single-word stimuli, we introduce unpredictable matrix sentences in German to simulate more realistic auditory environments. While the overall results were similar to previous versions, no strong effect of reorienting the auditory attention was found. Interaction effects in error rates still suggest that switching auditory attention is more demanding than remaining focused on the same target. The results further show a benefit in reaction of preparing attention, since reaction times were highest for target words at the beginning of the sentence and decreased for later target onsets. Findings further suggest an opposite trend in error rates, where target words in the beginning yield fewer errors than target words in the middle or end of a sentence in switch trials (8.9% vs 15.7% vs 14.7%), especially when the distractor is played later than the target. Taken together, this approach offers a paradigm for investigating auditory attention in more complex acoustic scenarios, advancing research on auditory perception in dynamic room acoustic environments.
 https://doi.org/10.1121/10.0039864
(Received 11 September 2024; revised 12 September 2025; accepted 16 October 2025; published online 14 November 2025)
[Editor: Jonas Braasch] Pages: 3805–3813

I. INTRODUCTION

Auditory selective attention (ASA) is known to be influenced by both the nature of the auditory stimuli and the surrounding acoustic conditions. A classic illustration of ASA is the "cocktail party effect," where multiple sound sources compete for attention, making it difficult to focus on a single target speaker.[1] In his investigations, Cherry[1] showed that participants were able to suppress information presented to one ear while focusing on the other ear. Many studies reproduced these findings and investigated how selective attention can be disturbed. One prominent stimulus to disrupt the attentional focus is one's own name.[2] This finding suggests that even the unattended acoustic stimuli are processed semantically, which requires involuntary attention switches to the unattended stimuli.[3,4] For detailed reviews on ASA, see Bronkhorst[5] or Shinn-Cunningham and Best.[6] However, the ASA can also be shifted voluntarily, e.g., by choosing to listen to a different talker at the cocktail party. Here, attention can be redirected to a specific speaker or location. For example, Kidd *et al.*[7] showed that when participants were asked to focus on a specific target location, they could identify target words more easily when listening to three simultaneously played sound sources, than just listening to the three sources without being cued.[7] However, the switching of ASA between target genders[8] or target locations[9–12] introduces switch costs in terms of longer reaction times (RTs) and higher error rates (ERs) in classifying target stimuli.

One paradigm to investigate voluntary shifts in ASA was developed by Koch *et al.*[8] In this dichotic paradigm, participants were presented with numbers between one and nine, excluding five. One target and one distracting number were played simultaneously but each to one ear only. In this experiment, female and male voices were used, and participants were cued which gender to attend to. The task was to classify whether the target played a number larger or smaller than five. Between trials, the target gender could either be repeated or switched. The latter was expected to induce auditory switch costs. Notably, both target and distractor stimuli could play congruent numbers belonging to the same category (e.g., both below five) or incongruent numbers (e.g., target below five and distractor above five). Koch *et al.*[8] found main effects of attention switch as well as congruence. Thus, the reorientation to a different target gender was more demanding than maintaining the focus. Further, the congruence effect suggests that the unattended distractor stimulus was processed. This paradigm was later enhanced by Fels *et al.*[11] and Oberem and Fels[13] introducing binaural spatial cues to more closely mimic real-world listening environments. Using binaural technology, they extended the paradigm to include eight spatial positions separated by 45° each. In this new version, the spatial location of the target was changed between trials rather than their gender. Fels *et al.*[11] found the same tendencies for attention shifts and congruence in the binaural version as Koch *et al.*[8] using the dichotic paradigm. Thus, spatial auditory attention shift in a more realistic context could be studied. Further, effects regarding reproduction accuracy introduced by state-of-the-art

[a]Email: carolin.breuer@akustik.rwth-aachen.de

acoustic reproduction, such as (non-) individualized binaural playback or head-tracking as opposed to real sound sources, were investigated.[14] Here, Oberem *et al.*[14] confirmed that the use of non-individualized binaural reproduction shows the same tendencies as individualized head-related transfer functions and real loudspeakers. This put the necessity of highly accurate acoustic simulations into the perspective of human perception.

To also include visual aspects, the described base paradigm was further extended to a virtual reality (VR) version, in which the participants are immersed in a VR classroom while performing the classification task.[15] Being motivated to investigate developmental effects, which are suspected to evolve during pre- and elementary school and reach an adult-like level at an age of eight to eleven years,[16–18] a child-appropriate version using animal names instead of number words was developed[19,20] and employed for the VR scenario. Here, animals need to be classified into flying (e.g., bee) and non-flying (e.g., cat) animals.

However, to create a realistic acoustic scenario, room acoustic properties also need to be included. Although several parameters can be used to characterize rooms, most often the acoustic properties are discussed in terms of the reverberation time. International recommendations for reverberation times in unoccupied rooms range from 0.3 s for good to 0.9 s for bad elementary school classrooms. Measured values in unoccupied classrooms range from 0.2 to 1.9 s.[21–23]

To investigate the influence of reverberation on ASA, Oberem and co-workers[24,25] extended the binaural paradigm by three reverberation conditions ($RT_{60} = 0$, 0.4, and 3 s). They found no influence of reverberation on ASA and hypothesized that the original stimuli described previously, with a length of approximately 730 ms for the digits and 600 ms for the animal names,[13,20] were too short. Such stimuli are not only artificial, but real-world artifacts, such as room acoustics, could not be addressed. Thus, longer should be used to investigate the influence of reverberation time on the voluntary switching of ASA.

Previous work to extend the stimulus lengths has been conducted by Fels *et al.*[11] where the number words were extended by direction words, i.e., "up" and "down," leading to a stimulus length of 1200 ms. Thus, the participants had to classify a more complex target, e.g., "up 3," leading to four different answer categories ("up/down" and "smaller/greater than 5").[11,26] Similarly, the child-appropriate paradigm used small and large animals to be categorized.[27,28] For the adults, the results showed higher RTs and ERs using the long stimuli than single-digit words. The difference can be explained by an increased task difficulty, since participants had to classify the stimuli into four instead of two categories. Still, the same trends in attention switch and congruence were found as in previous studies. Thus, the paradigm proved to be robust against the stimulus extension. Based on these findings, an investigation on reverberation time was conducted.[26,29] Here, Oberem *et al.*[29] used a low reverberation time of 0.8 s and a higher reverberation time of 1.75 s and found that ASA is impaired under high reverberation conditions. This was reflected in increased switch costs under the high reverberant condition, indicated by prolonged RTs. However, the RTs remained mostly unaffected in switch conditions, which indicates that redirecting the attention in itself is as demanding, so that reverberation did not increase task difficulty in RTs. Further, the congruence effect in ERs was higher under high reverberant conditions as opposed to the anechoic condition. This indicates that reverberation significantly impaired the ability to filter relevant information from the target.

A drawback to these approaches is the more complex task, i.e., classify within four instead of two categories, as well as the still rather artificial stimulus, which is not representative of speech occurring in real scenarios.

The current study addresses these limitations by employing full sentences containing target words. This offers a more naturalistic and ecologically valid stimulus set that mirrors real speech scenarios. By using matrix sentences, a format often used in speech intelligibility testing, such as the coordinate response measure[7,30–32] or the Oldenburg Sentence Test,[33] the semantic content of the sentences remains unpredictable and irrelevant to the task, maintaining focus on ASA. This innovation not only simplifies the classification task (reducing from four categories[11,26–28] back to two[15,19]) but also allows for investigating the impact of reverberation time in a more realistic auditory context. When introducing sentences, one design choice is where to place the target word within the sentence. Placing the target word directly at the beginning of the sentence would resemble the previous task of only presenting single words. An alternative would be to place the target word in the middle or end of a sentence, which would represent a longer cue-stimulus interval (CSI). Such a variation has previously been investigated by Nolden *et al.*,[34] who varied the CSI between 400 ms and 1200 ms in a similar ASA task. They further varied the stimulus onset asynchrony (SOA) between target and distractor so the distractor could be presented before or after the target. They found a general benefit of a longer CSI in a way that participants responded faster when the CSI was 1200 ms than for 400 ms. The effect of SOA further indicated that participants responded more slowly when the target and distractor were presented simultaneously. However, this benefit was largest when the distractor was presented before the target.

Moreover, the study incorporates gamification elements, building on prior work by Breuer *et al.*[15] The study extends the ASA task into a child-appropriate VR environment, paving the way for future investigations into the developmental aspects of ASA during early school years. By bridging the gap between controlled experimental conditions and real-world auditory scenarios, this study sets the foundation for deeper insights into how ASA functions in complex, dynamic acoustic environments.

II. METHODS

A. Participants

A sample size of 12 participants to detect a main effect of attention transition (AT) was calculated using G*Power[35]

https://doi.org/10.1121/10.0039864

using an effect size estimated from a prior study[15] of $f = 0.335$, $\alpha = 0.05$, and a power of $1 - \beta = 0.95$. Nevertheless, in accordance with previous investigations,[8,14,15] 20 adult participants [age = 18–47 years, mean (M) = 28 years, standard deviation (SD) = 8.3 years, ten female] were recruited for the current study. All participants had good German language proficiency, normal hearing between 250 Hz and 8 kHz at a maximum of 25 dB hearing level[36] according to a pure-tone audiometry and (corrected to) normal vision according to Snellen charts.[37] The study was performed in accordance with the Declaration of Helsinki.[38] A statement of non-objection was obtained from the Medical Ethics Committee at RWTH Aachen University with the protocol number EK 395-19. All participants gave informed written consent before the study.

B. Experimental task

The presented study is an extension of the paradigm presented by Breuer *et al.*[15] As described in the introduction, the overall task was to classify spoken animal names into flying and non-flying animals. A target and distractor stimuli were played simultaneously from different spatial directions [front, back, left, or right related to the participant, see Fig. 1(a)]. Each trial started with an acoustic cue indicating the target position. The cue was the sound of snapping fingers played from the respective target position. Following a cue-stimulus-interval of 500 ms, a target and a distractor stimulus were played simultaneously from different positions. Participants responded by pressing a trigger button on a hand-held controller [see Fig. 1(b)] either with the left or right index finger. The assignment of which controller belonged to which animal category was balanced across the test subjects. After responding, the participants received feedback on whether their answer was correct. The feedback was given by a happy or sad face and the statement "correct" (German: richtig) or "false" (German: falsch), which was displayed for 500 ms on a virtual blackboard in front of the participant. The next trial started after a 500 ms inter-trial interval. Participants were instructed to answer as quickly and as accurately as possible. They were allowed to answer during stimulus playback. Thus, the RTs were measured from stimulus onset.

In contrast to previous studies, the target and distractor stimuli were each a sentence instead of a single animal name. The sentences were constructed as matrix sentences including an object, verb, number word, adjective, and animal name as depicted in Table I. The words were chosen to be well-known and easy to understand. Also, the number of syllables was kept constant in each word category. Furthermore, the sentences were rearranged so the animal name could be positioned at the beginning (e.g., Seals have eleven small flowers. / Robben haben elf kleine Blumen.), in the middle (e.g., Eleven small seals have flowers. / Elf kleine Robben haben Blumen.), or at the end of the sentence (e.g., Flowers have eleven small seals. / Blumen haben elf kleine Robben.). This was done to investigate whether the task difficulty changed with different word positioning. The situation most comparable to previous studies, which used only one target word, was placing the animal name at the beginning of the sentence. However, to investigate the impact of room acoustics, it seemed more appropriate to have the target word positioned later in the sentence to have an impact of, e.g., reverberation on the actual target word. Given that each word had slightly different lengths, the duration of one sentence was M = 2.011 s, SD = 0.138 s.

C. Audiovisual reproduction

During the experiment, the participants were immersed in a virtual classroom introduced by Breuer *et al.*[15] [see Fig. 1(b)]. The model and experiment logic were implemented using the Unity game engine[39] and are available via Zenodo.[40,41] The virtual environment was displayed on an HTC Vive pro eye head-mounted display (HMD) (HTC Vive, Taoyuan, Taiwan), and the corresponding controllers were used for user input. Images of a paw and wing were displayed on the virtual controller models to indicate the response category of flying and non-flying animals.

The acoustic stimuli were produced using the software Voicemaker[42] and the German voices "Kerryl" and "Bruno" for a female and male version. To incorporate prosodic features, complete sentences were generated and played back. Six possible sentences were created per target word, position within the sentence, and voice. During the experiment, one of the six alternatives was chosen randomly for playback. All stimuli were downloaded at a sample rate of 48 kHz in the uncompressed .wav format. The male voice was adjusted to fit the speed of the female voice using the Voicemaker interface. Male and female voices were assigned to the target and distractor randomly per trial. Target and distractor were always played back with a different voice.

To benefit from the virtual environment and the binaural reproduction, the participants were allowed to move their heads during the experiment. Head position and rotation were obtained from the HMD and allowed for a real-time adjustment of the binaural rendering using the Virtual Acoustics auralization framework[43] and the respective Unity package.[44] The same generic head-related transfer function of the artificial head with a resolution of $5^\circ \times 5^\circ$ was used for all participants.[45] Given that all instructions were displayed in front of the participants and the lack of visual representation of the sound sources, a main orientation to the front was expected. Also, a previous study by Breuer *et al.*[15] found no head rotations larger than 5° using the same virtual environment and experimental setup. The acoustic stimuli were played over Sennheiser HD650 open headphones (Sennheiser, Wedemark, Germany). Headphone equalization filters were measured and calculated for each participant[46] using the ITAtoolbox for MATLAB.[47] All stimuli were played back at 60 dB(A) measured with an artificial head.[45] The playback did not include any room acoustic simulation.

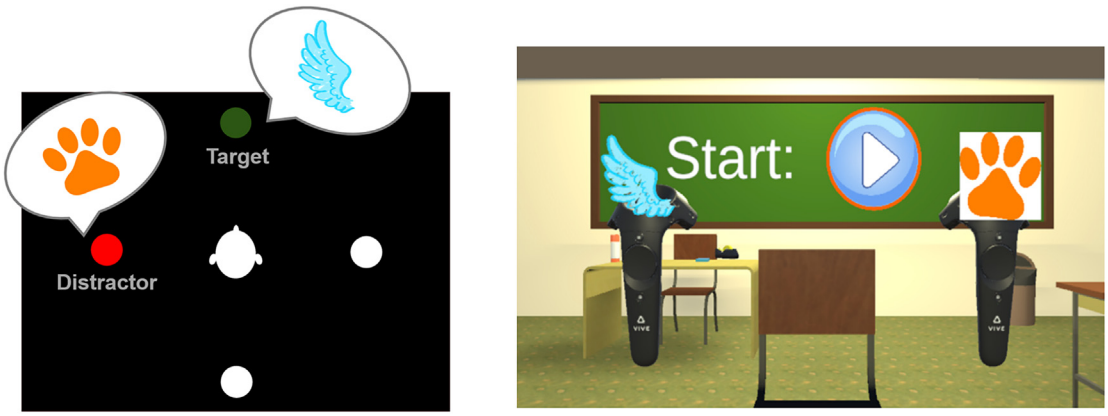

FIG. 1. (a) Graphical representation of a single trial. Overall, possible stimulus positions relative to the participant were front, back, left, or right. In this example, the target stimulus is placed in front of the participant and plays a flying animal name (e.g., bee). The distractor is located left of the participant and plays a non-flying animal name (e.g., cat). Thus, the depicted trial is incongruent. (b) virtual classroom from participant perspective, including virtual controller models. Reproduced from Breuer *et al.*, Int. J. Environ. Res. Public Health 19(24), 16569 (2022) licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (Ref. 15).

D. Experimental design

Given the spatial and content-related distribution of the stimuli presented in this paradigm, a number of variables were investigated. While the distractor position was changed each trial, the target stimulus could be presented from the same or a different position between trials. By changing the target position, an AT was provoked, which was expected to yield worse task performance than a repetition of the target position. The variable AT thus had two levels: switch and repetition. In line with previous studies,[11,15,19,29] the target-distractor position-combination, i.e., the spatial relation between the target and distractor stimulus, was altered trial-wise. Given the four spatial positions, three combinations were possible: front-back, left-right, and next-to. However, this variation was not further investigated, since it did not serve the present research question. In each trial, the content of the target and distractor stimulus, i.e., animal name category (flying or non-flying), could either be the same (congruent) or different (incongruent). This variable was called congruence (C). Based on the assumption that even task-irrelevant stimuli are processed, it was expected that task performance would be worse in incongruent than in congruent trials. Further, the position of the animal name within the target sentence (TS) and the distractor sentence was varied trial-wise. The positions could be beginning, middle, and end. Since participants were instructed to answer as quickly as possible, RTs are expected to be fastest for target positions in the beginning and longest for positions in the end. It was further hypothesized that the longer stimulus duration allows for an increased attentional focus, which results in lower ERs when the target word is positioned later in the sentence. The congruence between the target and distractor sentence position was manipulated [sentence congruence (SC)]. Thus, the target and distractor stimuli could either be placed at the same position or different positions in the sentence. A different positioning was expected to yield higher task performance, due to the attentional focus on the target stimulus.

Target performance was measured in terms of RTs (in ms) and ERs (in %).

Each variable was varied randomly per trial. All participants saw twelve trials per condition, and each participant received a unique experiment configuration. Before the experiment, all participants received a training of 32 trials to familiarize themselves with the task. The experiment was divided into six blocks with 48 trials each.

E. Data processing and statistical analysis

During the experiment, full sentences were played back in order to account for prosodic features. However, the RT

TABLE I. Words used to construct the sentences.

Object	Verb	Number	Adjective	Animal
Flowers (Blumen)	Have (haben)	Eleven (elf)	Small (kleine)	Seals (robben)
Mugs (Tassen)	Draw (malen)	Three (drei)	Expensive (teure)	Bees (bienen)
Cans (Dosen)	Catch (fangen)	Four (vier)	Pink (pinke)	Owls (eulen)
Socks (Socken)	Hunt (jagen)	Ten (zehn)	Yellow (gelbe)	Rats (ratten)
Pots (Töpfe)	Take (nehmen)	Nine (neun)	Pretty (schöne)	Ducks (enten)
Bags (Taschen)	Call (rufen)	Two (zwei)	Round (runde)	Doves (tauben)
Knives (Messer)	Love (lieben)	Five (fünf)	Wet (nasse)	Snakes (schlangen)
Shoes (Schuhe)	Like (mögen)	Six (sechs)	Old (alte)	Cats (katzen)

https://doi.org/10.1121/10.0039864

TABLE II. ANOVA (TS × SC × AT × C) for RT and ER.

	Reaction time				Error rate			
	df	*F*	*p*	η_p^2	*df*	*F*	*p*	η_p^2
TS	**(2, 38)**	**142.153**	**<0.001**	**0.882**	**(2, 38)**	**5.544**	**0.008**	**0.226**
SC	(1, 19)	1.343	0.261	0.066	(1, 19)	0.221	0.643	0.012
AT	(1, 19)	2.131	0.161	0.101	(1, 19)	0.027	0.871	0.001
C	(1, 19)	2.724	0.115	0.125	**(1, 19)**	**141.745**	**<0.001**	**0.882**
TS × SC	(2, 38)	2.306	0.113	0.108	(2, 38)	0.604	0.552	0.031
TS × AT	(1.557, 29.591)[a]	0.509	0.559	0.030	(2, 38)	0.910	0.411	0.046
TS × C	(2, 38)	2.317	0.112	0.109	(2, 38)	1.905	0.163	0.091
SC × AT	(1, 19)	2.205	0.145	0.108	(1, 19)	0.017	0.898	0.001
SC × C	(1, 19)	0.335	0.569	0.017	(1, 19)	0.899	0.355	0.045
AT × C	(1, 19)	2.312	0.145	0.108	(1, 19)	0.017	0.898	0.001
TS × AT × C	(2, 38)	1.025	0.369	0.051	(2, 38)	0.072	0.931	0.004
TS × SC × AT	(1.101, 20.924)[a]	0.140	0.736	0.007	**(2, 38)**	**3.568**	**0.038**	**0.158**
TS × SC × C	(2, 38)	0.809	0.453	0.041	(2, 38)	1.352	0.271	0.066
SC × AT × C	(1, 19)	0.159	0.694	0.008	(1, 19)	0.907	0.353	0.046
TS × SC × AT × C	(2, 38)	0.373	0.691	0.019	(2, 38)	0.852	0.434	0.043

[a]Greenhouse-Geisser correction applied due to sphericity. Significant main and interaction effects are highlighted.

was always measured starting from stimulus onset. Therefore, the measured RTs for target words in the middle and end of the sentence included the respective sentence duration until the target word. To account for this offset, the mean duration from stimulus onset to the target word was determined for each target word, position within the sentence, and voice. The calculated duration was subtracted from the measured RTs in the respective trials.

For the evaluation, normalized RTs in ms and ERs in % are considered. Training trials and the first trial of each block and trials following an error were removed. In accordance with previous studies,[8,11,19] trials with RTs below 50 ms and above 6000 ms were removed. Additionally, a Z-transformation of the RTs was performed and trials exceeding ±2 z were removed from the dataset as outliers (276 trials or 4.8%). For the RT evaluation, error trials were also removed. Repeated measures analyses of variance (ANOVAs) (TS × SC × AT × C) were performed for the RTs and ERs separately using IBM SPSS Statistics version 28.0.[48]

III. RESULTS

A full overview of the results is given in Table II, while only significant results are described in the text.

Regarding the RTs, the ANOVA showed only a significant main effect of TS, $F(2, 38) = 142.153$, $p < 0.001$, $\eta_p^2 = 0.882$ (see Fig. 2). Bonferroni adjusted *post hoc* tests indicated significantly higher RTs when the target word was in the beginning than in the middle or end of the sentence (1871.409 ms vs 1445.103 ms vs 985.009 ms, all $p < 0.001$). The difference between the target position in the middle and at the end was also significant ($p < 0.001$).

For the ERs, the ANOVA revealed a significant main effect of TS as well, $F(2, 38) = 5.544$, $p = 0.008$, $\eta_p^2 = 0.226$ (see Fig. 2). Bonferroni adjusted *post hoc* tests indicated significantly fewer errors when the target word was at the beginning than in the middle of the sentence (11.6% vs 15.0%, $p = 0.006$). The main effect of congruence was also significant, $F(1, 19) = 141.745$, $p < 0.001$, $\eta_p^2 = 0.882$, indicating significantly fewer errors in congruent than incongruent trials (4.9% vs 21.8%).

The three-way interaction of TS, SC, and transition in ERs was significant, $F(2, 38) = 3.568$, $p = 0.038$, $\eta_p^2 = 0.158$ (see Fig. 3). Bonferroni adjusted *post hoc* tests reveal that in switch trials, the difference between the target word at the beginning and middle (8.9% vs 15.7%, $p = 0.037$) as well as end position (8.9% vs 14.7%, $p = 0.021$) is significant for incongruent sentences. Further, in switch trials, the SC is significant when the target word is at the beginning. This is reflected in fewer errors in incongruent sentences than in congruent ones (8.9% vs 13.7%, $p = 0.006$).

IV. DISCUSSION

The present study extended an existing paradigm on ASA from using single-word stimuli to applying matrix sentences. Using a listening experiment with adult participants,

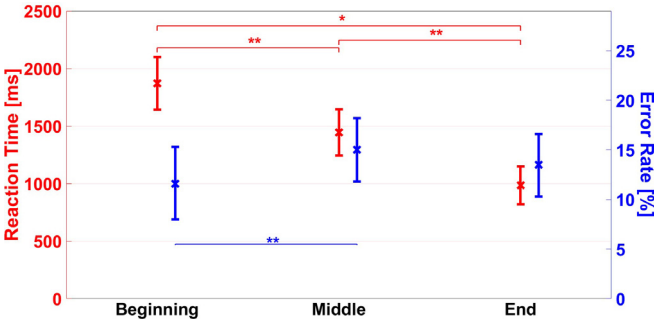

FIG. 2. Main effects by TS on ER (blue) and RT (red). Mean values and 95% confidence intervals are given. Significance levels are indicated by asterisks: ∗, $p \leqq 0.05$; ∗∗, $p \leqq 0.01$.

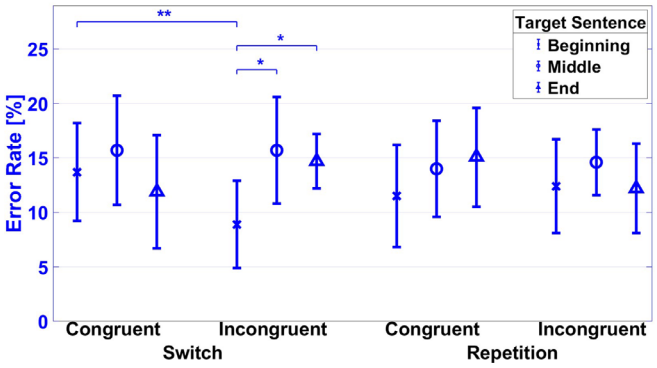

FIG. 3. Three-way interaction of TS, SC, and transition on ER (blue). Mean values and 95% confidence intervals are given. Significance levels are indicated by asterisks: $*, p \leqq 0.05$; $**, p \leqq 0.01$.

the impact of the longer stimulus material was investigated. It was expected that the overall trends of the original paradigm by Koch *et al.*[8] and the binaural version by Fels *et al.*[11] remained unaffected, while the target word position within the matrix sentence (beginning, middle, or end) as well as the position combination of target and distractor word within the sentence would influence the attentional preparation, thus simplifying the filtering of distracting words when the target word was positioned later in the sentence.

The main effect of the TS was present in RTs and ERs (see Fig. 2). The main effect in RTs reflects the onset of the target word within the sentence. Participants responded slowest for target words in the beginning and fastest for words in the end. This implies that the longer stimulus duration enhanced the attentional focus. Further, participants may have continued to listen to the sentences before responding when the target word was at the beginning or end. This is in line with previous results by Nolden *et al.*,[34] who also found a preparation effect for longer CSIs. Contrary to the hypothesis that a later position of the TS would lower task difficulty, the main effect in ERs only suggests that the difficulty is slightly increased when the target word is in the middle compared to in the beginning. This effect is not directly reflected in the three-way interaction of TS, SC, and AT (see Fig. 3). This interaction is only significant for switch trials, which are more challenging than repetition trials due to the reorientation of spatial attention. Here, trials in which the target word is placed in the beginning evoke significantly more errors when the distractor is also placed at the beginning than when the distractor is placed later in the sentence. This is in line with the auditory congruence effect also found in previous studies.[8,15] Interestingly, in switch trials, performance was best when the target was placed in the beginning and the distractor followed. However, no difference was found when the target was in the middle compared to at the end. Further, no difference between switch and repetition trials is indicated. Thus, only the reorientation of attention together with an early target presentation yielded better task performance than when the participants had more time to focus on the new spatial position.

In combination, this suggests that participants continued to listen even after the target word was presented. However, in switch trials, they answered more accurately when the target word was presented at the beginning of the sentence, especially when the distractor word was presented later in the sentence. The trend of fewer errors when the target is presented before the distractor is in line with previous findings by Nolden *et al.*[34] Also, the trend towards lower RTs for longer cue-target intervals is similar to the observations of Nolden *et al.*[34] This supports the hypothesis that the spatial attentional focus was overall stronger for longer stimulus exposure. Given the interaction with SC, an improved attentional focus was found when the target and distractor stimuli were played at different positions in the sentence. However, this effect was only found when the target was presented at the beginning of the sentence.

While the congruence effect regarding the stimulus content (flying vs non-flying) was found in the ERs as expected from previous studies,[8,15] the missing effect of AT, especially in RTs, is surprising. Although the switch cost introduced by rearranging the spatial attention is well-researched,[8–11] they were already only visible in RTs in previous studies using variations of the current paradigm.[15,19,29] Contrarily, the current study revealed only an interaction in ERs. This may be attributed to a floor effect introduced by the rather simple classification task. In the current study, no switch cost, i.e., no difference between switch and repetition trials, was found for RTs or ERs. However, a three-way interaction of TS, SC, and AT was revealed. Here, only significant differences within the switch condition were observed for target words at the beginning of a sentence. This is interpreted as the switch condition being more demanding and, thus, revealing even small effects on the auditory attention. Although smaller than expected, these findings together with the congruence effect are in line with Treisman's attenuation theory.[4]

Nevertheless, the question remains why the main effect of AT was not found. One possible reason is a lack of statistical power. Although the sample size was chosen according to an *a priori* power analysis using G*Power[35] and was in line with previous studies,[8,14,15] this approach does not capture the full complexity of the four-way repeated-measures ANOVA used here. Since G*Power does not support such designs, the analysis was necessarily based on a simplified model. More accurate estimations would require an *a priori* simulation-based approach that reflects the actual model structure and variance components, for example, using the simr package in R.[49,50] Still, the *a priori* estimation should have been sufficient to detect main effects such as AT. Future studies should nevertheless rely on simulation-based *a priori* power analyses to optimize sample size and repetitions for smaller effects. *Post hoc* power calculations were not performed, as they provide limited additional insight beyond the observed effect sizes and can be misleading regarding the true power of the study.[51] Importantly, the

effect sizes obtained here are comparable to previous investigations using the same ASA paradigm and virtual setup,[15] which supports the validity of the main conclusions.

At the same time, we consider it equally likely that the absence of a main AT effect is related to the paradigm itself. The use of sentences instead of single words, combined with the variable placement of target words within the sentences, may have masked switch costs by allowing participants more time to redirect attention. Thus, the mixture of stimulus types in the current design may have reduced the challenge of redirecting ASA. In addition, the use of longer stimuli and normalized RT measures may have further contributed to masking switch costs. To address this, future investigations considering different stimulus onsets should implement a trigger system or an adapted playback approach to measure RTs directly from the target rather than from the overall stimulus onset.

Another novelty introduced in this study was using synthetic rather than real speech. Synthetic speech could negatively influence speech intelligibility and, thus, worsen overall task performance. However, previous work by Nuesse *et al.*[52] suggests that matrix sentences created with synthetic speech can be used for speech recognition tests and should thus also be applicable for the presented study. Similar results were reported by Ibelings *et al.*[53] using the German Göttingen sentence test. Other concerns address the known training effects associated with matrix sentences, given the limited number of words within the speech corpus.[54] Since the original paradigm and variations[8,15,19] only use eight different target words (either numbers from one to nine excluding five or different animal names), additional training effects should not be relevant for solving this task. Further concerns could be the linguistic complexity introduced by placing the target at different positions within the sentence. Target positions at the beginning and end of a sentence are expected to have similar linguistic complexity, since two nouns change place.[55] In contrast to that, a target word in the middle also changes the sentence structure and can thus increase the complexity. The main effect of TS position on ERs would support this theory. However, the three-way interaction of the TS suggests that the linguistic complexity is not reflected in the ER results after all, since the interaction does not show a clear difference between beginning and middle position across conditions. Still, given that further experiments should use the most challenging stimulus version to detect even small impacts on ASA switches, the target position in the middle should be pursued.

V. CONCLUSION

The aim of the presented study was to extend a paradigm on the voluntary switching of ASA by longer stimuli in order to be able to investigate the impact of room acoustic parameters in upcoming studies. For that matter, unpredictable matrix sentences were introduced and tested in a listening experiment in a VR classroom and adult participants. In contrast to previous studies, the effect of AT introduced by a spatial reorientation of the selective attention was only found in an interaction with the positioning of the target word within a sentence, can the congruence of stimulus content. This interaction still indicated switch costs introduced by the voluntary switching of ASA. The effect of attention switches might be weakened by the different stimulus onsets and a too-long interval to reorient the attention between trials. For future investigations, the sentence configuration with the target word in the middle of the sentence is proposed, since this appears to be the most challenging version and is thus expected to be the most sensitive. Now, the next step towards creating a more close-to-real-life scenario is employing a room acoustic simulation in the classroom scenario and investigating the respective impact.

ACKNOWLEDGMENTS

The authors would like to thank Chenxin Ji for assisting in the paradigm design and data collection, as well as Cosima Ermert and Julia Seitz for their valuable discussions on the paradigm design and data evaluation. Parts of the research described in the paper was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Project ID No. 444697733 with the title "Evaluating cognitive performance in classroom scenarios using audiovisual virtual reality – ECoClass-VR." This project is part of the priority program "AUDICTIVE – SPP2236: Auditory Cognition in Interactive Virtual Environments."

AUTHOR DECLARATIONS

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethics Approval

The study was performed in accordance with the Declaration of Helsinki. A statement of non-objection was obtained from the Medical Ethics Committee at RWTH Aachen University with the protocol No. EK 395-19. All participants gave informed written consent before the study.

DATA AVAILABILITY

The software used to conduct the presented study, as well as the collected data are available in "Investigating the auditory selective attention switch using matrix sentences in VR" at https://zenodo.org/records/13736582 under the Creative Commons Attribution 4.0 International license.

[1]E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**(5), 975–979 (1953).
[2]N. Moray, "Attention in dichotic listening: Affective cues and the influence of instructions," Q. J. Exp. Psychol. **11**, 56–60 (1959).
[3]J. A. Deutsch and D. Deutsch, "Attention: Some theoretical considerations," Psychol. Rev. **70**(1), 80–90 (1963).
[4]A. M. Treisman, "Strategies and models of selective attention," Psychol. Rev. **76**(3), 282–299 (1969).

[5]A. W. Bronkhorst, "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," Atten. Percept. Psychophys. **77**(5), 1465–1487 (2015).
[6]B. Shinn-Cunningham and V. Best, "Auditory selective attention," in *The Handbook of Attention* (The MIT Press, Cambridge, MA, 2015), pp. 99–117.
[7]G. Kidd, Jr., T. L. Arbogast, C. R. Mason, and F. J. Gallun, "The advantage of knowing where to listen," J. Acoust. Soc. Am. **118**(6), 3804–3815 (2005).
[8]I. Koch, V. Lawo, J. Fels, and M. Vorländer, "Switching in the cocktail party: Exploring intentional control of auditory selective attention," J. Exp. Psychol.: Hum. Percept. Perform. **37**(4), 1140–1147 (2011).
[9]V. Best, E. J. Ozmeral, N. Kopčo, and B. G. Shinn-Cunningham, "Object continuity enhances selective auditory attention," Proc. Natl. Acad. Sci. U.S.A. **105**(35), 13174–13178 (2008).
[10]V. Best, B. G. Shinn-Cunningham, E. J. Ozmeral, and N. Kopčo, "Exploring the benefit of auditory spatial continuity," J. Acoust. Soc. Am. **127**(6), EL258–EL264 (2010).
[11]J. Fels, J. Oberem, and I. Koch, "Examining auditory selective attention in realistic, natural environments with an optimized paradigm," Proc. Mtgs. Acoust. **28**, 050001 (2016).
[12]A. Lavric and E. Schmied, "Preparatory switches of auditory spatial and non-spatial attention among simultaneous voices," J. Cognition **8**(1), 7 (2025).
[13]J. Oberem and J. Fels, "Speech material a paradigm intentional switching auditory selective attention," Technical Report No. RWTH-2020-02105, RWTH Publications (2020).
[14]J. Oberem, V. Lawo, I. Koch, and J. Fels, "Intentional switching in auditory selective attention: Exploring different binaural reproduction methods in an anechoic chamber," Acta Acust. united Ac. **100**(6), 1139–1148 (2014).
[15]C. Breuer, K. Loh, L. Leist, S. Fremerey, A. Raake, M. Klatte, and J. Fels, "Examining the auditory selective attention switch in a child-suited virtual reality classroom environment," Int. J. Environ. Res. Public Health **19**(24), 16569 (2022).
[16]A.-B. Doyle, "Listening to distraction: A developmental study of selective attention," J. Exp. Child Psychol. **15**(1), 100–115 (1973).
[17]P. R. Jones, D. R. Moore, and S. Amitay, "Development of auditory selective attention: Why children struggle to hear in noisy environments," Dev. Psychol. **51**(3), 353–369 (2015).
[18]J. Seitz, K. Loh, S. Nolden, and J. Fels, "Investigating intentional switching of spatial auditory selective attention in an experiment with preschool children," in *Proceedings of Jahrestagung Akustik, DAGA* (2023), Vol. 49.
[19]K. Loh, E. Fintor, S. Nolden, and J. Fels, "Children's intentional switching of auditory selective attention in spatial and noisy acoustic environments in comparison to adults," Dev. Psychol. **58**, 69–82 (2022).
[20]K. Loh and J. Fels, "ChildASA dataset: Speech and noise material for child-appropriate paradigms on auditory selective attention," Technical Report No. RWTH-2023-00740, RWTH Publications (2023).
[21]K. T. Mealings, "Classroom acoustic conditions: Understanding what is suitable through a review of national and international standards, recommendations and live classroom measurements," in *Proceedings Acoustics 2016: The Second Australasian Acoustical Societies Conference* (2016).
[22]A. Astolfi, G. E. Puglisi, S. Murgia, G. Minelli, F. Pellerey, A. Prato, and T. Sacco, "Influence of classroom acoustics on noise disturbance and well-being for first graders," Front. Psychol. **10**, 2736 (2019).
[23]K. Loh, M. Yadav, K. Persson Waye, M. Klatte, and J. Fels, "Toward child-appropriate acoustic measurement methods in primary schools and daycare centers," Front. Built Environ. **8**, 688847 (2022).
[24]J. Oberem, V. Lawo, I. Koch, and J. Fels, "Evaluation of experiments on auditory selective attention in an anechoic environmentand a reverberant room with nonindividual binaural reproduction," in *Proceedings of DAGA 2014, 40. Jahrestagung Akustik Fortschritte der Akustik* (2014).
[25]J. Oberem, "Examining auditory selective attention: From dichotic towards realistic environments," Ph.D. thesis, RWTH Aachen University, Aachen, Germany, 2020.
[26]J. Oberem, J. Seibold, I. Koch, and J. Fels, "Examining auditory selective attention in reverberant environments," J. Acoust. Soc. Am. **141**, 3691–3692 (2017).
[27]K. Loh, C. Hoog Antink, L. Mayer, and J. Fels, "Child-appropriate experiment on auditory selective attention in a virtual acoustic environment," in *Proceedings of Jahrestagung Akustik, DAGA 2020* (2020).
[28]K. Loh, C. Hoog Antink, S. Nolden, and J. Fels, "Combined assessment of cognitive and physiological parameters in child-appropriate listening experiments," in *Proceedings of Euronoise* (2021).
[29]J. Oberem, J. Seibold, I. Koch, and J. Fels, "Intentional switching in auditory selective attention: Exploring attention shifts with different reverberation times," Hear. Res. **359**, 32–39 (2018).
[30]R. S. Bolia, W. T. Nelson, M. A. Ericson, and B. D. Simpson, "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**(2), 1065–1066 (2000).
[31]D. S. Brungart, "Evaluation of speech intelligibility with the coordinate response measure," J. Acoust. Soc. Am. **109**(5), 2276–2279 (2001).
[32]W. Bologna, A. Carrillo, D. Clamage, L. Coco, Y. He, E. Lelo de Larrea Mancera, G. C. Stecker, F. Gallun, and A. Seitz, "Effects of gamification on assessment of spatial release from masking," Am. J. Audiol. **32**(1), 210–219 (2023).
[33]V. Kuehnel, B. Kollmeier, and K. Wagener, "Entwicklung und evaluation eines satztests für die deutsche sprache i: Design des oldenburger satztests (Development and evaluation of a sentence test for the German language I: Design of the Oldenburg sentence test)," Z. Audiologie **38**(1), 1–32 (1999).
[34]S. Nolden, C. N. Ibrahim, and I. Koch, "Cognitive control in the cocktail party: Preparing selective attention to dichotically presented voices supports distractor suppression," Atten. Percept. Psychophys. **81**(3), 727–737 (2019).
[35]F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," Behav. Res. Methods **39**, 175–191 (2007).
[36]World Health Organization, "Report of the informal working group on prevention of deafness and hearing impairment programme planning, Geneva, 18–21 June," https://iris.who.int/handle/10665/58839 (Last viewed 10 September 2024).
[37]H. Snellen, *Probebuchstaben Zur Bestimmung Der Sehschärfe* (*Sample Letters for Determining Visual Acuity*) (Van De Weijer, Utrecht, the Netherlands, 1862).
[38]World Medical Association, "Declaration of Helsinki - ethical principles for medical research involving human subjects," https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ (Last viewed 10 September 2024).
[39]Unity Technologies, "Unity 2019 long-term support release," (Last viewed 10 September 2024).
[40]C. Breuer, K. Loh, and J. Fels, (2022) "Auditory selective attention switch in a virtual reality classroom environment," Zenodo. https://doi.org/10.5281/zenodo.7248832
[41]C. Breuer and J. Fels, (2024) "Impact of realistic noise scenarios on auditory selective attention switch in a virtual classroom environment," Zenodo. https://doi.org/10.5281/zenodo.12688235
[42]Voicemaker Technologies Pvt. Ltd, "Voicemaker" (Last viewed 02 August 2024).
[43]Institute for Hearing Technology and Acoustics, RWTH Aachen University, (2022) "Virtual acoustics - a real-time auralization framework for scientific research," Zenodo. https://doi.org/10.5281/zenodo.13744523
[44]Institute for Hearing Technology and Acoustics, RWTH Aachen University, "Virtual acoustics unity package," https://git.rwth-aachen.de/ita/vaunity_package (Last viewed 10 September 2024).
[45]A. Schmitz, "Ein neues digitales kunstkopfmesssystem" ("A new digital measurement system for artificial heads") Acoustica **4**(81), 416–420 (1995).
[46]B. Masiero and J. Fels, "Perceptually robust headphone equalization for binaural reproduction," in *Proceedings of Audio Engineering Society Convention 130* (2011).
[47]P. Dietrich, M. Guski, M. Pollow, M. Müller-Trapet, B. Masiero, R. Scharrer, and M. Vorlaender, "Ita-toolbox - An open source MATLAB toolbox for acousticians," in *Proceedings of Fortschritte der Akustik: DAGA 2012* (2012).
[48]IBM Corp, "IBM SPSS statistics for Windows (version 28.0) [computer software]," https://www.ibm.com/products/spss-statistics (Last viewed 9 November 2025).
[49]P. Green and C. J. MacLeod, "SIMR: An R package for power analysis of generalised linear mixed models by simulation," Methods Ecol. Evol. **7**(4), 493–498 (2016).
[50]R Core Team, "R: A language and environment for statistical computing" (R Foundation for Statistical Computing, Vienna, Austria, 2021), https://www.R-project.org/ (Last viewed 9 November 2025).
[51]Y. Zhang, R. Hedo, A. Rivera, R. Rull, S. Richardson, and X. M. Tu, "*Post hoc* power analysis: Is it an informative and meaningful analysis?," Gen. Psych. **32**, e100069 (2019).

[52]T. Nuesse, B. Wiercinski, T. Brand, and I. Holube, "Measuring speech recognition with a matrix test using synthetic speech," Trends Hear. **23**, 2331216519862982 (2019).
[53]S. Ibelings, T. Brand, and I. Holube, "Speech recognition and listening effort of meaningful sentences using synthetic speech," Trends Hear. **26**, 23312165221130656 (2022).
[54]J. Heeren, T. Nuesse, M. Latzel, I. Holube, V. Hohmann, K. Wagener, and M. Schulte, "The concurrent OLSA test: A method for speech recognition in multi-talker situations at fixed SNR," Trends Hear. **26**, 23312165221108257 (2022).
[55]E. Gibson, "Linguistic complexity: Locality of syntactic dependencies," Cognition **68**(1), 1–76 (1998).