Title

Visibility of speech articulation enhances auditory phonetic convergence

Journal

Attention, Perception, & Psychophysics, 78(1)

ISSN

1943-3921

Authors

Dias, James W

Rosenblum, Lawrence D

Publication Date

2016

DOI

10.3758/s13414-015-0982-6

Peer reviewed

HHS Public Access

Author manuscript

Atten Percept Psychophys. Author manuscript; available in PMC 2021 April 29.

Published in final edited form as:

Atten Percept Psychophys. 2016 January ; 78(1): 317–333. doi:10.3758/s13414-015-0982-6.

Visibility of speech articulation enhances auditory phonetic convergence

James W. Dias[1], **Lawrence D. Rosenblum**[1]

[1]Department of Psychology, University of California, Riverside, 900 University Avenue, Riverside, CA 92521, USA

Abstract

Talkers automatically imitate aspects of perceived speech, a phenomenon known as *phonetic convergence.* Talkers have previously been found to converge to auditory and visual speech information. Furthermore, talkers converge more to the speech of a conversational partner who is seen and heard, relative to one who is just heard (Dias & Rosenblum *Perception*, *40*, 1457–1466, 2011). A question raised by this finding is what visual information facilitates the enhancement effect. In the following experiments, we investigated the possible contributions of visible speech articulation to visual enhancement of phonetic convergence within the noninteractive context of a shadowing task. In Experiment 1, we examined the influence of the visibility of a talker on phonetic convergence when shadowing auditory speech either in the clear or in low-level auditory noise. The results suggest that visual speech can compensate for convergence that is reduced by auditory noise masking. Experiment 2 further established the visibility of articulatory mouth movements as being important to the visual enhancement of phonetic convergence. Furthermore, the word frequency and phonological neighborhood density characteristics of the words shadowed were found to significantly predict phonetic convergence in both experiments. Consistent with previous findings (e.g., Goldinger *Psychological Review*, *105*, 251–279, 1998), phonetic convergence was greater when shadowing low-frequency words. Convergence was also found to be greater for low-density words, contrasting with previous predictions of the effect of phonological neighborhood density on auditory phonetic convergence (e.g., Pardo, Jordan, Mallari, Scanlon, & Lewandowski *Journal of Memory and Language*, *69*, 183–195, 2013). Implications of the results for a gestural account of phonetic convergence are discussed.

Keywords

Auditory noise; Audiovisual speech; Phonetic convergence; Phonological neighborhood density; Speech alignment; Speech articulation; Word frequency

Human perceivers exhibit a pervasive nonconscious inclination to spontaneously imitate the subtle nuances of articulated speech produced by conversational partners. This *phonetic convergence* (also known as *phonetic accommodation*, *speech accommodation*, and *speech alignment*), has been found to manifest in convergence along acoustical speech characteristics, including speech rate (e.g., Street, 1984), vocal intensity (e.g., Natale, 1975),

✉ Lawrence D. Rosenblum, rosenblu@citrus.ucr.edu.

and vowel spectra (e.g., Pardo, Gibbons, Suppes, & Krauss, 2012). Although phonetic convergence typically occurs without intent, the degree to which perceivers converge to the speech of conversational partners can be influenced by the social dynamics of a conversational interaction (e.g., Pardo, 2006; Pardo, Jay, & Krauss, 2010).

However, phonetic convergence occurs not only within the interactive context of a live conversation between individuals, but also when individuals shadow prerecorded speech tokens (e.g., Babel, 2012; Goldinger, 1998; R. M. Miller, Sanchez, & Rosenblum, 2010; Nielsen, 2011; Shockley, Sabadini, & Fowler, 2004). Within the *shadowing paradigm,* perceivers are presented with the speech of a prerecorded model. Immediately following presentation of the speech stimulus, perceivers verbally say aloud the speech perceived. Despite the lack of a conversational context, shadowers typically show phonetic convergence, in that their shadowing utterances sound more like the model's than their preshadowing utterances (according to the judgments of naïve raters).

Using the shadowing paradigm, research has demonstrated that perceivers can converge to speech that is not only heard, but is also *lipread.* All of this research was conducted on normal-hearing participants with no formal lipreading experience. For example, R. M. Miller et al. (2010) had perceivers shadow tokens that were presented both auditory-only and visual-only (showing an articulating face). Naïve perceptual judges rated the shadowed utterances of *heard* speech as sounding more like the auditory utterances of the model. Moreover, these same judges also rated the shadowed utterances of *lipread* speech as sounding more like the *auditory* utterances of the model. In addition, raters were able to make crossmodal assessments of phonetic convergence when rating the similarity of perceivers' shadowed auditory utterances to the visible articulations of the model.

Research has also revealed that visible speech information can modulate convergence to featural characteristics of auditory speech. Sanchez, Miller, and Rosenblum (2010) found that the shadowed utterances of auditory /pa/ dubbed onto visible articulations of /pa/ are modulated by the visible rate of articulation. If the visible articulation is produced at a slow rate, then shadowed utterances of /pa/ are produced with longer voice onset times (VOTs). Conversely, if the visible articulation is produced at a fast rate, then shadowed utterances of/ pa/ are produced with shorter VOTs. The crossmodal influence of visible speech articulations on shadowed utterances of auditory stimuli suggest that the information to which perceivers converge may take an amodal form.

Phonetic convergence to both auditory *and* visual speech information is significant to understanding the mechanisms underlying speech processing. Some have argued that phonetic convergence may serve as evidence for a link between the perception and production mechanisms of speech, which may be suggestive of a *common currency* between the processes of speech perception and speech production (for a review, see Fowler, 2004). The crossmodal nature of the findings of R. M. Miller et al. (2010) and Sanchez et al. (2010) suggest that the information to which perceivers converge (and the information that raters use to assess convergence) is available across sensory modalities, perhaps taking the form of amodal articulatory gestures. As such, this articulatory information would serve as the

common currency between the processes of speech perception and speech production (Fowler, 2004; R. M. Miller et al., 2010; Shockley et al., 2004).

The results of R. M. Miller et al. (2010) and Sanchez et al. (2010) are consistent with evidence from the speech perception literature for the multimodal nature of speech perception (for a review, see Rosenblum, 2008). It has long been understood that visible speech information can change the perception of auditory speech (e.g., McGurk & MacDonald, 1976). Perceivers can also demonstrate an enhanced ability to identify degraded auditory speech when presented with the concurrent visual speech component of the auditory stimulus. This *visual enhancement* of auditory speech perception has been demonstrated when identifying speech presented in noise (e.g., Erber, 1975; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954) and acoustically transformed speech (e.g., Remez, Fellowes, Pisoni, Goh, & Rubin, 1998). The visibility of a speaker can also improve the comprehension of accented speech (e.g., Sueyoshi & Hardison, 2005) and speech that conveys complicated content (e.g., Arnold & Hill, 2001; Reisberg, McLean, & Goldfield, 1987).

As a result of this capacity for visual speech information to enhance the perception of auditory speech, the question arises of whether such visual speech enhancement can occur for *convergence* to auditory speech. In fact, we have previously demonstrated that perceivers will converge more to the auditory speech of their conversational partner if they can *see* their partner during an interaction (Dias & Rosenblum, 2011). To demonstrate this effect, pairs of participants worked together to complete an interactive puzzle task. Perceivers in different groups interacted either with full view of their conversational partner or with the visibility of their partner occluded by speaker-grill cloth (acoustically permeable but visibly impermeable). Naive judges then rated the similarity of preinteraction speech utterances to speech utterances made during or after the interaction within an AXB rating paradigm (e.g., Goldinger, 1998). The speech utterances of conversational partners produced during or after interaction were rated as being more similar for groups who interacted while their partner was visible.

Though the results of Dias and Rosenblum (2011) suggest that the visibility of a speaker can enhance convergence to auditory speech, the question of why such an effect occurs remains. It could be that visibility of the speech articulations of a conversational partner enhances convergence to auditory speech, which would be consistent with the previously discussed evidence for visible speech enhancement of auditory speech *perception*. However, it is also possible that the visibility of a live conversational partner could modulate convergence by changing the social dynamics of the conversational interaction (e.g., Gregory, Green, Carrothers, Dagan, & Webster, 2001) or by making available visible socially salient information (e.g., Babel, 2009), both of which are known to influence convergence.

In the following experiments, we examined the basis of visual enhancement of phonetic convergence (Dias & Rosenblum, 2011). In order to more specifically assess the possible contribution of visible *speech* information over interactive conversational influences to enhance auditory phonetic convergence, we chose to use a shadowing paradigm. As we described above, the shadowing paradigm provides a noninteractive context to induce

convergence; shadowers simply listen and/or watch a recorded talker articulate a word and are instructed to say the perceived word out loud. Thus, shadowers do not interact with an interlocutor and perform their task in isolation. If the visual enhancement of convergence observed by Dias and Rosenblum (2011) was dependent on the interactive context of a conversation between individuals, then a visual enhancement would not be expected for participants in a shadowing context. If, on the other hand, the observed visual enhancement of convergence is dependent, at least in part, on having access to visible *speech* information, then visual enhancement of convergence would be expected for speech shadowers.

Besides providing a noninteractive context to test convergence, the shadowing paradigm allows for greater experimental control over the manipulation and presentation of speech stimuli. For this reason, our experiment also tested how other stimulus dimensions bear on the audiovisual induction of phonetic convergence. For example, the first experiment also tested the influence of background noise on auditory and audiovisual convergence. To our knowledge, the influence of noise on convergence to auditory speech has not previously been explored. However, within the speech perception literature, noise has been used extensively as an investigative tool for degrading speech information and introducing variability within experimental designs (for a review, see Pisoni, 1996), allowing for more sensitive measurement of the subtler characteristics of speech. For example, auditory noise has been shown to modulate the identifiability of phonetic information (e.g., French & Steinberg, 1947; G. A. Miller, Heise, & Lichten, 1951; Sarampalis, Kalluri, Edwards, & Hafter, 2009) and can modulate the influence of speaker-familiarity effects on spoken word recognition (e.g., Mullennix, Pisoni, & Martin, 1989; Smith, 2007).

Auditory noise can also modulate the enhancing influence of visual speech on the perception of auditory speech, such that as auditory noise increases, perceivers rely more on available visible speech information (e.g., Erber, 1975; Ross et al., 2007; Sumby & Pollack, 1954). As we previously discussed, evidence already suggests that the information to which perceivers align is available in visual displays of speech (R. M. Miller et al., 2010; Sanchez et al., 2010). By introducing noise to the auditory modality, perceivers may be forced to rely more on the information available within the visual modality. The result would be a greater visual enhancement of phonetic convergence to auditory speech when shadowing speech presented in auditory noise. However, our goal in the present investigation was not to impede the identifiability of shadowed *words,* but to modulate the availability of the information to which perceivers may converge. As such, we chose to use a low level of auditory noise to minimally disrupt talker information without increasing errors in word identifiability.

Our experiments also tested word stimuli that systematically varied in both their word frequency and phonological neighborhood density characteristics, consistent with some previous convergence studies (e.g., Goldinger, 1998; Goldinger & Azuma, 2004; Pardo, Jordan, Mallari, Scanlon, & Lewandowski, 2013). Perceivers have previously been found to converge more to the spoken utterances of words with lower frequencies of occurrence within the English lexicon (Goldinger, 1998; Goldinger & Azuma, 2004). Goldinger (1998) ascribed word frequency effects in phonetic convergence to stored representations within an episodic lexicon: Words with a lower rate of occurrence within the lexicon have fewer representations in memory. As a result, when such a word is encountered, fewer traces are

activated in memory, allowing the most recent instance to have a greater influence over a perceiver's shadowed response. Conversely, words that occur more frequently within the lexicon have more representations in memory, resulting in more traces being activated and providing less influence for the most recently encountered instance of the word.

Goldinger's (1998) model accounts for the influence of word frequency on subsequent phonetic convergence. However, another important characteristic of word processing involves the number of other words within the lexicon that share similar phonetic structures. Each word has its own *phonological neighborhood density,* constituting the number of other words within the lexicon that differ in one phonetic characteristic from the target (e.g., Luce & Pisoni, 1998). It has been suggested that neighborhood density influences spoken word identifiability by acting as a metric of confusability between a target word and its phonological neighbors (Luce & Pisoni, 1998). As such, phonological neighbors act as phonological competitors for recovery of the phonetic details relevant to the target word within a speech signal, and words with a higher neighborhood density have more phonological competitors influencing their identifiability. As a result, words with high neighborhood densities require greater effort to resolve their phonetic structure from within a speech signal (e.g., Luce & Pisoni, 1998).

Pardo and colleagues (2013) have recently investigated the influence of neighborhood density, in conjunction with word frequency, on phonetic convergence to shadowed auditory speech. Pardo et al. (2013) proposed that should the lexical characteristics of word frequency and neighborhood density influence phonetic convergence, it would reflect the perceptual effort required to resolve the phonetic details available in the speech signal. This effort is reflected by the effects of word frequency and phonological neighborhood density on spoken word intelligibility: The most difficult words to identify are those with low frequencies and high phonological neighborhood densities (e.g., Bradlow & Pisoni, 1999; Luce & Pisoni, 1998). Pardo et al. (2013) proposed that because low-frequency, high-density words require more effort to resolve their phonetic details, more information would be available for convergence. Pardo et al. also proposed that because talkers tend to hyperarticulate the vowels of words that are harder to identify—which has been found to be particularly true for words with high neighborhood densities (e.g., Munson & Solomon, 2004; Scarborough, 2013)—the greater distinctiveness of the articulatory information could itself lead to greater phonetic convergence. As such, Pardo et al. (2013) predicted that the shadowed utterances of low-frequency, high-neighborhood-density words would show more phonetic convergence than the shadowed utterances of high-frequency, low-neighborhood-density words. However, Pardo et al. (2013) failed to find any influence of word frequency or neighborhood density on phonetic convergence.

Since such discrepant findings have emerged regarding the influence of word frequency on phonetic convergence, and since only one study thus far has investigated whether neighborhood density plays a role, we included an evaluation of the influences of these lexical characteristics within the present investigation of phonetic convergence. Consistent with Goldinger's (1998) account for an episodic lexicon discussed above, we predict that perceivers will demonstrate greater phonetic convergence when shadowing low-frequency words (see also Pierrehumbert, 2002). However, we propose that the influence of

phonological neighborhood density on phonetic convergence may *not,* as Pardo et al. (2013) predicted, depend on the perceptual effort required to resolve the phonetic detail of a spoken word from its phonological neighbors.

Many current models of lexical influences on speech perception and production incorporate mechanisms for the encoding of idiosyncratic elements (e.g., talker idiolect) at a sublexical phonological (e.g., phoneme, feature) level (e.g., Cutler, Eisner, McQueen, & Norris, 2010; Johnson, 1997; Pierrehumbert, 2002). Some have suggested (e.g., Pierrehumbert, 2002) that idiosyncratic elements associated with talkers are incorporated at a phonological level within an episodic lexicon that incorporates phonological information *prior* to lexical (i.e., word) information. This idiosyncratic information encoded at the phonological level can influence the production and perception of words that share those specific phonological elements across the lexicon (e.g., Nielsen, 2011; Pierrehumbert, 2002). Thus, general production of a word within a denser phonological neighborhood would be influenced by greater amounts of idiosyncratic information encoded across the neighbors (Pierrehumbert, 2002). As a result, for words in denser phonological neighborhoods, the idiosyncratic information available within a recently perceived utterance would, itself, have a diluted influence on subsequent production of that same word. We predict that, if the idiosyncratic information encoded at the phonological level influences speech production as discussed above, then phonetic convergence resulting from shadowing high-density words will be less than phonetic convergence resulting from shadowing low-density words.

With regard to the suggestion made by Pardo et al. (2013) that hyperarticulation associated with word frequency and phonological neighborhood density may increase phonetic convergence to specific talkers, another interpretation exists. There is evidence that hyperarticulation may affect speech in *consistent* ways across talkers (i.e., less talker-specific). Low-frequency words are typically found to be articulated more slowly (e.g., Wright 1979), high-density words are typically produced with greater coarticulation (e.g., Scarborough, 2003, 2013), and both low-frequency and high-density words are typically found to be articulated with more expanded vowels (e.g., Munson & Solomon, 2004). These patterns of hyperarticulation are thought to make productions of difficult words more distinct and identifiable (e.g., Munson & Solomon, 2004; Scarborough, 2013). However, these lexically dependent patterns of articulation appear to transcend talker idiolect and have been exhibited by diverse groups of talkers across multiple studies (e.g., Munson & Solomon, 2004; Scarborough, 2003, 2013; Wright, 1979). Thus, *all* productions of low-frequency and high-density words may be influenced in systematic ways by hyperarticulation similarly across talkers. This may suggest that any convergence to a talker's idiolect would occur irrespective of lexically dependent hyperarticulations.

In sum, Pardo et al.'s (2013) hypothesis and our present hypothesis predict that phonetic convergence will be greater when a participant shadows words that occur less often within the lexicon (e.g., Goldinger, 1998). However, our predictions diverge from those of Pardo with regard to the influence of neighborhood density on phonetic convergence: Pardo et al. (2013) predict greater convergence to high-density words, whereas we predict greater convergence to low-density words.

In the following experiments, female participants were asked to shadow a female talker. The selection of this gender-matched design was based on a number of considerations. First, the design is consistent with our original investigation of the visual enhancement of auditory phonetic convergence (Dias & Rosenblum, 2011), thereby allowing for a more straightforward comparison of results. Second, it allowed us to avoid other potential, nonlinguistic influences on the degree of convergence. The convergence literature has shown complicated interactions between model and shadower gender, as well as social-relationship factors. Where some studies have reported significantly greater phonetic convergence among female participants (e.g., R. M. Miller et al., 2010; Namy, Nygaard, & Sauerteig, 2002), others have reported greater phonetic convergence among male participants (e.g., Pardo, 2006; Pardo et al., 2014), and still others have reported no gender differences at all (e.g., Pardo et al., 2013). Even within each of the studies above, researchers have reported a great deal of variability in convergence among participants, depending on gender consistency between the model and the shadower, as well as the social purpose established between the interlocutors (e.g., R. M. Miller et al., 2010; Namy et al., 2002; Pardo et al., 2013). Because the goal of the present investigation was not to adjudicate the individual differences in phonetic convergence behavior, we employed a sample of female participants shadowing the spoken utterances of a single female model, on the basis of our own success with this method in prior work (e.g., Dias & Rosenblum, 2011; Sanchez et al., 2010; see also Delvaux & Soquet, 2007). We will address possible limitations in the generalizability of the results imposed by the sample in the General Discussion.

In Experiment 1, we investigated whether the visibility of a speaker can enhance phonetic convergence to auditory speech. Participants shadowed the utterances of a prerecorded model that were presented both auditory-alone and audiovisually. In addition, half of these shadowers shadowed utterances that were presented in low-level auditory noise. The results suggest that the visibility of a speaker can enhance alignment to shadowed auditory speech, but only when low-level noise reduces overall auditory phonetic convergence.

In Experiment 2, we subsequently investigated the basis of this visual-enhancement effect by manipulating the visibility of the articulating mouth. All shadowers again shadowed auditory-alone and audiovisual speech utterances. However, the nature of the visual component of the audiovisual stimuli was manipulated across three groups: One group shadowed audiovisual stimuli composed of a still image of a nonarticulating face; one shadowed audiovisual stimuli composed of an articulating face with visibility of the mouth occluded by Gaussian blurring; and one shadowed audiovisual stimuli composed of a fully visible articulating face (as in Exp. 1).

AXB ratings were used to evaluate phonetic convergence. Perceptual measures, such as the AXB rating paradigm, are widely used for the assessment of phonetic convergence (e.g., Dias & Rosenblum, 2011; Goldinger, 1998; Namy et al., 2002; Pardo, 2006) and serve as a perceptually valid measure of similarity across speech utterances among naive raters. Furthermore, unlike acoustical measures of convergence, ratings of similarity avoid the complex task of determining to which acoustical characteristics participants converge—dimensions that may vary depending on the model, shadower, utterance, and rater (e.g., Goldinger, 1998; Pardo, 2006). In fact, recent evidence has suggested that although

acoustical measures can provide a metric for phonetic convergence, many commonly used acoustic measures are inconsistent with, and poorly related to, perceptual measures, which seem to provide more consistent, reliable metrics (Pardo et al., 2012; Pardo et al., 2013).

Experiment 1

In Experiment 1, we attempted to replicate the visual enhancement of phonetic convergence observed between conversational partners (Dias & Rosenblum, 2011) within the controlled experimental conditions of a shadowing paradigm. Within the paradigm, participants were tasked with shadowing the spoken-word utterances of a prerecorded model that were presented both auditory-alone and auditorily with visibility of the speaker's articulating face. The words shadowed were varied by word frequency and phonological neighborhood density. Half of the participants shadowed words presented in noise, whereas the other half shadowed words presented without noise. Phonetic convergence was assessed by naive raters using the previously discussed AXB rating paradigm.

Method

Phase I: Convergence elicitation

Participants: A total of 32 female undergraduate students from the University of California, Riverside, Human Subjects Pool participated as shadowers in the shadowing task, consistent with the sample used by Dias and Rosenblum (2011). All shadowers were native speakers of American English with normal hearing and normal or corrected-to-normal sight.

Stimuli: A female model (age = 34 years, resident of Southern California) was audio–video recorded uttering 120 bisyllabic words. These recordings were digitized at 30 frames per second at a size of 640 × 480 pixels and a sample rate of 44 kHz, 16 bit. The speaker was visible in the videos from the points of her shoulders to the top of the head.

Words were collected from the Irvine Phonotactic Online Dictionary (www.iphod.com; Vaden, Halpin, & Hickok, 2009), counterbalanced for lexical characteristics along the dimensions of word frequency and neighborhood density. In all, 30 of the words were high-frequency–high-density, 30 were high-frequency–low-density, 30 were low-frequency–high-density, and 30 were low-frequency–low-density (see the Appendix). We conducted an item analysis to ensure that the levels of word frequency and phonological neighborhood density were not confounded across the word groups.[1]

[1]A multivariate analysis of variance testing the frequency and density word groups (high vs. low), based on the dependent variables frequency and phonological neighborhood density (values from the *Irvine Phonotactic Online Dictionary)*, revealed main effects of both the frequency, $F(2, 115) = 10.567$, $p < .001$, $\eta_p^2 = .155$, and density, $F(2, 115) = 5.650$, $p < .005$, $\eta_p^2 = .089$, word groups, but not an interaction, $F(2, 115) = 1.357$, $p = .262$, $\eta_p^2 = .023$. Frequency differences were found between the high-frequency ($M = 139.954$, $SE = 29.965$) and low-frequency ($M = 0.829$, $SE = 0.101$) word groups, $F(1, 116) = 21.336$, $p < .001$, $\eta_p^2 = .155$, but no density differences were found between these word groups ($M = 7.420$, $SE = 0.841$, vs. $M = 6.850$, $SE = 0.781$, respectively), $F(1, 116) = 2.414$, $p = .123$, $\eta_p^2 = .020$. Likewise, density differences were found between the high-density ($M = 13.050$, $SE = 0.351$) and low-density ($M = 1.220$, $SE = 0.101$) word groups, $F(1, 116) = 1{,}052.780$, $p < .001$, $\eta_p^2 = .901$, but no frequency differences ($M = 60.916$, $SE = 22.116$, vs. $M = 79.867$, $SE = 23.870$, respectively), $F(1, 116) = 0.396$, $p = .530$, $\eta_p^2 = .003$. There were no interactions between the frequency and density word groups with respect to frequency, $F(1, 116) = 0.397$, $p = .530$, $\eta_p^2 = .003$, or density, $F(1, 116) = 0.134$, $p = .715$, $\eta_p^2 = .001$.

Procedure: Shadower baseline tokens were first recorded from read-aloud utterances of the 120 words. For each trial, one of the 120 words was presented on the screen in text form. When the word appeared, shadowers were instructed to quickly and accurately read the word aloud.

Following the baseline recordings, shadowers were recorded shadowing the 120 single-word utterances of the prerecorded model. Each trial consisted of the presentation of a single stimulus, following which shadowers would quickly and accurately say aloud the word they heard the model say (e.g., Goldinger, 1998). For audioalone (AO) trials, a target (small cross, presented for 1 s) prompted the shadowers to focus attention on the screen before hearing the single-word utterance of the prerecorded model. For audiovisual (AV) trials, the target (presented for 1 s) prompted shadowers to focus attention on the screen (17-in. Apple iMac). When the AV stimulus began, the target was replaced by the video of the model speaking the auditorily presented utterance. For half of the shadowers, tokens were presented in white noise at a signal-to-noise ratio (SNR) of +10 dB (over Sony MDR-V600 headphones at a comfortable listening level). On the basis of informal pilot tests, this SNR was chosen for having the potential to mask some talker-specific information while having a negligible impact on the identifiability of the shadowed tokens (e.g., Erber, 1969; Ross et al., 2007). We did not want identification errors made in the shadowed responses to influence later judgments of similarity between the shadower and model utterances. Though half of the participants were presented the tokens in noise, the shadowed utterances were made in the clear for both the noise and no-noise groups.

Presentation of the AO and AV stimuli was block-ordered between groups: Half of the shadowers shadowed AO tokens prior to the AV tokens, and the other half of the shadowers shadowed AV tokens prior to the AO tokens. The 120 word utterances were randomly assigned to the AO and AV shadowing conditions, controlling for frequency and density characteristics: Each lexical group (high-frequency–high-density, high-frequency–low-density, low-frequency–high-density, and low-frequency–low-density) was randomly and evenly split between the AO and AV shadowing conditions.

In total, there were four presentation groups, each with eight shadowers: (1) AO followed by AV shadowing, without noise; (2) AO followed by AV shadowing, with noise; (3) AV followed by AO shadowing, without noise; and (4) AV followed by AO shadowing, with noise.

All verbal responses were digitally recorded with a Shure SM57 microphone and Amadeus II software (Hairer, 2007). The experimental procedure was executed using PsyScope software (Cohen, MacWhinney, Flatt, & Provost, 1993).

Phase II: Convergence assessment

Participants: A total of 160 undergraduate students (100 female, 60 male) from the University of California, Riverside, Human Subjects Pool served as naïve raters of phonetic convergence. All of the raters had normal hearing and normal or corrected-to-normal sight.

Stimuli: The shadowers' (from Phase I) audio-recorded utterances were digitally extracted at 44 kHz, 16 bit. Because shadowers heard the words and noise over headphones, their recorded utterances were not heard against noise by the raters. The isolated single-word baseline and shadowed utterances served as the stimuli for rating similarity by the naive raters. Editing of the stimuli was accomplished using Final Cut Pro 5 for Mac OS X.

Procedure: Convergence was assessed using an AXB rating paradigm (e.g., Goldinger, 1998; R. M. Miller et al., 2010; Namy et al., 2002; Pardo et al., 2013; Shockley et al., 2004): Raters were tasked with judging whether a participant's baseline or shadowed utterance sounded more like the utterance of the model presented in the shadowing task. For any one trial, ratings were based on the similarity of the utterances of only one specific word. For example, a rater was presented with a shadower's baseline utterance of "battle" (A), the model's utterance of "battle" (X), and the same shadower's shadowed utterance of "battle" (B). Following presentation of these three token utterances, raters were asked, "Which utterance, A or B, is pronounced more like utterance X?" The proportion of rater responses indicating a shadower's shadowed utterance as being more similar to the model's utterance served as a metric of phonetic convergence. Each rater assessed the phonetic convergence of only one of the shadowers. Assessments were made for all of a shadower's words twice, counterbalancing the order of A and B, for a total of 240 AXB trials per rater. The 160 raters were split so that each of the 32 shadowers was assessed by five raters.

Results and discussion

Mixed-effects binomial/logistic regression models were employed to assess the influences of acoustic noise, visible speech information, word frequency, and phonological neighborhood density on raters' assessments of speech similarity between a shadower and the model. The rater responses, judging a shadower's baseline or shadowed utterance as sounding more like the utterance of the shadowed model, served as the binomial dependent variable (e.g., Pardo et al., 2013).

First, a control model was constructed, including shadower, rater, and word as random effects. This model yielded a significant intercept, indicating that the rate at which raters identified shadowed responses as sounding more like the shadowed model ($M = .687$) was significantly greater than chance (.5), $\beta_0 = 0.925$, $SE = 0.108$, $Z = 8.554$, $p < .001$. This suggests that shadowers did converge to the speech of the shadowed model. Shadowing-modality block order (whether the shadower shadowed AO first or AV first), rater sex, and rating target (whether, on a given AXB trial, the shadowed utterance was A or B) were not found to be significant fixed-effects parameters, nor did their inclusion in the control model allow it to fit the data better. As such, these parameters were excluded from the following analyses.

Visual enhancement and noise effects on phonetic convergence—The control model was improved by adding shadowing modality (AO, AV) as a fixed effect, $\chi^2(1) = 7.027$, $p = .008$. Shadowers were found to converge more when shadowing audiovisual tokens ($M = .693$) than when shadowing audioalone tokens ($M = .680$), $\beta = 0.062$, $SE = 0.024$, $Z = 2.655$, $p = .008$.

To determine whether visual influences on phonetic convergence to auditory speech are modulated by acoustic noise, noise condition (noise, no noise) and the interaction of shadowing modality and noise were added to the model as fixed effects. The resulting modality–noise model provided a better fit than the previous model that had only included shadowing modality as a fixed effect, $\chi^2(2) = 18.665$, $p < .001$. This improved model reveals that shadowers converged *marginally* less when shadowing speech presented in acoustic noise ($M = .670$) than when shadowing speech not presented in noise ($M = .703$), $\beta = -0.321$, $SE = 0.191$, $Z = -1.684$, $p = .092$. The reduced convergence exhibited when shadowing speech presented in auditory noise is consistent with findings in which speech identifiability has been reduced when identifying speech presented in auditory noise (e.g., G. A. Miller et al., 1951; Song, Skoe, Banai, & Kraus, 2011). The results are also consistent with findings illustrating reduced talker facilitation of speech perception when identifying speech in noise (e.g., Mullennix et al., 1989). However, the noise level used in the present experiment (+10-dB SNR) was not sufficient to influence the identifiability of shadowed speech tokens: Participants identified shadowed tokens at ceiling—98 % or better for all noise and modality conditions. We will address possible explanations for the influence of auditory noise on phonetic convergence in the General Discussion.

Adding the modality–noise interaction term pushed the main effect of shadowing modality to nonsignificance, $\beta = -0.039$, $SE = 0.034$, $Z = -1.151$, $p = .250$. However, the interaction term itself was significant, $\beta = 0.196$, $SE = 0.047$, $Z = 4.174$, $p < .001$. We explored the interaction of shadowing modality and noise by constructing separate models for the noise and no-noise groups that included shadowing modality as a fixed effect along with the same random effects from the control model (shadower, rater, and word). For the no-noise group, no significant difference in ratings of convergence was apparent when shadowing AV ($M = .699$) versus AO ($M = .708$) tokens, $\beta = -0.034$, $SE = 0.034$, $Z = -0.975$, $p = .329$. However, when shadowing speech in noise, a significant difference in ratings of convergence did appear when shadowing AV ($M = .688$) versus AO ($M = .652$) tokens, $\beta = 0.157$, $SE = 0.033$, $Z = 4.736$, $p < .001$. Shadowers converged more to auditory speech presented in noise when they could see the articulating face of the speaker (see Fig. 1). In general, these results suggest that convergence to auditory speech can be enhanced by visual information in a noninteractive context.

Lexical influences—Word frequency and phonological neighborhood density were added as fixed effects to the control model. This lexical model provided a better fit to the data than the random-effects model, $\chi^2(2) = 12.963$, $p < .002$. However, the frequency–density interaction term did not improve model fit, $\chi^2(1) = 0.197$, $p = .657$, and was excluded.

The lexical model revealed ratings of convergence to be greater for low-frequency ($M = .707$) than for high-frequency ($M = .666$) words, $\beta = 0.202$, $SE = 0.093$, $Z = 2.175$, $p < .05$. This effect of word frequency on rated convergence is consistent with the past literature (e.g., Goldinger, 1998; Goldinger & Azuma, 2004; but see Pardo et al., 2013). The lexical model also revealed ratings of convergence to be greater for low-density ($M = .710$) than for high-density ($M = .664$) words, $\beta = 0.280$, $SE = 0.093$, $Z = 3.019$, $p < .003$. This effect suggests that the phonetic similarity of words stored within the mental lexicon can

influence phonetic convergence. Specifically, our results suggest that perceivers converge more to words with fewer phonological neighbors.

Experiment 1 revealed several compelling influences on phonetic convergence. We found that phonetic convergence is greater when shadowing low-frequency words, replicating previous findings (Goldinger, 1998; Goldinger & Azuma, 2004; but see Pardo et al., 2013). However, we also found that perceivers converge more to words with smaller phonological neighborhoods. We also found that phonetic convergence is reduced when shadowing words that are presented in background noise. However, the noise used in Experiment 1 did not interfere with the identifiability of the shadowed words. The observed influences of neighborhood density and auditory noise on phonetic convergence have not previously been demonstrated. The implications of all of these effects will be discussed in more detail in the General Discussion.

Experiment 1 also revealed that the visibility of a speaker's face can enhance phonetic convergence to auditory speech when shadowing speech that is presented in noise. These results suggest that the visual enhancement of phonetic convergence observed during a conversational interactive context (Dias & Rosenblum, 2011) can be achieved within a nonconversational context. This might mean that the important visual information for enhancing phonetic convergence takes the form of visible articulatory information, as has been shown for the enhancement of speech perception (e.g., Rosenblum, 2005, 2008).

However, although the shadowing paradigm does not involve interpersonal interaction, socially relevant information may still influence phonetic convergence. Preserved within a visible face is information relating to gender, age, ethnicity, attractiveness, and emotion, which can all provide socially relevant information, some of which has been found to influence phonetic convergence (e.g., Babel, 2009, 2012). It is quite possible that simply seeing this socially relevant information in the face of a speaker can change the degree to which perceivers converge, irrespective of the visible speech information. This issue will be addressed in Experiment 2.

Experiment 2

The purpose of Experiment 2 was to evaluate the specific contribution of visible speech-relevant information available in the face to enhancing phonetic convergence to shadowed auditory speech. For this purpose, the visibility of the speech information was manipulated while the visibility of the socially relevant information in the speaker's face was maintained.

Within the speech perception literature, extensive evidence suggests that visibility of the dynamic movements of the mouth is important for lipreading (e.g., Greenberg & Bode, 1968; IJsseldijk, 1992; Jackson, Montgomery, & Binnie, 1976) and audiovisual speech perception (e.g., Rosenblum, Johnson, & Saldana, 1996; Thomas & Jordan, 2004). In fact, when identifying audiovisual speech presented in auditory noise, perceivers will spend more time gazing toward the mouth than toward other areas of the visible face (e.g., Pare, Richler, ten Hove, & Munhall, 2003; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998).

It should be noted that extra-oral information in the face has also been found to provide linguistically relevant information for the identification of words (e.g., Davis & Kim, 2006; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). Explanations of these findings often attribute the results to expressions of prosody, via head and eyebrow movements, which cue the recognition of relevant associative word productions. In other words, it is likely that the influence of extra-oral information in the face to enhanced identification of words is largely attributable to cues not related to phonetic or idiolectic speech information (e.g., Munhall et al., 2004). However, it has also been suggested that, due to their association with articulations of the mouth, extra-oral movements may directly provide linguistically relevant information, though not as effectively as visible articulations of the mouth itself (Thomas & Jordan, 2004).

In Experiment 2, we attempted to explore the importance of speech-salient articulatory information in the visible mouth, in the context of a fully visible face, to enhancing auditory phonetic convergence. To examine this question, we retested the audiovisual stimuli of Experiment 1, along with two new types of stimuli, which each removed articulatory mouth movements. Removing the visibility of the articulating mouth was accomplished by (1) occluding the visible mouth articulations within a dynamic face and (2) presenting still facial images (mouth visible but nonarticulating). Although articulatory information was deleted from these stimuli, both manipulations allowed us to preserve the socially relevant information visible in the speaker's face (e.g., gender, ethnicity, and attractiveness; Babel, 2012). The first type of stimulus (dynamic face with mouth occluded) maintained the extra-oral movements that might be useful for conveying socially relevant information, and the second type (still face) maintained any socially relevant information available in the structure of the (still) mouth.

Since Experiment 1 revealed that the salience of visible information to enhancing auditory phonetic convergence was only found for speech presented in auditory noise, all of the shadowers in Experiment 2 shadowed speech presented in noise. If the visibility of the articulatory speech gestures of the mouth is important to the visual enhancement of phonetic convergence to auditory speech found in Experiment 1, then perceivers who shadowed audiovisual speech that included visibility of the articulating mouth should demonstrate greater visual enhancement than would perceivers who shadowed audiovisual speech that did not provide visible articulations of the mouth (i.e., the mouth-occluded and static-face conditions).

The same spoken-word shadowing stimuli used in Experiment 1 were again used in Experiment 2. As such, the effects of word frequency and neighborhood density on phonetic convergence were again tested for the purposes of replicating their observed effects in Experiment 1. However, we chose to present audio and audiovisual tokens randomly across trials in Experiment 2, forgoing the blocked design employed in Experiment 1. Blocked and randomized designs have both been employed in the audiovisual speech literature, and we wanted to examine whether the effects observed in the blocked design of Experiment 1 would replicate when audio and audiovisual presentations were randomized in Experiment 2.

Method

Phase I: Convergence elicitation

Participants: A total of 24 female undergraduate students from the University of California, Riverside, Human Subjects Pool participated in the shadowing task. As with Experiment 1, all of the participants were native speakers of American English with normal hearing and normal or corrected-to-normal sight.

Stimuli: The audio and audiovisual stimuli constructed for Experiment 1 were again used for Experiment 2. However, the visual component of the audiovisual stimuli was manipulated to produce two more types of audiovisual conditions.

The first manipulation involved the application of a Gaussian blur over the mouth of the dynamically articulating face for each of the original audiovisual stimuli. The blur was applied using Final Cut Pro X for Mac OS at a maximum intensity (radius = 100 pixels). For each stimulus, the blur was applied to the circular area over the mouth: between the left and right ears and between the bridge of the noise and the bottom of the chin (see Fig. 2). Gaussian blurring has previously been found to effectively occlude the visibility of facial articulations, which can cause perceivers to rely more on available auditory information when identifying audiovisual speech stimuli (Thomas & Jordan, 2002).

The second manipulation involved pairing auditory tokens with static facial images of the model. To achieve this, the visual component of each original audiovisual stimulus was replaced with a single video frame taken from the original video component of each stimulus. As a result, each static-face audiovisual stimulus contained a different static facial image taken from the original video component of the respective audiovisual stimulus. The frames were selected from each video to depict the face of the model in a nonarticulating state. We did not want articulatory information to provide information in any still image regarding the auditory speech component of the complete audiovisual stimuli (e.g., Campbell, 1996; Irwin, Whalen, & Fowler, 2006). The use of these static-image audiovisual stimuli ensured the availability of socially relevant information, such as gender, age, and attractiveness, within the context of a fully visible face. However, the static images provided no dynamic articulatory information, either oral or extra-oral.

Including these two new manipulations, three different types of audiovisual stimuli, differing in visual information, were used in Experiment 2: audiovisual with full view of the dynamic face (AF), audiovisual with visibility of the mouth occluded (blurred) in the dynamic face (AB), and audiovisual with a static image of a nonarticulating face (AS; see Fig. 2).

Procedure: The procedure was the same in Experiment 2 as in Experiment 1. However, instead of block-ordering the presentation of shadowed audio-only and audiovisual tokens, presentation modality was randomized across the 120 shadowing trials. As in Experiment 1, the 120 word utterances were randomly assigned to the AO and AV shadowing conditions, controlling for frequency and density characteristics. Each lexical group (high-frequency–high-density, high-frequency–low-density, low-frequency–high-density, and low-frequency–low-density) was evenly split between the AO and AV shadowing conditions. Since visual

enhancement of convergence to auditory speech was only observed in the noise condition of Experiment 1, all speech tokens presented in Experiment 2 were presented in noise (+10-dB SNR).

Groups were divided on the basis of the type of audiovisual stimuli they shadowed. As a result, there were three shadowing groups: (1) shadowing audio-only and AF audiovisual tokens; (2) shadowing audio-only and AB audiovisual tokens; and (3) shadowing audio-only and AS audiovisual tokens.

As with Experiment 1, all verbal responses were digitally recorded with a Shure SM57 microphone and Amadeus II software (Hairer, 2007), and the experimental procedure was executed using PsyScope software (Cohen et al., 1993).

Phase II: Convergence assessment

Participants: A total of 120 undergraduate students (61 female, 59 male) from the University of California, Riverside, Human Subjects Pool served as naive raters of phonetic convergence. As in Experiment 1, all of the raters had normal hearing and normal or corrected-to-normal sight.

Stimuli: The participants from Phase I's audio-recorded utterances were digitally extracted following the same procedure used in Experiment 1. The resulting single-word baseline and shadowed utterances served as the stimuli for comparisons of similarity.

Procedure: Phonetic convergence was assessed using the same AXB rating paradigm as in Experiment 1.

Results and discussion

Prior to analyzing the ratings of phonetic convergence, we again found the shadowing-word response accuracy to be at ceiling (above 98 %) for the audiovisual and audio-only shadowing conditions across all experimental groups. Thus, participants were again nearly perfect in identifying the shadowing words, despite the background noise.

Mixed-effects binomial/logistic regression models were again employed for assessing the influence of visible speech information, word frequency, and phonological neighborhood density on the raters' assessments of similarity between a shadower and the model. As in Experiment 1, the rater responses (whether or not a shadower's shadowed utterance was judged as sounding more like the utterance of the shadowed model) served as the binomial dependent variable (e.g., Pardo et al., 2013).

Visual enhancement—Control models were constructed for each of the Experiment 2 shadowing groups (AF, AB, and AS). Similar to the Experiment 1 control model, the Experiment 2 control models included shadower, rater, and word as random effects. Each model yielded a significant intercept: Raters judged shadowed utterances as sounding more like the shadowed model for the AF ($M = .693$, $\beta_0 = 0.981$, $SE = 0.219$, $Z = 4.490$, $p < .001$), AB ($M = .619$, $\beta_0 = 0.549$, $SE = 0.129$, $Z = 4.260$, $p < .001$), and AS ($M = .622$, β_0

= 0.553, SE = 0.092, Z = 5.992, p < .001) groups, suggesting that shadowers converged to the speech of the shadowed model.

Shadowing modality (AO, AV) was added to the control model for each shadowing group. For the AF group, adding modality provided a better model fit to the data, $\chi^2(1) = 4.170$, p = .041: Ratings of phonetic convergence were greater for AV (M = .708) than for AO (M = .679) shadowed utterances, β = 0.102, SE = 0.050, Z = 2.049, p = .041, replicating the visual enhancement of convergence to auditory speech presented in noise observed in Experiment 1. However, for the AB group, adding modality did not provide a better model fit, $\chi^2(1) = 1.509$, p = .219: Ratings of phonetic convergence did not differ between the AV (M = .625) and AO (M = .613) shadowed utterances, β = 0.057, SE = 0.046, Z = 1.230, p = .219. Adding modality did not provide a better model fit for the AS group, either, $\chi^2(1) = 0.639$, p = .424: Ratings of phonetic convergence did not differ between the AV (M = .625) and AO (M = .619) shadowed utterances, β = −0.037, SE = 0.046, Z = −0.800, p = .424. These models suggest that visibility of the articulating mouth (AF group) is necessary for visual enhancement of phonetic convergence (see Fig. 3).

The results suggest that, at least within a shadowing context, the visible articulations of the mouth can account for visual enhancement of phonetic convergence over other (e.g., social) information available in the visible face of a speaker. Furthermore, these results are consistent with evidence from the speech perception literature suggesting that visible articulations of the mouth are important to audiovisual speech perception (e.g., Rosenblum et al., 1996; Thomas & Jordan, 2004; Vatikiotis-Bateson et al., 1998).

The obvious differences in phonetic convergence for *auditory-alone* shadowed tokens between the three groups (see Fig. 3) compelled us to construct a model evaluating the effect of group on ratings of AO phonetic convergence. Shadower, rater, and word again served as random effects within the model, and shadowing-group condition (AF, AB, and AS) was added to the model as a fixed effect. Relative to the AB group, ratings of phonetic convergence for AO shadowed tokens did not increase for the AS group, β = 0.053, SE = 0.204, Z = 0.261, p = .794. However, as compared to the AS group, ratings of phonetic convergence were marginally greater for the AF group, β = 0.384, SE = 0.205, Z = 1.869, p = .062. These results may suggest that when shadowing audiovisual tokens that provide salient visible articulatory information for enhanced phonetic convergence, the enhancement affect transfers to instances when shadowing auditory-alone tokens. The implications of this finding will be discussed in more detail in the General Discussion.

Lexical influences—The same lexical model from Experiment 1 was constructed for Experiment 2, including ratings of phonetic convergence from all three shadowing groups. Shadowing group itself was included as a control parameter to account for possible differences between the groups resulting from modality influences (discussed above). As such, the Experiment 2 lexical model included shadower, rater, and word as random effects, and word frequency (high/low), phonological neighborhood density (high/low), and shadowing group (AF, AB, AS) as fixed effects. The effect of shadowing group proved to be a significant parameter: Relative to the AB group (M = .619), ratings of phonetic convergence (across both modalities) did not increase for the AS group (M = .622), β =

0.023, $SE = 0.203$, $Z = 0.112$, $p = .911$. However, ratings of phonetic convergence for the AF group ($M = .693$) were greater than ratings of phonetic convergence for the AS group, $\beta = 0.426$, $SE = 0.203$, $Z = 2.099$, $p = .036$.

Consistent with the results of Experiment 1, word frequency and phonological neighborhood density were again found to be significant parameters. Ratings of convergence were greater for low-frequency ($M = .664$) than for high-frequency ($M = .626$) words, $\beta = 0.192$, $SE = 0.092$, $Z = 2.083$, $p = .037$, and ratings of convergence were also greater for low-density ($M = .681$) than for high-density ($M = .608$) words, $\beta = 0.364$, $SE = 0.092$, $Z = 3.957$, $p < .001$. Replication of these effects in Experiment 2 further substantiates the influences of word frequency and phonological neighborhood density on phonetic convergence.

General discussion

The aim of the present investigation was to evaluate the contribution of visible speech information to the enhancement of auditory phonetic convergence. Though we previously demonstrated a visual enhancement of convergence to auditory speech between conversational partners (Dias & Rosenblum, 2011), the contributions of the conversational-context and visible-speech factors were unclear. To eliminate the potential influence of conversational context on visually enhanced phonetic convergence, we attempted to replicate the visual enhancement of convergence within the noninteractive context of a shadowing paradigm. The results suggest that the visibility of a talker's face can enhance convergence to auditory speech, but only when participants are shadowing speech presented in noise. Experiment 2 revealed that the visibility of the articulating mouth is necessary for the visual enhancement of phonetic convergence to auditory speech presented in noise. In addition, the results showed that when auditory speech is presented in noise, phonetic convergence is reduced, as well as some interesting influences of word frequency and neighborhood density on the degree of phonetic convergence. Each of these issues will be addressed in the following sections.

Visible articulation enhances phonetic convergence

Similar to our previous finding that the visibility of a speaker can enhance convergence to auditory speech during a conversational interaction (Dias & Rosenblum, 2011), convergence to shadowed auditory speech was enhanced in the present experiment by the availability of visual information, at least when the auditory speech was embedded in low-level noise. Moreover, it seems that the increased phonetic convergence was based on the visibility of articulatory information. The manipulations of Experiment 2 were designed to preserve the visible socially salient information in the face while varying the visibility of the articulatory information provided by the speaker's mouth. These results suggest that the visibility of the dynamic articulations of the mouth, in the context of a full face, is most salient to the observed enhancement effects. In fact, neither visibility of a static face (with visibility of the mouth) nor visibility of a dynamic face with the mouth occluded produced any significant visual enhancement to auditory phonetic convergence. These results are consistent with evidence from the speech perception literature illustrating the importance of the visibility of

mouth articulations to audiovisual speech perception (e.g., Rosenblum et al., 1996; Thomas & Jordan, 2004; Vatikiotis-Bateson et al., 1998).

A surprising finding accompanied the increased convergence induced by visible speech. Although the visibility of the articulating mouth provided enhanced convergence over auditory-alone stimuli, we observed a carryover effect, so that auditory-alone tokens that were intermixed with the visible articulations showed greater convergence than did auditory-alone tokens intermixed with the control audiovisual conditions. Thus, shadowers who shadowed audiovisual tokens that allowed visibility of the articulating mouth converged more to auditory-only tokens than did shadowers who were presented audiovisual tokens without visible mouth articulations. This carryover effect may suggest that the talker-specific articulatory information available in the visible mouth can transfer to enhance phonetic convergence to that talker's auditory-only speech. This interpretation would be consistent with previous evidence from our lab illustrating that the learning involved in talker facilitation effects can transfer between modalities (Rosenblum et al., 2007; Sanchez et al., 2013). If visual influences can carry over to affect phonetic convergence to auditory-only stimuli, then we would expect to see such influences in the Experiment 1 noise group that shadowed audiovisual speech prior to shadowing auditory-only speech. However, a reevaluation of the Experiment 1 data revealed no difference in ratings of phonetic convergence for auditory-only speech between the noise group that shadowed AO prior to AV tokens ($M = .675$) and the noise group that shadowed AV prior to AO tokens ($M = .629$), $\beta = -0.239$, $SE = 0.212$, $Z = -1.124$, $p = .261$, suggesting that no carryover effects were observed in Experiment 1. Finding carryover effects only in Experiment 2 may be a consequence of randomizing AO and AV trials, as opposed to the Experiment 1 blocked ordering of AO and AV trials.

Although visible speech enhancement of phonetic convergence was observed in both experiments, it must be acknowledged that convergence occurred only when some background noise was present in the auditory signal. We can only speculate as to why background noise was required in order to observe visual enhancement of convergence. Recall that in our previous report (Dias & Rosenblum, 2011), visual enhancement of phonetic convergence was found between conversational partners interacting in an environment *without* an auditory noise manipulation, as such. However, it is possible that some ambient noise inherent to the conversational environment (acoustically typical lab rooms) used in our prior study created noise conditions similar to those in the present experiments. Participants in the present investigation shadowed speech presented over headphones while seated in a sound-attenuated booth. These environmental controls practically eliminated the potential influences of ambient sounds during the experimental procedure, so that when noise was added, it may have been functionally comparable to the background noise conditions of our prior study (Dias & Rosenblum, 2011).

It is also possible that the social or attentional nuances inherent to the interactive task of Dias and Rosenblum (2011) may have provided enough variability in the perceptual salience of the auditory speech to observe enhancing effects of talker visibility. For example, attention paid to the interactive puzzle task, and not the speech of the conversational partner, may have modulated phonetic convergence. The visibility of the conversational partner may

have drawn more attention to, or changed the dynamics of, the conversational interaction, as well as possibly providing redundant speech information, subsequently enhancing auditory phonetic convergence (for details, see Dias & Rosenblum, 2011). This explanation would suggest that within the context of a shadowing task, the visibility of a speaker's articulators provides visible speech information that is important to enhancing phonetic convergence, but when interacting within a live conversational context, the visibility of a speaker could also change the dynamics of the conversational interaction in a way that enhances phonetic convergence.

Regardless, the results of Experiments 1 and 2 illustrate the salience of visible articulations to phonetic convergence. The multimodal nature of talker-specific speech information for convergence is consistent with other findings suggesting that both speech and speaker information can take an amodal, gestural form (for reviews, see Fowler, 2004; Rosenblum, 2008).

Low-level auditory noise suppresses phonetic convergence

Besides interacting with modality, we observed that adding auditory noise had an overall effect of reducing phonetic convergence (Exp. 1). To the authors' knowledge, this is the first time that an effect of speech in noise on phonetic convergence has been reported.

It may not be surprising that auditory noise would reduce speech convergence, since it is known to reduce performance with phonetic identification. However, it should be noted that the low level of noise (+10-dB SNR) used in our experiments was *not* sufficient to reduce the identifiability of the shadowed words (Exp. 1). This could mean that phonetic convergence is more fragile to noise than is phonetic identification. Why might this be the case? One possibility is that different information and processes are used for convergence and speech identification, and that they are differentially influenced by noise. During phonetic convergence, perceivers may converge to the idiolectic characteristics (speaker-specific articulatory style) of the perceived speaker. In fact, noise has previously been found to reduce idiolectic influences on spoken word recognition (Mullennix et al., 1989; Smith, 2007). Thus, noise could inhibit phonetic convergence by masking the idiolectic information available in a phonetic signal, while sparing the phonetic information needed to identify a spoken word. As a result, less idiolectic information is perceptible to which perceivers can converge.

Another possibility is that despite the lack of a reduction in phonetic identification, noise may still have impacted the information and processes used in the recovery of phonetic information (e.g., French & Steinberg, 1947; G. A. Miller et al., 1951; Sumby & Pollack, 1954). A low level of noise may have induced a reallocation of resources to the recovery of phonetic information for word identification, thereby leaving fewer resources for the processes underlying phonetic convergence.

More recent conceptualizations have questioned the separability of the information used for talker and phonetic recovery. It has been suggested that phonetic and idiolectic information may possess a common form, defined by the speaker-specific phonetically relevant articulatory information preserved in a speech signal (for reviews, see Pardo & Remez,

2006; Sheffert, Pisoni, Fellowes, & Remez, 2002). For example, Remez, Fellowes, and Rubin (1997) proposed that the isolated phonetic information preserved in sine-wave speech provides both phonetic and talker information. Remez et al. (1997) suggested that this information may be preserved within segment-level articulatory assimilations, which are considered to be idiosyncratic to a speaker's talker-specific articulatory style of speech. If phonetic and talker-specific information overlap within a speech signal, then noise may inhibit phonetic convergence by modulating the availability of both the phonetic and idiolectic information preserved in the speech signal. In the present experiments, the level of noise may have been great enough to degrade how this information was used for convergence, but too low to influence the recovery of this same information for purposes of phonetic identification.

Lexical characteristics influence phonetic convergence

Spoken word frequency and phonological neighborhood density were found to influence phonetic convergence to shadowed speech in both Experiments 1 and 2. We found that phonetic convergence was greater for low-frequency than for high-frequency words, consistent with previous findings (Goldinger, 1998; Goldinger & Azuma, 2004; but see Pardo et al., 2013). The reported effects are also consistent with the observed influences of word frequency on speech *identification.* When identifying spoken words, high-frequency words are generally found to be easier to identify than low-frequency words (e.g., Catlin, 1969; Nakatani, 1973; Rosenzweig & Postman, 1957; Savin, 1963). Within an exemplar model of memory (e.g., Goldinger, 1998), the results from the phonetic convergence and speech identification literatures complement nicely. High-frequency words have more representations in memory, making them easier to identify, but also resulting in subsequent productions reflecting the mean characteristics of more exemplars. On the other hand, low-frequency words have fewer representations in memory, making them harder to identify, and causing subsequent productions to reflect more closely the characteristics of a recently experienced exemplar.

With regard to our neighborhood density effects, we found that shadowers converged more to words with fewer phonological competitors in the mental lexicon. To our knowledge, no influences of neighborhood density on phonetic convergence have previously been reported. As we discussed, a recent attempt by Pardo et al. (2013) to evaluate the influences of lexical characteristics on phonetic convergence to shadowed speech failed to find any reliable influence of neighborhood density on convergence across acoustic and perceptual measures. However, Pardo et al. (2013) also failed to find any reliable influence of spoken word frequency on phonetic convergence, contrasting with the results of this study and of previously mentioned studies (Goldinger, 1998; Goldinger & Azuma, 2004). One methodological difference that could account for the differences between Pardo et al. (2013) and those investigations that have reported lexical influences (the present investigation; Goldinger, 1998; Goldinger & Azuma, 2004) could be that Pardo et al. (2013) used monosyllabic words, whereas others have used multisyllabic words (two or more syllables per word utterance). Multisyllabic words have been found to produce stronger lexical activations than monosyllabic words (e.g., Pitt & Samuel, 2006; Samuel, 1981, 1996;

Strauss & Magnuson, 2008), which could potentially account for the lack of lexical influences observed by Pardo et al. (2013).

Our observed effects of neighborhood density suggest that the influence of lexical factors (word frequency and neighborhood density) on phonetic convergence is not entirely determined by the perceptual effort required to identify the spoken word. Instead, phonetic convergence seems to be influenced by the amount of idiosyncratic information previously encoded at the lexical (word) and sublexical phonological (phoneme, feature) levels (e.g., Goldinger, 1998; Pierrehumbert, 2002). At the lexical level, words that occur more frequently within the lexicon have more idiosyncratic information (including talker idiolect) encoded within their many lexical episodes. This encoded idiosyncratic information dilutes convergence to the idiosyncratic information available within a recently experienced exemplar. As such, perceivers converge less to high-frequency words (e.g., Goldinger, 1998, Goldinger & Azuma, 2004). At the phonological level, the idiosyncratic information encoded within phonological episodes modulates production of that phonological element across the lexicon. The idiosyncratic information associated with phonological elements that are shared with more words (i.e., denser phonological neighborhoods) will dilute the influence of the idiosyncratic information available in a recently experienced exemplar. As a result, perceivers converge less to words with high phonological neighborhood densities.

The present set of experiments revealed an enhancement of phonetic convergence to shadowed speech that was facilitated by visible dynamic mouth articulation, consistent with previously reported visual enhancement effects of phonetic convergence within a live conversational setting (Dias & Rosenblum, 2011). The observed visual enhancement effects were observed only when shadowing speech presented in auditory noise, suggesting that visibility of speech articulation can compensate for the masking effect that noise has on the availability of the information to which perceivers phonetically converge. That visible speech articulation can compensate for the reduction in phonetic convergence when participants shadow speech in noise suggests that the information to which perceivers converge is available across sensory modalities, perhaps taking the form of gestural speech information.

We must point out some limitations imposed by using a sample of female participants shadowing the speech of a single female model. As we previously discussed, we chose to use this design in part to be consistent with the gender-matched design of our original investigation of the visual enhancement of auditory phonetic convergence (Dias & Rosenblum, 2011). However, as we discussed in the introduction, contrasting findings pertaining to gender differences in phonetic convergence (e.g., R. M. Miller et al., 2010; Namy et al., 2002; Pardo, 2006; Pardo et al., 2014; Pardo et al., 2013) and the variability in phonetic convergence among individual perceivers (e.g., R. M. Miller et al., 2010; Namy et al., 2002; Pardo et al., 2013) were also determining factors in choosing the present design. It was not our intent in the present investigation to address the issue of individual differences in phonetic convergence. Instead, our goal was to examine whether visual speech influences on auditory phonetic convergences (previously observed between female conversational partners; Dias & Rosenblum, 2011) can be explained, at least in part, by visible speech information.

However, we cannot discount the possibility that our observed visual, lexical, and phonetic convergence effects are the result of idiosyncratic characteristics associated with our female model. Nor can we discount the possibility that these effects would be different for male shadowers. However, there is plenty of precedence to suggest that our observed effects would generalize to other diverse samples of models and shadowers of both genders. Within the speech literature, robust visual (for reviews, see Fowler, 2004; Rosenblum, 2008) and lexical (for a review, see Luce & McLennan, 2005) influences on auditory speech perception are reported across diverse samples of models and participants, with many studies employing stimuli derived from a single model talker (e.g., Brancazio, 2004; Erber, 1971; Luce, Pisoni, & Goldinger, 1990; Remez et al., 1998; Rosenblum et al., 1996; Ross et al., 2007; Tye-Murray, Sommers, & Spehar, 2007). Within the phonetic convergence literature (much of which is cited in this article), automatic imitation of perceived speech has been observed across studies employing diverse samples of models, shadowers, and conversational partners. In fact, many of these studies also employed test stimuli derived from a single model talker (e.g., Babel, 2010; Babel & Bulatov, 2012; Honorof, Weihing, & Fowler, 2011; Nielsen, 2011; Sanchez et al., 2010) and measured samples composed only of male (e.g., Pardo et al., 2012) or female (e.g., Delvaux & Soquet, 2007; Sanchez et al., 2010) participants. Still, future work should investigate whether the visual enhancement effects on auditory phonetic convergence observed here will generalize across different samples of talkers and perceivers.

Acknowledgments

Author note This research was supported by NIDCD Grant Number 1R01DC008957–01.

Appendix

Table 1

List of words used for the present study

		Frequency (occurrence per million)			
		LOW ($M_{frequency}$ = 0.829)		High ($M_{frequency}$ = 139.954)	
Density (number of words differnting in 1 phonetic unit from target)	High ($M_{density}$ = 13.050)	babble	Huddle	battle	keeping
		barren	Lemon	boring	leaving
		basin	Mettle	buried	looking
		baton	Muddle	button	mister
		beret	Mussel	calling	picking
		callous	Peddle	cases	pieces
		choral	Peril	common	reason
		fable	rabble	dealing	riding
		feted	ripple	double	saving
		fickle	Rubble	faces	selling
		Hassel	Ruffle	feeling	settle
		hasten	Sated	getting	table
		heron	Sickle	hated	waited
		herring	Tattle	hearing	willing
		huckle	Whittle	hitting	written
	LOW ($M_{density}$ = 1.220)	ballade	Motif	became	marriage
		benign	Natal	because	message
		bovine	Nugget	before	minute
		cajole	Pecan	began	moving
		chiffon	Pipette	cannot	naked
		coupon	Rotate	Chinese	nowhere
		debit	Rural	damage	police
		facade	Sedate	female	purpose
		feline	Sonar	figure	research
		genome	Tepid	given	social
		golem	Typhoon	hotel	toilet
		humid	Vapid	knowledge	tonight

Atten Percept Psychophys. Author manuscript; available in PMC 2021 April 29.

lapel	Vegan	ladies	weapon
lipid	Woven	learned	within
methane	Zealot	magic	without

All words were disyllabic and were acquired from the Irvine Phonotactic Online Dictionary.

References

Arnold P, & Hill F (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. British Journal of Psychology, 92, 339–355.

Babel M (2009). Phonetic and social selectivity in phonetic accommodation (PhD dissertation). University of California, Berkeley, CA.

Babel M (2010). Dialect divergence and convergence in New Zealand English. Language in Society, 39, 437–456.

Babel M (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. Journal of Phonetics, 40, 177–189. doi: 10.1016/j.wocn.2011.09.001

Babel M, & Bulatov D (2012). The role of fundamental frequency in phonetic accommodation. Language and Speech, 55, 231–248. [PubMed: 22783633]

Bradlow AR, & Pisoni DB (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. Journal of the Acoustical Society of America, 106, 2074–2085.

Brancazio L (2004). Lexical influences in audiovisual speech perception. Journal of Experimental Psychology: Human Perception and Performance, 30, 445–463. [PubMed: 15161378]

Campbell R (1996). Dissociating face processing skills: Decisions about lip read speech, expression, and identity. Quarterly Journal of Experimental Psychology, 49A, 295–314. doi:10.1080/713755619

Catlin J (1969). On the word-frequency effect. Psychological Review, 76, 504–506.

Cohen J, MacWhinney B, Flatt M, & Provost J (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. Behavior Research Methods, Instruments, & Computers, 25, 257–271. doi:10.3758/BF03204507

Cutler A, Eisner F, McQueen JM, & Norris D (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. Laboratory Phonology, 10, 91–111.

Davis C, & Kim J (2006). Audio–visual speech perception off the top of the head. Cognition, 100, B21–B31. doi:10.1016/j.cognition.2005.09.002 [PubMed: 16289070]

Delvaux V, & Soquet A (2007). The influences of ambient speech on adult speech productions through unintentional imitation. Phonetica, 64, 145–173. [PubMed: 17914281]

Dias JW, & Rosenblum LD (2011). Visual influences on interactive speech alignment. Perception, 40, 1457–1466. [PubMed: 22474764]

Erber NP (1969). Interaction of audition and vision in the recognition of oral speech stimuli. Journal of Speech and Hearing Research, 12, 423–425. [PubMed: 5808871]

Erber NP (1971). Auditory and audiovisual reception of words in low-frequency noise by children with normal hearing and by children with impaired hearing. Journal of Speech and Hearing Research, 14, 496–512. [PubMed: 5163883]

Erber NP (1975). Auditory–visual perception of speech. Journal of Speech and Hearing Disorders, 40, 481–492.

Fowler CA (2004). Speech as a supramodal or amodal phenomenon. In Calvert GA, Spence C, & Stein BE (Eds.), The handbook of multisensory processing (pp. 189–202). Cambridge, MA: MIT Press.

French NR, & Steinberg JC (1947). Factors governing the intelligibility of speech sounds. Journal of the Acoustical Society of America, 19, 90–119.

Goldinger SD (1998). Echoes of echoes? An episodic theory of lexical access. Psychological Review, 105, 251–279. doi:10.1037/0033-295X.105.2.251 [PubMed: 9577239]

Goldinger SD, & Azuma T (2004). Episodic memory reflected in printed word naming. Psychonomic Bulletin & Review, 11, 716–722. doi:10.3758/BF03196625 [PubMed: 15581123]

Greenberg HJ, & Bode DL (1968). Visual discrimination of consonants. Journal of Speech and Hearing Research, 11, 869–874. [PubMed: 5719244]

Gregory SWJ, Green BE, Carrothers RM, Dagan KA, & Webster SW (2001). Verifying the primacy of voice fundamental frequency in social status accommodation. Language & Communication, 21, 37–60.

Hairer M (2007). Amadeus II (Version 3.8.7). Kenilworth, UK: HairerSoft. Retrieved from www.hairersoft.com/Amadeus.html

Honorof DN, Weihing J, & Fowler CA (2011). Articulatory events are imitated under rapid shadowing. Journal of Phonetics, 39, 18–38. [PubMed: 23418398]

IJsseldijk FJ (1992). Speechreading performance under different conditions of video image, repetition, and speech rate. Journal of Speech and Hearing Research, 35, 466–471. [PubMed: 1573883]

Irwin JR, Whalen DH, & Fowler CA (2006). A sex difference in visual influence on heard speech. Perception & Psychophysics, 68, 582–592. doi:10.3758/BF03208760 [PubMed: 16933423]

Jackson PL, Montgomery AA, & Binnie CA (1976). Perceptual dimensions underlying vowel lipreading performance. Journal of Speech and Hearing Research, 19, 796–812. [PubMed: 1003957]

Johnson K (1997). Speech perception without speaker normalization: An examplar model. In Johnson K & Mullennix JW (Eds.), Talker variability in speech processing (pp. 145–166). San Diego, CA: Academic Press.

Luce PA, & McLennan CT (2005). Spoken word recognition: The challenge of variation. In Pisoni D & Remez R (Eds.), The handbook of speech processing (pp. 591–609). Malden, MA: Blackwell.

Luce PA, & Pisoni DB (1998). Recognizing spoken words: The neighborhood activation model. Ear and Hearing, 19, 1–36. [PubMed: 9504270]

Luce PA, Pisoni DB, & Goldinger SD (1990). Similarity neighborhoods of spoken words. In Altmann G (Ed.), Cognitive models of speech processing: Psycholinguistics and computation perspectives (pp. 122–147). Cambridge, MA: MIT Press.

McGurk H, & MacDonald J (1976). Hearing lips and seeing voices. Nature, 264, 746–748. doi:10.1038/264746a0 [PubMed: 1012311]

Miller GA, Heise GA, & Lichten W (1951). The intelligibility of speech as a factor of the context of the test materials. Journal of Experimental Psychology, 41, 329–335. [PubMed: 14861384]

Miller RM, Sanchez K, & Rosenblum LD (2010). Alignment to visual speech information. Attention, Perception, & Psychophysics, 72, 1614–1625. doi:10.3758/APP.72.6.1614

Mullennix JW, Pisoni DB, & Martin CS (1989). Some effects of talker variability on spoken word recognition. Journal of the Acoustical Society of America, 85, 365–378.

Munhall KG, Jones JA, Callan DE, Kuratate T, & Vatikiotis-Bateson E (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. Psychological Science, 15, 133–137. [PubMed: 14738521]

Munson B, & Solomon NP (2004). The effect of phonological neighborhood density on vowel articulation. Journal of Speech, Language, and Hearing Research, 47, 1048–1058.

Nakatani LH (1973). On the evaluation of models for the word-frequency effect. Psychological Review, 80, 195–202.

Namy LL, Nygaard LC, & Sauerteig D (2002). Gender differences in vocal accommodation: The role of perception. Journal of Language and Social Psychology, 21, 422–432.

Natale M (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. Journal of Personality and Social Psychology, 32, 790–804.

Nielsen K (2011). Specificity and abstractness of VOT imitation. Journal of Phonetics, 39, 132–142.

Pardo JS (2006). On phonetic convergence during conversational interaction. Journal of the Acoustical Society of America, 119, 2382–2393.

Pardo JS, Gash H, Urmanche A, Decker A, Francis K, Wiener J, & Parker S (2014). Effects of talker sex on phonetic convergence to shadowed speech. Journal of the Acoustical Society of America, 135, 2420.

Pardo JS, Gibbons R, Suppes A, & Krauss RM (2012). Phonetic convergence in college roommates. Journal of Phonetics, 40, 190–197.

Pardo JS, Jay IC, & Krauss RM (2010). Conversational role influences speech imitation. Attention, Perception, & Psychophysics, 72, 2254–2264. doi:10.3758/BF03196699

Pardo JS, Jordan K, Mallari R, Scanlon C, & Lewandowski E (2013). Phonetic convergence in shadowing speech: The relation between acoustic and perceptual measures. Journal of Memory and Language, 69, 183–195.

Pardo JS, & Remez RE (2006). The perception of speech. In Traxler M & Gernsbacher MA (Eds.), The handbook of psycholinguistics (2nd ed., pp. 201–248). New York, NY: Academic Press.

Paré M, Richler RC, ten Hove M, & Munhall KG (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. Perception & Psychophysics, 65, 553–567. doi:10.3758/BF03194582 [PubMed: 12812278]

Pierrehumbert JB (2002). Word-specific phonetics. Laboratory Phonology, 7, 101–139.

Pisoni DB (1996). Word identification in noise. Language and Cognitive Processes, 11, 681–688. doi:10.1080/016909696387097 [PubMed: 21687807]

Pitt MA, & Samuel AG (2006). Word length and lexical activation: Longer is better. Journal of Experimental Psychology: Human Perception and Performance, 32, 1120–1135. doi:10.1037/0096-1523.32.5.1120 [PubMed: 17002526]

Reisberg D, McLean J, & Goldfield A (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In Dodd B & Campbell R (Eds.), Hearing by eye: The psychology of lip-reading (pp. 97–113). Hillsdale, NJ: Erlbaum.

Remez RE, Fellowes JM, Pisoni DB, Goh WD, & Rubin PE (1998). Multimodal perceptual organization of speech: Evidence from tone analogs of spoken utterances. Speech Communication, 26, 65–73. [PubMed: 21423823]

Remez RE, Fellowes JM, & Rubin PE (1997). Talker identification based on phonetic information. Journal of Experimental Psychology: Human Perception and Performance, 23, 651–666. doi: 10.1037/0096-1523.23.3.651 [PubMed: 9180039]

Rosenblum LD (2005). Primacy of multimodal speech perception. In Pisoni D & Remez R (Eds.), Handbook of speech perception (pp. 51–78). Malden, MA: Blackwell.

Rosenblum LD (2008). Speech perception as a multimodal phenomenon. Current Directions in Psychological Science, 17, 405–409. [PubMed: 23914077]

Rosenblum LD, Johnson JA, & Saldaña HM (1996). Point-light facial displays enhance comprehension of speech in noise. Journal of Speech and Hearing Research, 39, 1159–1170. [PubMed: 8959601]

Rosenblum LD, Miller RM, & Sanchez K (2007). Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects. Psychological Science, 18, 392–396. doi:10.1111/j.1467-9280.2007.01911.x [PubMed: 17576277]

Rosenzweig MR, & Postman L (1957). Intelligibility as a function of frequency of usage. Journal of Experimental Psychology, 54, 412–422. [PubMed: 13491767]

Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, & Foxe JJ (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cerebral Cortex, 17, 1147–1153. [PubMed: 16785256]

Samuel AG (1981). Phonemic Restoration: Insights from a new methodology. Journal of Experimental Psychology: General, 110, 474–494. doi: 10.1037/0096-3445.110.4.474 [PubMed: 6459403]

Samuel AG (1996). Does lexical information influence the perceptual restoration of phonemes? Journal of Experimental Psychology: General, 125, 28–51.

Sanchez K, Dias JW, & Rosenblum LD (2013). Experience with a talker can transfer across modalities to facilitate lipreading. Attention, Perception, & Psychophysics, 75, 1359–1365. doi:10.3758/s13414-013-0534-x

Sanchez K, Miller RM, & Rosenblum LD (2010). Visual influences on alignment to voice onset time. Journal of Speech, Language, and Hearing Research, 53, 262–272.

Sarampalis A, Kalluri S, Edwards B, & Hafter E (2009). Objective measures of listening effort: Effects of background noise and noise reduction. Journal of Speech, Language, and Hearing Research, 52, 1230–1240.

Savin HB (1963). Word-frequency effect and errors in the perception of speech. Journal of the Acoustical Society of America, 35, 200–206.

Scarborough RA (2003). Lexical confusability and degree of coarticulation. In Proceedings of the Twenty-Ninth Annual Meeting of the Berkeley Linguistics Society: General session and parasession on phonetic sources of phonological patterns. Synchronic–diachronic explanations *(pp.* 367–378*).* Berkeley, CA: Berkeley Linguistics Society.

Scarborough R (2013). Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation. Journal of Phonetics, 41, 491–508. doi:10.1016/j.wocn.2013.09.004

Sheffert SM, Pisoni DB, Fellowes JM, & Remez RE (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. Journal of Experimental Psychology: Human Perception and Performance, 28, 1447–1469. doi:10.1037/0096-1523.28.6.1447 [PubMed: 12542137]

Shockley K, Sabadini L, & Fowler CA (2004). Imitation in shadowing words. Perception & Psychophysics, 66, 422–429. doi: 10.3758/BF03194890 [PubMed: 15283067]

Smith R (2007, 8). The effect of talker familiarity on word segmentation in noise. Paper presented at the Meeting of the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany.

Song J-H, Skoe E, Banai K, & Kraus N (2011). Perception of speech in noise: Neural correlates. Journal of Cognitive Neuroscience, 23, 2268–2279. doi:10.1162/jocn.2010.21556 [PubMed: 20681749]

Strauss T, & Magnuson JS (2008). Beyond monosyllables: Word length and spoken word recognition. In Love BC, McRae K, & Sloutsky VM (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society (pp. 1306–1311). Austin, TX: Cognitive Science Society.

Street RLJ (1984). Speech convergence and speech evaluation in factfinding interviews. Human Communication Research, 11, 139–169.

Sueyoshi A, & Hardison DM (2005). The role of gestures and facial cues in second language listening comprehension. Language Learning, 55, 661–699.

Sumby WH, & Pollack I (1954). Visual contribution of speech intelligibility in noise. Journal of the Acoustical Society of America, 26, 212–215.

Thomas SM, & Jordan TR (2002). Determining the influence of Gaussian blurring on inversion effects with talking faces. Perception & Psychophysics, 64, 932–944. [PubMed: 12269300]

Thomas SM, & Jordan TR (2004). Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. Journal of Experimental Psychology: Human Perception and Performance, 30, 873–888. doi:10.1037/0096-1523.30.5.873 [PubMed: 15462626]

Tye-Murray N, Sommers M, & Spehar B (2007). Auditory and visual lexical neighborhoods in audiovisual speech perception. Trends in Amplification, 11, 233–241. doi:10.1177/1084713807307409 [PubMed: 18003867]

Vaden KI, Halpin HR, & Hickok GS (2009). Irvine Phonotactic Online Dictionary, Version 2.0. [Data file]. Available from www.iphod.com

Vatikiotis-Bateson E, Eigsti I-M, Yano S, & Munhall KG (1998). Eye movement of perceivers during audiovisual speech perception. Perception & Psychophysics, 60, 926–940. doi:10.3758/BF03211929 [PubMed: 9718953]

Wright CE (1979). Duration differences between rare and common words and their implications for the interpretation of word frequency effects. Memory & Cognition, 7, 411–419. doi:10.3758/BF03198257 [PubMed: 542114]

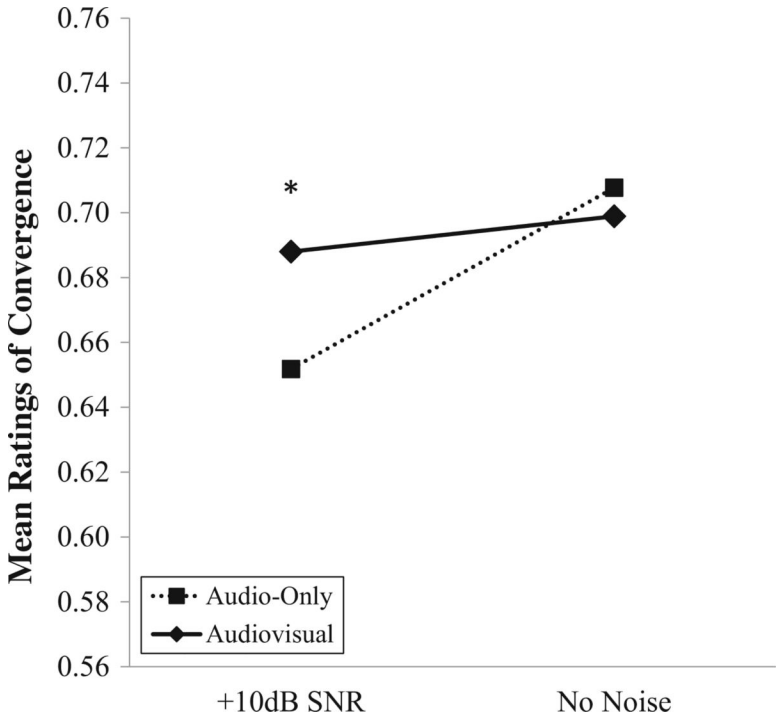

Fig. 1.

Experiment 1: Mean ratings of phonetic convergence for audio-only and audiovisual shadowed tokens between the noise (+10-dB signal-to-noise ratio [SNR]) and no-noise groups. The asterisk marks a significant difference

Fig. 2.

Examples of the visual component of the different audiovisual stimuli used in Experiment 2. AS: Static image of a nonarticulating face. AB: Dynamically articulating face with the visibility of the mouth occluded. AF: Fully visible dynamically articulating face

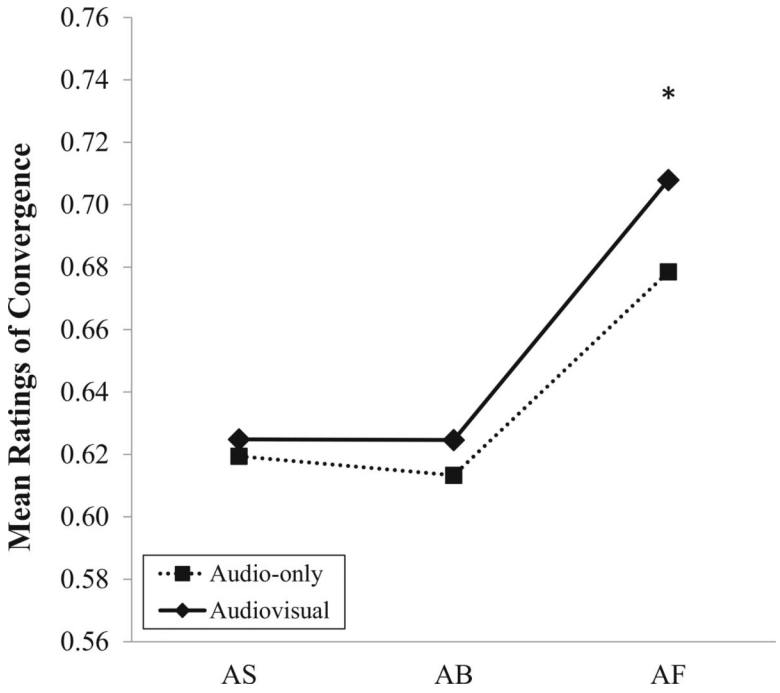

Fig. 3.

Experiment 2: Mean ratings of phonetic convergence for audio-only and audiovisual shadowed words for each of the experimental groups, differing in the types of audiovisual tokens shadowed. AS: Static image of a nonarticulating face. AB: Dynamically articulating face with the visibility of the mouth occluded. AF: Fully visible dynamically articulating face. The asterisk marks a significant difference.