Article

PERCEPTION

Perception
2019, Vol. 48(8) 649–667

DOI: 10.1177/0301006619858076
journals.sagepub.com/home/pec
SAGE

The Stolen Voice Illusion

David Brang
Department of Psychology, University of Michigan, Ann Arbor, MI, USA

Abstract
Visual cues facilitate speech perception during face-to-face communication, particularly in noisy environments. These visual-driven enhancements arise from both automatic lip-reading behaviors and attentional tuning to auditory-visual signals. However, in crowded settings, such as a cocktail party, how do we accurately bind the correct voice to the correct face, enabling the benefit of visual cues on speech perception processes? Previous research has emphasized that spatial and temporal alignment of the auditory-visual signals determines which voice is integrated with which speaking face. Here, we present a novel illusion demonstrating that when multiple faces and voices are presented in the presence of ambiguous temporal and spatial information as to which pairs of auditory-visual signals should be integrated, our perceptual system relies on identity information extracted from each signal to determine pairings. Data from three experiments demonstrate that expectations about an individual's voice (based on their identity) can change where individuals perceive that voice to arise from.

Keywords
multisensory, cross-modal, auditory-visual, speech, congruity, gender, identity

Date Received: 28 September 2018; accepted: 25 May 2019

Introduction

Informative visual cues from a speaking individual (e.g., lip movements and gestures) are often present during face-to-face conversation, facilitating speech perception both in terms of speed and accuracy (Ghazanfar, Maier, Hoffman, & Logothetis, 2005; Rosenblum, 2008; Sumby & Pollack, 1954), particularly in the presence of environmental noise or hearing impairments (Grant & Seitz, 2000; Kim & Davis, 2003, 2004; Rouger et al., 2007). In natural settings, such as a crowded party, multiple competing signals are present, challenging our perceptual system to determine which auditory and visual signals refer to the same multisensory object (e.g., which speaker's voice arises from which speaking individual). While

Corresponding author:
David Brang, Department of Psychology, University of Michigan, 530 Church Street, Ann Arbor, MI 48109, USA.
Email: djbrang@umich.edu

substantial research has demonstrated that spatial and temporal coincidence are important factors in our determination of when to integrate auditory and visual signals and when to separate them into different objects or events (Radeau & Bertelson, 1987; Stein, Meredith, & Wolf, 1993; Stein & Stanford, 2008), these studies have largely examined single pairs of stimuli (one face and one voice). However, in natural settings, we may experience several overlapping voices arising from several speaking individuals. When competing signals are present, how do we accurately bind the associated voice to the correct face enabling speech perception to benefit from visual information? Here, we present a novel illusion demonstrating that expectations about what a speaker sounds and looks like can modulate which auditory and visual signals are bound into unified multisensory percepts, overriding ambiguous information from spatial and temporal cues.

Previous research investigating why only some combinations of auditory and visual signals are integrated together into unified percepts has given rise to Unity Assumption models of perception (Vatakis & Spence, 2008; Welch & Warren, 1980), which emphasize that different sensory signals are perceived as arising from the same multisensory object when there is statistical coincidence among the sensory signals. This unification of perceptual information, such as when auditory-visual signals (including speech—Vatakis & Spence, 2008—and nonspeech signals—Chuen & Schutz, 2016) are integrated into a single object, significantly benefits our ability to make perceptual inferences about the natural environment. That is, when our sensory systems bind separate signals together, we establish cause-and-effect predictions that allow generalization of the information from one sensory modality to the other for that specific object. While this unification is most effective when multisensory cues are precisely aligned on both spatial and temporal dimensions (Stein et al., 1993), these are not absolute requirements for multisensory integration to occur (Magnotti, Ma, & Beauchamp, 2013; Spence, 2013; Wallace et al., 2004). In the temporal domain, auditory speech occurring within approximately 200 ms of visual speech is typically perceived as a single synchronous event (Dixon & Spitz, 1980; Grant, Wassenhove, & Poeppel, 2003) and demonstrates greater multisensory integration than for auditory-visual signals presented outside of this window (Munhall, Gribble, Sacco, & Ward, 1996). Indeed, individuals' tendency to integrate auditory-visual speech even if the signals are not temporally synchronized has led researchers to argue that a temporal window of integration exists for auditory-visual speech perception (Lewkowicz, 1996). In the spatial domain, an auditory signal that is temporally, but not spatially, aligned with a speaking face will be perceived as originating from the location of the visual source (Pick, Warren, & Hay, 1969; Welch & Warren, 1980), described as the ventriloquist illusion.

In addition to spatial and temporal congruence, several studies have demonstrated an important role of sematic congruence for the integration of auditory-visual signals. In particular, auditory-visual speech integration is reduced when the gender of the auditory-visual signals is mismatched (e.g., a male voice paired with a female face; Vatakis & Spence, 2007) even if the signals are spatially and temporally aligned (Walker, Bruce, & O'Malley, 1995). However, as this past research examined multisensory integration only in terms of gender congruity, it is unclear how specific this semantic feature is. For example, is the perceptual system willing to integrate a male voice with any male face or is the identity of the individual important to perceptually bind these signals together? In natural settings, it may be more important to search for a specific individual in crowded settings than simply anyone who is of that gender.

One additional concern of many prior studies, however, is that the influence of different forms of multisensory congruity (e.g., spatial, temporal, semantic) has typically been examined in the context of isolated signals (e.g., one visual signal and one auditory signal), which

may artificially increase the statistical likelihood of auditory-visual integration since no alternative candidate sources remain. Conversely, in the natural environment, individuals can experience multiple overlapping auditory-visual signals that occur close in time and space (e.g., in the cocktail party effect, listeners must segregate one voice from a myriad of overlapping speakers; Moray, 1959). Given that multisensory integration may occur even if only one dimension (e.g., temporal synchrony) is shared across the modalities, what information determines whether the signals are integrated or separated when multiple auditory and visual cues are presented in close temporal proximity (e.g., two speakers talking over one another)? As our perceptual system must evaluate which auditory stimulus is associated with which visual stimulus, one possibility is that the auditory-visual pair that is closest in time will be integrated, regardless of spatial or semantic congruity between the two. For example, when presented with the video of a male speaker along with both male and female voices, this model would predict that the only determining factor about which voice will be integrated with the face is temporal proximity (which voice occurred closer in time to the lip movements?). Alternatively, since research has demonstrated that individuals possess a flexible ability to integrate auditory-visual stimuli even at temporal delays of up to several hundred milliseconds (Dixon & Spitz, 1980), it is possible that temporal simultaneity only matters up to a point (e.g., within the temporal window of integration) and if both auditory stimuli are presented within that window, an additional dimension needs to be consulted in order to decide which voice is integrated with the face. Here, we present a novel multisensory illusion demonstrating a limitation in the factor of temporal simultaneity. Specifically, in support of this latter model, we demonstrate that semantic information (visual identity information) can modulate which sensory signals are integrated in the presence of competing auditory-visual cues.

Experiment 1

Throughout the experiment, participants were presented with videos of a female face and a male face (shown side-by-side) as they articulated the same phoneme (e.g.,/ba/) (Figure 1 and Supplemental Movies). Each speaker's voice was always either temporally synchronized with the movements of the face of the matching gender (Gender Synchronous, Target Present, and Absent conditions) or mismatching gender (Gender Asynchronous [Target Absent and Illusion] condition). For example, in the Gender Asynchronous (Target Absent and Illusion) condition, a female voice was synchronized with the timing of the male facial movements, and the male voice synchronized with the timing of the female facial movements. Across a trial, the videos were repeated, and the face-voice pairs spoke at increasing delays. When the voices were presented close in time, the ambiguous temporal information enabled speaker gender to influence the spatial position of each voice to be perceived as spatially congruent with the gender-matched face across all conditions. Throughout the experiment, participants were instructed to simply respond when they detected a change in the spatial position of the voices. Critically, the only real changes in spatial position occurred during the Gender Synchronous (Target Present) trials. However, during the Gender Asynchronous (Target Absent and Illusion) trials, we expected participants to experience an illusory change in the spatial position of the voices at some point during the trial, when the temporal delay between the matching voice and face pair became too large to sustain a unified multisensory experience, leading to the voice moving to the mismatched-gender face that it was temporally synchronized with.

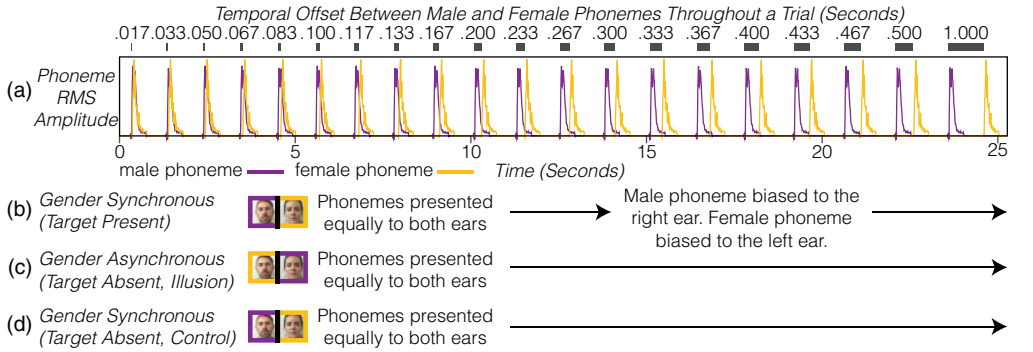

Figure 1. (a) Schematic of a typical trial in Experiment 1. Traces reflect auditory speech envelopes for a representative male (purple) and female (yellow) phoneme/ba/. Phoneme dyads were repeated 20 times on each trial with increasing temporal offset between the two audio streams. Each audio stream was always temporally synchronized to one of the visual streams (either a male or female speaking face). In this example, the male face was presented on the left side of the screen and the female face on the right side of the screen. (b) Gender Synchronous (Target Present) trials. The male phoneme was synchronized with the male face and female phoneme was synchronized with female face. At some point during the trial (between Repetitions 11 and 20), the audio streams change from being balanced across the ears to spatially biased; subjects were instructed to respond when they heard a change in the spatial location of the sounds. (c) Gender Asynchronous (Target Absent and Illusion) trials. The female phoneme was synchronized with male face and male phoneme was synchronized with female face. The audio streams were balanced across the ears throughout the trial, thus no response was required. Nevertheless, subjects experienced an illusory change in the spatial position of the voices (from the gender-matched face to the temporally synchronized face) when the temporal discrepancy became too large. (c) Gender Synchronous (Target Absent and Control) trials. The male phoneme was synchronized with the male face and the female phoneme was synchronized with the female face. Audio streams were balanced across the ears throughout the trial, thus no response was required.

Methods

Participants. Data were collected from 34 fluent English-speaking undergraduate students at the University of Michigan ($M = 18.76$ years old, $SD = 0.74$; 15 females; 33 right-handed). Data from five additional participants were excluded for failing to follow task instructions (0% of changes were detected). Power analysis (paired t test, effect size = 0.5, power = .80, $\alpha = .05$) indicated a minimum sample size of 34 participants, and additional participants were recruited to ensure that we would reach this target number after the exclusion of participants. All participants gave informed consent prior to the experiment and were given course credit for their participation. This study was approved by the institutional review board (IRB) at the University of Michigan.

Design and procedure. Video stimuli were recorded from two native English-speaking individuals (one male and one female) in a well-lit room against a white background at 59.94 frames per second using a Nikon D3200 camera and were separated offline into audio and visual streams. Audio files were amplitude normalized, and video frames were cropped and centered on the speakers' faces. Speech onset-time was extracted at the first consonantal sound in each audio file for subsequent processing using Audacity. Visual movies were edited to begin before speech-related lip movements and end following the completion of the speech sound (57 frames for each of the four stimuli).

Stimuli were presented using PsychToolbox in MATLAB using Sennheiser HD 280 Pro headphones. On each trial, participants were presented with both a male and female face,

presented adjacent to one another on the screen (each video subtending 21 horizontal and 35 vertical degrees visual angle with no gap between the videos) along with one of two spoken phonemes (/ba/,/pa/) produced by each speaker; both speakers voiced the same phoneme on a given trial and auditory-visual stimuli were always congruent in terms of the auditory phoneme and visual movie. Two exemplars for each phoneme were chosen for each speaker, yielding eight pairs of auditory and visual stimuli.

The positions of the videos (male on left side of screen, female on right, and vice versa) were counterbalanced within trial conditions. The two videos were temporally offset from one another ranging from 1 frame (16.67 ms) to 60 frames (1,000 ms) (see Figure 1). The order of the videos (whether the male or female face began first) was counterbalanced within trial conditions. Matching auditory phonemes from the male and female speakers were temporally aligned with each of the faces, depending on the condition. Within each trial, videos were repeated 20 times with systematically increasing onset delays between the two faces (Figure 1). A constant ISI of 1 second was maintained between repetitions.

Three conditions were included: (a) Gender Synchronous (Target Present), (b) Gender Asynchronous (Target Absent and Illusion), and (c) Gender Synchronous (Target Absent and Control). In the two Gender Synchronous conditions, the phonemes were temporally synchronized to the video of the matching gender face (e.g., the male phoneme synchronized with the male face and the female phoneme synchronized with the female face). In the Gender Asynchronous (Target Absent and Illusion) condition, the phonemes were temporally aligned to the video of the mismatching gender face (e.g., the female phoneme synchronized with the male face and the male phoneme synchronized with the female face). In all conditions, participants were initially presented with both voices equally mixed into the left and right ears of the headphones, providing no information about the spatial position of the voices. In the Gender Synchronous (Target Present) condition, however, at a random interval between the 11th and 20th repetition during a trial, the weighting of the auditory stimulus changed from equal weighting across the ears to a 60%/40% biased weighting, shifting the perceived position of the sounds to be more spatially aligned with the face of the mismatched gender (e.g., the male voice was biased to the spatial location of the female face). Participants were instructed to respond with a button press when they detected this change in the spatial position of the voices and withhold responses on trials without spatial changes (i.e., Target Absent trials). However, during Gender Asynchronous (Target Absent and Illusion) trials, many participants perceived an illusory change in the position of the voices (as confirmed through debriefing) and reported the experience of a change even though one had not occurred.

Responses were made using a Cedrus Response Box model RB-834. As soon as participants made a response, they were presented with a blank screen for 2 seconds, which then advanced to the next trial without providing feedback on their performance. Participants completed six practice trials from the Gender Synchronous (Target Present) condition, with feedback. Specifically, to ensure that participants could identify when a spatial change occurred, the word "Change" appeared on the screen when the spatial location of the sounds changed during these trials, to acclimate the participants to the task of identifying changes in the spatial position of the voices. Participants were instructed that a change in the spatial position of the sounds would be present on approximately 50% of the trials. Participants completed one block containing 16 trials for each of the three conditions, resulting in 48 trials total and a total time of approximately 16 minutes.

Data analysis. Data were analyzed using SPSS 23, and power calculations were conducted using G*Power 3. The percentage of change detections reported showed nonnormality according to the Shapiro–Wilk test and so were rank-transformed (Wobbrock, Findlater,

Gergle, & Higgins, 2011) prior to being subjected to repeated measures analyses, with condition (Gender Synchronous [Target Present], Gender Asynchronous [Target Absent and Illusion], and Gender Synchronous [Target Absent, Control]) as the independent variable. Although the original degrees of freedom are reported here for clarity, p values were subjected to Greenhouse–Geisser correction where appropriate (Greenhouse & Geisser, 1959). Given a sample size of 34, α rate of .05, 80% power, one group, three measurements, and a nonsphericity correction of .814, the minimal effect size this test can detect is $\eta_p^2 = .151$. Wilcoxon Signed Ranks Test was used in follow-up analyses to the repeated measures analysis of variance (ANOVA). To control for multiple comparisons among the triad of Wilcoxon Signed Ranks Test computed for relevant measures, we only considered results to be significant if the p value did not exceed the Bonferroni correction ($p < .0167$). Given a sample size of 34, α rate of .0167, and 80% power, the minimal effect size these data can detect is $d = 0.595$. Reaction times for responses during the first half of the illusion trials were compared with those in the second half using a two-sided paired t test.

Results

Examining the frequency at which participants reported a change in the location of the voices across the three experimental conditions, results from the ANOVA revealed significantly different patterns of performance, $F(2, 66) = 23.31$, $p < .0001$, $\eta_p^2 = .414$, (Figure 2 (a)). Follow-up paired Wilcoxon Signed Ranks Tests demonstrated that participants reported a significantly greater number of spatial changes in the Gender Asynchronous (Target Absent and Illusion) condition compared with the Gender Synchronous (Target Absent and Control) condition ($Z = 4.704$, $p < .0001$, $d = 1.255$). Critically, no spatial change occurred in either of these conditions on any trial, but the gender-mismatch effect in the Gender Asynchronous (Target Absent and Illusion) condition produced an illusory change, indicating the illusion deteriorated at some point during the trial resulting in participants' experience of a change in the position of where each voice originated. A similar magnitude difference was observed between the Gender Synchronous (Target Present)

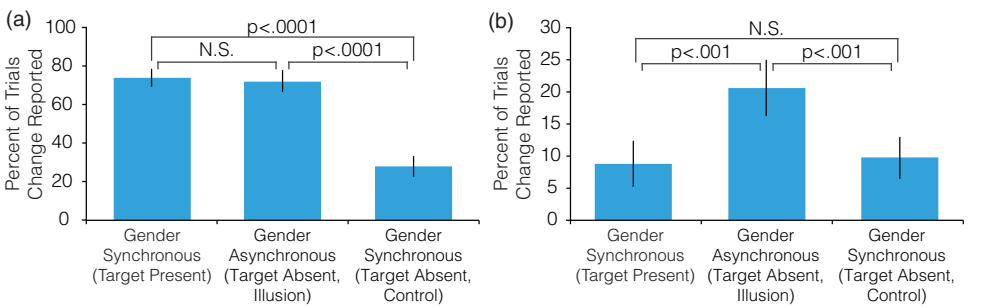

Figure 2. Experiment 1. Percent of trials in each of the conditions in which a change in the spatial location of the auditory phonemes was reported (a) anytime during the trial or (b) during the first half of the trial (during which, unknown to the subject, no change ever occurred). (a) Subjects reported a significantly greater number of changes during the Gender Asynchronous (Target Absent and Illusion) condition compared with the Gender Synchronous (Target Absent and Control) condition, even though no change was present in either of these conditions, differing only in terms of which voice was temporally synchronized to which face. (b) Even though no change occurred during this time period, subjects reported a significantly greater number of changes during the Gender Asynchronous (Target Absent and Illusion) condition. Error bars reflect standard error of the mean.

condition compared with the Gender Synchronous (Target Absent and Control) condition ($Z = 4.491$, $p <. 0001$, $d = 1.147$), demonstrating that participants could in fact discriminate between present and absent changes in the spatial location of the voices. Finally, there was no significant difference between the Gender Synchronous (Target Present) condition and the Gender Asynchronous (Target Absent and Illusion) condition ($Z = 0.089$, $p = .923$, $d = 0.039$), with this null effect demonstrating that participants were equally likely to respond to a real change in the spatial position of the voices, as they were to respond to an illusory shift in the spatial location.

Trials were structured so that no real change in the spatial position of the voices ever occurred in the first 10 repetitions during a trial. This allowed us to compare participants' likelihood to report a change when one was never present across the three experimental conditions. Of note, as no changes occurred during this period of the trial, the Gender Synchronous (Target Present) and Gender Synchronous (Target Absent and Control) conditions were identical, with the Gender Asynchronous (Target Absent and Illusion) condition differing only in terms of temporal synchronicity between the matching voice and face. Nevertheless, even though no real spatial change occurred during this time period in any condition, participants may still have experienced a shift in the spatial position in the Gender Asynchronous (Target Absent and Illusion) condition if the illusion deteriorated at some point during the trial and the voices changed from the spatial position of the expected gender toward the spatial position of the temporally synchronized (mismatched) gender. Consistent with our expectations, participants' pattern of responses significantly differed across the conditions, $F(2, 66)=12.29$, $p < .0001$, $\eta_p^2 = .271$ (Figure 2(b)). Follow-up paired Wilcoxon Signed Ranks Tests demonstrated that participants reported a significantly greater number of spatial changes in the Gender Asynchronous (Target Absent and Illusion) condition compared with the Gender Synchronous (Target Absent and Control) condition ($Z = 3.579$, $p < .001$, $d = 0.721$) as well as the Gender Synchronous (Target Present) condition ($Z = 3.478$, $p < .001$, $d = 0.660$), with no difference between the Gender Synchronous (Target Present) condition and the Gender Synchronous (Target Absent and Control) condition ($Z = 0.752$, $p = .452$, $d = 0.138$).

While it is possible that participants reported the experience of spatial changes in the Gender Asynchronous (Target Absent and Illusion) condition randomly, we predicted that a greater number of spatial changes would be reported in the latter half a trial (during the 11th–20th repetitions) due to the larger temporal discrepancy between the speaking face-voice pairs. Comparing the number of changes reported in this condition between the first half of the trial ($M = 20.6\%$ of trials, $SD = 25.2\%$) and second half ($M = 51.5\%$ of trials, $SD = 28.6\%$) of the trial indicated that the illusion was more likely to deteriorate in the latter half, $t(33) = 4.186$, $p < .001$, $d = 0.718$, when the temporal discrepancy was larger, indicating that participants were not reporting spatial changes at random. In terms of the temporal discrepancy between the speaking face-voice pairs, participants reported a spatial change (indicative of the illusion deteriorating) when the gender-congruent face-voice pairs were presented on average 304 ms apart ($SD = 92$), outside of the typical temporal window of integration (Dixon & Spitz, 1980; Grant et al., 2003).

Experiment 2

Experiment 1 demonstrated that participants relied on gender information to spatially and temporally bind a voice to a specific face, even when both voices were presented without clear spatial information (equal weighting across each ear of the headphones). We interpret this result to mean that the expected face-voice pairings (based on gender or identity

information) had a prominent effect on perceived spatial position when there was no competing spatial information. Experiment 2 served to replicate the three conditions present in Experiment 1 and additionally add in a factor of competing spatial information. Specifically, would the illusion persist in the presence of *both* competing temporal and spatial information, for example, if presented with a male face on the left and a female face on the right, would a left lateralized female voice that is temporally synchronized with the male face persist in being localized to the female face on the right side of the screen (due to expectation of the speaker's face-voice identity-congruence)? This would tell us if our expectation about face-voice pairings is only relevant when there is no information about the spatial source, or if this information biases perception even when it conflicts with both spatial and temporal cues.

Methods

Experimental methods were matched to those in Experiment 1, except where noted below.

Participants. Data were collected from 26 fluent English-speaking undergraduate students at the University of Michigan ($M = 18.77$ years old, $SD = 0.65$; 14 females; 24 right-handed). Data from one additional participant were excluded for failing to follow task instructions. Power analyses indicated that the minimum sample size required to achieve 80% power for the smallest Wilcoxon Signed Ranks Test effect size reported in Experiment 1 (effect size of $d = 0.660$, α rate of .05, 80% power) was 21 participants, and additional participants were recruited to ensure that we would reach this target number after the exclusion of participants. All participants gave informed consent prior to the experiment and were given course credit for their participation. This study was approved by the IRB at the University of Michigan.

Design and procedure. The same stimuli and procedure in Experiment 1 were used in Experiment 2. A total of six conditions were examined in Experiment 2. First, we included the three conditions described in Experiment 1 (Gender Synchronous [Target Present], Gender Asynchronous [Target Absent and Illusion], and Gender Synchronous [Target Absent and Control]) in which the voices were always equally weighted across the ears at the start of the trial (such that no spatial information was initially present; i.e., the Spatially Unbiased conditions). Second, we included three additional matched conditions differing from the initial ones only in that the voices were initially spatially biased (60%/40%) to the spatial location of the face that they were temporally synchronized with (see Supplemental Movies) (i.e., the Spatially Biased conditions). Accordingly, in the Gender Asynchronous (Target Absent and Illusion) condition, this resulted in the male voice being both temporally synchronized and spatially aligned with the female face. As in Experiment 1, the voices in the Gender Synchronous (Target Present) conditions changed their spatial position at a random interval between the 11th and 20th repetition; in the Spatially Biased variant, the weighting of the auditory stimulus was switched at this time point (e.g., a male voice that was initially spatially biased to the left would change to a rightward spatial position). Participants completed one block containing 12 trials for each of the six conditions, resulting in 72 trials total and a total time of approximately 24 minutes.

Data analysis. The percentage of change detections reported and reaction time were subjected to repeated measures analyses, with Trial Type (Gender Synchronous [Target Present], Gender Asynchronous [Target Absent and Illusion], Gender Synchronous [Target Absent

and Control]) and Initial Spatial Bias (Spatially Biased or Spatially Unbiased) as the independent variables. The percentage of change detections reported showed nonnormality according to the Shapiro–Wilk test and so were rank-transformed (Wobbrock et al., 2011) prior to being subjected to repeated measures analyses. In the analysis of reaction time data, some conditions in some participants did not have any reaction times due to a 0% response rate for those conditions; these missing values were excluded. Although the original degrees of freedom are reported here for clarity, p values were subjected to Greenhouse–Geisser correction where appropriate (Greenhouse & Geisser, 1959). Given a sample size of 26, α rate of .05, 80% power, one group, six measurements, and a nonsphericity correction of .8, the minimal effect size this test can detect is $\eta_p^2 = .111$. The percentage of change detections reported showed nonnormality according to the Shapiro–Wilk test, and so the Wilcoxon Signed Ranks Test was used in follow-up analyses to the repeated measures ANOVAs. To control for multiple comparisons among the six Wilcoxon Signed Ranks Test computed for relevant measures, we only considered results to be significant if the p value did not exceed the Bonferroni correction ($p < .0083$). Given a sample size of 26, α rate of .0083, and 80% power, the minimal effect size these data can detect is $d = 0.754$.

Results

As can be seen in Figure 3(a), participants' reports of the number of changes in the spatial location of the voices differed across the conditions, yielding a significant main effect of Trial Type, $F(2, 50) = 19.64$, $p < .0001$, $\eta_p^2 = .440$, no main effect of Initial Spatial Bias, $F(1, 25) = 1.053$, $p = .315$, $\eta_p^2 = .040$, and no interaction between the two, $F(2, 50) = 1.149$, $p = .320$, $\eta_p^2 = .044$. Follow-up paired Wilcoxon Signed Ranks Tests examining the driving features of this main effect of Trial Type replicated the findings of Experiment 1, such that participants reported a significantly greater number of spatial changes in the Gender

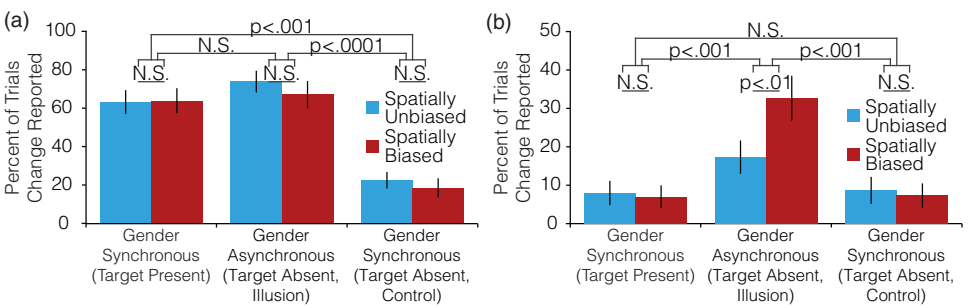

Figure 3. Experiment 2. Percent of trials in each of the conditions in which a change in the spatial location of the auditory phonemes was reported (a) anytime during the trial or (b) during the first half of the trial. Blue bars reflect the same trial conditions used in Experiment 1 in which the male and female voices were initially balanced across the ears at the start of the trial. Red bars reflect trials in which the male and female voices were spatially biased to one ear at the start of the trial. (a) Subjects reported a significantly greater number of changes during the Gender Asynchronous (Target Absent and Illusion) condition compared with the Gender Synchronous (Target Absent and Control) condition, even though no change was present in either of these conditions and differed only in which voice was temporally synchronized to which face. (b) Even though no change occurred during this time period, subjects reported a significantly greater number of changes in the Gender Asynchronous (Target Absent and Illusion) condition. In addition, a significantly greater number of changes were detected in the Initial Spatial Bias condition, indicating that the presence of competing spatial information caused the illusion to deteriorate more quickly. Error bars reflect standard error of the mean.

Asynchronous (Target Absent and Illusion) condition compared with the Gender Synchronous (Target Absent and Control) condition ($Z = 4.108$, $p < .0001$, $d = 1.499$) with a similar magnitude difference observed between the Gender Synchronous (Target Present) condition compared with the Gender Synchronous (Target Absent and Control) condition ($Z = 3.862$, $p < .001$, $d = 1.219$). As in Experiment 1, no significant difference was observed between the Gender Synchronous (Target Present) condition and the Gender Asynchronous (Target Absent and Illusion) condition ($Z = 0.829$, $p = .407$, $d = 0.133$). This latter null effect demonstrates that participants were equally likely to respond to a real change in the spatial position of the voices as they would the illusory shift in the spatial location.

The absence of an interaction in the aforementioned data suggest that participants experienced an equal strength of the illusion throughout the entire trial period in both the presence and absence of competing auditory spatial information. However, to more directly examine whether the illusion was weakened by the introduction of auditory spatial cues, we examined two measures of when the illusion deteriorated for participants. First, similar to the analysis reported in Experiment 1, we examined false alarms within the first half of the trial period when no real spatial change occurred. Consistent with this prediction of a weaker illusion in the presence of incongruent auditory spatial information, the repeated measures ANOVA demonstrated significant main effects of Trial Type, $F(2, 50) = 18.59$, $p < .0001$, $\eta_p^2 = .427$, Spatial Location, $F(1, 25) = 13.85$, $p < .01$, $\eta_p^2 = .356$, and an interaction between the two, $F(2, 50) = 18.89$, $p < .0001$, $\eta_p^2 = .430$. Follow-up paired Wilcoxon Signed Ranks Tests demonstrated that participants reported a significantly greater number of spatial changes in the Gender Asynchronous (Target Absent and Illusion) condition compared with both the Gender Synchronous (Target Absent and Control) condition ($Z = 3.431$, $p < .001$, $d = 0.817$) and the Gender Synchronous (Target Present) condition ($Z = 3.657$, $p < .001$, $d = 0.845$), with no difference between the Gender Synchronous (Target Present) condition and the Gender Synchronous (Target Absent and Control) condition ($Z = 0.370$, $p = .711$, $d = 0.078$). Critically, the effect of Spatial information (Spatially Biased vs. Spatially Unbiased) was only significant between the two Illusion conditions ($Z = 2.993$, $p < .01$, $d = 0.656$), and not either the Gender Synchronous (Target Present) ($Z = 0.667$, $p = .505$, $d = 0.116$) or Gender Synchronous (Target Absent and Control) condition ($Z = 1.138$, $p = .255$, $d = 0.275$).

Second, we examined response times across the trial conditions as trial-by-trial measures of when participants experienced the real or illusory changes (data from two participants who made no responses in at least one of these conditions were excluded). Results showed that response times for reporting an illusory change were significantly faster in the Spatially Biased Gender Asynchronous (Target Absent and Illusion) condition ($M = 11.1$ seconds, $SD = 2.7$) than the Spatially Unbiased Gender Asynchronous (Target Absent and Illusion) condition ($M = 14.1$ seconds, $SD = 3.0$), $t(23) = 5.944$, $p < .0001$, $d = 1.213$. Examining these same data instead in terms of the lag at which participants reported the spatial change during the illusion, participants experienced the illusory spatial shift on average at 330 ms ($SD = 116$) in the Spatially Unbiased condition and on average at 237 ms ($SD = 70$) in the Spatially Biased condition. These data demonstrate that the introduction of incongruent spatial information does not prevent the occurrence of the illusion but simply weakens its persistence.

Experiment 3

Experiments 1 and 2 demonstrate that ambiguous spatial and temporal information can be overridden by gender information. However, it remains unclear whether this effect is due to

binary assignment based on gender cues (i.e., integrate the male voice with the male face) or if this bias extends to the specific identity of the speaker (i.e., integrate this voice with the specific face of the associated individual). Consistent with this latter view, research has demonstrated that individuals are capable of matching a face to the voice of an unknown person (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004) indicating the maintenance of statistical priors about the congruency between face-voice pairs. Indeed, correlations have been observed between the perceived masculinity/femininity of individuals faces and voices (Smith, Dunn, Baguley, & Stacey, 2016), based in part on increased testosterone levels in men resulting in increased perceived facial masculinity (Pisanski, Mishra, & Rendall, 2012) and lower fundamental frequencies (Dabbs & Mallinger, 1999; Fant, 1971). If the expectations that lead to the demonstrated illusion are based on gender information, then trials using same-gender pairs of individuals would significantly reduce the strength of the illusion (i.e., participants will integrate the male voice to the male face that is most temporally congruent with the lip movements regardless of who is the real owner of the voice). Conversely, if this illusion occurs from expectations about *identity* information, then subjects would persist in experiencing the illusion even when two male faces and voices are presented.

Methods

Experimental methods were matched to those in Experiment 1, except where noted below.

Participants. Data were collected from 27 fluent English-speaking undergraduate students at the University of Michigan ($M = 18.89$ years old, $SD = 0.96$; 17 females; 26 right-handed). Data from one additional participant were excluded for failing to follow task instructions. Power analyses indicated that the minimum sample size required to achieve 80% power for the smallest Wilcoxon Signed Ranks Test effect size reported in Experiment 1 (effect size of $d = 0.660$, α rate of .05, 80% power) was 21 participants, and additional participants were recruited to ensure that we would reach this target number after the exclusion of participants. All participants gave informed consent prior to the experiment and were given course credit for their participation. This study was approved by the IRB at the University of Michigan.

Design and procedure. In Experiment 3, the stimuli were changed from the single male speaker and single female speaker used in Experiments 1 and 2 to four new speakers (two men and two women) to accomplish two goals. First, we sought to compare gender and identity information. Second, we included new speakers to ensure that the present illusion could be observed across several speaking individuals and did not reflect idiosyncrasies specific to the single pair of individuals from the first two experiments. We selected the four speakers from a set of 20 stimulus sets made in our lab with the goal of producing stimuli that included a wide range of differences in fundamental frequencies ($f0$) to ensure a generality of results. Each speakers' average $f0$ throughout the utterance was extracted using Praat: Male 1 (93.1 Hz), Male 2 (135.9 Hz), Female 1 (165.1 Hz), and Female 2 (196.4 Hz). Same-Gender trials paired speakers Male 1/Male 2 (f0 difference = 42.8 Hz) and Female 1/Female 2 ($f0$ difference = 31.3 Hz). Different-Gender trials paired speakers Male 2/Female 1 ($f0$ difference = 29.2 Hz) and Male 1/Female 2 ($f0$ difference = 103.3 Hz). Results did not differ as a function of the difference in $f0$ between the speakers.

The same procedure in Experiments 1 and 2 was used in Experiment 3. A total of six conditions were examined in Experiment 3. First, we included the three main conditions

described in Experiments 1 and 2: (Identity Synchronous [Target Present], Identity Asynchronous [Target Absent and Illusion], and Identity Synchronous [Target Absent and Control]); note the change in terminology from "Gender" used in the previous experiments to "Identity." Second, we included an additional factor of Gender Matching, such that the two speakers presented on a trial were either the Same Gender or Different Gender (Same-Gender and Different-Gender trials). In the Same-Gender, Identity Asynchronous (Target Absent and Illusion) condition, two male voices would each be synchronized with the other individual's face (and on separate trials, two female voices would be synchronized with the other individual's face). If participants' expectations based on the identity of the speaker drive the observed illusion, then participants will initially bind the voices to the correct faces based on the identity of the speaker, regardless of the genders of the individuals presented on the screen. Then, as in the previous illusion conditions, as the trial progresses, the temporal discrepancy between these natural pairings would become too great, and the voices would change to be integrated with the other speaker. Participants completed one block containing 12 trials for each of the six conditions, resulting in 72 trials total and a total time of approximately 24 minutes.

Data analysis. The percentage of change detections reported and RT were subjected to repeated measures analyses, with factors of Trial Type (Identity Synchronous [Target Present], Identity Asynchronous [Target Absent and Illusion], and Identity Synchronous [Target Absent and Control]) and Gender Matching (Same Gender and Different Gender) as the independent variables. The percentage of change detections reported showed nonnormality according to the Shapiro–Wilk test and so were rank-transformed (Wobbrock et al., 2011) prior to being subjected to repeated measures analyses. In the analysis of reaction time data, some conditions in some participants did not have any reaction times due to a 0% response rate for those conditions; these missing values were excluded. Although the original degrees of freedom are reported here for clarity, p values were subjected to Greenhouse–Geisser correction where appropriate (Greenhouse & Geisser, 1959). Given a sample size of 27, α rate of .05, 80% power, one group, six measurements, and a nonsphericity correction of .8, the minimal effect size this test can detect is $\eta_p^2 = .107$. The percentage of change detections reported showed nonnormality according to the Shapiro–Wilk test, and so the Wilcoxon Signed Ranks Test was used in follow-up analyses to the repeated measures ANOVAs. To control for multiple comparisons among the six Wilcoxon Signed Ranks Test computed for relevant measures, we only considered results to be significant if the p value did not exceed the Bonferroni correction ($p < .0083$). Given a sample size of 27, α rate of .0083, and 80% power, the minimal effect size these data can detect is $d = 0.738$.

Results

Data in Experiment 3 yielded a significant main effect of Trial Type, $F(2, 52) = 22.90$, $p < .0001$, $\eta_p^2 = .468$, but no main effect of Gender Matching, $F(1, 26) = 0.264$, $p = .612$, $\eta_p^2 = .010$, or significant interaction between the two, $F(2, 52) = 0.01$, $p = .979$, $\eta_p^2 = .001$ (Figure 4). Follow-up paired Wilcoxon Signed Ranks Tests examining the driving features of this main effect of Trial Type replicated the findings of Experiments 1 and 2, such that participants reported a significantly greater number of spatial changes in the Identity Asynchronous (Target Absent and Illusion) condition compared with the Identity Synchronous (Target Absent and Control) condition ($Z = 4.204$, $p < .0001$, $d = 1.020$) with a similar magnitude difference observed between the Identity Synchronous (Target Present)

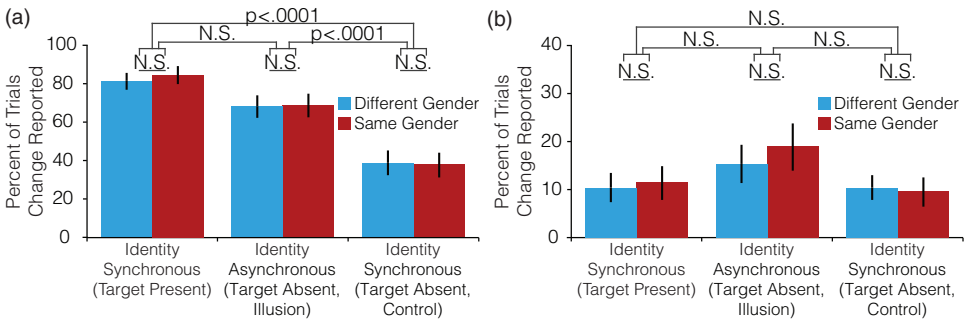

Figure 4. Experiment 3. Percent of trials in each of the conditions in which a change in the spatial location of the auditory phonemes was reported (a) anytime during the trial or (b) during the first half of the trial. Blue bars reflect trials in which one male and one female face-voice pairs were presented (as in Experiments 1 and 2), whereas red bars reflect trials in which two faces and voices of the Same Gender were presented (e.g., two male faces and voices). (a) Subjects reported a significantly greater number of changes during the Identity Asynchronous (Target Absent and Illusion) condition compared with the Identity Synchronous (Target Absent and Control) condition regardless of whether same- or different-gender pairs were presented.

condition compared with the Identity Synchronous (Target Absent and Control) condition ($Z = 4.026$, $p < .0001$, $d = 1.262$). As in Experiment 1, no significant difference was observed between the Identity Synchronous (Target Present) condition and the Identity Asynchronous (Target Absent and Illusion) condition ($Z = 1.842$, $p = .066$, $d = 0.408$). This latter null effect demonstrates that participants were similarly likely to respond to a real change in the spatial position of the voices as they would to the illusory shift in the spatial location. As is apparent in Figure 4, there was no difference between the Different-Gender and Same-Gender trials in any of the three Trial Types (Target Absent, Illusion, and Control): all p values $> .346$.

As in Experiment 2, we additionally compared the number of changes reported during the first half of the trial to examine whether the illusion deteriorated more quickly in Same-Gender or Different-Gender trials. Repeated measures ANOVA revealed a nonsignificant trend of Trial Type, $F(2, 52) = 3.056$, $p = .065$, $\eta_p^2 = .105$, a significant effect of Gender Matching, $F(1, 26) = 5.056$, $p = .033$, $\eta_p^2 = .163$, and a nonsignificant trend of an interaction between the two, $F(2, 52) = 2.273$, $p = .116$, $\eta_p^2 = .080$. No differences were observed in follow-up paired Wilcoxon Signed Ranks Tests (all p greater than .088 uncorrected for multiple comparisons).

To further examine whether any difference in Gender Matching was observed, we compared reaction times for the critical Identity Asynchronous (Target Absent and Illusion) conditions as trial-by-trial measures of when participants' experienced the illusion deteriorate (i.e., how big of a temporal discrepancy could the illusion overcome?); data from one participant who made no responses in at least one of these conditions were excluded. However, no significant difference between the Same-Gender ($M = 14.6$ seconds, $SD = 0.70$) and the Different-Gender ($M = 14.5$ seconds, $SD = 0.65$) conditions was observed, $t(25) = 0.249$, $p = .805$, $d = 0.049$. In line with the aforementioned data, there is no evidence that this effect is driven by gender information alone. Instead, these data are largely consistent with a model in which participants use identity information to decide which voice should be perceptually integrated with which face.

Discussion

Past research has demonstrated that temporal and spatial coincidence are critical features in determining whether multisensory signals are unified into either a single or separate perceptual objects (Radeau & Bertelson, 1987), with perceptual unification leading to maximal benefits of multisensory integration (Vatakis & Spence, 2007). Here, we show that a third feature, namely, multisensory identity information, can modulate which auditory and visual signals are bound into unified multisensory percepts, particularly when spatial and temporal cues are ambiguous. This study examined the extent to which expectations about what a speaker should sound like (based on gender information in Experiments 1 and 2, and identity information in Experiment 3) would cause the identity-matched voice to override temporal (Experiment 1) and spatial (Experiment 2) cues, persisting in being integrated into a unified multisensory object. Results demonstrated that even in the presence of competing information (i.e., the voice of another individual that was more temporally and spatially aligned with the face), participants' expectations about the identities of the individuals' faces and voices were responsible for deciding which face-voice pairs were bound in space and time.

While many features can determine which auditory signals are bound to which visual signals, here we demonstrate that there is no strict hierarchy in terms of information (i.e., while temporal and spatial congruency are important, when these signals are ambiguous additional information is needed to decide which signals should be integrated). Impressively, this illusion highlights that multisensory identity information can overcome as much as 300 ms of auditory-visual temporal discrepancy, which is outside of the typical temporal window of integration for auditory-visual signals (Dixon & Spitz, 1980; Grant et al., 2003). However, as soon as the temporal offset between the competing auditory-visual signals increases past ~300 ms, identity information is no longer used, and participants' perception was biased more strongly by temporal coincidence (the auditory signal was bound to whichever visual signal had the highest temporal correlation). The demonstration of this result using a perceptual illusion makes it appropriate for testing multisensory processing in a wide range of individuals (e.g., children and patients) and experimental contexts, particularly given the large effect size of the illusion relative to the control condition ($d = 1.255$) and short testing duration.

In the ventriloquist illusion, temporally synchronized auditory-visual speech signals can override spatial cues about where the voice originates (Pick et al., 1969; Welch & Warren, 1980). However, ventriloquist illusion-like effects are typically examined with a single pair of auditory and visual stimuli, leaving little ambiguity about which two signals should be bound together even in presence of large spatial (~15°) and temporal (800 ms) disparities across the auditory-visual signals (Wallace et al., 2004). In the context of multiple competing auditory and visual signals, the present data show that expectations about the identity-congruence of the stimuli (i.e., which face is expected to match with which voice) can override weak spatial and temporal cues that would otherwise indicate the identity-incongruent voice and face signals should be perceptually integrated. However, this illusion failed to persist beyond an average delay of 300 ms between the identity-matched stimuli, demonstrating an upper limit to the window of integration enabled by identity-matched face-voice pairs. After this time point, the perceptual binding of the identity-matched auditory-visual speech cues became untenable as the temporal discrepancy was too large, and the sound appeared to change spatial position (reverting to a classic ventriloquism effect).

This illusion shows that our expectations about the source of an auditory-visual object can override weak temporal and spatial cues, which is a phenomenon that is likely

commonplace in the natural environment. Even in the absence of spatial information or temporal synchrony, we can estimate the source of objects in the world based on statistical or contextual associations. In the most extreme case, these types of identity effects should hold for static speakers whose motion may be impoverished (e.g., watching the back of male and female individuals' heads as they speak). Research examining the effects of gender-congruity on multisensory integration has generally demonstrated that congruent auditory-visual signals are better unified as multisensory objects. Most relevant to this study, Vatakis and Spence (2007) demonstrated that participants showed larger temporal windows of integration for auditory-visual stimuli that were gender-congruent (e.g., a male voice paired with a male face) relative to gender-incongruent (e.g., a male voice paired with a female face). This finding likely contributes to the generation of the present illusion. However, here participants are presented with multiple overlapping auditory-visual signals, introducing an extra dimension of ambiguity; in prior research utilizing a single voice with a single face, there would have been little ambiguity in which auditory signal should be bound to which visual target. Furthermore, we extend these results from simple Gender Matching to show that speaker-specific identity information can modulate these multisensory processes as well. Thus, while past work has shown the ability for gender information to modulate temporal integration, here we demonstrate that in the presence of ambiguous temporal and spatial signals, identity information determines which auditory-visual signals are bound into unitary percepts. Independent of identity-congruity effects, research has demonstrated that the perceived spatial position of a voice is modulated by the congruity of auditory and visual phonemes (Kanaya & Yokosawa, 2011). As this study utilized congruent auditory and visual phonemes (e.g., the male voice and male face articulated the same phoneme on any given trial), future research will be required to understand the relative contributions of identity-congruity and auditory-visual phoneme congruity.

The generalization of these findings from gender information to identity information in Experiment 3 clarifies the functional significance of these results. In Experiments 1 and 2, we relied on the associations between acoustical signals (vocal pitch and timbre) and visual gender information in the elicitation of these gender-congruent associations. Research has demonstrated that both pitch (the fundamental frequency of a sound) and timbre (harmonic frequencies and how the sound changes over time) are both important cues in extracting gender information from voices (Pernet & Belin, 2012). In adults, pitch is strongly predictive of a speakers' gender (e.g., Dabbs & Mallinger, 1999), and pitch is a cue used to identify the perceived gender of adult individuals (e.g., Davies & Goldberg, 2006). Of note, vocal pitch is less predictive of the gender of children, indicating that a weaker effect may persist for stimuli that include children. As the same acoustic cues that segregate gender can be used to segregate individual identity information (Kamachi et al., 2003; Pernet & Belin, 2012), in Experiment 3, we demonstrated that the identity information could drive the illusion as well, consistent with a model in which an expectation about what a speaker should sound like can be the driving factor in which voice binds to which face. We suggest that future research should examine the relative contributions of pitch and timbre in integrating faces to voices to better understand how this occurs in the natural environment.

Importantly, participants were not trained on which voice was associated with which face in any experiment. Nevertheless, participants demonstrated an association between these voices and faces in their experience of the illusion. These associations were likely due to statistical expectations about acoustical information (pitch and timbre) associated with facial features, as several studies have demonstrated that participants can readily match faces and voices based on this information (Kamachi et al., 2003; Lachs & Pisoni, 2004). While these statistical expectations likely served as the initial factor in determining pairings,

it is possible that these priors were updated throughout the task based on the frequency at which each voice was temporally synchronized with each face. Specially, throughout the task, participants viewed matching faces and voices in a temporally synchronized manner in two thirds of the trials, which could potentially strengthen the relationship further. While we anticipate that each of these factors was present in this study, without additional research they remain speculative predictions. Future research should examine the role of statistical priors and learning in this illusion to better understand the malleability of these associations in our assessment of multisensory identity-congruity. Indeed, one real-world consequence of these expectations is negative societal repercussions for individuals whose face and voice fail to be perceived as gender-congruent, particularly for those within the transgender community who must alter their voice along with their external appearance to experience a congruent and societally acceptable identity (Adler, Hirsch, & Mordaunt, 2012).

These data do not demonstrate that identity information facilitates speech; however, other studies have reported such effects based on gender-congruity (Walker et al., 1995). Instead, the present design demonstrates that identity information modulates selection processes about which multisensory signals are bound into the same object. While this is a necessary first step to understanding the phenomenon more broadly, future research should examine these effects in more natural contexts to examine their practical relevance, including attentional selection in noisy environments. For example, when one is scanning the room at a crowded party, searching for the source of a familiar voice, identity information will likely facilitate spatial binding of the voice to the face, speeding speaker localization and identification.

Several studies have failed to find the effects of contextual congruity on perceptual unity when using nonspeech stimuli (such as an individual smashing a block of ice with a hammer, a bouncing ball, and musical instruments being played; Vatakis & Spence, 2008), suggesting that variants of this illusion using nonspeech stimuli may show significantly reduced effects (but see Chuen & Schutz, 2016). While it is possible that effects of contextual congruity on perceptual unity are limited to speech stimuli due to speech being a unique stimulus to human observers (Chen & Spence, 2017; Vatakis, Ghazanfar, & Spence, 2008), these effects may alternatively be due to our extensive exposure to auditory speech in natural contexts (Lee & Noppeney, 2014) or more simply due to the idiosyncratic features of auditory-visual speech. Specifically, data from past studies indicate that auditory and visual speech signals are correlated with one another (e.g., changes in the acoustic speech envelope covary with the area of the mouth that is open during speech) and auto-regressive in nature, such that these signals operate according to slow time-varying functions (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). Indeed, substantial research has demonstrated that auditory-visual speech signals occur at a slow, rhythmic rate of approximately 5hz ($\sim$200 ms) (Crystal & House, 1982; Greenberg, Carvey, Hitchcock, & Chang, 2003), allowing substantial temporal ambiguity to exist regarding which speech signals correspond to which fluctuations in the visual stimulus. The nature of auditory-visual speech as slow fluctuations in the time-varied signal is in contrast to the dynamic transients (onset and offset sounds) that occur with most nonspeech events, such as the collision between two objects. Indeed, simple auditory-visual stimuli such as beeps and flashes of light show maximal effects of multisensory integration when the stimuli are transient in nature and not graded over time or time-varied (Keetels & Vroomen, 2005; Van der Burg, Cass, Olivers, Theeuwes, & Alais, 2010), in contrast to the effects on speech. In the present illusion, the lack of dynamic transients in the auditory-visual speech signals likely engender the perceptual ambiguity of which voice matches with which face necessary for gender information to play a modulatory role.

In summary, this study demonstrates a novel illusion that expands on previous research examining the auditory-visual integration of speech signals to highlight the useful role of identity information in the presence of ambiguous spatial and temporal information. Given that the natural environment typically contains several overlapping and ambiguous auditory-visual cues, these data indicate that expectation about identity-congruity may play a strong role in deciding what signals are bound into multisensory objects in everyday settings.

Acknowledgements

The authors are grateful to Marcia Grabowecky and Satoru Suzuki for earlier discussions on this project and Jessica Creery for aiding generation of the stimuli.

Author Contributions

D. B. wrote the main manuscript text, analyzed the data, and prepared the figures.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by NIH Grant R00 DC013828.

ORCID iD


David Brang https://orcid.org/0000-0002-2706-6777

Supplemental Material

Supplemental material for this article is available online.

References

Adler, R. K., Hirsch, S., & Mordaunt, M. (2012). *Voice and communication therapy for the transgender/transsexual client: A comprehensive clinical guide*. San Diego, CA: Plural Publishing.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*, e1000436. doi:10.1371/journal.pcbi.1000436

Chen, Y. C., & Spence, C. (2017). Assessing the role of the 'unity assumption' on multisensory integration: A review. *Frontiers in Psychology*, *8*, 445.

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

Crystal, T. H., & House, A. S. (1982). Segmental durations in connected speech signals: Preliminary results. *Journal of the Acoustical Society of America*, *72*, 705–716. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7130529

Dabbs, J. M., & Mallinger, A. (1999). High testosterone levels predict low voice pitch among men. *Personality and Individual Differences*, *27*, 801–804.

Davies, S., & Goldberg, J. M. (2006). Clinical aspects of transgender speech feminization and masculinization. *International Journal of Transgenderism*, *9*, 167–196.

Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception*, *9*, 719–721. doi:10.1068/p090719

Fant, G. (1971). *Acoustic theory of speech production: With calculations based on X-ray studies of Russian articulations* (vol. 2). Berlin, Germany: Walter de Gruyter.

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience*, *25*, 5004–5012. doi:10.1523/JNEUROSCI.0799-05.2005

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, *108*, 1197–1208. doi:10.1121/1.1288668

Grant, K. W., Wassenhove, V. V., & Poeppel, D. (2003, September). *Discrimination of auditory-visual synchrony*. Paper presented at the AVSP 2003-International Conference on Audio-Visual Speech Processing, St. Jorioz, France.

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. Y. (2003). Temporal properties of spontaneous speech - a syllable-centric perspective. *Journal of Phonetics*, *31*, 465–485. doi:10.1016/j.wocn.2003.09.005

Greenhouse, S. W., & Geisser, S. (1959). On Methods in the Analysis of Profile Data. *Psychometrika*, *24*, 95–112. doi:10.1007/Bf02289823

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice': Matching identity across modality. *Current Biology*, *13*, 1709–1714. doi:10.1016/j.cub.2003.09.005

Kanaya, S., & Yokosawa, K. (2011). Perceptual congruency of audio-visual speech affects ventriloquism with bilateral visual stimuli. *Psychonomic Bulletin & Review*, *18*, 123–128. doi:10.3758/s13423-010-0027-z

Keetels, M., & Vroomen, J. (2005). The role of spatial disparity and hemifields in audio-visual temporal order judgments. *Experimental Brain Research*, *167*, 635–640. doi:10.1007/s00221-005-0067-1

Kim, J., & Davis, C. (2003). Hearing foreign voices: Does knowing what is said affect visual-masked-speech detection? *Perception*, *32*, 111–120.

Kim, J., & Davis, C. (2004). Investigating the audio-visual speech detection advantage. *Speech Communication*, *44*, 19–30. doi:10.1016/j.specom.2004.09.008

Lachs, L., & Pisoni, D. B. (2004). Crossmodal Source Identification in Speech Perception. *Ecological Psychology*, *16*, 159–187. doi:10.1207/s15326969eco1603_1

Lee, H., & Noppeney, U. (2014). Music expertise shapes audiovisual temporal integration windows for speech, sinewave speech, and music. *Frontiers in Psychology*, *5*, 868. doi:10.3389/fpsyg.2014.00868

Lewkowicz, D. J. (1996). Perception of auditory-visual temporal synchrony in human infants. *Journal of Experimental Psychology*. *Human Perception and Performance*, *22*, 1094–1106. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8865617.

Magnotti, J. F., Ma, W. J., & Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech. *Frontiers in Psychology*, *4*, 798. doi:10.3389/fpsyg.2013.00798

Moray, N. (1959). Attention in dichotic-listening—Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, *11*, 56–60. doi:10.1080/17470215908416289

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, *58*, 351–362. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8935896

Pernet, C. R., & Belin, P. (2012). The role of pitch and timbre in voice gender categorization. *Frontiers in Psychology*, *3*, 23.

Pick, H. L., Warren, D. H., & Hay, J. C. (1969). Sensory conflict in judgments of spatial direction. *Perception & Psychophysics*, *6*, 203. doi:10.3758/Bf03207017

Pisanski, K., Mishra, S., & Rendall, D. (2012). The evolved psychology of voice: Evaluating interrelationships in listeners' assessments of the size, masculinity, and attractiveness of unseen speakers. *Evolution and Human Behavior*, *33*, 509–519. doi:10.1016/j.evolhumbehav.2012.01.004

Radeau, M., & Bertelson, P. (1987). Auditory-visual interaction and the timing of inputs. Thomas (1941) revisited. *Psychological Research*, *49*, 17–22.

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, *17*, 405–409. doi:10.1111/j.1467-8721.2008.00615.x

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., & Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 7295–7300. doi:10.1073/pnas.0609419104

Smith, H. M. J., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016). Concordant cues in faces and voices: Testing the backup signal hypothesis. *Evolutionary Psychology*, *14*, 1474704916630317. doi:10.1177/1474704916630317

Spence, C. (2013). Just how important is spatial coincidence to multisensory integration? Evaluating the spatial rule. *Annals of the New York Academy of Sciences*, *1296*, 31–49.

Stein, B. E., Meredith, M. A., & Wolf, S. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, *9*, 255–266. doi:10.1038/nrn2331

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America*, *26*, 212–215. doi:10.1121/1.1907309

Van der Burg, E., Cass, J., Olivers, C. N., Theeuwes, J., & Alais, D. (2010). Efficient visual search from synchronized auditory signals requires transient audiovisual events. *PLoS One*, *5*, e10664. doi:10.1371/journal.pone.0010664

Vatakis, A., Ghazanfar, A. A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision*, *8*, 14–14.

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Attention, Perception, & Psychophysics*, *69*, 744–756.

Vatakis, A., & Spence, C. (2008). Evaluating the influence of the 'unity assumption' on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica*, *127*, 12–23. doi:10.1016/j.actpsy.2006.12.002

Walker, S., Bruce, V., & O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics*, *57*, 1124–1133. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8539088

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research*, *158*, 252–258. doi:10.1007/s00221-004-1899-9

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, *88*, 638–667. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7003641

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011, May). *The aligned rank transform for nonparametric factorial analyses using only anova procedures*. Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems, Vancouver, BC, Canada.