Spatial alignment between faces and voices improves selective attention to audio-visual speech

Justin T. Fleming, Ross K. Maddox and Barbara G. Shinn-Cunningham

Citation: The Journal of the Acoustical Society of America **150**, 3085 (2021); doi: 10.1121/10.0006415
View online: https://doi.org/10.1121/10.0006415
View Table of Contents: https://asa.scitation.org/toc/jas/150/4
Published by the Acoustical Society of America

JASA ARTICLE

Spatial alignment between faces and voices improves selective attention to audio-visual speech

Justin T. Fleming,[1,a)] Ross K. Maddox,[2,b)] and Barbara G. Shinn-Cunningham[3,c)]

[1]*Speech and Hearing Bioscience and Technology Program, Harvard University, 243 Charles Street, Boston, Massachusetts 02114, USA*

[2]*Department of Biomedical Engineering, University of Rochester, 430 Elmwood Avenue, Rochester, New York 14620, USA*

[3]*Neuroscience Institute, Carnegie Mellon University, 4825 Frew Street, Pittsburgh, Pennsylvania 15213, USA*

ABSTRACT:

The ability to see a talker's face improves speech intelligibility in noise, provided that the auditory and visual speech signals are approximately aligned in time. However, the importance of spatial alignment between corresponding faces and voices remains unresolved, particularly in multi-talker environments. In a series of online experiments, we investigated this using a task that required participants to selectively attend a target talker in noise while ignoring a distractor talker. In experiment 1, we found improved task performance when the talkers' faces were visible, but only when corresponding faces and voices were presented in the same hemifield (spatially aligned). In experiment 2, we tested for possible influences of eye position on this result. In auditory-only conditions, directing gaze toward the distractor voice reduced performance, but this effect could not fully explain the cost of audio-visual (AV) spatial misalignment. Lowering the signal-to-noise ratio (SNR) of the speech from +4 to −4 dB increased the magnitude of the AV spatial alignment effect (experiment 3), but accurate closed-set lipreading caused a floor effect that influenced results at lower SNRs (experiment 4). Taken together, these results demonstrate that spatial alignment between faces and voices contributes to the ability to selectively attend AV speech.

 https://doi.org/10.1121/10.0006415

(Received 27 April 2021; revised 30 August 2021; accepted 1 September 2021; published online 26 October 2021)

[Editor: Matthew J. Goupell] Pages: 3085–3100

I. INTRODUCTION

The ability to see a person's face as they are speaking leads to a well-established improvement in speech recognition accuracy, particularly in high levels of background noise (Crosse *et al.*, 2016; Erber, 1975; MacLeod and Summerfield, 1987; Ross *et al.*, 2006; Schwartz *et al.*, 2004; Sumby and Pollack, 1954). This benefit critically depends on the temporal relationship between the auditory and visual signals (Grant and Greenberg, 2001; Grant and Seitz, 2000). For both speech and nonspeech stimuli, temporal coherence between cross-modal features drives binding of the sensory inputs into a single perceptual object (Maddox *et al.*, 2015). If the signals are offset beyond the limits of a temporal binding window, benefits of audio-visual (AV) binding are abolished; for AV speech, this window spans an auditory offset of roughly −40 to 200 ms relative to the visual input (Conrey and Pisoni, 2006; van Wassenhove *et al.*, 2007). Electroencephalography studies have shown that event-related potentials (ERPs) elicited by AV speech stimuli have reduced latencies and amplitudes compared to their auditory-only counterparts, suggesting that vision can provide anticipatory information that facilitates auditory processing (Besle *et al.*, 2004; Peelle and Sommers, 2015; van Wassenhove *et al.*, 2005). However, as with the behavioral benefits of integration, introducing temporal offsets between the auditory and visual stimuli systematically reduces the strength of these AV ERP modulations (Simon and Wallace, 2018).

Consensus has not been reached regarding the practical role that AV *spatial* alignment plays in integration. On one hand, several cross-modal illusions are known to be unaffected by spatial misalignment between their unisensory components. For instance, the McGurk effect, in which visual articulator movements influence auditory perception of syllables, occurs even with a large spatial disparity between the talker's video and voice (Bertelson *et al.*, 1994; Jones and Munhall, 1997). The same is true of the sound-induced flash illusion; briefly presented auditory stimuli can influence the number of perceived visual stimuli even if the cross-modal signals are spatially misaligned (DeLoss and Andersen, 2015; Innes-Brown and Crewther, 2009). However, in a study that introduced a more complex version of the sound-induced flash paradigm with multiple competing streams of auditory and visual stimuli, the strength of the effect was in fact modulated by spatial alignment within each AV stream (Bizley *et al.*, 2012). Similarly, in the pip

[a)]Present address: Department of Speech-Language-Hearing Sciences, University of Minnesota, 164 Pillsbury Dr. SE, Minneapolis, MN 55455, USA. Electronic mail: jtf@umn.edu ORCID: 0000-0001-7070-1695.

[b)]Also at: Department of Neuroscience, the Del Monte Institute for Neuroscience, and the Center for Visual Science at the University of Rochester, Rochester, NY 14620, USA, ORCID: 0000-0003-2668-0238.

[c)]ORCID: 0000-0002-5096-5914.

and pop effect, in which visual search is facilitated by a tone played in synchrony with a visual target (Van der Burg *et al.*, 2008), behavioral and electrophysiological signatures of integration were observed regardless of AV spatial alignment between visual targets and tones. However, in a version of the task with two competing AV stimuli, search benefits and ERP signatures of integration did depend on hemifield alignment between the auditory and visual components of each stimulus (Fleming *et al.*, 2020). Taken together, these findings indicate that in relatively simple scenes lacking in multisensory competition, temporal coherence alone is sufficient to drive AV integration. When the sensory environment becomes more complex, however, AV spatial alignment may represent an important secondary cue to aid selective integration of the correct inputs.

Some previous studies have investigated visual facilitation of auditory selective attention using speech stimuli in "cocktail party" listening environments. For instance, visual input that is temporally coherent with one stream in an auditory mixture improves perceptual and neural tracking of that stream (Atilgan *et al.*, 2018; Zion Golumbic *et al.*, 2013). However, to our knowledge, no existing studies have examined whether spatial alignment (or explicit misalignment) between faces and voices influences this AV selective attention advantage. In investigating this, target and distractor voices would necessarily be spatially separated from one another, providing another auditory selective attention benefit in the form of spatial release from masking (SRM) (Litovsky, 2012). Acoustic advantages arising from the relative positions of the target and masker contribute to SRM, but their contribution is relatively minor when competing speech is present. Instead, the main benefits of spatially separating the competing talkers arise from facilitating perceptual segregation of the voices, thereby allowing listeners to focus attention selectively on the target talker based on its location (Durlach *et al.*, 2003; Watson, 1987; Wu *et al.*, 2005). Given that AV binding can also improve perceptual segregation, the current work aimed to tease apart these benefits. In one previous study that combined SRM with the ability to see a target talker's face, the presence of visual input was found to provide a greater speech recognition benefit when the target and masker speech signals were spatially coincident (Helfer and Freyman, 2005). However, as with most multisensory selective attention paradigms using speech stimuli, this study focused on how a *single* visual stimulus can perceptually highlight a target speech stream in an auditory mixture. The presence of competing sensory inputs in both audition and vision may provide a closer approximation of real-world communication challenges.

In the present study, we aimed to determine whether AV spatial alignment improves selective attention to speech in the presence of speech-shaped noise and a competing AV talker. Participants performed a speech selective attention task, which required them to pay attention to a cued target talker while ignoring a distractor talker, and then indicate which of four words was spoken by the target talker. Importantly, multiple cues were available to separate the target and distractor speech streams (e.g., differences in pitch, vowel spaces, and other talker-specific characteristics), so attention to spatial features was not explicitly required to perform the task.

In experiment 1, we examined whether spatial alignment between corresponding faces and voices affected the magnitude of AV benefits in speech attention. Aligned and misaligned AV conditions were compared against auditory-only conditions with spatially separated or spatially coincident speech streams. In experiment 2, we examined gaze fixation effects on auditory spatial attention, and whether they may have influenced the results of experiment 1. In experiment 3, we tested the effect of AV spatial alignment at increasingly challenging SNRs. Finally, in experiment 4 we measured lipreading performance on this closed-set task using a visual-only condition. Across these experiments, spatial alignment between corresponding faces and voices provided a consistent selective attention benefit. While gaze position alone had an effect, it could not account for the full effect of AV spatial alignment (experiment 2). The effect of spatial alignment between faces and voices may be magnified in noisier settings (experiment 1 and experiment 3), though a future open-set version of this task is needed to confirm this (experiment 4).

II. GENERAL METHODS

There were several common aspects across the four experiments presented in this study. These general procedures will first be described, with experiment-specific methods covered in subsequent subsections. All experiments were presented online using the Gorilla Experiment Builder (Gorilla, 2021), and subject recruitment was conducted using the Prolific online recruitment service (Prolific, 2021). Participants were required to use a laptop or desktop computer and either the Microsoft Edge or Google Chrome web browser, due to known issues with media autoplay in other web browsers. To be included in the study, participants were required to be 18–55 years old, have learned English as their first language, and have no known hearing loss. Participants were allowed to complete only one of the four experiments. Participants provided informed consent, and all study procedures were approved by the Carnegie Mellon University Institutional Review Board.

A. Screening and pre-experiment tasks

Before participants were allowed to start the main experiment, they had to complete several screening and setup tasks designed to ensure their web browser settings and audio equipment were suitable to perform the study. First, a questionnaire at the end of the consent form asked participants to confirm the first language and hearing status they reported in Prolific; those who failed to confirm this information were rejected. Next, a brief piece of music was automatically played to ensure that participants had autoplay enabled in their web browser. If they could not hear the music, instructions for enabling autoplay were provided.

https://doi.org/10.1121/10.0006415

If they did not want to change their browser settings, participants were also given the option to withdraw from the study at this point.

Next, participants completed an illusory pitch detection task based on the Huggins pitch phenomenon to check that they were using headphones. In the Huggins pitch effect, identical white noise is played to the two ears, except that the noise is phase-shifted by 180° in a narrow frequency band in one ear. Monaurally, this phase shift is undetectable, but when presented dichotically, participants perceive a pitch corresponding to the narrowband noise that is phase-inverted between the ears (Chait *et al.*, 2006; Cramer and Huggins, 1958). Since free-field interference disrupts this interaural phase offset, screening tasks based on Huggins' pitch are highly selective for participants who are using headphones (Milne *et al.*, 2020). On each trial, three binaural noise stimuli were presented sequentially, one of which contained a Huggins' pitch stimulus. Participants made a three-alternative forced choice judgment about which interval contained the "hidden tone." Participants were first given an example trial, on which they were told which interval contained the tone. They then performed six trials of this task without feedback. All trials needed to be answered correctly to proceed to the main task, but participants were allowed one retry with six new trials if they did not pass on the first attempt.

Participants were next asked to set their computer volume in preparation for the main experiment. An audio-only speech stimulus resembling those used in the main study (two female talkers embedded in speech-shaped noise) was played, and participants were asked to turn up their computer volume until the stimulus was as loud as possible without becoming uncomfortable. Finally, participants performed a brief spatial hearing task using similar audio-only speech stimuli. Prior to starting this task, participants were instructed to check that their headphones were not on backwards. On each trial, one of the talkers from the main experiment spoke a five-word sentence, which was spatialized using non-individualized head-related transfer functions (HRTFs) to either −15° or 15° azimuth. These sentences were not repeated by this talker in the actual experiment. Participants' task was to judge whether the speech came from the left or the right. After two practice trials with feedback, participants were required to answer five out of six trials correct without feedback to advance to the main experiment.

B. Trial structure and task

During the main experiment, participants were asked to pay attention to one talker's speech while ignoring the other talker. The general timeline of a trial is illustrated in Fig. 1(A). Each trial started with 1 s of fixation, with the fixation position changing based on the target talker and experimental condition, but remaining constant throughout the trial. Next, the word "nine" was spoken by the target talker to inform participants which talker to attend for the upcoming trial. The stimulus configuration was also present in the cue (e.g., video on or off, spatial location of the target talker's voice). This was followed by another 1.5 s of fixation, with a separate token of speech-shaped noise (SSN) coming on in each ear for the last 500 ms. The SSN continued throughout presentation of the two competing speech stimuli. Stimulus presentation was followed by another 500 ms of fixation, and finally the participant response screen.

At the response screen, participants were shown four words, one of which had been spoken by the target talker. Participants' task was to identify this target word and click on it using their mouse. Among the non-target word options, one or two (chosen randomly on each trial) always came from the distractor talker's stream. The other one or two word options were selected randomly from words present in the stimulus corpus, but not spoken by either talker on the current trial. The instructions stated that participants should respond as quickly as possible without sacrificing accuracy.

C. Stimuli

The original speech stimuli were high-definition AV recordings of two female talkers saying short sentences in a neutral affect, drawn from the Sensimetrics Speech Test Video Corpus (Sensimetrics, 2021). The video components of these stimuli are limited to the talkers' shoulders and above and do not include hand gestures. The sentences are constrained to a syntactic structure of Name–Verb–Number–Adjective–Plural Noun (for example, "Peter gives nine red tables"). One of 10 interchangeable words can appear in each position. The corpus includes each talker saying 500 unique sentences of this structure; a subset of 160 of these sentences were chosen randomly from each talker for this experiment. On each trial, stimulus pairings from the two talkers were restricted such that all five words were different between the two talkers. Within each talker, no sentences were repeated during the experiment.

On each trial, the auditory signals from the two talkers underwent the following processing steps: amplitude normalization, onset and offset ramping, time alignment, spatialization, and the addition of speech-shaped noise. The speech envelopes were first approximated separately for each stimulus by lowpass filtering the signals (third order Butterworth filter, 8 Hz cutoff frequency). The speech-on portions were estimated by finding the first and last points at which the envelope crossed an arbitrary threshold value. The root mean square (RMS) amplitude of each stimulus was calculated within the speech-on portion, and then the entire signal was scaled to an RMS level such that no clipping would occur across the entire stimulus set. This procedure ensured that the stimuli were amplitude-normalized across talkers and trials; however, given that the actual stimulus level was set by each participant in this online study, we did not control absolute sound levels.

The first and last 30 ms of the stimuli were cosine-ramped to avoid onset and offset artifacts. Within each pair of stimuli, the auditory signals were then time-aligned such that the speech-on portions overlapped maximally.

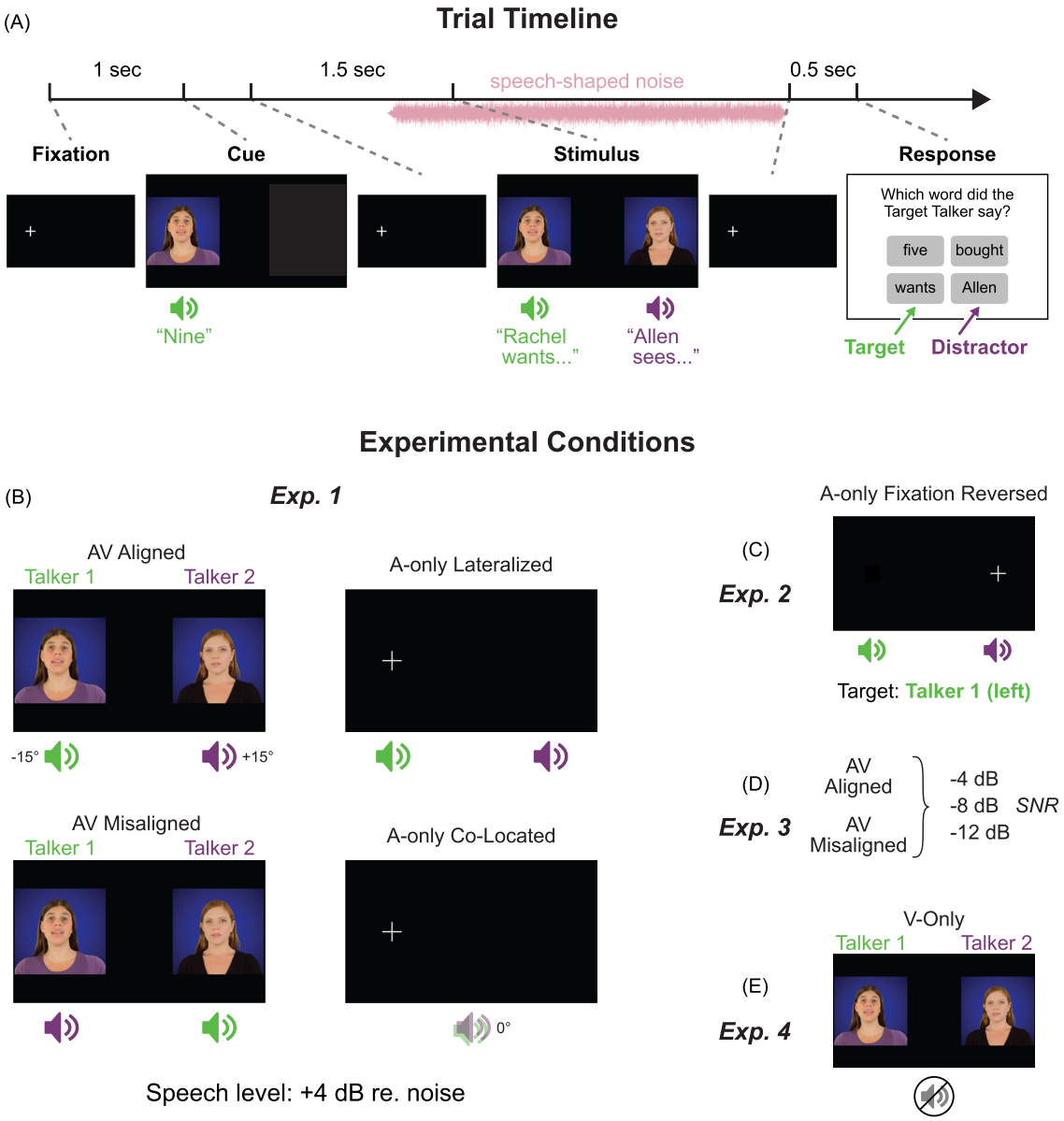

FIG. 1. (Color online) Trial timeline and illustration of experimental conditions. All panels show talker 1 (green, left) as the target talker. (A) The timeline of trial events is shown for an AV trial in which the faces and voices were spatially aligned. Audio-visual configurations for the four conditions of experiment 1 are shown in (B), and the new conditions added in the subsequent experiments are shown in (C)–(E). Auditory stimulus locations are shown beneath the corresponding video snapshots. In auditory-only (A-only) conditions, the fixation cross-remained on the screen throughout the trial.

Whichever signal had their earlier onset was delayed such that the midpoints of the speech-on portions matched, with the amount of time shift rounded to an integer multiple of the video frame rate to preserve AV alignment. Zeros were appended to the other signal to match stimulus lengths. The stimuli were then separately spatialized to −15°, 0°, or 15° azimuth (depending on the experimental condition) and 0° elevation by convolving them with non-individualized HRTFs from the CIPIC database (Algazi *et al.*, 2001).

Speech-shaped noise was generated based on a random subset of 60 experimental stimuli (30 samples from each of the talkers). To generate the noise, we randomized and concatenated the 60 stimuli, computed the discrete fast Fourier transform (FFT), randomized the phases of the frequency components, and then returned the stimulus to the time domain with the inverse FFT. The resulting noise stimulus was approximately 186 s long; a random time segment of this signal was extracted for each stimulus. The level of the noise was measured by RMS, and depending on the condition, scaled to an amplitude of −4, 4, 8, or 12 dB relative to the speech signal at the louder ear (speech was embedded in noise in all experiments). Two separate noise tokens were drawn on each trial, one of which was spatialized to −15° and the other to +15° azimuth, matching the locations of the speech signals. Each noise stimulus was then added to the corresponding spatialized speech signal.

We performed minimal additional processing on the video components of the stimuli, but to align them with the auditory stimuli, we duplicated the first or last frame until the video and audio lengths matched. Finally, the two

auditory and visual (if present) stimulus components were combined in Adobe Premiere Pro. Fixation and cue trial phases were also added at this stage, and complete trials were exported as MP4 video files. To prevent lagging or freezing in a web browser-based experimental environment, the stimuli were compressed using default online video settings in the Handbrake Open Source Video Transcoder software (The HandBrake Team, 2021).

D. Data analysis

Proportion correct and response time (RT) data were analyzed using mixed effects models. For the accuracy data, binomial models (using the logit link function) were employed at the level of individual trials. For the RT data, the median RT was first calculated for each participant and condition to limit the influence of anomalously slow RTs. These median RTs were then analyzed using linear mixed effects models. In both cases, the significance of fixed effects terms and their interactions (where appropriate) were assessed by subjecting the model to a type III analysis of variance (ANOVA), with p-values based on the Satterthwaite approximation for degrees of freedom. *Post hoc* comparisons between each relevant pair of factor levels were made by extracting model terms for these contrasts, then cycling which level was treatment-coded as baseline until all the necessary contrasts were computed. To account for multiple comparisons, the p-value criterion for assessing the significance of these model terms was adjusted using the Bonferroni-Holm correction, assuming a starting alpha level of $p = 0.05$.

For plotting, each participant's median RT data were centered on the average of their median RTs across conditions due to the large inter-subject variability in RT. Stimulus processing and data organization were performed in PYTHON, while statistical analyses (using the lme4 and lmerTest packages) and figure generation were conducted in R.

III. EXPERIMENT 1

A. Experiment 1 methods

1. Participants

One hundred participants completed all experimental procedures for experiment 1. Ten additional participants completed some but not all components; two failed to confirm either their first language or hearing status as reported in Prolific, six failed one of the headphone checks and were rejected, and two started the main experiment but did not finish it. Of the 100 complete datasets, four were rejected because participants performed below chance in one or more of the experimental conditions, making it difficult to verify that they were consistently attending the correct talker. Thus, the final dataset comprised 96 participants (mean age = 30.1 years, SD = 8.8; 57 female, 38 male, 1 non-binary). Participants who failed headphone screenings or did not complete the task due to technical issues were awarded partial compensation. Participants who completed all study components were paid a flat amount of $5.50, corresponding to an average pay rate of $7.77/h.

2. Experiment design

Participants performed the speech selective attention task in four different conditions [Fig. 1(B)]. In the *AV aligned* condition (top-left panel), participants heard two competing speech streams and saw corresponding video of the talkers' faces. The videos were positioned on opposite sides of the screen, with each talker's video appearing in the same location on all trials. The auditory signals were spatialized to −15° and 15° such that each voice was presented in the same hemifield as the corresponding face. In the *AV misaligned* condition (bottom-left panel), the video components were structured identically to the AV aligned condition (with the same talker's face always appearing on the left), but the voice spatialization was reversed, such that each voice was presented in the opposite hemifield as the corresponding face. In both AV conditions, participants were instructed to look at the target talker's face in whatever way felt natural.

Two auditory-only (A-only) control conditions were also included so we could test for perceptual advantages of being able to see the talkers' faces. In the *A-only lateralized* condition [Fig. 1(B), top-right panel], voices were spatialized in the same was as in the AV aligned condition, but the video components were removed and replaced with a constant fixation cross. Fixation was lateralized to the same hemifield as the target voice to control for possible eye position effects on auditory localization and attention (Maddox *et al.*, 2014; Reisberg *et al.*, 1981). Finally, in the *A-only co-located* condition (bottom-right panel), both voices were spatialized to 0° azimuth, which removed the benefit of SRM (Litovsky, 2012). In this condition, the fixation cross was presented in the "stereotyped" hemifield for the target talker (i.e., where their voice was presented in the AV aligned and A-only lateralized conditions) such that fixation would be lateralized similarly across conditions.

In all conditions of experiment 1, the speech-shaped noise was set to be 4 dB less intense than the speech. As with all experiments, the noise was spatialized to the same azimuthal locations as the two speech signals. Prior to starting the main experiment, participants were given one practice trial from each condition with feedback. If they answered a practice trial incorrectly, they were asked to repeat the trial until they correctly chose the target word. Trials from the four conditions were randomly intermixed throughout the experiment. Participants performed 40 trials of each condition (160 trials total) without feedback. An opportunity to take a break was provided every 10 trials.

B. Experiment 1 results

1. Speech attention task accuracy

A binomial mixed effects model was used to analyze task accuracy. The model had one fixed effect term, condition, with the four experimental conditions as levels. Random effects included participant-specific intercepts for each

condition and a stimulus term. A fuller version of the model also included a random effect term for participant age, but removing this term did not result in a significantly worse fit to the data as measured by the Akaike or Bayesian information criterion, and so the simpler model was favored. The structure of the final model was as follows, in Wilkinson notation,

$$\text{logit}(Correct) \sim Condition + (1 + Condition \,|\, ID) + (1 \,|\, Stimulus).$$

An ANOVA conducted on this model revealed a main effect of condition [$X^2 = 204.58$, df $= 3$, $p = 4.33 \times 10^{-44}$; Fig. 2(A)]. *Post hoc* comparisons (Wald's z tests on model contrasts, followed by Bonferroni-Holm correction of the alpha criterion) revealed that participants performed significantly worse on the A-only co-located condition than any of the other three (vs AV aligned, $z = -13.70$, $p = 1.03 \times 10^{-42}$; vs AV misaligned, $z = -8.96$, $p = 3.02 \times 10^{-19}$; vs A-only lateralized, $z = -9.60$, $p = 7.78 \times 10^{-22}$). The difference between the two A-only conditions validates a strong benefit of SRM, on the order of a 30% improvement in target speech recognition, using an online platform. To assess AV benefits beyond spatial separation of the two voices, each

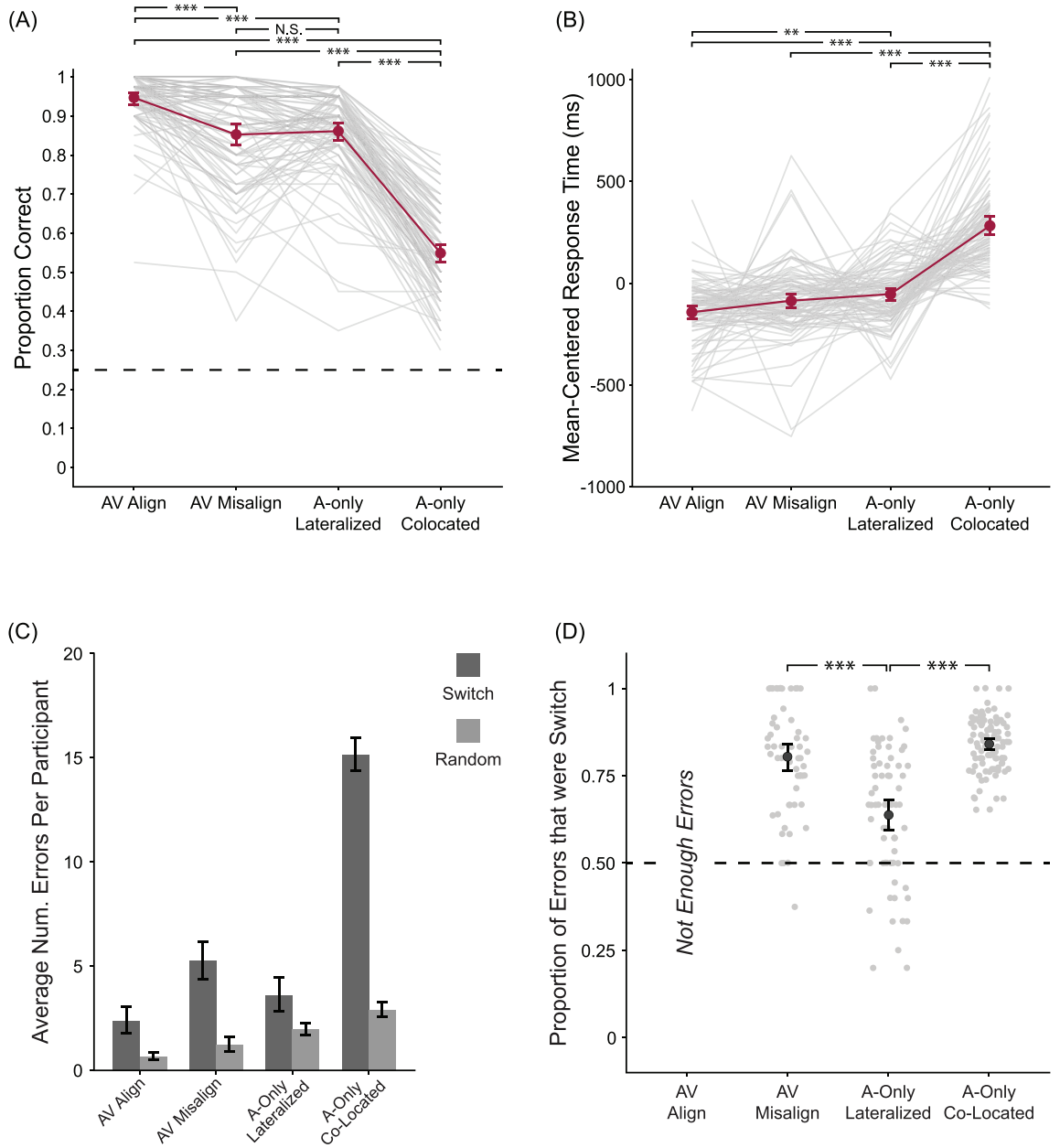

FIG. 2. (Color online) Task performance in experiment 1. (A) Proportion correct. Chance performance is indicated by the dashed line. (B) Response time, with each participant's average RT across conditions subtracted from their RT in each condition. (C) Average counts of Switch errors (the participant chose a word from the distractor stream) and Random errors (the participant chose a word spoken by neither talker). (D) The proportion of errors that were of the Switch type in each condition. Because half of the incorrect words came from the distractor stream on average, random guessing on each error trial would yield a proportion of 0.5 (dashed line). All gray lines and dots represent individual participants, error bars represent 95% confidence intervals, and asterisks indicate statistical significance in *post hoc* comparisons: ** = $p < 0.01$, *** = $p < 0.001$, N.S. = not significant.

https://doi.org/10.1121/10.0006415

AV condition was compared against the A-only lateralized condition. We observed an additional performance benefit of seeing the talkers' faces when each voice was presented in the same hemifield as the corresponding face (AV aligned vs A-only lateralized, $z = 4.72$, $p = 2.39 \times 10^{-6}$). On the other hand, speech recognition accuracy did not differ significantly between the AV misaligned and A-only lateralized conditions ($z = -0.40$, $p = 0.68$). A direct comparison between the two AV conditions revealed that accuracy was significantly higher when the unisensory components of AV speech were spatially aligned than when they were misaligned (AV aligned vs AV misaligned, $z = 4.87$, $p = 1.09 \times 10^{-6}$). All significant p-values survived Bonferroni-Holm adjustment of the alpha criterion for significance.

2. Response time

Median RTs within each condition were widely variable across participants, from a minimum value of 1224 ms to a maximum of 3911 ms. Nonetheless, the pattern of RTs varied systematically across conditions in a matter consistent with the accuracy data. Since the RT data were first reduced to the median across stimuli within each condition, these data were modelled using a linear mixed effects model with no random effects term for stimulus. The structure of this model was

$$RT \sim Condition + (1 + Condition \,|\, ID).$$

An ANOVA conducted on this model again revealed a significant main effect of condition [$X^2 = 235.73$, $df = 3$, $p = 2.97 \times 10^{-51}$; Fig. 2(B)]. *Post hoc* testing showed that participants were slower to respond in the A-only co-located condition than any of the other three, reflecting elevated task difficulty when neither spatial separation nor visual information were available to help segregate the speech streams (vs A-only lateralized, $t = 10.99$, $p = 1.17 \times 10^{-23}$; vs AV misaligned, $t = 12.06$, $p = 2.49 \times 10^{-27}$; vs AV aligned, $t = 13.88$, $p = 8.47 \times 10^{-34}$). A modest difference in RT was also observed between the AV aligned and A-only lateralized conditions (mean RTs of 1999 and 2087 ms, respectively), which was statistically significant ($z = -2.88$, $p = 0.0042$, Bonferroni-Holm adjusted p-value criterion: $p = 0.017$). This AV facilitation of RT did not reach significance when comparing the AV misaligned and A-only lateralized conditions ($t = 1.07$, $p = 0.29$), hinting at a particular AV RT benefit when faces and voices were spatially aligned. However, the RT difference between the AV aligned and AV misaligned conditions approached but did not reach significance in this experiment ($t = 1.82$, $p = 0.07$, adjusted p-value criterion: $p = 0.025$).

3. Error types

In all conditions, participants more commonly made Switch errors (in which they chose a word spoken by the distractor talker) than Random errors [in which they chose a word spoken by neither talker; Fig. 2(C)]. Figure 2(D) shows the proportion of errors that were of the switch type, in all conditions except AV aligned, in which there were not enough error trials to reliably compute these proportions. Data were excluded if a participant made fewer than three errors in a given condition, although that participant's data could be included in other conditions where this error threshold was met. These error data were analyzed using a model similar to that used for the RT data,

$$Prop.Switch \sim Condition + (1 + Condition \,|\, ID).$$

Participants committed a significantly higher proportion of switch errors in the AV misaligned condition than in the A-only lateralized condition ($t = 6.73$, $p = 2.76 \times 10^{-10}$), although average proportion correct scores did not differ between these conditions. This suggests that when the faces and voices were spatially misaligned, errors were higher relative to the AV aligned condition because the distractor stream was processed to a greater extent. When visual information was removed altogether (A-only lateralized) on the other hand, the decrement in proportion correct was caused by reduced intelligibility of the target stream, as well as potentially the distractor stream. Switch errors were also relatively more common in the A-only co-located condition than in the A-only lateralized condition ($t = 9.25$, $p = 1.48 \times 10^{-16}$), indicating that across-stream confusions were the dominant error type when the streams could not be segregated on the basis of spatial separation or AV coherence.

IV. EXPERIMENT 2

A. Experiment 2 methods

In the previous experiment, performance differences between conditions may have been partially caused by eye position, which has been shown to affect auditory localization (Cui *et al.*, 2010; Razavi *et al.*, 2007) and spatial discrimination (Maddox *et al.*, 2014). In the AV aligned and A-only lateralized conditions, participants gaze was held in the same hemifield to which they were listening. On the other hand, in the A-only co-located condition, participants fixated laterally but listened to a target voice presented at the midline, and in the AV misaligned condition, participants looked at the target talker's face in the opposite hemifield as the corresponding voice. Directing gaze toward an auditory distractor stream in this manner can reduce participants' ability to selectively attend and remember information in an auditory target stream (Reisberg *et al.*, 1981).

To account for these issues, we conducted a follow-up experiment in which the A-only co-located condition was replaced with an A-only fixation reversed condition. On these trials, target and distractor voices were spatialized as in the A-only lateralized condition, but participants were asked to fixate in the opposite hemifield as the target voice. Poorer performance in this condition than the A-only lateralized condition would provide evidence that eye position effects may have contributed the difference between the AV aligned and AV misaligned conditions observed in experiment 1. Much of the methodology was the same as in experiment 1, and so this section will focus on differences between the two experiments.

1. Participants

There were 109 participants included in the final dataset for experiment 2 (mean age = 30.0 years, SD = 8.8; 58 female, 49 male, 2 non-binary), none of whom had participated in experiment 1. A total of 52 additional participants were rejected for the following reasons: four failed to verify their language and hearing information from Prolific, 17 failed the headphone screening task, eight started but did not complete the main task, seven were rejected due to performance that fell anywhere below chance in one or more experimental condition (6 in the new A-only fixation reversed condition, 1 in the AV misaligned condition), and 16 failed a new fixation check sub-task implemented in this experiment (more below). Including these rejected participants, 161 individuals were originally recruited into the experiment. Below-chance performers were paid in full, and the remaining rejected participants received partial compensation, with a maximum partial payment of $3 for those who completed all study procedures but failed the fixation check. Full payment was a flat $5.50, yielding an average pay rate of $7.53 per hour.

2. Experiment design

The AV aligned, AV misaligned, and A-only lateralized conditions from experiment 1 were left largely unchanged in experiment 2. In the new A-only fixation reversed condition, the talkers' voices were again spatialized horizontally to −15° and 15°, but the fixation cross was positioned in the hemifield opposite the target talker's voice [see Fig. 1(C)]. In all four conditions, catch trials were included to ensure that participants were maintaining fixation in the correct location. On these trials, a small green dot was briefly presented at the center of the fixation cross (A-only conditions) or the target talker's video (AV conditions). The dot was presented for 300 ms at a random time when both talkers were speaking. When participants saw this dot, they were instructed to ignore the normal speech attention task and click a separate "catch" button on the response screen. Four catch trials replaced actual trials in each condition (16 catch trials overall), so participants performed 36 actual trials of each experimental condition. Participants were required to detect at least 60% of the catch trials overall, and at least 2 out of the 4 in each condition, to be included in the dataset and receive full payment.

B. Experiment 2 results

1. Speech attention task accuracy

A binomial mixed effects model was again used to analyze task accuracy. The conditions in this experiment were separated into two factors: Modality (AV or A-only) and fixation (aligned or reversed). Note that in the AV conditions, this fixation term also captured effects of AV spatial alignment. The model included these factors and their interaction as fixed effects, and random effect terms for participant-specific intercepts and individual stimuli,

$$\begin{aligned}\mathrm{logit}(Correct) \sim\ & Modality^{*}Fixation \\ & + (1 + Modality + Fixation \,|\, \mathrm{ID}) \\ & + (1 \,|\, Stimulus).\end{aligned}$$

Since this experiment had a two-by-two factorial design, the model was first computed with sum-coded contrasts, such that the significance of fixed effect model terms could be interpreted in a similar fashion to ANOVA results. The modality term was significant ($z = 4.96$, $p = 7.22 \times 10^{-7}$), reflecting better performance in the AV than the A-only conditions. The fixation term was also significant ($z = 6.69$, $p = 2.24 \times 10^{-11}$), indicating that participants generally performed better when fixating in the same hemifield as they were listening [Fig. 3(A)].

Importantly, the interaction term between modality and fixation was also significant ($z = 2.63$, $p = 0.009$), indicating that the effect of spatial alignment between faces and voices in the AV conditions was larger than the gaze position effect in the A-only conditions [Fig. 3(B)]. However, *post hoc* testing revealed significant effects of gaze position in both the AV (AV aligned vs AV misaligned, $z = 6.51$, $p = 7.62 \times 10^{-11}$) and A-only (A-only lateralized vs A-only fixation reversed, $z = 3.07$, $p = 0.002$) conditions. Thus, eye position effects may have contributed to the performance difference between the AV aligned and AV misaligned conditions, but the interaction term indicates an added detrimental effect of having to attend auditory and visual speech signals across hemifields.

2. Response time

RTs were modeled using a linear mixed effects model with modality, fixation, and their interaction included as the fixed effect terms,

$$\begin{aligned}RT \sim\ & Modality^{*}Fixation \\ & + (1 + Modality + Fixation \,|\, \mathrm{ID}).\end{aligned}$$

This model revealed that RTs were significantly impacted by modality ($t = -5.14$, $p = 1.27 \times 10^{-6}$), with faster RTs in the AV than the A-only conditions, and fixation ($t = -3.49$, $p = 6.95 \times 10^{-4}$), with generally faster RTs when fixation was aligned with the target voice than when it was misaligned, both in accord with the accuracy data [Fig. 3(C)]. The interaction term also reached marginal significance ($t = 2.02$, $p = 0.046$), motivating *post hoc* tests. These tests revealed that RTs were significantly faster in the AV aligned condition than the AV misaligned condition ($t = -3.94$, $p = 1.09 \times 10^{-4}$), but that fixation did not significantly affect RTs in the A-only conditions ($t = -1.22$, $p = 0.22$). Thus, RTs in this speech selective attention task were speeded by the ability to see the talkers' faces, particularly when the cross-modal components of the target and distractor streams were spatially aligned. The significant RT difference in the AV conditions, but not the A-only conditions, suggests an attentional cost when AV speech is

https://doi.org/10.1121/10.0006415

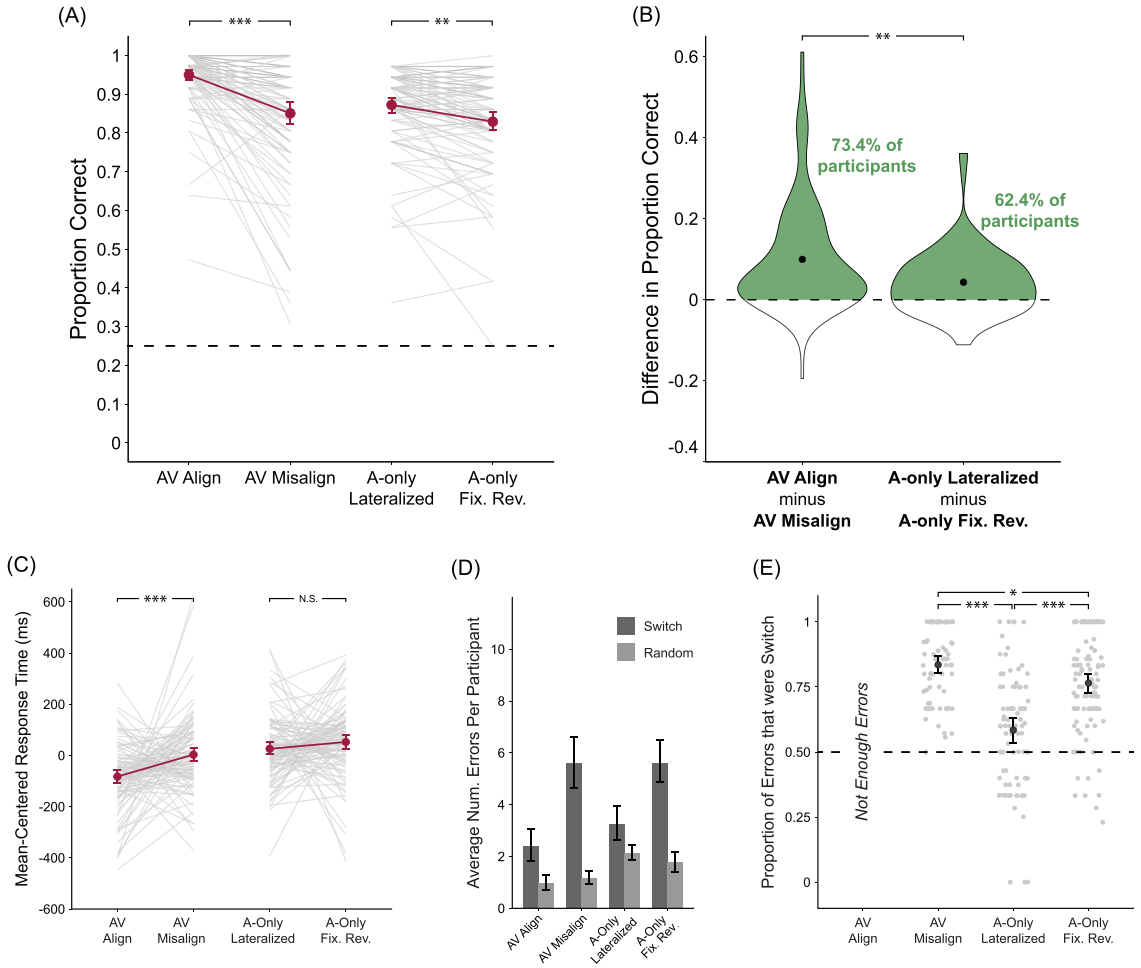

FIG. 3. (Color online) Task performance in experiment 2. (A) Proportion correct. Chance performance is indicated by the dashed line. (B) To illustrate the interaction between the modality and fixation terms, the difference in proportion correct between the two fixation levels is shown for the AV and A-only conditions. Black dots represent means of the distributions. (C) Response time, with each participant's average RT across conditions subtracted from their RT in each condition. (D) Average counts of switch and random errors. (E) The proportion of errors that were of the switch type. Random guessing on each error trial would yield a proportion of 0.5 (dashed line). All gray lines and dots represent individual participants, error bars represent 95% confidence intervals, and stars indicate statistical significance in *post hoc* comparisons: * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, N.S. = not significant.

separated across hemifields, which cannot be fully explained by gaze differences.

3. Error types

Replicating the results of experiment 1, switches with the distractor stream represented the majority of error trials across conditions [Fig. 3(D)]. Switch error proportions were again analyzed using a statistical model similar to that used for the RT data,

$$Prop.Switch \sim Modality^{*}Fixation + (1 + Modality + Fixation \,|\, \mathrm{ID})$$

As in Experiment 1, Switch errors were relatively more common in the AV Misaligned condition than in the A-only Lateralized condition [$t = 8.37$, $p = 8.37 \cdot 10^{-14}$; Fig. 3(E)]. Switch error proportions were also higher in the A-only Fixation Reversed condition than the A-only Lateralized condition ($t = 6.54$, $p = 8.14 \cdot 10^{-10}$), demonstrating that gazing in the direction of the distractor speech increased the confusability between streams. Additionally, Switch error proportions were slightly but significantly higher in the AV Misaligned condition than the A-only Fixation Reversed condition ($t = 2.42$, $p = 0.017$). This effect suggests that gazing toward a distractor speech stream and integrating spatially misaligned AV speech can each increase confusability between streams, but that integrating misaligned speech comes at an especially high attentional cost. This result is corroborated by patterns in the overall performance [Fig. 3(B)] and RT data [Fig. 3(C)].

V. EXPERIMENT 3

A. Experiment 3 methods

In experiments 1 and 2, near-ceiling performance was observed in the AV aligned condition. Significant effects of AV spatial alignment were found in both experiments in spite of this ceiling effect, but we reasoned that more

dramatic effects of alignment between faces and voices would be observed if the task were made more difficult. Such an effect could also be interpreted as an example of inverse effectiveness, a principle of multisensory integration which states that the greatest multisensory gain is observed when the unisensory stimulus components elicit weak behavioral or neuronal responses (Crosse *et al.*, 2016; Senkowski *et al.*, 2011; Stevenson *et al.*, 2012; Stevenson and James, 2009). As more intense noise corrupts the envelope of the speech signals, which likely contains the temporal information that binds each voice to the corresponding face, we expected that AV spatial alignment would take on greater importance as a secondary cue to AV speech integration. To this end, in experiment 3 we tested effects of AV spatial alignment at varying SNRs, all of which were more difficult than those used in experiments 1 and 2.

1. Participants

One hundred participants were included in the final dataset for experiment 3 (mean age = 31.8 years, SD = 10.2; 52 female, 47 male, 1 non-binary), none of whom had participated in the previous experiments. A total of 26 additional participants were rejected for the following reasons: nine failed the headphone screening task, four started but did not complete the main task, and 13 failed the fixation check sub-task. Since more difficult conditions were being introduced, we no longer excluded data on the basis of below-chance performance in any of the experimental conditions. As with the previous experiments, rejected participants received partial compensation and the full payment amount was $5.50, yielding an average pay rate in this experiment of $8.37 per hour.

2. Experiment design

This experiment used only the AV aligned and AV misaligned conditions. Whereas the noise level was previously set to be 4 dB less intense than the speech, in this experiment we tested three lower SNRs of −4, −8, and −12 dB. The different SNRs were achieved by varying the noise level while keeping the speech level constant. Prior to the experiment, an extra trial from the −12 dB condition (in which the noise level was highest) was used to let participants set their system volume. Fixation catch trials, as introduced in experiment 2, were also used here. Nine catch trials were included in both the AV aligned and AV misaligned conditions (18 overall, evenly distributed across SNRs). Only sessions in which participants got six out of nine catch trials correct in each AV spatial alignment condition were included in the final dataset. In addition to the catch trials, participants performed 23 trials of each combination of AV spatial alignment and SNR, making for a total of 156 trials. Conditions were randomly intermixed throughout the experiment.

B. Experiment 3 results

1. Speech attention task accuracy

A binomial mixed effects model was used to analyze task accuracy. The conditions in this experiment were separated into two factors: AV spatial alignment (aligned or misaligned) and SNR (−4, −8, or −12). The model included these factors and their interaction as fixed effects, and random effect terms for participant-specific intercepts and individual stimuli,

$$\begin{aligned}\text{logit}(Correct) \sim\ & Alignment^{*}SNR \\ & + (1 + Alignment + SNR \,|\, \text{ID}) \\ & + (1 \,|\, Stimulus).\end{aligned}$$

The significance of these fixed effects was assessed by passing the model to an ANOVA, with p-values based on the Satterthwaite approximation for degrees of freedom. This ANOVA revealed significant main effects of AV spatial alignment ($X^2 = 15.41$, $df = 1$, $p = 8.65 \times 10^{-5}$), with better performance in the AV aligned condition, and SNR ($X^2 = 10.95$, $df = 2$, $p = 0.004$), with performance worsening as the SNR was lowered. There was also a marginally significant interaction between AV spatial alignment and SNR [$X^2 = 5.97$, $df = 2$, $p = 0.051$; Fig. 4(A)].

Post hoc testing was restricted to comparisons between the AV aligned and AV misaligned conditions at each SNR, as we were interested in how this effect changed as a function of SNR. Performance on the speech attention task was significantly better in the AV aligned than the AV misaligned condition at −4 dB SNR ($z = 3.84$, $p = 1.24 \times 10^{-4}$) and −8 dB SNR ($z = 2.62$, $p = 0.009$), but not at the most difficult SNR, −12 dB ($z = 0.51$, $p = 0.61$). This pattern ran counter to our hypothesis, with effects of AV spatial alignment appearing to weaken as the task was made more difficult; a likely explanation for this will be discussed below.

At the highest SNR (−4 dB), Switch errors accounted for a greater proportion of error trials than Random errors in the AV misaligned condition, but not the AV aligned condition (supplemental[1] Fig. 1). At the lowest SNR (−12 dB), Random guessing became more common as energetic masking substantially degraded both speech streams.

2. Response time

RTs were modeled using a linear mixed effects model with AV spatial alignment, SNR, and their interaction included as fixed effect terms,

$$RT \sim Alignment^{*}SNR + (1 + Alignment + SNR \,|\, \text{ID}).$$

Similar to the accuracy data, submitting this model to an ANOVA revealed main effects of AV spatial alignment ($X^2 = 27.57$, $df = 1$, $p = 1.51 \cdot 10^{-7}$), SNR ($X^2 = 87.26$, $df = 2$, $p = 2.2 \times 10^{-16}$), and an interaction between these factors [$X^2 = 24.80$, $df = 2$, $p = 4.11 \times 10^{-6}$; Fig. 4(B)]. *Post hoc* testing revealed significantly faster response times

https://doi.org/10.1121/10.0006415

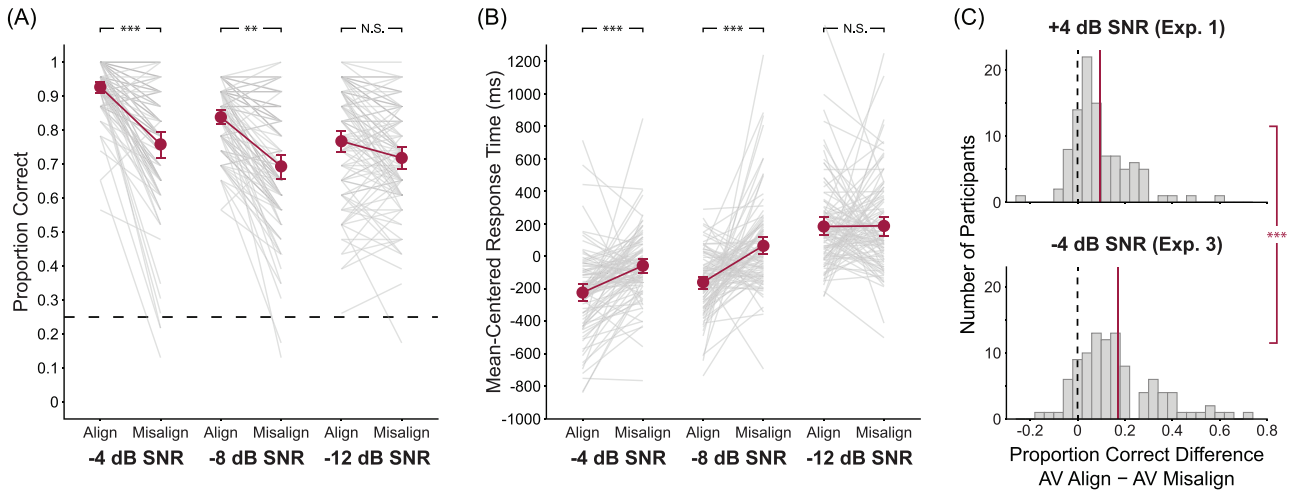

FIG. 4. (Color online) Task performance in experiment 3 and comparison to experiment 1. (A) Proportion correct. Chance performance is indicated by the dashed line. (B) Response time, with each participant's average RT across conditions subtracted from their RT in each condition. Grey lines represent individual participants, error bars represent 95% confidence intervals, and stars indicate statistical significance in *post hoc* comparisons: ** = $p < 0.01$, *** = $p < 0.001$, N.S. = not significant. (C) Histograms of the difference in proportion correct between the AV aligned and AV misaligned conditions at+4 dB SNR (data from experiment 1) and−4 dB SNR. The vertical dashed lines indicate no difference between these conditions, while each red line indicates the average difference across the participant population. Asterisks indicate significance in a Wilcoxon rank-sum test: *** = $p < 0.001$.

in the AV aligned than the AV misaligned condition at−4 dB SNR ($t = -4.54$, $p = 8.02 \times 10^{-6}$) and−8 dB SNR ($t = -6.16$, $p = 2.41 \times 10^{-9}$), both of which survived Bonferroni-Holm adjustment of the significance criterion, but not at−12 dB SNR ($t = -0.09$, $p = 0.93$). Thus, the RT results mirrored the accuracy data, with AV spatial alignment significantly speeding responses at the relatively high SNRs, but not at the lowest SNR, when responses were slowest overall.

3. Comparison of SNRs across experiment 1 and experiment 3

In experiment 3, participants performed at well above-chance levels in all conditions, including the AV misaligned condition at the most difficult SNR (a potential reason for this high effective performance floor will be examined in experiment 4). This limited our ability to measure the influence of AV spatial alignment at the lower SNRs, as the lowest average performance levels were around 70% correct. However, at +4 dB SNR (experiment 1) and −4 dB SNR (experiment 3), average performance in both alignment conditions was above the effective performance floor reached at lower SNRs. Thus, we next compared the effect of AV spatial alignment between these two SNR levels across the experiments [Fig. 4(C)]. At both SNRs, proportion correct was computed for each participant in the AV aligned and AV misaligned conditions. These proportion correct scores were then subtracted, yielding the individual benefit (in terms of a proportion correct difference) of spatial alignment between corresponding faces and voices. Shapiro-Wilk tests revealed that these difference scores were not normally distributed across the population at either SNR; thus, the effect of AV spatial alignment was compared between the +4 dB and −4 dB SNR conditions using a two-sided Wilcoxon rank-sum test. This test showed a greater benefit of AV spatial alignment at the lower SNR from experiment 3 ($W = 3480$, $p = 8.74 \times 10^{-4}$). This result should be interpreted carefully, as there was a potential ceiling effect in the AV aligned condition in experiment 1, although average performance in the AV aligned condition was similar between these two SNRs (94.7% correct at +4 dB SNR, 92.7% correct at −4 dB SNR). This provides some evidence that effects of AV spatial alignment may indeed become stronger as the auditory speech signal is degraded. However, this should be validated using a paradigm in which performance is not compressed within the upper range of percent correct scores.

VI. EXPERIMENT 4

A. Experiment 4 methods

The floor effect observed in experiment 3 could have resulted from the closed-set nature of the stimuli and response method used in the current study. Although neither talker ever repeated the same sentence during the experiment, the words in the corpus were all repeated, opening the possibility that participants learned the set of possible words during the experiment. Since participants had to select the target word rather than repeating what they heard, they essentially received feedback on this learning at the response stage. While introducing a different stimulus set or response methodology were beyond the scope of the current study, in experiment 4 we aimed to assess the extent to which these closed-set factors influenced our results. One way that this could manifest is in an increased ability to lipread the stimuli given prior knowledge of the possible words in the corpus. If this were the case, participants may have been encouraged to rely more on lipreading at lower SNRs, rendering spatial alignment between the faces and voices irrelevant. Here, we performed a version of experiment 3

with the addition of a visual-only (V-only) condition to determine the performance level participants were able to achieve using lipreading alone.

1. Participants

a. Experiment 4 methods. There were 39 participants included in the dataset for experiment 4 (mean age = 34.7 years, SD = 10.1; 15 female, 24 male, 0 non-binary), none of whom had participated in the previous experiments. A total of 17 additional participants were rejected for the following reasons: two failed to confirm their native language information from Prolific, 11 failed the headphone screening task, two started but did not complete the main task, and two failed the fixation check sub-task. Data were not excluded on the basis of below-chance performance. As with the previous experiments, rejected participants received partial compensation and the full payment amount was $5.50, yielding an average pay rate in this experiment of $8.56 per hour.

2. Experiment design

The AV aligned and AV misaligned conditions were again used, keeping the −4 and −12 dB SNR conditions from experiment 3. In place of the −8 dB SNR condition, we added V-only trials on which participants attempted to lipread the target talker's speech with no voice. Fixation catch trials were included as in experiments 2 and 3. Three catch trials were included in each combination of AV spatial alignment and SNR, as well as the V-only condition (15 catch trials overall). Only sessions in which participants got two out of three catch trials correct in each condition were included in the final dataset. In addition to the catch trials, participants performed 23 trials of each condition, making for a total of 130 trials. To maximize potential learning of the words in the corpus, participants first performed intermixed AV aligned and AV misaligned trials in the −4 dB SNR condition. We reasoned that participants would be most likely to hear all the words spoken by the target talker—and possibly some spoken by the distractor talker—at this relatively high SNR. Participants then completed the remaining AV aligned and misaligned trials in the −12 dB SNR condition, as well as the V-only trials, all randomly intermixed.

B. Experiment 4 results

1. Speech attention task accuracy

Performance was quite high on average in the V-only condition (66.8% correct, SD = 18.4%). Previous studies have shown that, in normal-hearing participants, open-set word recognition performance using lipreading is typically between 10% and 15% correct (Altieri *et al.*, 2011; Grant and Seitz, 2000). A binomial mixed effects model was used to analyze differences between experimental conditions, which were all coded as a single condition factor. The model included condition as the only fixed effect, and random effect terms for participant-specific intercepts and individual stimuli,

$$\text{logit}(Correct) \sim Condition + (1 + Condition\,|\,ID) + (1\,|\,Stimulus).$$

Significance of the condition effect was assessed by passing the model to an ANOVA, with p-values based on the Satterthwaite approximation for degrees of freedom. A significant effect of condition was found ($X^2 = 20.67$, df = 4, $p = 3.68 \times 10^{-4}$, Fig. 5). *Post hoc* tests compared the V-only condition to each of the other four, as well as the AV aligned vs AV misaligned conditions within each SNR level (six total comparisons). The latter comparisons replicated the results of experiment 3, with performance significantly improved when faces and voices were spatially aligned at −4 dB SNR ($z = 2.97$, $p = 0.003$; adjusted p-value criterion = 0.01), but not at −12 dB SNR ($z = 0.75$, $p = 0.45$). When the auditory components of AV speech stimuli were presented at −12 dB SNR, task performance was not significantly different than when participants used only lipreading, regardless of AV spatial alignment. The same was true of the AV misaligned speech at −4 dB SNR; only AV aligned task performance at this SNR was significantly better than the V-only condition ($z = 4.36$, $p = 1.32 \times 10^{-5}$; adjusted p-value criterion = 0.008). In sum, this experiment revealed that participants could achieve a high-level of lipreading accuracy, setting a floor level of performance that participants approached when the auditory stimuli were embedded in noise at −12 dB SNR.

Error types were also analyzed (supplemental[1] Fig. 2), but it should be noted that in experiments 3 and 4, AV

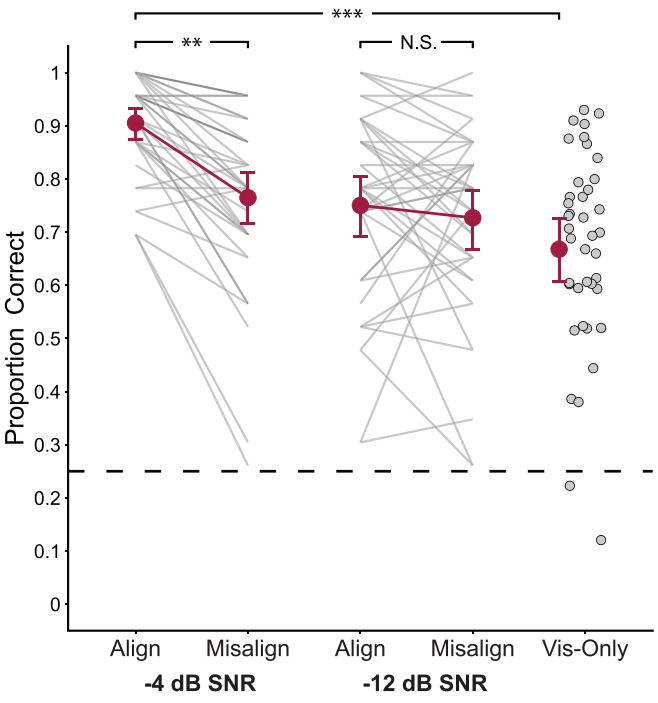

FIG. 5. (Color online) Task performance in experiment 4. Proportion correct scores are shown, with chance performance indicated by the dashed line. Grey lines and dots represent individual participants, error bars represent 95% confidence intervals, and asterisks indicate statistical significance in *post hoc* comparisons: ** = $p < 0.01$, *** = $p < 0.001$, N.S. = not significant.

https://doi.org/10.1121/10.0006415

spatial misalignment may have produced competing effects on the prevalence of switch errors: AV misalignment could increase confusability between streams, leading to more switch errors, but also decrease the intelligibility of the target stream (particularly at low SNRs), leading to more random errors. These data should therefore be interpreted with caution. Nonetheless, the results were generally consistent with error types in experiment 3, but also showed a surprising prevalence of switch errors in the V-only condition (about 66% of V-only errors).

VII. GENERAL DISCUSSION

Across four online experiments, we evaluated the extent to which AV speech perception benefits depend on spatial alignment between faces and voices, using a paradigm that featured both acoustic noise and a competing talker. In experiment 1, performance was worst overall by far in the A-only co-located condition, in which neither visual information nor spatial separation of the talkers were available to help participants segregate the speech streams. This indicates that the benefits of SRM in a multi-talker environment were preserved in the online experiment format (Kidd *et al.*, 1998; Marrone *et al.*, 2008; Shinn-Cunningham *et al.*, 2005). Beyond this benefit of spatial separation, the ability to see the talkers' faces further improved task performance, but critically, this was only true when the corresponding faces and voices for each talker were aligned in the same hemifield. In experiment 2, we examined the possibility that fixating in one hemifield while listening to a target in the other led to the performance decrement in the AV misaligned condition. Indeed, auditory-only task performance was worse when participants fixed their gaze in the hemifield of the distractor talker, but this effect was smaller than the effect of spatial alignment between faces and voices in the AV conditions. The AV spatial alignment effect decreased as the SNR was reduced from −4 to −12 dB (experiment 3), but participants may have hit the performance floor at the lowest SNR, as lipreading performance was quite good given the closed-set nature of the stimuli (experiment 4). Nonetheless, comparison of the alignment effect in the +4 dB SNR condition of experiment 1 and the −4 dB condition of experiment 3 revealed a stronger effect at the lower SNR. This suggests that AV spatial alignment may become a more relevant cue as noise degrades the speech envelope, which provides temporal information linking the voice to the corresponding face.

Despite quite large inter-subject differences in overall response times, consistent patterns emerged across conditions in line with the performance accuracy results. The large degree of inter-subject differences was likely caused by a combination of variability in participants' personal hardware and the response method of clicking on the word spoken by the target talker; the starting position of the participant's mouse and the order in which they examined the possible words both could introduce noise into RT measures. In the first two experiments, RTs were reliably faster in the AV than the A-only conditions, in general agreement with task accuracy. Faster responses to multisensory as opposed to unisensory stimulation is a hallmark of multisensory integration, at least when simple stimuli are used (Colonius and Diederich, 2006). However, when more temporally complex AV stimuli (including speech) are used, it has been reported that AV RTs are actually slower than those to A-only stimuli (Fraser *et al.*, 2010; Strand *et al.*, 2020). Thus, the speeded AV responses we observed probably do not reflect differences in the speed of early sensory processing; a more likely explanation is that AV integration made the task easier by facilitating the allocation of selective attention to the target talker.

In experiment 2, RTs were significantly faster in the AV aligned than the AV misaligned condition, but RTs did not differ significantly between these same conditions in experiment 1. The fixation check sub-task introduced in experiment 2 may explain this result; without fixation control in experiment 1, participants may have adopted a different strategy in the AV misaligned condition, such as fixating centrally rather than on the target face. Other factors, however, such as the size of the participants' computer display and their distance from it, are more difficult to control in an online format. If using a smaller monitor or seated farther away, participants could more clearly see both talkers' faces simultaneously, which may contribute to the large inter-subject variability in task performance in the AV misaligned condition. The use of newly developed online eye tracking tools (Semmelmann and Weigelt, 2018) could partially mitigate these issues in future online AV studies.

A. Spatial attention in audio-visual processing

A broad literature has converged to demonstrate that top-down attention influences the strength of multisensory integration (Talsma *et al.*, 2010). This has been frequently shown in the realm of AV speech using McGurk effect manipulations (McGurk and Macdonald, 1976). If, instead of the standard single face, two lateralized faces accompany a single auditory stimulus, the face to which covert attention is directed has greater influence over auditory syllable perception than the unattended face (Andersen *et al.*, 2009). Similar results were found when the long-term focus of spatial attention was shifted by reliably presenting auditory stimuli from a particular location (e.g., presenting stimuli from −90° azimuth on 90% of trials); the McGurk percept was strengthened at this attended location (Tiippana *et al.*, 2011). Neurally, deploying top-down spatial attention to AV stimuli modulates ERPs elicited by them in several time ranges, indicating effects at multiple processing stages (Talsma and Woldorff, 2005). Similarly, fMRI activation differs across a wide network depending on whether visuospatial attention is directed toward speaking lips that are matched or unmatched to an auditory speech stimulus (Fairhall and MacAluso, 2009). These studies, among others, demonstrate an unequivocal link between whether a

cross-modal stimulus is attended and whether signatures of multisensory integration are observed.

In the AV misaligned condition of the present study, the "spotlight" of top-down spatial attention had to be either divided or broadened across hemifields in order to successfully integrate the target talker's face and voice. In vision, the spotlight of spatial attention can be efficiently divided into multiple locations across or within hemifields (Malinowski *et al.*, 2007; McMains and Somers, 2005; Müller *et al.*, 2003). Dividing auditory spatial attention, on the other hand, has been shown to come at a processing cost (Parasuraman, 1978). Thus, it is possible that dividing cross-modal spatial attention in the AV misaligned condition decreased participants' ability to track the target talker's speech. In experiment 2, such a division of cross-modal attention was required in both the AV misaligned and A-only fixation reversed conditions; in the latter, participants had to listen to speech in one hemifield while visually monitoring the other to detect the fixation catch trials. This form of divided spatial attention did cause a performance decrement compared to the A-only lateralized condition (in which participants visually monitored a fixation cross-in the same hemifield as the target speech), but this effect was smaller than the difference between the AV aligned and AV misaligned conditions. Thus, dividing cross-modal spatial attention appeared to have an especially deleterious effect on the selective integration of AV speech. This is consistent with previous studies demonstrating that top-down attention is required for many forms of multisensory integration (see Talsma *et al.*, 2010 for review), as well as studies showing a reduction in behavioral (Alsius *et al.*, 2005) and neural (Alsius *et al.*, 2014) signatures of multisensory integration under a high degree of attentional load.

B. When do we make use of auditory spatial information?

In auditory selective attention tasks, participants do not obligatorily rely on spatial features to separate component sounds from an auditory mixture. Top-down attention can be volitionally directed toward other sound features, such as pitch, depending on task demands and participant goals (Maddox and Shinn-Cunningham, 2012). Further, even when a target is defined only by a cued location, neural signatures of spatial attention appear at first, but are not sustained past initial selection of the relevant auditory target as long as the competing streams can be segregated on the basis of pitch (Bonacci *et al.*, 2020). Both competing talkers were female in the present study, and so their voices were somewhat similar (although clearly differentiable) in fundamental frequency. Beyond pitch though, many other features were available to separate the two voices, including talker-specific characteristics (e.g., speech rate, vowel space, etc.), and in the AV conditions, the fact that each voice was temporally coherent with a separate face. Thus, cross-modal spatial features were by no means required to segregate the target and distractor talkers, and yet we observed strong benefits of AV alignment nonetheless.

This discrepancy may trace its roots to the high degree of spatial reliability of the visual component of AV speech. The multisensory perceptual system is highly sensitive to the reliability of individual cues, such that vision dominates in instances of cross-modal spatial conflict (i.e., ventriloquism) as long as it remains more spatially reliable than audition (Alais and Burr, 2004; Recanzone, 1998; Wozny and Shams, 2011). This visual dominance is distinct from auditory spatial information being ignored altogether. For instance, ventriloquism breaks down if the auditory and visual signals become too far separated in space, indicating that the auditory spatial information is still being encoded (Bosen *et al.*, 2016). Because part of the AV stimulus is in general spatially reliable, cross-modal spatial alignment may play a more obligatory role in AV than auditory-only selective attention, even if—as in the current study—spatial information is not explicitly task-relevant.

C. Eye position effects on auditory spatial attention

The direction of eye gaze influences auditory responses in many stations along the processing hierarchy, including the inferior colliculus (Groh *et al.*, 2001), superior colliculus (Jay and Sparks, 1984), and primary auditory cortex (Fu *et al.*, 2004; Werner-Reiss *et al.*, 2003). These effects are mirrored by improved auditory selective attention when fixating in the direction of an auditory target (Maddox *et al.*, 2014; Reisberg *et al.*, 1981), which was also found in the A-only conditions of experiment 2. The A-only fixation reversed condition created a particularly challenging scenario, as participants were asked to fix their gaze in the direction of the distractor talker. This has been shown to reduce the difference between target and distractor ERPs—a measure of successful deployment of attention—relative to fixation on the auditory target or a neutral location (Okita and Wei, 1993).

Importantly, however, our ability to claim that participants were fixated on the exact location of the auditory source is limited because of the methodologies employed. First, sounds were spatialized with non-individualized HRTFs and presented through unknown headphones, so the degree to which participants externalized the sounds to the intended locations likely varied widely. Second, the use of HRTFs in may have led to a weaker auditory spatial percept than free-field sound sources would have (Brungart and Simpson, 2001). Finally, without the use of eye-tracking we cannot be completely sure that participants maintained correct fixation throughout the experiment, although our fixation check task excluded several participants who likely failed to do so. Given these factors, the effects of eye position in this study may reflect the relationship between the hemifield of fixation and the hemifield of the target talker's voice, more so than effects of looking at the auditory target *per se*. Similarly, it is impossible to control factors such as display size or participant distance from the display in an online experiment format. Thus, effects of AV spatial alignment in this study should be interpreted at the hemifield

level; further research using free-field sound sources, preferably in a laboratory setting so that eye gaze can be more reliably monitored, could distinguish the effects of integrating cross-modal information across hemifields from subtler forms of AV spatial misalignment.

VIII. CONCLUSIONS

Participants exhibited improved speech selective attention performance when the target and distractor talkers' faces were visible, as compared to auditory-only conditions. However, this was only true when each talker's voice was spatially aligned with their face; spatially misaligning voices and faces disrupted the AV benefit. This spatial dependence of AV benefits was found despite the presence of alternative features that could be used to separate the competing speech streams, such as voice pitch and other talker-specific characteristics. Taken together, these data provide evidence that cross-modal spatial alignment provides an important cue to the integration of AV speech stimuli in an acoustically and attentionally challenging environment.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Tyler Perrachione for advice on mixed model construction and interpretation, and Audra Irvine for logistical support in collecting the online data. This work was supported by the Office of Naval Research (Grant No. N000141812069). J.T.F., R.K.M., and B.G.S.C. all contributed to conceptualization and design of the experiments. Additionally, J.T.F. implemented the experiments, collected, analyzed, and interpreted the data, and wrote the manuscript. R.K.M. and B.G.S.C. also contributed to analysis and edited the manuscript. The authors declare no competing interests.

[1]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0006415 for supplemental figures and captions.

Alais, D., and Burr, D. (**2004**). "The ventriloquist effect results from near-optimal bimodal integration," Curr. Biol. **14**(3), 257–262.
Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendaño, C. (**2001**). "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, Cat. No.01TH8575, pp. 99–102.
Alsius, A., Möttönen, R., Sams, M. E., Soto-Faraco, S., and Tiippana, K. (**2014**). "Effect of attentional load on audiovisual speech perception: Evidence from ERPs," Front. Psychol. **5**, 00727.
Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (**2005**). "Audiovisual integration of speech falters under high attention demands," Curr. Biol. **15**(9), 839–843.
Altieri, N. A., Pisoni, D. B., and Townsend, J. T. (**2011**). "Some normative data on lip-reading skills (L)," J. Acoust. Soc. Am. **130**(1), 1–4.
Andersen, T. S., Tiippana, K., Laarni, J., Kojo, I., and Sams, M. (**2009**). "The role of visual spatial attention in audiovisual speech perception," Speech Commun. **51**(2), 184–193.
Atilgan, H., Town, S. M., Wood, K. C., Jones, G. P., Maddox, R. K., Lee, A. K. C., and Bizley, J. K. (**2018**). "Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding," Neuron **97**, 640–655.
Bertelson, P., Vroomen, J., Wiegeraad, G., and de Gelder, B. (**1994**). "Exploring the relation between McGurk interference and ventriloquism," in *Proceedings of the Third International Congress on Spoken Language Processing*, Yokohama, Japan (September 18–22), pp. 559–562.
Besle, J., Fort, A., Delpuech, C., and Giard, M. H. (**2004**). "Bimodal speech: Early suppressive visual effects in human auditory cortex," Eur. J. Neurosci. **20**(8), 2225–2234.
Bizley, J. K., Shinn-Cunningham, B. G., and Lee, A. K. C. (**2012**). "Nothing is irrelevant in a noisy world: Sensory illusions reveal obligatory within-and across-modality integration," J. Neurosci. **32**(39), 13402–13410.
Bonacci, L. M., Bressler, S., and Shinn-Cunningham, B. G. (**2020**). "Nonspatial features reduce the reliance on sustained spatial auditory attention," Ear Hear. **41**(6), 1635–1647.
Bosen, A. K., Fleming, J. T., Brown, S. E., Allen, P. D., O'Neill, W. E., and Paige, G. D. (**2016**). "Comparison of congruence judgment and auditory localization tasks for assessing the spatial limits of visual capture," Biol. Cybern. **110**(6), 455–471.
Brungart, D. S., and Simpson, B. D. (**2001**). "Auditory localization of nearby sources in a virtual audio display," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, Cat. No.01TH8575, pp. 107–110.
Chait, M., Poeppel, D., and Simon, J. Z. (**2006**). "Neural response correlates of detection of monaurally and binaurally created pitches in humans," Cerebral Cortex **16**(6), 835–848.
Colonius, H., and Diederich, A. (**2006**). "The race model inequality: Interpreting a geometric measure of the amount of violation," Psychol. Rev. **113**(1), 148–154.
Conrey, B., and Pisoni, D. B. (**2006**). "Auditory-visual speech perception and synchrony detection for speech and nonspeech signals," J. Acoust. Soc. Am. **119**(6), 4065–4073.
Cramer, E. M., and Huggins, W. H. (**1958**). "Creation of pitch through binaural interaction," J. Acoust. Soc. Am. **30**(5), 413–417.
Crosse, M. J., Di Liberto, G. M., and Lalor, E. C. (**2016**). "Eye can hear clearly now: Inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration," J. Neurosci. **36**(38), 9888–9895.
Cui, Q. N., Razavi, B., O'Neill, W. E., and Paige, G. D. (**2010**). "Perception of auditory, visual, and egocentric spatial alignment adapts differently to changes in eye position," J. Neurophys. **103**(2), 1020–1035.
DeLoss, D. J., and Andersen, G. J. (**2015**). "Aging, spatial disparity, and the sound-induced flash illusion," PLOS One **10**(11), e0143773.
Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). "Note on informational masking (L)," J. Acoust. Soc. Am. **113**(6), 2984–2987.
Erber, N. P. (**1975**). "Auditory-visual perception of speech," J. Speech Hear. Disord. **40**(4), 481–492.
Fairhall, S. L., and MacAluso, E. (**2009**). "Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites," Eur. J. Neurosci. **29**(6), 1247–1257.
Fleming, J. T., Noyce, A. L., and Shinn-Cunningham, B. G. (**2020**). "Audio-visual spatial alignment improves integration in the presence of a competing audio-visual stimulus," Neuropsychologia **146**, 107530.
Fraser, S., Gagné, J.-P., Alepins, M., and Dubois, P. (**2010**). "Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues," J. Speech Lang. Hear. Res. **53**(1), 18–33.
Fu, K.-M. G., Shah, A. S., O'Connell, M. N., McGinnis, T., Eckholdt, H., Lakatos, P., Smiley, J., and Schroeder, C. E. (**2004**). "Timing and laminar profile of eye-position effects on auditory responses in primate auditory cortex," J. Neurophysiology **92**(6), 3522–3531.
Gorilla (**2021**). "Gorilla Experiment Builder," https://www.gorilla.sc (Last viewed 9/17/2021).
Grant, K. W., and Greenberg, S. (**2001**). "Speech intelligibility derived from asynchronous processing of auditory-visual information," in *International Conference on Auditory-Visual Speech Processing (AVSP)*, Aalborg, Denmark (September 7).
Grant, K. W., and Seitz, P.-F. (**2000**). "The use of visible speech cues for improving auditory detection of spoken sentences," J. Acoust. Soc. Am. **108**(3), 1197–1208.
Groh, J. M., Trause, A. S., Underhill, A. M., Clark, K. R., and Inati, S. (**2001**). "Eye position influences auditory responses in primate inferior colliculus," Neuron **29**(2), 509–518.
Helfer, K. S., and Freyman, R. L. (**2005**). "The role of visual speech cues in reducing energetic and informational masking," J. Acoust. Soc. Am. **117**(2), 842–849.

Innes-Brown, H., and Crewther, D. (**2009**). "The impact of spatial incongruence on an auditory-visual illusion," PLoS One **4**(7), e6450.
Jay, M. F., and Sparks, D. L. (**1984**). "Auditory receptive fields in primate superior colliculus shift with changes in eye position," Nature **309**(5966), 345–347.
Jones, J. A., and Munhall, K. G. (**1997**). "Effects of separating auditory and visual sources on audiovisual integration of speech," Can. Acoust. **25**(4), 13–19.
Kidd, G., Mason, C. R., Rohtla, T. L., and Deliwala, P. S. (**1998**). "Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns," J. Acoust. Soc. Am. **104**(1), 422–431.
Litovsky, R. Y. (**2012**). "Spatial release from masking," Acoust. Today **8**(2), 18–25.
MacLeod, A., and Summerfield, Q. (**1987**). "Quantifying the contribution of vision to speech perception in noise," Brit. J. Audiol. **21**(2), 131–141.
Maddox, R. K., Atilgan, H., Bizley, J. K., and Lee, A. K. (**2015**). "Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus human listeners," ELife **4**, e04995.
Maddox, R. K., Pospisil, D. A., Stecker, G. C., and Lee, A. K. C. (**2014**). "Directing eye gaze enhances auditory spatial cue discrimination," Curr. Biol. **24**(7), 748–752.
Maddox, R. K., and Shinn-Cunningham, B. G. (**2012**). "Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention," J. Assoc. Res. Otolaryngol. **13**(1), 119–129.
Malinowski, P., Fuchs, S., and Müller, M. M. (**2007**). "Sustained division of spatial attention to multiple locations within one hemifield," Neurosci. Lett. **414**(1), 65–70.
Marrone, N., Mason, C. R., and Kidd, G. (**2008**). "The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms," J. Acoust. Soc. Am. **124**(5), 3064–3075.
McGurk, H., and Macdonald, J. (**1976**). "Hearing lips and seeing voices," Nature **264**(5588), 746–748.
McMains, S. A., and Somers, D. C. (**2005**). "Processing efficiency of divided spatial attention mechanisms in human visual cortex," J. Neurosci. **25**(41), 9444–9448.
Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., and Chait, M. (**2020**). "An online headphone screening test based dichotic pitch," Behav. Res. Methods. **53**, 1551–1562.
Müller, M. M., Malinowski, P., Gruber, T., and Hillyard, S. A. (**2003**). "Sustained division of the attentional spotlight," Nature **424**(6946), 309–312.
Okita, T., and Wei, J.-H. (**1993**). "Effects of eye position on event-related potentials during auditory selective attention," Psychophysiology **30**(4), 359–365.
Parasuraman, R. (**1978**). "Auditory evoked potentials and divided attention," Psychophysiology **15**(5), 460–465.
Peelle, J. E., and Sommers, M. S. (**2015**). "Prediction and constraint in audiovisual speech perception," Cortex **68**, 169–181.
Prolific (**2021**). "Online participant recruitment," https://www.prolific.co (Last viewed 9/17/2021).
Razavi, B., O'Neill, W. E., and Paige, G. D. (**2007**). "Auditory spatial perception dynamically realigns with changing eye position," J. Neurosci. **27**(38), 10249–10258.
Recanzone, G. H. (**1998**). "Rapidly induced auditory plasticity: The ventriloquism aftereffect," Proc. Natl. Acad. Sci. **95**(3), 869–875.
Reisberg, D., Scheiber, R., and Potemken, L. (**1981**). "Eye position and the control of auditory attention," J. Exp. Psychol.: Hum. Percept. Perform. **7**(2), 318–323.
Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (**2006**). "Do you see what I Am saying? Exploring visual enhancement of speech comprehension in noisy environments," Cerebral Cortex **17**(5), 1147–1153.
Schwartz, J. L., Berthommier, F., and Savariaux, C. (**2004**). "Seeing to hear better: Evidence for early audio-visual interactions in speech identification," Cognition **93**, B69–B78.
Semmelmann, K., and Weigelt, S. (**2018**). "Online webcam-based eye tracking in cognitive science: A first look," Behav. Res. Methods **50**(2), 451–465.
Senkowski, D., Saint-Amour, D., Höfle, M., and Foxe, J. J. (**2011**). "Multisensory interactions in early evoked brain activity follow the principle of inverse effectiveness," NeuroImage **56**(4), 2200–2208.
Sensimetrics (**2021**). "STEVI speech test video corpus," https://www.sens.com/products/stevi-speech-test-video-corpus (Last viewed 9/17/2021).
Shinn-Cunningham, B. G., Ihlefeld, A., Satyavarta, and Larson, E. (**2005**). "Bottom-up and top-down influences on spatial unmasking," Acta Acust. Acust. **91**(6), 967–979.
Simon, D. M., and Wallace, M. T. (**2018**). "Integration and temporal processing of asynchronous audiovisual speech," J. Cogn. Neurosci. **30**(3), 319–337.
Stevenson, R. A., Bushmakin, M., Kim, S., Wallace, M. T., Puce, A., and James, T. W. (**2012**). "Inverse effectiveness and multisensory interactions in visual event-related potentials with audiovisual speech," Brain Topography **25**(3), 308–326.
Stevenson, R. A., and James, T. W. (**2009**). "Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition," NeuroImage **44**(3), 1210–1223.
Strand, J. F., Brown, V. A., and Barbour, D. L. (**2020**). "Talking points: A modulating circle increases listening effort without improving speech recognition in young adults," Psychonomic Bull. Rev. **27**(3), 536–543.
Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**(2), 212–215.
Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (**2010**). "The multifaceted interplay between attention and multisensory integration," Trends Cognitive Sci. **14**(9), 400–410.
Talsma, D., and Woldorff, M. G. (**2005**). "Selective attention and multisensory integration: Multiple phases of effects on the evoked brain activity," J. Cognitive Neurosci. **17**(7), 1098–1114.
The HandBrake Team (**2021**). "Handbrake open source video transcoder software," https://www.handbrake.fr (Last viewed 9/17/2021).
Tiippana, K., Puharinen, H., Möttönen, R., and Sams, M. (**2011**). "Sound location can influence audiovisual speech perception when spatial attention is manipulated," Seeing Perceiving **24**(1), 67–90.
Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., and Theeuwes, J. (**2008**). "Pip and pop: Nonspatial auditory signals improve spatial visual search," J. Exp. Psychol.: Human Percept. Perform. **34**(5), 1053–1065.
van Wassenhove, V., Grant, K. W., and Poeppel, D. (**2005**). "Visual speech speeds up the neural processing of auditory speech," Proc. Natl. Acad. Sci. **102**, 1181–1186.
van Wassenhove, V., Grant, K. W., and Poeppel, D. (**2007**). "Temporal window of integration in auditory-visual speech perception," Neuropsychologia **45**, 598–607.
Watson, C. S. (**1987**). "Uncertainty, informational masking, and the capacity of immediate auditory memory," in *Auditory Processing Complex Sounds* (Lawrence Erlbaum Associates, Mahwah, NJ), pp. 267–277.
Werner-Reiss, U., Kelly, K. A., Trause, A. S., Underhill, A. M., and Groh, J. M. (**2003**). "Eye position affects activity in primary auditory cortex of primates," Curr. Biol. **13**(7), 554–562.
Wozny, D. R., and Shams, L. (**2011**). "Recalibration of auditory space following milliseconds of cross-modal discrepancy," J. Neurosci. Official J. Soc. Neurosci. **31**(12), 4607–4612.
Wu, X., Wang, C., Chen, J., Qu, H., Li, W., Wu, Y., Schneider, B. A., and Li, L. (**2005**). "The effect of perceived spatial separation on informational masking of Chinese speech," Hear. Res. **199**(1–2), 1–10.
Zion Golumbic, E., Cogan, G. B., Schroeder, C. E., and Poeppel, D. (**2013**). "Visual input enhances selective speech envelope tracking in auditory cortex at a 'cocktail party,' " J. Neurosci. **33**(4), 1417–1426.