JASA Express Letters

OCTOBER 27 2025

Visual induction of spatial release from masking during speech perception in noise

Sarah Knight; Charlotte Levy; Sven Mattys

JASA Express Lett. 5, 105204 (2025)
https://doi.org/10.1121/10.0039627

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Visual induction of spatial release from masking during speech perception in noise

Sarah Knight,[1,2,a)] Charlotte Levy,[2] and Sven Mattys[2]
[1]School of Psychology, Newcastle University, Newcastle-upon-Tyne, NE1 7RU, United Kingdom
[2]Department of Psychology, University of York, York, YO10 5DD, United Kingdom
sarah.knight@newcastle.ac.uk, charlottelevy08@gmail.com, sven.mattys@york.ac.uk

Abstract: Spatially separating target and masker talkers improves speech perception in noise, an effect known as spatial release from masking (SRM). Independently, the perceived location of a sound can erroneously shift towards an associated but spatially displaced visual stimulus (the "ventriloquist effect"). This study investigated whether SRM can be induced by spatially separating visual stimuli associated with a target and masker without separating the sound sources themselves. Results showed that SRM was not induced by spatially separated visual stimuli, but collocated visual stimuli reduced the benefit of auditory SRM. There was no influence of individual differences in auditory localization ability on effects related to the visual stimuli.

[Editor: Charles C. Church] https://doi.org/10.1121/10.0039627

Received: 11 July 2025 **Accepted:** 30 September 2025 **Published Online:** 27 October 2025

1. Introduction

Speech perception in noise (SPiN) refers to any situation in which listeners must selectively attend to a target talker whilst ignoring irrelevant background sound. In many cases, SPiN is challenging due to energetic masking (EM)—spectro-temporal overlap between target and masker, which results in direct competition at the cochlea.[1] Target intelligibility depends on how much of the target can be "glimpsed" through the masker.[2] However, SPiN is also cognitively challenging, requiring listeners to not only successfully parse the auditory scene into separate streams, but also allocate attention to the target stream and inhibit the masker.[3]

Spatial separation of target and masker(s) has been shown to improve SPiN performance—a phenomenon known as spatial release from masking (SRM).[4] Separation leads to an increase in the number and duration of spectro-temporal regions in which the target energy exceeds that of the masker at a given ear,[5] thus affording more glimpses. Separation also provides strong location-based cues to auditory object identity (provided the target and masker are consistently associated with different positions in space), allowing for more successful streaming and deployment of selective attention.[6,7] Taken together, these effects can give rise to large improvements in performance, with some studies reporting as much as a 12 dB reduction in the signal-to-noise ratio (SNR) required to achieve a set performance level.[8]

In a separate literature, it has been demonstrated that the apparent spatial location of sound can be affected by visual cues, with sound localization biased towards the location of a concurrent visual stimulus. This has been termed the "ventriloquist effect" (VE)[9,10] and appears to occur even when participants are primed to believe that the audio and visual stimuli are unrelated.[11]

Recent studies have explored the effect on target intelligibility of illusorily separating competing auditory streams through the use of associated visual information. Specifically, these studies have investigated two questions: First, does visual-only separation improve SPiN? We term this positive effect "vSRM+." Second, does visual-only collocation impair the benefits of genuine auditory SRM? We term this negative effect "vSRM−."

The existence and extent of both vSRM+ and vSRM− are likely to depend critically on which mechanism underlies the auditory SRM benefit. If this benefit derives primarily from a reduction in EM (i.e., increased glimpses of the target), then incongruent visual cues seem unlikely to help or hinder, since such cues do not affect the availability of the target signal. However, if the benefit derives primarily from the enhancement of streaming and improved selective attention, then visual cues affecting the perceived locations of target and masker may assist (or hinder) the formation, maintenance of, and attention to separate auditory streams.

Evidence for either mechanism is mixed. Driver found that recall accuracy for target speech was higher when target visuals (but not target audio) were displaced from the location of the competing stream.[12] This apparent vSRM benefit

a)Author to whom correspondence should be addressed.

(vSRM+) was confirmed in one study,[13] but either it failed to replicate in others,[14,15] or the results were mixed.[16] One possible reason for these discrepancies is a failure to account for the strength of the VE experienced by participants. It has been suggested that vSRM+ may be affected by the magnitude of individual participants' VE,[17] which may in turn be affected by individual differences in localization ability.[18] By failing to measure localization ability, previous studies may therefore have diluted vSRM+ effects by including participants for whom the VE was only weakly experienced. A second issue with existing studies concerns the nature of their stimuli. Most prior studies have used facial or lip movements congruent with the target, thus conflating the VE with the effects of, and individual differences in, lip-reading. One study has attempted to address this limitation: Valzolgher *et al.* used a synchronized pulsing circle as the target-associated visual rather than dynamic facial information, and investigated both vSRM+ and vSRM−.[19] Despite showing the existence of the VE, they found no evidence for vSRM+ or vSRM−. However, further work examining vSRM in the absence of associated facial movements is warranted.

To address these gaps, the current study seeks to examine both vSRM+ and vSRM− for speech stimuli in the absence of associated facial/lip movements, whilst simultaneously measuring sound localization ability. This allows us to assess both positive and negative potential consequences of illusory spatial manipulation of sounds during SPiN without any confounds related to lip-reading ability; it also allows us to explore, and account for, any relationship between vSRM effects and localization ability.

2. Methods

Ethical approval was granted by the local departmental ethics committee (Reference No. 232469). All participants provided informed consent, and all procedures were performed in compliance with relevant laws and institutional guidelines.

2.1 Participants

Participants ($N = 60$) were native English speakers 18–30 years of age, with normal hearing and normal/corrected-to-normal vision. They self-reported as follows: woman = 41, man = 15, other = 3, and prefer not to say = 1. Two participants had an average score in the main sentence transcription task that fell over 2 standard deviations from the overall task mean. Three further participants were identified as having persistent left/right confusions in the headphone check (see below for details). Data from these participants were not included in the analyses.

2.2 Materials

Stimuli were constructed from 160 sentences drawn from the IEEE corpus[20] and modified to fit modern British English. Eighty audio files were produced: Each consisted of two simultaneous sentences, one spoken by a male talker of Southern Standard British English (SSBE) (target) and one by a female talker of SSBE (masker), mixed at 0 dB SNR. Each sentence contained five keywords. In order to avoid ceiling effects, the sentence pairs were mixed with 12-talker babble (6 male, 6 female), presented diotically at a SNR of −1 dB relative to the combined target and masker streams. All files were then equalised to 0.06 Pa [∼69.5 dB sound pressure level (SPL)].

The target (male) sentences and babble noise were presented diotically (i.e., creating a perceived location of approximately 0° azimuth), while the masker (female) sentences were presented at one of three positions. For half (i.e., 40) of the trials, the masker was presented at 0° azimuth (i.e., collocated with the target), for a quarter (20) of the trials, the masker was presented to the left of the target (at approximately −30° azimuth), and for the remaining 20 trials, the masker was presented to the right of the target (at approximately +30° azimuth). Perceived location was manipulated by changing the relative intensities in each stereo channel. This was achieved using the panning function from the audio manipulation module PyDub in PYTHON (see Ref. 41). This technique, which relies on interaural level differences, has previously been used to successfully generate perceived approximate azimuthal offset of similar auditory stimuli.[21]

For the visual stimuli, two cartoon images were sourced from the Free Clipart Library (see Ref. 42): one of a man (target-associated) and one of a woman (masker-associated). On each trial, the target image was located at the centre of the screen, while the masker image appeared either directly below the target image (i.e., visual collocated) or at the far left/right of the screen (i.e., visual separated).

The auditory stimuli and possible visual arrangements were combined to form four conditions, each containing 20 trials: 1, AC-VC (auditory collocated, visual collocated); 2, AS-VS (auditory separated, visual separated); 3, AC-VS (auditory collocated, visual separated); 4, AS-VC (auditory separated, visual collocated). Better performance under condition 3 compared to 1 would therefore represent a vSRM+ effect, while worse performance under condition 4 compared to 2 would represent a vSRM− effect. A schematic illustrating the different audiovisual conditions is presented in Fig. 1.

Half of the auditory trials with the masker stream at each displaced location (left or right) were used for each of the AS conditions. Conditions with separated auditory or visual stimuli (AS-VS, AC-VS, AS-VC) had ten trials for each direction of separation. For AS-VS trials, the direction of separation (i.e., masker offset) was always the same for the audio and visual stimuli.

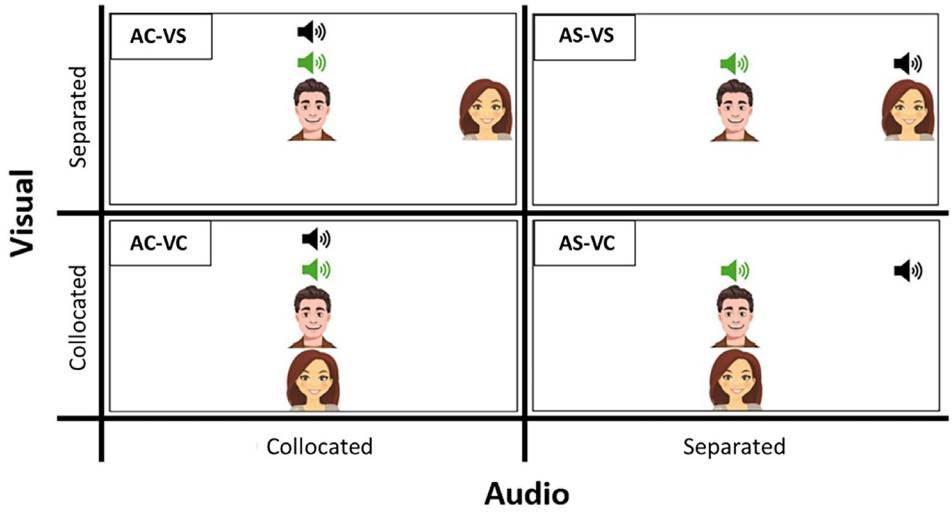

Fig. 1. Schematic illustrating the four audiovisual conditions used in the study. AC, auditory collocated; AS, auditory separated; VC, visual collocated; VS, visual separated.

2.3 Procedure

The study was run using the online testing platform Gorilla (gorilla.sc).[22] After giving informed consent and providing demographic details, participants were presented with a snippet of white noise at the same root-mean-square (RMS) level as the experimental stimuli and asked to adjust their volume to a comfortable listening level. Participants then completed a validated headphone check,[23] to ensure they were wearing stereo headphones. Participants who failed the check twice were prevented from continuing.

Before performing the main tasks, participants were introduced to the two talkers used in the experiment. For each talker, they were presented with one sentence in quiet, presented with the relevant associated image.

Participants then performed the sentence transcription task. Before starting the main task, they were presented with one example of a full audiovisual stimulus under the AS-VS condition. This was presented alongside a transcript of the correct response. Participants then completed three full practice trials under the AS-VS, AC-VS, and AS-VC conditions (i.e., one using each possible audio masker location). On each trial, participants saw a 500 ms central fixation cross, after which they were presented with an auditory stimulus with its associated visual images. After a 500 ms blank screen, they were given a text-entry box and asked to transcribe the sentence spoken by the male voice. They then proceeded to the main transcription task. The procedure was identical to the practice trials. The 80 experimental trials were presented in a random order, with the opportunity for a break every 20 trials.

Finally, participants completed the sound localization task. This used a subset of 30 stimuli from the sentence transcription task (10 from each audio masker location) with three practice trials (one from each audio masker location). On each trial, a 500 ms central fixation cross preceded a blank screen and the presentation of one auditory stimulus. Participants were then prompted to indicate where the masker (female) voice came from by selecting one of three onscreen buttons for centre, left, and right. Trials were presented in a single block in a random order.

Since the headphone check only determined whether participants wore headphones and whether the headphones were stereo, not whether the left and right channels were correctly assigned, the sound localization task was also used to exclude participants who may have had their headphones the wrong way round. Three participants were identified who responded "right" to left trials on four or more trials (out of ten) and who also responded "left" to right trials on four or more occasions (out of ten). These participants were removed from further analysis.

3. Results

All analyses were performed in R (v4.4.0) using R STUDIO (v2024.04.0) and the packages *lme4*, *emmeans*, and *ggplot2*.[24–26]

3.1 Sentence transcription task

The proportion of keywords correctly transcribed was calculated for each trial. Average performance is shown in Fig. 2.

Data were analysed at the trial level using generalized linear mixed models (GLMMs) with a binomial distribution and a logit link. The outcome variable was the proportion of correct keywords recorded. The model included fixed effects of auditory location (collocated, separated), visual location (collocated, separated), and their interaction, with a random intercept for participants. Fixed effects were treatment coded with the collocated conditions as the reference levels. Significance was assessed via likelihood ratio tests (LRTs) followed by Tukey-corrected *post hoc* pairwise comparisons.

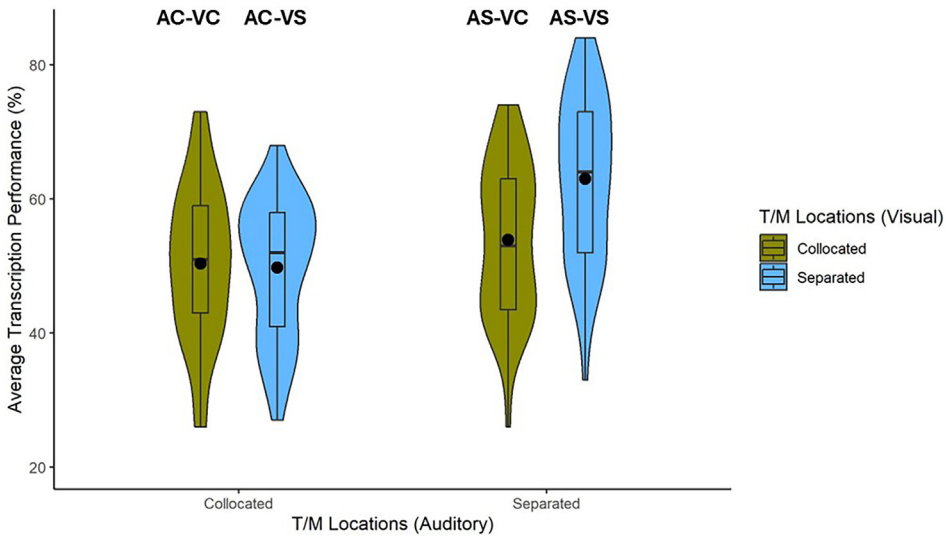

Fig. 2. Average transcription performance (percentage of correctly transcribed keywords) for each condition in the sentence transcription task. T/M, target/masker; AC, auditory collocated; AS, auditory separated; VC, visual collocated; VS, visual separated.

Odds ratios (ORs) are reported as a measure of effect size; these were derived by exponentiating the log odds (i.e., model coefficients) of the relevant fixed effects.

The model showed a significant main effect of auditory location, with better performance when the target and masker were auditorily separated than collocated [$X^2(1) = 14.40$, $p < 0.001$, OR = 1.16]. There was also a significant interaction of auditory × visual location [$X^2(1) = 55.31$, $p < 0.001$, OR = 1.51]. *Post hoc* tests (Tukey-corrected) showed that there were significant differences between all conditions ($p \leq 0.001$), except between AC-VC and AC-VS, where the difference was not significant ($p > 0.5$). There was no significant main effect of visual location ($p > 0.5$).

3.2 Sound localization task

Responses were coded as correct or incorrect on a trial-by-trial basis. Overall localization ability performance was relatively high (71.6% correct). The poorest performance was for the collocated masker (63.8%), with similar performance levels for the left-offset and right-offset maskers (74.9% and 76.0%, respectively). To assess the effect of localization ability on the sentence transcription task, a GLMM was run on the transcription scores with each participant's overall sound localization score as an additional predictor. Similar to the first GLMM, this model showed a significant main effect of auditory location [$X^2(1) = 14.60$, $p < 0.001$, OR = 1.16] and a significant interaction of auditory × visual location [$X^2(1) = 56.55$, $p < 0.001$, OR = 1.52]. It also showed a significant interaction of auditory location × localization score [$X^2(1) = 5.68$, $p < 0.05$, OR = 1.10]. *Post hoc* analysis using the *emtrends* function revealed a stronger positive relationship between localization scores and sentence transcription scores for the AS conditions compared to the AC conditions (estimated trends = 0.20 and 0.06, respectively). In other words, auditory localization ability was a stronger predictor of transcription performance for trials during which the auditory stimuli were separated as opposed to collocated. However, there was no significant main effect of sound localization score and there was no significant three-way interaction between localization score, auditory location, and visual location; there was also no significant effect of visual location (all $p > 0.05$).

4. Discussion

SPiN performance has been shown to improve when target and masker(s) are spatially separated (SRM). SRM is thought to be due to a combination of (1) reduced EM, allowing for more glimpses of the target through the masker and (2) location-based cues to auditory object identity, allowing for more successful streaming and deployment of selective attention. Separately, it has been shown that the perceived location of a sound can shift towards an associated but spatially displaced visual stimulus. It has therefore been suggested that it may be possible to induce SRM by spatially separating only the associated visual stimuli (vSRM+) and/or that genuine auditory SRM may be reduced if the associated visual stimuli are not spatially separated (vSRM−). However, existing findings are mixed and may have been conflated with individual differences in both lip-reading ability and sound localization ability. In the current study, we therefore used static images of faces as the talker-associated visuals during a speech-on-speech (one talker, one masker) listening task; we also obtained a measure of sound localization ability for each participant using the stimuli from the speech-on-speech task.

A classic auditory SRM effect was observed, with higher performance when the target and masker were spatially separated in the auditory domain than when they were collocated. Importantly, we did not observe a vSRM+ effect—that is, performance was not improved by spatially separating the visual stimuli associated with the target and masker while the

auditory stimuli remained collocated. However, we did observe a vSRM− effect: When the auditory stimuli were spatially separated, there was a reduction in auditory SRM when the associated visual stimuli were collocated. Nevertheless, performance was still higher in this condition than when both auditory and visual stimuli were collocated.

Taken together, these findings support an account of SRM in which both EM reduction and streaming cues play a role. Importantly, they also suggest that the presence or absence of streaming cues becomes more important in situations in which EM is already relatively low. This pattern can be understood in terms of the data- vs resource-limit framework proposed by Norman and Bobrow.[27] In this framework, a task is said to be data-limited if performance is determined primarily by the quantity of data available—in our case, the amount of the target that can be glimpsed through the masker. In contrast, a task is said to be resource-limited if performance is determined primarily by the processing resources allocated to the available data. In our case, these resources include the cognitive processes involved in streaming and attentional control. Central to Norman and Bobrow's account is the claim that the allocation of additional processing resources to data-limited tasks is unlikely to lead to performance improvements.

In the case of SPiN, cognitive abilities have been shown to predict speech-on-speech listening performance most robustly when EM is low,[21] which is consistent with Norman and Bobrow's resource-limited scenario. This pattern has also been shown when the target speech is unfiltered (as opposed to filtered)[28] or played at a slow rate (as opposed to time-compressed).[29] In other words, when target data availability is relatively high, performance is more strongly determined by the application of relevant processing resources. When target data availability is relatively low (high EM), however, performance is determined primarily by the amount of target data available, with processing resources playing a more minor role.[30]

In the current study, the AC-VS condition (collocated auditory stimuli with spatially separated visual stimuli) created a situation in which no additional data from the auditory target signal were available compared to AC-VC, but which provided visual cues to assist in the deployment of the relevant cognitive resources related to streaming and attentional control. Yet, performance was no higher under this condition compared to a condition under which both auditory and visual stimuli were collocated—that is, we did not observe a vSRM+ benefit. This therefore suggests that the basic SRM effect is underpinned primarily by EM reduction: When performance improves during genuine auditory SRM, it does so largely because more of the target stream is available to glimpse. In other words, performance is limited by target data availability rather than a lack of cues to aid the deployment of relevant cognitive processes. Since the ability to allocate additional processing resources is not leading to a performance improvement, the task can be thought of as falling towards the data-limit end of the spectrum. This characterization of auditory SRM accords with the various studies that have failed to replicate Driver's original demonstration of a vSRM+ benefit.[14,15]

We did, however, observe a vSRM− effect. Under the AS-VC condition (collocated visual stimuli with spatially separated auditory stimuli), performance was poorer than under the condition in which both auditory and visual stimuli were spatially separated: That is, collocated visual presentation of the target and masker reduced the auditory SRM benefit. In other words, although additional auditory information was available, participants were less able to use it when mismatched visual stimuli were present, suggesting that there is interference with the deployment of processes associated with streaming and attentional control. This in turn implies at least some role for cognitive processes in the overall SRM effect. Nevertheless, performance was still significantly better under the AS-VC condition compared to a condition in which both auditory and visual stimuli were collocated. This points to a two-step process, in which a basic SRM benefit can be derived from reductions in EM, but with an additional role for cognitive resources once more of the target data are available. Once the constraints on data availability are lifted, performance is limited by constraints on the deployment of cognitive resources—the task is now further towards the resource-limit end of the spectrum.

The precise mechanism underlying the vSRM− effect remains to be elucidated. However, three possibilities can be proposed. First, an account based on the VE suggests that the spatially separated auditory masker stimuli are pulled perceptually closer to the target when the visual masker and target stimuli are collocated. This perceived proximity may in turn interfere with listeners' ability to make use of the genuine spatial separation cues available from the audio. This account makes the strong prediction that the vSRM− effect should occur primarily when the visual target and masker stimuli are closer together than the auditory stimuli. Exactly how such a perceptual illusion might cause interference with cognitive processes, though, is unclear.

The second account is similar to the first, but relies only on the visual stimuli being in a different location than their associated auditory stimuli. Driver and Spence (2004) found poorer SPiN performance when an auditory target and its associated visual information (in this case, matched lip-read information) were presented in different places—an effect they attributed to the challenges of cross-modal division of attention and stimulus integration across different spatial locations.[31] In our case, the mismatched locations of the auditory and visual masker stimuli may have interfered with listeners' ability to segregate and stream—and thus selectively ignore—the interfering voice. This account therefore does not rely on the visual stimuli being closer together than their associated auditory stimuli; rather, any location-related audiovisual mismatch would be sufficient. This implies that the vSRM− effect would be observed if, for example, the target and masker visual stimuli were even further apart than their associated (spatially separated) auditory stimuli, or offset in a different plane.

A third possibility is that simply becoming aware of the mismatch between the auditory and visual locations is enough to disrupt cognitive processes through attentional capture (e.g., Ref. 32) This account does not assume any specific locations for the auditory vs visual stimuli or even that processes related to streaming and/or attentional control are uniquely vulnerable. Instead, the assumption is that any salient mismatch between associated auditory and visual stimulus locations is sufficient to disrupt various types of stimulus processing.

Note that all three proposed mechanisms involve a resource-limit scenario, since in all cases auditory spatial separation removes the data limit created by the initial auditory overlap. It is therefore the listener's ability to appropriately apply cognitive resources that is affected. The difference between the three accounts instead rests in the nature of the resource limit and the stimuli required to generate it. Note also that we are conceptualizing the role of cognitive resources as the second step in a two-stage process, as discussed above: That is, these resource limits become apparent once the basic auditory SRM benefit is already in place. Thus, our second and third potential mechanisms do not, for example, predict poorer performance under the AC-VS condition compared to the AC-VC position purely on the basis of the audiovisual location mismatch; rather, both conditions are predicted to be equally poor due to the collocated audio. Only when the audio is separated do we expect to see the vSRM− effect (i.e., poorer performance for the AS-VC condition compared to the AS-VS condition). Future work should aim to disambiguate these potential accounts by manipulating both the type of audiovisual location mismatch and the specific cognitive resources required to complete the task.

Importantly, these accounts do not rely on audiovisual integration [i.e., they do not require the associated auditory and visual stimuli to be perceptually grouped into a single unified multisensory object (e.g., Ref. 33)]. As mentioned above, the VE can occur even when participants believe that the relevant auditory and visual stimuli are unrelated,[11] and detrimental effects of audiovisual mismatches on processing can be observed using coincident but unrelated stimuli [as in an audiovisual Stroop paradigm (e.g., Ref. 34)]. However, it seems plausible to assume that any audiovisual effects will become larger as the perceived connection between the audio and visual stimuli grows stronger. Furthermore, during real-world audiovisual speech perception, auditory and visual information does indeed appear to be perceived as emanating from a single multisensory event (i.e., audiovisual integration takes place).[35] Full exploration of the nature of this integration is beyond the scope of this article; however, it is worth noting that connections between stimuli presented across different modalities are likely to be stronger when those stimuli share temporal correspondences.[36] In the current study, static images were used to avoid any potential confounds with lip-reading ability; this also, however, removed the potential for matched audiovisual temporal dynamics, thus likely resulting in a weaker audiovisual connection. Future studies should use both non-facial dynamic visual stimuli (such as pulsing circles)[19] and moving faces (both with and without lip movements) in order to explore this issue further. Only once results are obtained using a range of stimuli featuring a variety of more and less naturalistic audiovisual correspondences can the implications of the current findings for models of audiovisual speech perception be properly understood.

If the static visual stimuli led to a relatively weak connection between the auditory and visual components, this may in turn have led to a weak or absent VE: That is, although audiovisual integration is not required to generate a VE *per se*, a perceptual connection between the different voices and their visual representations would have been necessary in the current study in order for an individual voice to be perceptually pulled towards its associated image. If the VE was weak or absent, this could explain why no vSRM+ was observed. Again, this suggests that future studies should use a range of visual stimuli incorporating dynamic temporal changes. Future work should also attempt to assess the strength of the experienced VE, despite the practical difficulties of doing so.[17]

However, even studies which used temporally-aligned dynamic visual stimuli did not always find a vSRM+ effect. Indeed, one such study (Valzolgher *et al.*)[19] demonstrated neither a vSRM+ nor a vSRM− effect. This disparity with our findings may be due to differences in the stimuli and task between our study and theirs. In Valzolgher *et al.*, it was the position of the target speech that was manipulated relative to a masker that was always offset to the side. In the current study, we manipulated the position of the masker relative to a target always presented at 0° azimuth. Furthermore, Valzolgher *et al.* used streams of digits as their targets (and maskers in experiment 2) rather than meaningful sentences. Whether or not these methodological differences reliably affect vSRM remains an open question. However, it is worth noting that overall performance in the study by Valzolgher *et al.* was higher than in the current study (around 3.5/5 correct digits, or 70% correct, compared to around 54% correct here). This may suggest that segregation of the target and masker streams was generally easier in the work by Valzolgher *et al.*, thus rendering it less vulnerable to interference from incongruent visual information.

Individual differences in auditory localization ability did not impact overall performance. However, there was an interaction between localization ability and auditory separation, indicating that localization ability was more strongly related to performance when the auditory stimuli were separated. In other words, auditory SRM benefits were larger for those listeners with better auditory localization skills. This finding is in line with studies showing that knowing "where to listen" can improve performance when target and masker are spatially separated[37] and suggests that—at an individual level—having a more precise sense of target location allows for more focused spatial auditory attention, thus improving target perception. However, this pattern is not always observed, with several previous studies finding no relationship between localization abilities and SRM.[38–40] In the current study, localization ability was measured to investigate its effect on vSRM specifically; nevertheless, it is worth noting its relationship to auditory SRM in light of these mixed previous results.

We found no effect of auditory localization ability on the influence of the visual stimuli (i.e., no relationship between auditory localization and vSRM effects). This was unexpected given the conclusion of Jack *et al.*[17] that vSRM effects may rely on the strength of the experienced VE, which in turn seems likely to rely on relatively good auditory localization abilities. However, although it seems reasonable to assume that good auditory localization skills are necessary for experiencing the VE, they are not necessarily sufficient.

More broadly, the generalizability of the findings reported here is limited by the nature of the online setup. Although the results from the localization task confirm the overall success of the auditory spatial manipulation, it cannot be guaranteed that the perceived degree of separation was identical across listeners, given the likely variability in hardware, software, and listening environments. Similarly, although the visual stimuli were clearly either collocated or separated, we could not control for screen size or viewing distance/angle. Future studies should attempt to better control these aspects of stimulus presentation, either through a more rigorous online implementation (incorporating, e.g., verification of screen size and angle) or by using a laboratory implementation in which visual stimuli are physically collocated with loudspeakers.

In conclusion, these results suggest that it is possible to influence performance on speech-perception-in-noise tasks through the use of talker-associated visual representations, even when the visual representations contain no lip movements or other cues to speech. However, it was only possible to *disrupt* existing auditory SRM: There was no *enhancement* of auditory SRM through visual spatial separation. This is in line with an account of SPiN in which cognitive processes—such as those related to streaming and attentional control—play only a limited role when data availability is poor, such as when there is no auditory SRM to increase target glimpsing. In contrast, the role of cognitive resources becomes larger as the target signal becomes more available, explaining the detrimental role of misleading visual cues to streaming in the presence of auditory SRM. These effects should be further investigated using laboratory-based tasks and a range of alternative stimuli.

Acknowledgments

The data for this experiment were collected by Charlotte Levy as part of her B.Sc. dissertation at the University of York, York, UK. Further support was provided by the Economic and Social Research Council (ESRC).

Author Declarations

Conflict of Interest

The authors have no conflicts to disclose.

Ethics Approval

Ethical approval was granted by the local departmental ethics committee (Reference No. 232469). All participants provided informed consent, and all procedures were performed in compliance with relevant laws and institutional guidelines.

Data Availability

The data that support the findings of this study are openly available via the Open Science Framework at https://osf.io/fkupy/.

References

[1]D. S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**(3), 1101–1109 (2001).
[2]M. Cooke, "Glimpsing speech," J. Phon. **31**(3–4), 579–584 (2003).
[3]B. G. Shinn-Cunningham, "Object-based auditory and visual attention," Trends Cogn. Sci. **12**(5), 182–186 (2008).
[4]R. Y. Litovsky, "Spatial release from masking," Acoust. Today **8**(2), 18–25 (2012).
[5]B. A. Edmonds and J. F. Culling, "The spatial unmasking of speech: Evidence for better-ear listening," J. Acoust. Soc. Am. **120**(3), 1539–1545 (2006).
[6]K. Allen, D. Alais, B. Shinn-Cunningham, and S. Carlile, "Masker location uncertainty reveals evidence for suppression of maskers in two-talker contexts," J. Acoust. Soc. Am. **130**(4), 2043–2053 (2011).
[7]G. Kidd, C. R. Mason, and F. J. Gallun, "Combining energetic and informational masking for speech identification," J. Acoust. Soc. Am. **118**(2), 982–992 (2005).
[8]R. L. Freyman, K. S. Helfer, D. D. McCall, and R. K. Clifton, "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**(6), 3578–3588 (1999).
[9]L. K. Canon, "Intermodality inconsistency of input and directed attention as determinants of the nature of adaptation," J. Exp. Psychol. **84**(1), 141–147 (1970).
[10]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**(3), 967–979 (1973).
[11]M. Radeau and P. Bertelson, "The after-effects of ventriloquism," Q. J. Exp. Psychol. **26**(1), 63–71 (1974).
[12]J. Driver, "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading," Nature **381**(6577), 66–68 (1996).
[13]P. F. Heard and P. Webb, "The ventriloquist effect occurs with auditory and visual delay," paper presented at the *24th European Conference on Visual Perception*, Kusadasi, Turkey (2001) (as cited in Ref. 17).

[14]D. S. Brungart, A. J. Kordik, and B. D. Simpson, "Audio and visual cues in a two-talker divided attention speech-monitoring task," Hum. Factors **47**, 562–573 (2005).
[15]D. S. Rudmann, J. S. McCarley, and A. F. Kramer, "Bimodal displays improve speech comprehension in environments with multiple speakers," Hum. Factors **45**, 329–336 (2003).
[16]S. M. Leech, "The effect on audiovisual speech perception of auditory and visual source separation," Ph.D. thesis, University of Sussex, UK (2001) (as cited in Ref. 17).
[17]B. N. Jack, R. P. O'Shea, D. Cottrell, and W. Ritter, "Does the ventriloquist illusion assist selective listening?," J. Exp. Psychol. Human **39**(5), 1496–1502 (2013).
[18]S. Savel, "Individual differences and left/right asymmetries in auditory space perception. I. Localization of low-frequency sounds in free field," Hear. Res. **255**(1–2), 142–154 (2009).
[19]C. Valzolgher, E. Giovanelli, R. Sorio, G. Rabini, and F. Pavani, "Can visual capture of sound separate auditory streams?," Exp. Brain Res. **240**(3), 813–824 (2022).
[20]E. H. Rothauser, "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**(3), 225–246. (1969).
[21]S. Knight, L. Rakusen, and S. Mattys, "Conceptualising acoustic and cognitive contributions to divided-attention listening within a data-limit versus resource-limit framework," J. Mem. Lang. **131**, 104427 (2023).
[22]A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder," Behav. Res. **52**(1), 388–407 (2020).
[23]K. J. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," Atten. Percept. Psychophys. **79**(7), 2064–2072 (2017).
[24]D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," J. Stat. Soft. **67**(1), 1–48 (2015).
[25]R. Lenth, "emmeans: Estimated marginal means, aka least-squares means," R package version 1.10.1, https://rvlenth.github.io/emmeans/.
[26]H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, New York, 2016).
[27]D. A. Norman and D. G. Bobrow, "On data-limited and resource-limited processes," Cogn. Psychol. **7**(1), 44–64 (1975).
[28]E. Janse and S. J. Andringa, "The roles of cognitive abilities and hearing acuity in older adults' recognition of words taken from fast and spectrally reduced speech," Appl. Psycholinguist. **42**(3), 763–790 (2021).
[29]R. M. O'Leary, J. Neukam, T. A. Hansen, A. J. Kinney, N. Capach, M. A. Svirsky, and A. Wingfield, "Strategic pauses relieve listeners from the effort of listening to fast speech: Data limited and resource limited processes in narrative recall by adult users of cochlear implants," Trends Hear. **27**, 23312165231203514 (2023).
[30]S. L. Mattys, R. M. O'Leary, R. A. McGarrigle, and A. Wingfield, "Reconceptualizing cognitive listening," Trends Cogn. Sci. in press (2025).
[31]J. Driver and C. J. Spence, "Spatial synergies between auditory and visual attention," in *Attention and Performance: Conscious and Nonconscious Information Processing*, edited by C. Umiltà and M. Moscovitch (MIT Press, Cambridge, MA, 1994), Vol. 15, pp. 311–331.
[32]H. Krause, T. R. Schneider, A. K. Engel, and D. Senkowski, "Capture of visual attention interferes with multisensory speech processing," Front. Integr. Neurosci. **6**, 67 (2012).
[33]J. K. Bizley, R. K. Maddox, and A. K. Lee, "Defining auditory-visual objects: Behavioral tests and physiological mechanisms," Trends Neurosci. **39**(2), 74–85 (2016).
[34]S. E. Donohue, L. G. Appelbaum, C. J. Park, K. C. Roberts, and M. G. Woldorff, "Cross-modal stimulus conflict: The behavioral effects of stimulus input timing in a visual-auditory Stroop task," PLoS One **8**(4), e62802 (2013).
[35]A. Vatakis and C. Spence, "Crossmodal binding: Evaluating the 'unity assumption' using audiovisual speech stimuli," Percept. Psychophys. **69**(5), 744–756 (2007).
[36]C. Spence, "Audiovisual multisensory integration," Acoust. Sci. Technol. **28**(2), 61–70 (2007).
[37]G. Kidd, T. L. Arbogast, Jr., C. R. Mason, and F. J. Gallun, "The advantage of knowing where to listen," J. Acoust. Soc. Am. **118**, 3804–3815 (2005).
[38]R. Drullman and A. W. Bronkhorst, "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," J. Acoust. Soc. Am. **107**, 2224–2235 (2000).
[39]M. L. Hawley, R. Y. Litovsky, and H. S. Colburn, "Speech intelligibility and localization in a multi-source environment," J. Acoust. Soc. Am. **105**, 3436–3448 (1999).
[40]A. M. Rothpletz, F. L. Wightman, and D. J. Kistler, "Informational masking and spatial hearing in listeners with and without unilateral hearing loss," J. Speech Lang. Hear. Res. **55**(2), 511–531 (2012).
[41]PyDub, "PyDub module in PYTHON," https://pydub.com/.
[42]Free Clipart Library, "Free Clipart Library," https://clipart-library.com/.