Journal of Perceptual Imaging® 5: 000404-1–000404-10, 2022.

The Ventriloquist Effect is not Consistently Affected by Stimulus Realism†

Thirsa Huisman[1], Torsten Dau[1], Tobias Piechowiak[2], Ewen MacDonald[3]

[1]*Hearing Systems Section, Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark;* [2]*GN Hearing, GN ReSound, Region Hovedsteden, Denmark;* [3]*Department of Systems Design Engineering, University of Waterloo, Waterloo, Canada*
E-mail: thuis@dtu.dk

Abstract. *Despite more than 60 years of research, it has remained uncertain if and how realism affects the ventriloquist effect. Here, a sound localization experiment was run using spatially disparate audio-visual stimuli. The visual stimuli were presented using virtual reality, allowing for easy manipulation of the degree of realism of the stimuli. Starting from stimuli commonly used in ventriloquist experiments, i.e., a light flash and noise burst, a new factor was added or changed in each condition to investigate the effect of movement and realism without confounding the effects of an increased temporal correlation of the audio-visual stimuli. First, a distractor task was introduced to ensure that participants fixated their eye gaze during the experiment. Next, movement was added to the visual stimuli while maintaining a similar temporal correlation between the stimuli. Finally, by changing the stimuli from the flash and noise stimuli to the visuals of a bouncing ball that made a matching impact sound, the effect of realism was assessed. No evidence for an effect of realism and movement of the stimuli was found, suggesting that, in simple scenarios, the ventriloquist effect might not be affected by stimulus realism.* © *2022 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.Percept.Imaging.2022.5.000404]

1. INTRODUCTION

In our everyday lives, our senses are continuously stimulated. While our sensory systems (e.g., auditory, visual, or tactile) receive their input separately, it is known that the speed of processing through the neural pathways and the precision and accuracy of our sensory perception is enhanced through the integration of information across the sensory modalities [13, 16, 29, 33, 34]. As the brain cannot "know" with certainty which sensory inputs belong together, since processing times and neural representations vary across modalities, it must estimate which sensory inputs should be integrated. This means that it is possible that sensory inputs from stimuli that originated from a different location (and potentially a different source) are integrated into a common percept. In such a situation, the location of the combined percept is determined through statistically optimal integration of, for example, an auditory and a visual percept [2]. By weighting the auditory and visual percept relative to their reliability (i.e., the inverse of the localization variance), the variance of the combined percept is minimized. As the spatial resolution of the visual system is higher than that of the auditory system, the auditory percept is generally strongly biased toward the visual percept, although studies have also shown that the bias can be shifted toward the auditory percept by reducing the reliability of the visual percept [2]. This effect, where spatially disparate audio-visual stimuli are integrated, resulting in a shift of the perceived location, is called the (spatial) ventriloquist effect [19].

Many studies have investigated aspects of audio-visual integration through this ventriloquist effect. However, most studies have used relatively simple stimuli, such as noise bursts and light flashes or white circles (e.g., [2, 6, 44]). While these studies provide insights into fundamental features of audio-visual integration, it is unclear to what extent the results obtained with these laboratory stimuli generalize toward real-world scenarios, as natural audio-visual stimuli share, besides temporal and spatial alignment, also contextual and semantic features, which are associated with those stimuli based on prior experience [24].

As the shift in the perceived location of the auditory stimulus has been shown to arise relatively late in the neural processing [5], top-down processes likely can influence the biasing effect of visual information. Indeed, top-down influences, like semantic congruence, attention, and motivation, have recently been shown to be able to influence audio-visual integration [8, 11, 23, 24, 36, 37, 39, 43, 45], see [7] for an overview. However, not in all scenarios [4, 22, 32, 35, 38, 40, 41]. Specifically, in the case of the ventriloquist effect, the influence of stimulus realism remains unclear.

Jackson [21] found that participants responded to (audio-) visual information over far larger ranges of spatial separation for realistic stimuli (kettle blowing steam with a whistling noise) than for artificially matched stimuli (light and bell). However, it is unclear whether the observed visual bias was due to audio-visual integration or due to a response bias [39, 40], i.e., people might have adjusted their response to match their expectation as they assumed that audio and

†Special Issue on Multisensory & Crossmodal Interactions.

Received Nov. 30, 2020; accepted for publication July 13, 2021; published online Sep. 20, 2021. Associate Editor: Fang Jiang.

2575-8144/2022/5/000404/10/$00.00

visual information would belong together, or their decision might have been based on the increased temporal correlation between the steam and the whistle [10]. Similar effects of realism on the probability of audio-visual integration were found by Warren et al. [42], who used synchronized and desynchronized audio-visual speech stimuli. While they attributed the difference in the visual bias found between these stimuli to the difference in realism, the temporal correlation could also have accounted for this effect. Thurlow and Jack [38] investigated various facilitators of the ventriloquist effect. In various experiments, using both speech and non-speech signals, they found more audio-visual integration when the stimuli were more realistic. However, they did not differentiate between movement (i.e., facial movements of a puppet) and realism (facial features). Indeed, using a similar experimental setup [20], the same authors found a significant effect of the movement, but no effects of realism. However, again, the stimuli varied not only in movement and realism but also with respect to their temporal correlation. Hence, it remains unclear if the increased realism of the moving stimuli or the increased temporal correlation between the visual and auditory stimuli was the facilitative factor.

Radeau [32], using a voice in combination with a modulated light or an image of the speaker, found that audio-visual adaptation (an after effect of the ventriloquist effect) was unaffected by semantic congruence and was only due to the temporal synchronization and Parise et al. [30] demonstrated that temporal correlation, rather than temporal alignment, facilitated integration. These results further support the hypothesis that effects of realism and movement were mainly driven by an increased temporal correlation of the auditory and visual stimuli. Thus, the ventriloquist effect could be dominated, in some scenarios, solely by "low-level" factors, such as the spatial and temporal alignment and the temporal correlation.

With the recent rise of virtual reality, it has become easier to create and manipulate the realism of the stimuli. In the present study, a ventriloquist experiment was designed where the realism of the stimuli was varied stepwise to investigate the effect of realism, while maintaining a similar temporal correlation between the stimuli. Starting from the baseline condition using noise burst and light flashes, three factors were introduced: attention (through a distractor task), movement (through movement of the visual stimulus), and realism (through a change of the stimuli). To maintain the similar temporal correlation, movement was added only to the visual stimulus. The distractor task was introduced to ensure that participants were focused on the intended location, as it has been shown that eye movements can influence audio-visual integration [31]. As the ventriloquist effect has been shown to be unaffected by attention in similar conditions [4, 41], no effect of attention was expected. However, based on most previous findings, increased stimulus realism was hypothesized to facilitate audio-visual integration over longer ranges of spatial disparity between the auditory and visual stimuli.

2. METHODS

2.1 *Participants*

21 participants (11 female, 10 male; age 29 ± 10 years) were recruited from the Hearing Systems Section's volunteers' database and from the Technical University of Denmark (DTU) student community for this experiment. All participants reported normal vision and normal hearing. This was confirmed with standard clinical tests. All participants had normal hearing thresholds at octave frequencies between 125 Hz and 8 kHz and all scored a visual acuity rating of at least 0 on a LogMAR visual acuity chart [15]. Data from participant 15 were excluded from the analysis based on extreme outliers in the unimodal visual and pointing conditions (20 datasets remained). The procedure was approved by the local ethical committee "Videnskabsetiske Komitéer for Region Hovedstaden" (H-16036391), and all participants provided written, informed consent. The participants were compensated with an hourly rate of 122 DKK.

2.2 *Apparatus*

The experiment took place in the audio visual immersion lab (AVIL) of the DTU. Auditory stimuli were presented using seven loudspeakers (KEF LS50, KEF, Maidstone, UK) that were part of a 64-loudspeaker array. The loudspeakers used were evenly positioned between ±45 degrees azimuth at a distance of 2.4 m. In the center of the loudspeaker array was a height adjustable chair. The chair was adjusted such that the height of the participants' ears was aligned with the centers of the loudspeakers.

For the presentation of the visual stimuli, an HTC Vive HMD (Head Mounted Display; HTC Corporation) was used. This HMD was run with a separate computer, which was controlled by the computer that ran both the experiment and the loudspeaker array. A 1:1 model of the experimental room, created in UNITY3D (Unity Technologies), was used for the virtual environment. Calibration was done as in [1], ensuring spatial alignment between the real and the virtual environment. For the calibration, three HTC Vive Trackers were placed at known positions and were tracked during the experiment. A shift of more than 1 cm in the position of one of the trackers, or the HMD losing tracking, resulted in a recalibration of the virtual world.

In the virtual environment, a virtual loudspeaker array was not included until the last part of the experiment. Instead, a gray ring (10 cm in height) was used to indicate the height of the loudspeaker array. At 0 degrees azimuth, just below this ring, a white square was placed that served as a focus point during the experiment (see Figure 1). The virtual environment was continuously visible during the experiment and did not change, except in the last task where the ring was replaced by the loudspeaker array. Between trials, a small sphere was used to help participants with the visual alignment process. This sphere was positioned at about eye height at a distance of 2.4 meters straight ahead of the participants and moved synchronously with their head movements. At the start of each trial, this sphere disappeared. Only after the trial was finished

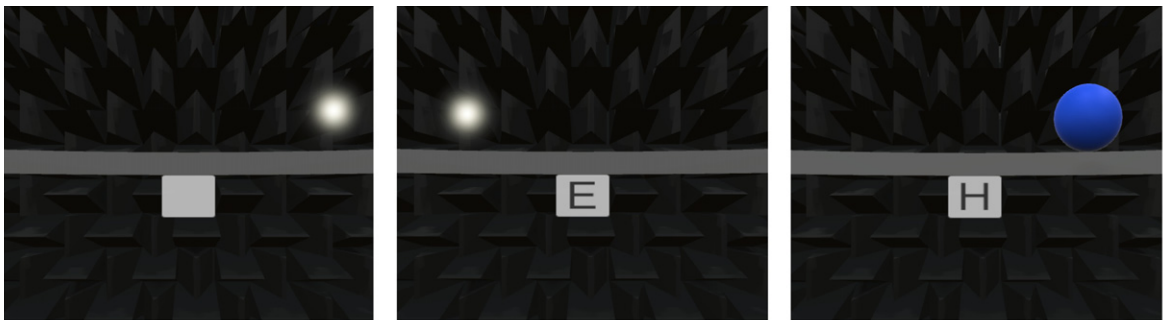

Figure 1. The different stimulus conditions and the experimental setup. The ring is visible in gray with the focus point below it. The stimuli shown here, from the left to the right, represent the baseline visual stimulus, the attention visual stimulus, and the congruence stimulus. The middle and the right panels also illustrate the distractor task stimulus.

did it reappear. To proceed through the experiment and record their localization judgements, the participants used a handheld HTC VIVE controller. In the virtual environment, a thin red rod was attached to simulate a laser pointer to help the participants point toward the perceived auditory stimuli. This "laser" disappeared at the start of a new trial and reappeared when a response from the participant was requested.

2.3 *Stimuli*

Three different visual stimuli and two different auditory stimuli were used in this experiment. However, not all combinations were tested. The experiment was designed such that each bimodal condition, containing a single set of stimuli, added one new factor. The baseline condition represented the commonly used laboratory conditions in audio-visual experiments, i.e., flashes and noise bursts. For these baseline stimuli, the magnitude spectrum of the realistic sound was combined with a randomized phase to obtain a noise with the same loudness as that of the original recordings of the real handball impact stimulus. The visual baseline stimulus was a 20-ms light blur that appeared synchronously with the auditory stimulus above the loudspeaker ring. The light blur was 33.56 cm in diameter, corresponding to an 8-degree visual angle, as indicated in Fig. 1. The Gaussian blur had a standard deviation of approximately 5.5 cm (standard Gaussian blur scaled to the size of the visual stimulus).

The second bimodal set consisted of the same baseline stimuli, but it introduced a distractor task where a letter was shown on the white screen in the center at the same time as the other stimuli, as indicated in the middle panel of Fig. 1 for the flash and distractor stimulus. The purpose of the distractor task was to ensure that participants were fixating straight ahead during each trial (which is particularly important for later conditions involving moving visual stimuli). As no effect of attention was expected, this condition was included as a control. The letter remained visible for only 200 ms. After the participants had finished the localization task of the auditory stimuli, they were shown a matrix of 16 letters and had to select the letter that had appeared during the collision. If the participant was incorrect, the trial was repeated at a later time chosen at random. This process was repeated if the participant continued to indicate an incorrect letter.

The third set of stimuli again used the baseline stimuli (with the distractor task) but introduced movement. The visual stimulus appeared above the ring at the start of the trial, fell for half a second, bounced once on the ring and then disappeared 20 ms after bouncing. The bouncing on the ring was the trigger for the audio stimulus.

The realistic stimuli consisted of the sound and visuals of a dropping ball. The auditory stimulus was a 20-ms recording of the impact of a handball landing on a carpeted floor, presented at a peak equivalent (pe) sound pressure level (SPL) of 65 dB. As illustrated in the right panel of Fig. 1, the visual stimulus was a blue ball, in the same size as the flash stimulus (33.56 cm, or 8-degrees visual angle in diameter). As with the moving flash stimuli, the ball appeared above the ring, fell down, bounced once on the ring, triggering the auditory stimulus, and disappeared.

Due to a miscorrected latency in the system, the audio was played, on average, 105 ms after the visual stimulus. There was a variation of ± 13 ms due to the frame rate of the HMD and a variation in the communication speed between the computers running the virtual environment and the audio system. This asynchrony was the same across all conditions. Furthermore, as the visual stimuli appeared slightly above the ring, there was a slight elevation difference of 3 degrees between the auditory and the center of the visual stimulus. However, due to the low sensitivity to incongruities in elevation [17], this should not affect the integration.

2.4 *Conditions*

The main task of the experiment consisted of a localization task, using only auditory or only visual (i.e., unimodal) stimulation, or a combination of both (i.e., bimodal stimulation). In total, the experiment consisted of ten conditions which were divided into four blocks (see Table I). The block order was fixed: unimodal audio, bimodal, unimodal visual, pointing task. However, within each block, the conditions were presented in a counterbalanced manner across participants.

Table I. The experiment consisted of ten conditions presented in four blocks. Blocks were presented in a fixed order, but within a block, conditions were counterbalanced across participants. Each bimodal condition added a new factor.

Block	Stimuli	Distractor task	Movement	Realism
1. Audio	Noise burst	Yes	No	No
	Ball impact sound	Yes	No	Yes
2. Audio-visual	Baseline (noise + flash)	No	No	No
	Attention (noise + flash)	Yes	No	No
	Moving (noise + moving flash)	Yes	Yes	No
	Realism (ball + moving ball)	Yes	Yes	Yes
3. Visual	Flash	Yes	No	No
	Moving flash	Yes	Yes	No
	Moving ball	Yes	Yes	Yes
4."Pointing"	Loudspeaker targets	No	No	Yes

The first two conditions were unimodal audio conditions. Here, the sounds were presented randomly from one of the seven loudspeakers and each position was repeated five times resulting in 35 trials each. As the HMD has a limited field of view (110 degrees), the visual stimuli were limited to a maximum eccentricity of ±45 degrees. Because of this, the two outer loudspeakers were not used in the bimodal conditions. For each of the five loudspeakers used to present sound in the bimodal conditions, visual stimuli were presented in a 30-degree range around that loudspeaker in 3-degree steps and also at the other six loudspeaker locations. As a result, the densest sampling occurred between 15 and -30 degrees audio-visual disparity, and the maximum disparity was up to ±75 degrees. The sampling is also shown in Figure 2, which shows for each auditory stimulus position, all tested visual positions. Each combination was presented three times, leading to a total of 322 trials per bimodal condition.

The next block consisted of the three unimodal visual conditions. At this point in the experiment, the task changed from localizing sound to localizing visual stimuli. These conditions included stimuli at all seven loudspeakers and each position was tested three times, resulting in 21 trials per condition.

To account for potential biases in the pointing response [1], a "pointing" condition was included where the participants had to point at a continuously present static visual stimulus. In this task, no distractor ring was used and the gray ring was replaced by a model of the loudspeaker array. Participants were then shown a number and had to point with the "laser pointer" at the center of the loudspeaker with that number. This was the final task of the experiment. As in the unimodal visual conditions, this task used all seven loudspeaker positions and three repetitions were conducted, resulting in 21 trials. The conditions and stimuli are summarized in Table I.

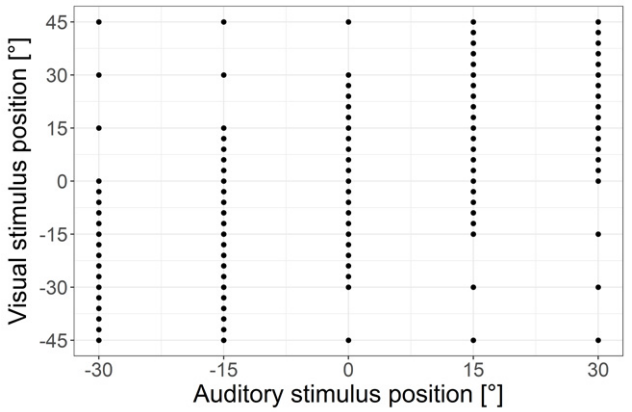

Figure 2. The left panel shows all unique combinations of auditory (abscissa) and visual (ordinate) positions in the bimodal conditions. Each combination (indicated by a dot) was repeated three times.

2.5 *Analysis*

Data were analyzed using the statistical software *R* [12]. The unimodal data were analyzed using Levene's test to evaluate differences in the variance and an ANOVA was applied to the localization data. To analyze the bimodal results, the localization error was calculated per participant, condition, auditory stimulus position, and angle by subtracting the position of the auditory stimulus from the response. This localization error was corrected by subtracting the mean localization error in the congruent trials at each loudspeaker location to account for angle-dependant localization biases. This corrected error was then divided by the spatial audio-visual disparity to calculate the visual bias. The spatial audio-visual disparity itself was calculated as the position of the visual stimulus minus the position of the auditory stimulus, with positive disparities indicating that the visual stimulus occurred more to the right compared to the auditory stimulus. An ANOVA analysis compared the visual bias with the absolute spatial audio-visual disparity, condition, absolute auditory stimulus position, and the relative stimuli positioning as potential predictors. The relative stimuli positioning refers here to if the visual stimulus occurred outwards compared to the auditory stimulus or if it occurred closer to the center. To investigate how the various factors affected the results, a Bonferroni-corrected post hoc within factor comparison analysis was used. For this analysis, the disparities larger than 30 degrees were not included (i.e., only the densely tested area was included), as the initial analysis revealed interactions which could not be explored when these data points were included. Dropping these specific points from the analysis did not affect the results.

3. RESULTS

3.1 *Pointing Bias*

Figure 3 shows the localization error as a function of the stimulus position when pointing at a continuously present visual target. This task was included to measure the error in pointing. As a "laser pointer" was included, the accuracy and precision of pointing is very high. As can be seen in

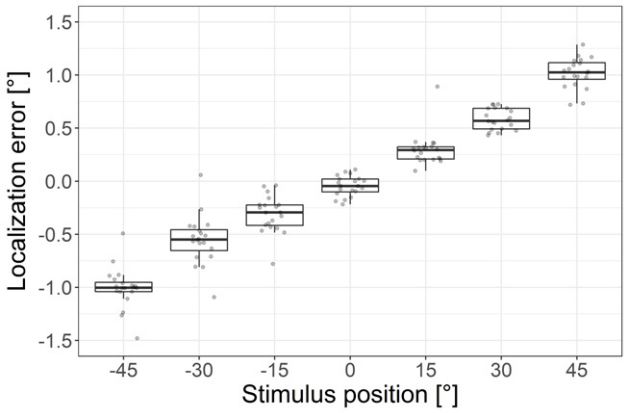

Figure 3. The median error in pointing per participant and angle. Each data point corresponds to the median error for one participant at a stimulus position. The boxes extend from the first to the third quartile, the line shows the median perceived response. Outliers are indicated separately.

Fig. 3, the maximum median localization error was around one degree, and the variance of the error was below one degree. There was a small dependency (i.e., bias) of the localization on the stimulus position, with slightly increased errors at higher eccentricities (±1 degree at ±45 degrees azimuth, $F_{1,6} = 3.234$, $p < 0.01$). The variance did not vary significantly with stimulus position [$F_{1,6} = 3.234$, $p = 0.631$].

3.2 *Unimodal Conditions*

The unimodal conditions were used to test if there were differences in localization between the stimuli that were used. Figure 4 shows the localization error for the auditory (left panel) and the visual stimuli (right panel) as a function of the presentation angle. For the auditory stimuli, the baseline stimulus (noise burst) is indicated in light blue, and the congruent stimulus (ball impact audio) is indicated in dark blue. Levene's tests showed that, for the auditory stimuli, the variance varied significantly only with angle [$F_{1,6} = 2.662$, $p < 0.05$], but not with condition [$F_{1,6} = 1.341$, $p = 0.2470$]. Similarly, the localization error also varied with angle [$F_{1,6} = 51.035$, $p < 0.001$] and not with condition [$F_{1,1} = 0.251$, $p = 0.6162$]. However, an interaction between the stimulus angle and condition was found [$F_{1,6} = 2.645$, $p < 0.05$].

The right panel of Fig. 4 shows the localization data for the visual stimuli, with the static flash data shown in light red, moving flash data shown in red and the ball stimulus data shown in dark red. For the visual stimuli, the localization error and variance were much smaller than for the auditory stimuli. Moreover, a clear trend can be seen in the visual responses, where responses were closer to the center (positive errors at negative angles and vise versa) when the stimulus was presented more laterally. Additionally, the variance also increased with presentation angle. This was confirmed by the statistical analysis, which revealed an effect of stimulus position on both the variance [$F_{1,6} = 4.6560$, $p < 0.001$] and the localization error [$F_{1,6} = 55.935$, $p < 0.0001$], but no effect of the stimulus used on either the localization error [$F_{1,2} = 0.305$, $p = 0.7375$] or the variance [$F_{1,2} = 0.721$, $p = 0.1520$], respectively.

3.3 *Bimodal Condition*

Figure 5 shows the localization error as a function of the spatial disparity between the auditory and the visual stimuli in the four bimodal conditions for participant 4. A bias, where responses are shifted toward the position of the visual stimulus, can be seen in all conditions. However, comparing the average (dashed line) responses, no clear effect of condition is visible for this participant.

The visual bias, averaged across participants, per condition is shown in Figure 6. The left and right panels show the relative stimuli position. In the left panels, A-V-A, the visual stimulus occurred inwards relative to the auditory stimuli, whereas in the right panels it is instead the auditory stimulus that occurs inwards relative to the visual stimuli. The upper panels show results for when the auditory stimulus was positioned at 0 degrees azimuth, the middle panels show the results for when the auditory stimulus was presented at ±15 degrees, and the bottom panels show the results for when auditory stimuli were presented at ±30 degrees. In the upper panels, the auditory stimulus is presented at

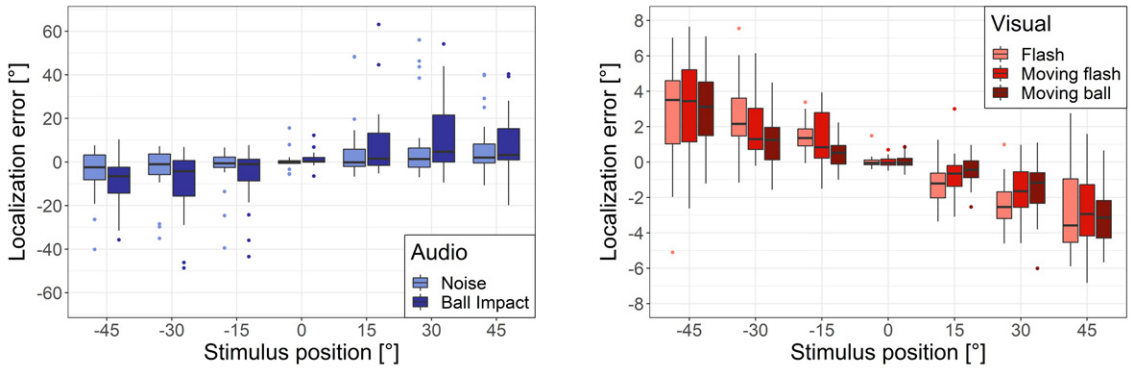

Figure 4. The left panel shows the localization error for the two different auditory stimuli. The right panel shows the localization error for the three different visual stimuli. The boxplot extends from the first to the third quartile, with the median across all participants shown in black. The whiskers extend to 1.5 times the interquartile range. Outliers beyond this range are indicated separately. Note the different ordinate scales for each figure.

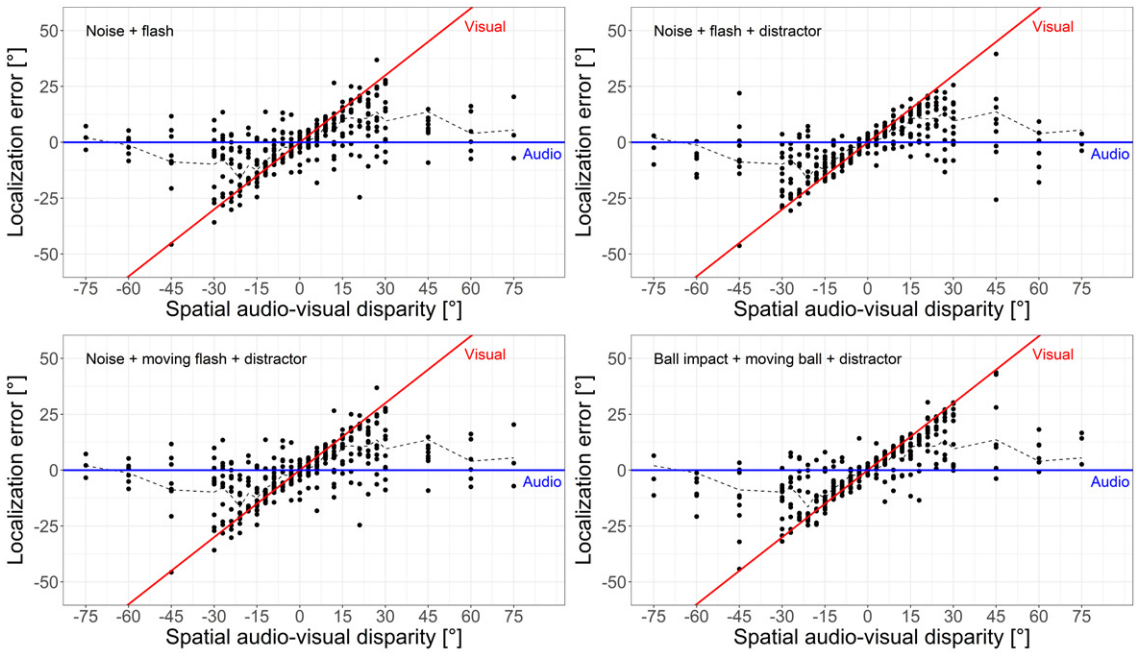

Figure 5. Bimodal responses of a representative participant, participant 4. The four panels show the results for each of the bimodal conditions. The conditions are indicated in the top-left corner of the panel. Perfect visual localization is indicated by a red line and perfect auditory localization is indicated with a blue line. The dashed curve (black) shows the mean response as a function of audio-visual disparity. When the spatial disparity was small, participant 4 showed a visual bias on most trials, that is, most responses shifted away from auditory localization toward the visual localization line. No clear difference in either the range or the strength of the visual bias was found between the four conditions.

0 degrees, as such, the left panel show responses where the visual stimulus was presented in the left hemisphere and the right panel shows responses for visual stimuli presented in the right hemisphere. As no visual bias can be calculated at 0 degrees spatial disparity, where the auditory and visual position overlap, the curve is interrupted at this position. Since the visual bias is calculated by dividing the localization error by the spatial audio-visual disparity, similar errors in localization cause much larger changes in the bias at small disparities. This results in a steep increase of the standard deviation as the disparity decreases.

As visible in Fig. 6, the visual bias was found to decrease, in most cases, significantly with increasing absolute spatial audio-visual disparity [$F_{9,6980} = 5.513$, $p < 0.0001$] and varied depending on both the relative positioning of the stimuli [$F_{1,6980} = 5.034$, $p < 0.05$] and the absolute position of the auditory stimulus [$F_{2,6980} = 10.317$, $p < 0.0001$]. However, significant interactions between these factors were found, namely an interaction between the effect of the relative and auditory stimulus positioning [$F_{2,6980} = 18.373$, $p < 0.0001$], an interaction between the spatial disparity and the relative stimulus positioning [$F_{9,6980} = 2.187$, $p < 0.05$] and a three-way interaction [$F_{13,6980} = 4.427$, $p < 0.0001$]. In the A-V-A stimulus setup, at small disparities (<15 degrees), the visual bias was larger when the auditory stimulus was presented at 0 degrees compared to ±15 and ±30 degrees [3 degrees, [0–15: $t_{6980} = 5.657$, $p < 0.0001$; 3 degrees, 0–30: $t_{6980} = 6.105$, $p < 0.0001$]. On the contrary, in the V-A-V setup, the visual bias was lower when the auditory stimuli were presented at 0 degrees [0–15 : $t_{6980} = -5.278$, $p < 0.0001$].

The results for the various conditions (see Fig. 6, upper panels) were very similar at larger audio-visual disparities. However, when the stimuli were close together, the visual bias increased, and some differences appeared between the conditions. Although no main effect of condition was found [$F_{3,6980} = 0.4372$, $p = 0.4372$], there was a significant interaction between the relative stimuli positioning and the condition [$F_{3,6980} = 10.108$, $p < 0.001$] and a three-way interaction between the auditory stimulus position, the relative stimuli positioning and the conditions [$F_{13,6980} = 4.427$, $p < 0.001$]. As can be seen in Fig. 6, upper-left panel, when the visual stimulus occurred in the left hemisphere with the auditory stimulus at 0 degrees azimuth (denoted, A–V–A, but since the auditory stimulus occurred at the center, this corresponds to stimulus occurring left), the realistic stimuli produced a significantly smaller visual bias than the noise and flash (baseline condition) stimuli [$t_{6980} = 3.354$, $p < 0.01$]. The other combinations did not reach statistical significance.

In contrast, when the visual stimulus was presented in the V–A–V setup, these realistic stimuli evoked a much more similar visual bias, and it was instead the second set of stimuli (noise and flash with a distractor) that produced a lower visual bias. Both at 0 [$t_{6980} = -3.372$, $p < 0.01$] and ±15 degrees [$t_{6980} = 3.034$, $p < 0.05$], the difference between the second and fourth (realistic stimuli) condition was significant. Curiously, a negative visual bias can be

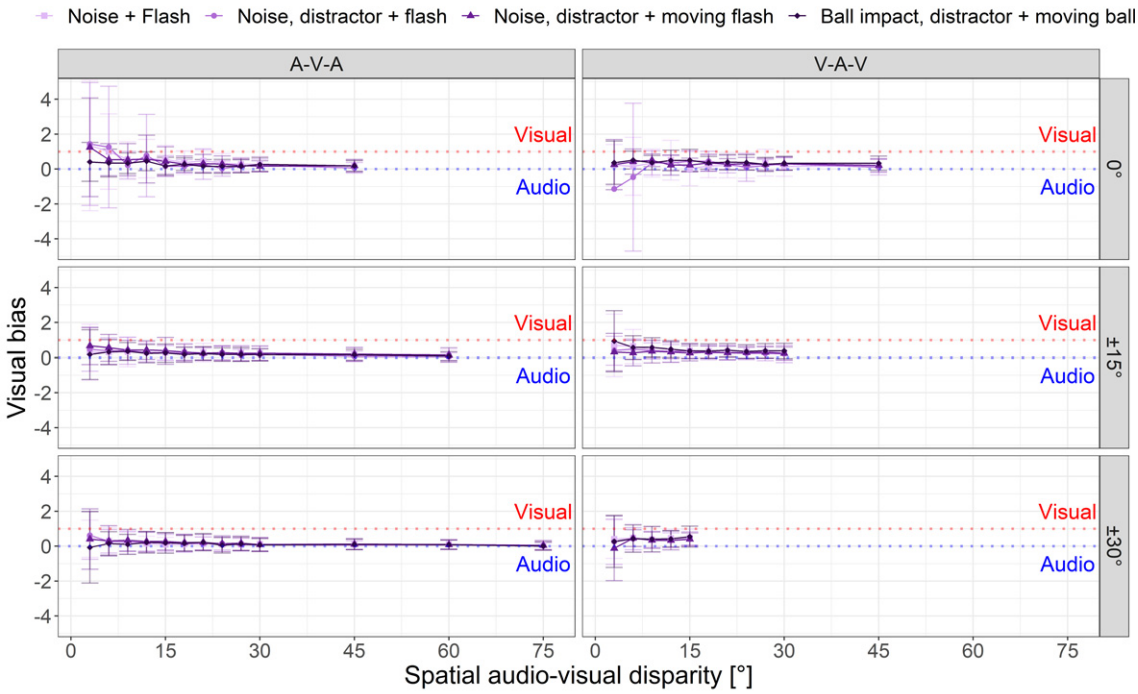

Figure 6. Average visual bias as a function of spatial audio-visual disparity per condition. The left and right panels show the relative positioning of the stimuli. The A-V-A panels show, when the visual stimulus occurred, left (0 degrees) or more toward the center (±15 and ±30 degrees) compared to the auditory stimulus. For example, for 30 degrees disparity, visual stimuli occurred toward the left in this setup, whereas for −30 degrees, the stimuli occurred toward the right. The V-A-V panels show the visual bias for when the stimuli occurred right (0 degrees) or further outwards (±15 and ±30 degrees), compared to the auditory stimulus. The horizontal panels show the results per (absolute) auditory stimulus position. The red dotted line indicates a complete visual bias, where localization responses are completely shifted toward the visual stimulus, whereas the blue dotted line indicates pure auditory localization, without any visual bias. The various conditions are indicated by different shapes and purple shades. Due to limitations of the field of view of the HMD, not all disparities could be tested for all angles, hence the difference in start and end points. For one point in the top-right panel, the standard deviation is not included as the difference in conditions cannot be assessed on the required scale. This point is the noise + flash with distractor condition at 3 degrees spatial disparity (-1.14 ± 7.90).

seen for the noise and flash with distractor stimuli in the upper-right panel of Fig. 6 indicating that participants perceived the auditory stimulus to be further away from the visual stimulus. Again, the stepwise comparison between the first and second, second and third, and third and fourth conditions was not significant.

To see if introducing the additional factor (attention, movement, realism) improved the model, equality constraints were tested using a Bayes factor test [27]. The fully unconstrained model performed worse than the combined noise and flash with and without distractor model, indicating that this factor indeed did not improve the model ($\mathrm{BF} = 2.3501e^{-11}$). Similarly, the fully unconstrained model performed worse than the model with the combined noise and flash with the distractor, and moving noise and flash with the distractor stimuli ($BF = 9.6465e^{-7}$), and the combined moving noise and flash with the distractor and the realistic stimuli ($BF = 1.7162e^{-7}$).

4. DISCUSSION

The present study investigated if the movement and realism of the stimuli influence the spatial ventriloquist effect. Starting from stimuli that are commonly used in experiments (noise burst and light flash stimuli), new factors were added to the stimuli in a stepwise fashion to be able to differentiate between the effects of movement and realism, while maintaining a similar temporal correlation between stimuli. The results of this study showed no consistent effects of the studied factor. However, some differences in specific stimulus combinations were found.

In the V-A-V stimulus setup, where the visual stimulus occurred at increased eccentricities, or when the auditory stimulus was presented exactly in the center, right compared to the auditory stimulus, the realistic stimuli evoked a significantly larger visual bias compared to the flash and noise stimuli with distractors. However, this occurred only at small spatial audio-visual disparities and the difference between the other stimuli was not significant, that is, the stepwise comparison between the first and second, second and third condition etc., was not significant. Moreover, in the A-V-A stimulus setup, where it was instead the visual stimulus that occurred more toward the center (or to the left), the realistic stimuli evoked the smallest visual bias. However, again no stepwise comparison was significant. Thus, no consistent effect of any of the factors by itself was found, and at most audio-visual disparities, no effect was found at all.

Although the results are inconclusive with regards to the effect of realism, the similar results that were found with the various stimuli do call into question the size of the effect that realism could have. The Bayesian model comparison showed no improvement with any of the factors that were included, although realism was the closest to improving the model. Studies such as by Jackson [21] found large facilitative effect of stimulus realism. The lesser to no effect found in the present study could indicate, as hypothesized, that the temporal correlation between stimuli in other studies [20, 21, 38, 42] facilitated at least part of the effect of realism. These results are in line with Radeau et al., [32], who compared continuous speech with either a face or a modulated light and found a significant effect of synchronization, but not realism. However, besides the same temporal correlation in the various conditions, there are some alternative explanations for the smaller/lack of results found in the present study, and there are limitations to the present study that warrant discussion.

First, as effects of top-down influences have been shown in some, but not all, cases of audio-visual integration, it is possible that there are specific experimental setups where these effects become relevant. For example, it has been suggested that attention only affects audio-visual integration when the stimulus salience is low [35]. As the stimuli were presented well above threshold levels in the present study, the stimulus salience was high. As such, the lack of a strong influence of high-level factors on the ventriloquist effect could be the result of the high stimulus salience.

Similarly, it is possible that contextual factors contribute to deciding which stimuli to integrate when there are several competing stimuli. This has been supported by a study by Bailey et al. [3], where it was shown that realism of the stimuli facilitated integration, but only in a cue-rich environment. As such, the simple setup used in the experiment could contribute to the lack of a consistent effect of realism. However, this explanation cannot fully account for the discrepancy between the results from the present and previous studies. For example, Jackson [21] used a similarly simple setup, but still found a large facilitative effect of realism.

Third, as mentioned also in the introduction, a common problem with the ventriloquist paradigm is a response bias [39, 40]. Since audio-visual and visual responses are very similar, it can be difficult to differentiate between true integrative responses and a response bias. Since audio-visual integration decreases the response times, response times are generally used to confirm integration, through a violation of the race model [26]. However, both the response method and the delay in the auditory stimuli added substantial variation to reaction times (and the localization responses). Thus, in the present study, it was not possible to test for a violation of the race model to confirm that integration occurred. While the biases in the localization results for bimodal stimuli compared to unimodal stimuli indicate that integration occurred, it cannot be fully ascertained that the visual bias is not, at least partially, due to response biases toward the visual stimulus.

Finally, the visual bias at small disparities was smaller than anticipated and substantial variation in the visual bias was found. The smaller visual bias is likely due to temporal delay. Studies on the optimal temporal window for audio-visual integration have found varying results. While some studies found integration windows that would still support integration in the present study [14, 25, 28], it is possible that for some participants the temporal asynchrony disrupted integration. However, as the temporal disparity was present in all conditions, this disruption should lower the visual bias equally in all conditions. The large variation can be attributed to the response method. As shown in Fig. 4, the variance in localization of unimodal stimuli was quite large. Especially at small audio-visual disparities, such variation can strongly influence the calculation of the visual bias. The use of discrete response options could largely reduce such variance.

Much stronger than the effect of condition, was the effect of the relative positioning of the stimuli, which was dependent on the angle of the auditory stimulus. At $\pm$15 degrees and $\pm$30 degrees, the visual bias was larger in the V-A-V setup. This is similar to the results from Hairston et al., [18] and Charbonneau et al., [9], where centrally positioned (visual) stimuli evoked a greater bias than more peripheral (visual) stimuli did. Alternatively, the increased bias could also be a result of perceptually closer stimuli. As visual localization shows a bias toward the center and auditory localization tends to show a bias away from the center, these biases might counteract each other and reduce the perceived disparity when the visual stimulus is positioned further outwards compared to the auditory stimulus. This hypothesis was tested, but not supported in the study of Godfroy et al. [17]. Either way, the large difference that occurred already at small angles of spatial disparity could warrant a study that further investigates the effect of relative positioning on the ventriloquist effect, as it could provide further insight in how the biases in unimodal localization affect integration.

When the auditory stimulus was presented at 0 degrees azimuth, there was still an effect of relative stimuli positioning. In this case, the A-V-A and V-A-V setup correspond to whether the visual stimulus occurred left or right to the auditory stimulus, respectively. Curiously, a much stronger visual bias was found when the visual stimuli were presented to the left. It is possible that the response method contributed to this. However the results from Fig. 3 make this less likely as pointing responses were similar both in the left and right hemisphere. Thus, an increased variability in pointing results increasing the visual bias by chance is not a likely explanation. Alternatively, it could indicate a mismatch in the virtual and real world. Although care was taken to calibrate these, there could still be small differences. Such a hypothetical mismatch, if consistent across participants, could favor integration in one direction as the stimuli line up better similar to the effect of relative stimulus positioning

at ±15 degrees and ±30 degrees. In this case, results could indicate a shift of the VR world to the left. However the calibration did not indicate the existence of such a shift.

Overall the results of the present study could be valuable for studies using the less natural noise burst and light flash stimuli. These stimuli are much easier to create in laboratory settings but, based on the results here, should still generalize well to more ecologically valid stimuli. At the same time, the discrepancy in the various experiments investigating the influence of realism on audio-visual integration suggests that top-down factors influence integration only in more complex experimental settings. To further investigate how well these studies generalize also to real-world settings, future studies could explore in which environments high-level features become impactful.

5. CONCLUSION

The present study investigated the influence of realism on the ventriloquist effect. No consistent evidence for an effect of movement or realism on the visual bias was found, as in one particular stimulus setup, realistic stimuli evoked a slightly stronger visual bias, whereas in another setup they evoked a slightly weaker visual bias. Either way, the results indicate that the effect of realism, if present, is minor at best. While other studies have observed a more noticeable effect of realism, the more realistic conditions involved audio-visual stimuli with higher temporal correlation than the less realistic conditions. The present study suggests that it was the temporal correlation between the auditory and visual stimuli, rather than realism per se, that more strongly facilitated integration in previous studies. As such, previous studies on the ventriloquist effect, which used the less natural noise burst and light flash stimuli, should generalize to more realistic stimuli. However, the present study presented only a simple environment. It is possible that high-level factors such as attention and realism influence integration more strongly only in complex settings with competitive stimuli. To differentiate between these factors, future studies might investigate the influence of realism in more complex settings. In simple settings the effect of stimulus realism facilitates integration only to a minor or no extent.

ACKNOWLEDGMENT

This work was carried out as part of the research activities at the Centre for Applied Hearing Research (CAHR) at DTU, supported by GN Hearing, Oticon, and WSA. Audiograms were measured by audiologist Rikke Skovhøj Sørensen. The model of the room and loudspeaker array used in the experiment was created by Kasper Duemose Lund.

Funding

This research was supported by the Centre for Applied Hearing Research (CAHR) through a research consortium agreement with GN Resound, Oticon, and Widex.

Declaration of Conflicting Interests

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

REFERENCES

[1] A. Ahrens, K. D. Lund, M. Marschall, and T. Dau, "Sound source localization with varying amount of visual information in virtual reality," PLoS One **14**, 1–19 (2019).

[2] D. Alais and D. Burr, "Ventriloquist effect results from near-optimal bimodal integration," Curr. Biol. **14**, 257–262 (2004).

[3] H. D. Bailey, A. B. Mullaney, K. D. Gibney, and L. D. Kwakye, "Audiovisual integration varies with target and environment richness in immersive virtual reality," Multisens. Res. **31**, 689–713 (2018).

[4] P. Bertelson, J. Vroomen, B. D. Gelder, and J. Driver, "The ventriloquist effect does not depend on the direction of deliberate visual attention," Percept. Psychophys. **62**, 321–332 (2000).

[5] B. Bonath, T. Noesselt, A. Martinez, J. Mishra, K. Schwiecker, H.-J. Heinze, and S. A. Hillyard, "Neural basis of the ventriloquist illusion," Curr. Biol. **17**, 1697–1703 (2007).

[6] A. K. Bosen, J. T. Fleming, S. E. Brown, P. D. Allen, W. E. O'Neill, and G. D. Paige, "Comparison of congruence judgment and auditory localization tasks for assessing the spatial limits of visual capture," Biol. Cybern. **110**, 455–471 (2016).

[7] P. Bruns, "The ventriloquist illusion as a tool to study multisensory processing: An update," Frontiers Integrative Neurosci. **13** (2019).

[8] P. Bruns, M. Maiworm, and B. Röder, "Reward expectation influences audiovisual spatial integration," Attention, Perception, Psychophys. **76**, 1815–1827 (2014).

[9] G. Charbonneau, M. Veronneau, C. Boudrias-Fournier, F. Lepore, and O. Collignon, "The ventriloquist in periphery: Impact of eccentricity-related reliability on audio-visual localization," J. Vis. **13**, 1–14 (2013).

[10] Y. C. Chen and C. Spence, "Assessing the role of the 'unity assumption' on multisensory integration: A review," Frontiers Psychol. **8** (2017).

[11] L. Chuen and M. Schutz, "The unity assumption facilitates cross-modal binding of musical, non-speech stimuli: The role of spectral and amplitude envelope cues," Attention, Perception, Psychophys. **78**, 1512–1528 (2016).

[12] R. Core Team, "R: A Language and Environment for Statistical Computing," Vienna, Austria, 2020 [Online]. Available https://www.r-project.org/.

[13] A. Diederich and H. Colonius, "Bimodal and trimodal multisensory enhancement: Effects of stimulus onset and intensity on reaction time," Percept. Psychophys. **66**, 1388–1404 (2004).

[14] S. E. Donohue, M. G. Woldorff, and S. R. Mitroff, "Video game players show more precise multisensory temporal processing abilities," Attention, Perception, Psychophys. **72**, 1120–1129 (2010).

[15] D. B. Elliott, "The good (logMAR), the bad (Snellen) and the ugly (BCVA, number of letters read) of visual acuity measurement," Ophthalmic Physiol. Opt. **36**, 355–358 (2016).

[16] L. C. A. Freeman, K. C. Wood, and J. K. Bizley, "Multisensory stimuli improve relative localisation judgments compared to unisensory auditory or visual stimuli," J. Acoust. Soc. Am. **143**, 6 (2018).

[17] M. Godfroy, C. Roumes, and P. Dauchy, "Spatial variations of visual – auditory fusion areas," Perception **32**, 1233–1245 (2003).

[18] W. D. Hairston, M. T. Wallace, J. W. Vaughan, B. E. Stein, J. L. Norris, and J. A. Schirillo, "Visual localization ability influences cross-modal bias," J. Cogn. Neurosci. **15**, 20–29 (2003).

[19] I. P. Howard and W. B. Templeton, *Human Spatial Orientation* (Wiley, New York, 1966).

[20] C. E. Jack and W. R. Thurlow, "Effects of degree of visual association and angle of displacement on the ventriloquism effect," Percept. Mot. Skills **37**, 967–979 (1973).

[21] C. V. Jackson, "Visual factors in auditory localization," Q. J. Exp. Psychol. **5**, 52–65 (1953).

[22] C. Koppen, A. Alsius, and C. Spence, "Semantic congruency and the colavita visual dominance effect," Exp. Brain Res. **184**, 533–546 (2008).

[23] A. Kramer, B. Röder, and P. Bruns, "Feedback modulates audio-visual spatial recalibration," Front. Integr. Neurosci. **13**, 1–15 (2020).

[24] P. J. Laurienti, R. A. Kraft, J. A. Maldjian, J. H. Burdette, and M. T. Wallace, "Semantic congruence is a critical factor in multisensory behavioral performance," Exp. Brain Res. **158**, 405–414 (2004).
[25] J. Lewald and R. Guski, "Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli," Cogn. Brain Res. **16**, 468–478 (2003).
[26] J. Miller, "Divided attention: Evidence for coactivation with redundant signals," Cogn. Psychol. **14**, 247–279 (1982).
[27] R. D. Morey and J. N. Rouder, "BayesFactor: computation of bayes factors for common designs." 2018. [Online]. Available https://cran.r-project.org/package=BayesFactor.
[28] J. P. Noel, K. Modi, M. T. Wallace, and N. Van der Stoep, "Audiovisual integration in depth: multisensory binding and gain as a function of distance," Exp. Brain Res. **236**, 1939–1951 (2018).
[29] B. Odegaard, D. R. Wozny, and L. Shams, "Biases in visual, auditory, and audiovisual perception of space," PLoS Comput. Biol. **11**, 1–23 (2015).
[30] C. V. Parise, C. Spence, and M. O. Ernst, "When correlation implies causation in multisensory integration," Curr. Biol. **22**, 46–49 (2012).
[31] U. Pomper and M. Chait, "The impact of visual gaze direction on auditory object tracking," Sci. Rep. **7**, 1–16 (2017).
[32] M. Radeau and P. Bertelson, "Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations," Percept. Psychophys. **22**, 137–146 (1977).
[33] E. Schröger and A. Widmann, "Speeded responses to audiovisual signal changes result from bimodal integration," Psychophysiology **35**, 755–759 (1998).
[34] B. E. Stein, M. A. Meredith, W. S. Huneycutt, and L. McDade, "Behavioral indices of multisensory integration: Orientation to visual cues is affected by auditory stimuli," J. Cogn. Neurosci. **1**, 12–24 (1989).
[35] D. Talsma, D. Senkowski, S. Soto-Faraco, and M. G. Woldorff, "The multifaceted interplay between attention and multisensory integration," Trends Cogn. Sci. **14**, 400–410 (2010).
[36] K. I. Taylor, H. E. Moss, E. A. Stamatakis, and L. K. Tyler, "Binding crossmodal object features in perirhinal cortex," Proc. Natl. Acad. Sci. USA **103**, 8239–8244 (2006).
[37] J. P. Thomas and M. Shiffrar, "Meaningful sounds enhance visual sensitivity to human gait regardless of synchrony," J. Vis. **13**, 1–13 (2013).
[38] W. R. Thurlow and C. E. Jack, "Certain determinants of the ventriloquism effect," Percept. Mot. Skills **36**, 1171–1184 (1973).
[39] A. Vatakis and C. Spence, "Crossmodal binding: Evaluating the 'unity assumption' using audiovisual speech stimuli," Percept. Psychophys. **69**, 744–756 (2007).
[40] A. Vatakis and C. Spence, "Evaluating the influence of the 'unity assumption' on the temporal perception of realistic audiovisual stimuli," Acta Psychol. (Amst). **127**, 12–23 (2008).
[41] J. Vroomen, P. Bertelson, and B. De Gelder, "The ventriloquist effect does not depend on the direction of automatic visual attention," Percept. Psychophys. **63**, 651–659 (2001).
[42] D. H. Warren, R. B. Welch, and T. J. McCarthy, "The role of visual-auditory 'compellingness' in the ventriloquism effect: Implications for transitivity among the spatial senses," Percept. Psychophys. **30**, 557–564 (1981).
[43] V. van Wassenhove, K. W. Grant, and D. Poeppel, "Temporal window of integration in auditory-visual speech perception," Neuropsychologia **45**, 598–607 (2007).
[44] D. R. Wozny, U. R. Beierholm, and L. Shams, "Probability matching as a computational strategy used in perception," PLoS Comput. Biol. **6**, 8 (2010).
[45] B. Zierul, J. Tong, P. Bruns, and B. Röder, "Reduced multisensory integration of self-initiated stimuli," Cognition **182** (2019).