Attention, Perception, & Psychophysics
2010, 72 (6), 1614-1625
doi:10.3758/APP.72.6.1614

Alignment to visual speech information

Rachel M. Miller, Kauyumari Sanchez, and Lawrence D. Rosenblum
University of California, Riverside, California

Speech alignment is the tendency for interlocutors to unconsciously imitate one another's speaking style. Alignment also occurs when a talker is asked to shadow recorded words (e.g., Shockley, Sabadini, & Fowler, 2004). In two experiments, we examined whether alignment could be induced with visual (lipread) speech and with auditory speech. In Experiment 1, we asked subjects to lipread and shadow out loud a model silently uttering words. The results indicate that shadowed utterances sounded more similar to the model's utterances than did subjects' nonshadowed read utterances. This suggests that speech alignment can be based on visual speech. In Experiment 2, we tested whether raters could perceive alignment across modalities. Raters were asked to judge the relative similarity between a model's visual (silent video) utterance and subjects' audio utterances. The subjects' shadowed utterances were again judged as more similar to the model's than were read utterances, suggesting that raters are sensitive to cross-modal similarity between aligned words.

Starting from infancy, humans show an amazing ability to imitate one another (see Meltzoff & Moore, 1997, for a review). As adults, we unconsciously imitate facial expressions, body posture, and mannerisms of a conversational partner in a social context (e.g., Chartrand & Bargh, 1999; Shockley, Santana, & Fowler, 2003). Chartrand and Bargh suggest that imitation is often passive and can occur without volition. They propose the *chameleon effect*, an unconscious tendency toward mimicking facial expressions, body posture, and mannerisms of another person. Although this imitation is typically unintentional, it can be influenced by multiple factors, including the social relationship between the conversational partners.

Imitation also occurs in speech communication. During conversational interaction, interlocutors subtly align to each other's speech rate, intonation, and vocal intensity (Giles, Coupland, & Coupland, 1991; Natale, 1975). This alignment is considered to have important linguistic and social functions, allowing interlocutors to be more effectively and efficiently understood (Giles et al., 1991; and see Chartrand & Bargh, 1999). But even outside of a social setting, talkers will imitate aspects of the speech of a recorded model producing individual words (Goldinger, 1998; Goldinger & Azuma, 2004; Namy, Nygaard, & Sauerteig, 2002; Pardo, 2006; Shockley, Sabadini, & Fowler 2004). Goldinger implemented a shadowing paradigm in which talkers uttered isolated words immediately after a recorded model. In the shadowing paradigm, talkers are asked to say the words they hear out loud quickly, but clearly. Talkers are never instructed to imitate what they hear. In the typical shadowing experiment, subjects first read a series of words off a computer monitor. These read words act as baseline stimuli for later perceptual ratings of alignment. The subjects then perform the shadowing task.

In order to assess imitation of a model's speech, the baseline (control) and shadowed words of each subject are typically compared with the model's words in an AXB perceptual matching task (Goldinger, 1998). The presence of imitation is indicated when raters choose the shadowed words as sounding more similar to the model's words than do the baseline words. The results of Goldinger's experiment indicated that immediate shadowing produced greater perceived imitation than delayed shadowing; that over the two conditions, low-frequency words were considered better imitations than high-frequency words; and that the strength of perceived imitation for raters increased with the number of repetitions that the shadower heard of the model's utterances. According to Goldinger, this evidence suggests that episodic traces of words that we hear are present and accessible in lexical memory. Alignment during shadowing emerges as a byproduct of how words are accessed from memory. Alignment also shows that perceivers are sensitive to a talker's articulatory style and unconsciously incorporate that style into their own speech productions. In this sense, speech alignment phenomena are consistent with other results showing that perceivers are sensitive to talker-specific phonetic information and use this talker information to facilitate later speech perception (for a review, see Nygaard, 2005).

In order to evaluate possible acoustic dimensions imitated during shadowing, Shockley et al. (2004) digitally extended voice onset time (VOT) durations in the initial consonants of a model's words before presentation to the shadowers. The results of an AXB rating task showed that the shadowers tacitly imitated the lengthened VOTs at better-than-chance levels. An acoustical analysis also revealed that the VOTs of the subjects' shadowed tokens were significantly longer than the VOTs of their baseline

L. D. Rosenblum, rosenblu@citrus.ucr.edu

tokens. The fact that the talkers show evidence of alignment to VOT in a shadowing task is important in providing evidence that phonetically relevant dimensions of speech are imitated (see also Pardo, 2004).

In summary, speech alignment occurs on both phonetic and extraphonetic levels, with conversational interaction or without. As with other types of chameleon effects, speech alignment, although typically unconscious, can be influenced by outside social factors (e.g., Namy et al., 2002; Pardo, 2006). One missing piece of alignment research is the determination of whether auditory speech is the only type of speech information that can induce alignment. The next section discusses whether *visual* speech information may also have this ability.

Visual Speech Information for Talker-Specific Characteristics

It is well known that visual speech information plays a vital role in face-to-face communication (see Rosenblum, 2005, for a review). When the auditory signal is degraded either by hearing loss or by a noisy environment, individuals are aided by seeing the articulating face of a talker (e.g., Grant & Seitz, 2000; Sumby & Pollack, 1954). Even when the auditory signal is clear, visual speech information can help perceivers recover a complicated message or understand messages spoken with a heavy foreign accent (e.g., Arnold & Hill, 2001; Reisberg, McLean, & Goldfield, 1987). Visual speech information also facilitates language acquisition in infants (e.g., Mills, 1987). In fact, blind infants show a delay in acquiring certain phonetic distinctions that are acoustically similar but visually distinct (e.g., /m/ vs. /n/). Visual speech information can facilitate second language perception and learning as well (Davis & Kim, 2001; Navarra & Soto-Faraco, 2007).

Notably, visual speech also influences the perception of heard syllables when discrepant auditory and visual syllables are presented synchronously (i.e., the McGurk effect; McGurk & MacDonald, 1976). The automatic and ubiquitous nature of audiovisual speech perception has led some theorists to argue that the primary mode of speech perception is multimodal, typically relying on both auditory and visual input. Spoken communication may in fact have evolved to take advantage of visuofacial, as well as auditory, sensitivities (e.g., Rosenblum, 2005). This perspective is consistent with neurophysiological findings suggesting that visual speech information modulates auditory cortex activity as if the brain is responding to heard speech (e.g., Calvert et al., 1997; MacSweeney et al., 2000; MacSweeney et al., 2002).

Given the importance of visual speech information, the question arises whether it can induce the unconscious imitation, or alignment, that has been shown for auditory speech. For visual speech to do so, it must convey information about a talker's speaking style to the perceiver. In fact, there is evidence that visual speech can provide talker-specific characteristics. For example, perceivers can recognize talkers from simply seeing their isolated speech movements. Speech movements can be isolated by using a point-light technique, in which only moving dots placed on the face are seen against a dark background. From these stimuli, talkers can be recognized in both matching and identification contexts (Rosenblum, Niehus, & Smith, 2007; Rosenblum et al., 2002). Point-light research has also shown that a talker's isolated speech movements can be matched to their voice at better-than-chance levels (e.g., Lachs & Pisoni, 2004c; Rosenblum, Smith, Nichols, Hale, & Lee, 2006). These findings suggest that observers are sensitive to the articulatory style of a talker as it is reflected in both auditory and visual modalities.

In summary, research suggests that the visual speech signal provides not only phonetic information, but also information about the talker-specific articulatory style—or *idiolect*—of a talker. If talker-specific articulatory information is conveyed in visual speech, visual speech stimuli could have the potential to induce the type of speech alignment shown for auditory speech. Indeed, Gentilucci and Bernardis (2007) recently reported initial evidence that visual speech information might have the potential to induce speech alignment. These researchers asked women to lipread and shadow two male and two female talkers silently uttering /aba/ bisyllables. Kinematic and acoustic analyses of the women's utterances showed that their lip movements were larger and their voice spectra were lower when shadowing the male than when shadowing the female talkers. Gentilucci and Bernardis suggested that this would be expected from (what they claim) are the known differences in articulatory movements between the genders (with male talkers having larger excursions). These results suggest that the visual information for the male talker's utterances induced the female subjects to produce shadowed utterances that were more male-like in their movements: The women aligned to talker gender.

The research by Gentilucci and Bernardis (2007) provided initial evidence that perceivers can align to some aspects of visible speech utterances, but a number of important questions about visual speech alignment remain. For example, although Gentilucci and Bernardis used a single /aba/ stimulus to induce alignment, auditory alignment researchers have typically used word lists (e.g., Goldinger, 1998; Namy et al., 2002; Shockley et al., 2004). It has proven important to test words in auditory alignment research in order to examine the role of lexical access (e.g., word frequency, neighborhood density) in speech alignment (e.g., Goldinger, 1998; Goldinger & Azuma, 2004; Shockley et al., 2004). This makes it essential to determine whether alignment to visual speech can occur with words, as well as with the bisyllables /aba/ tested by Gentilucci and Bernardis.

Gentilucci and Bernardis (2007) tested only female subjects in their experiment, whereas in most of the auditory alignment research both male and female shadowers have been tested. In fact, there is some evidence that male and female subjects do align differently. For example, Namy et al. (2002) found that female shadowers tended to align more than male shadowers (but see Pardo, 2006). Namy et al. attributed the finding to gender differences in perceptual sensitivity. They speculated that women may be more sensitive to talker-specific information than men and that this information influences their own productions. If this is true, the visual alignment reported by Gentilucci

and Bernardis may have been a result of the fact that only female shadowers were tested. The putatively less sensitive male observers may not align to visual speech. This makes it critical to test visual speech alignment with both male and female subjects.

A third question arising from the research of Gentilucci and Bernardis (2007) is whether visual alignment will occur with shadowers not asked to repeat the utterances that they perceive. A majority of auditory alignment researchers have intentionally instructed subjects to say the perceived utterances out loud, thereby avoiding any suggestion that the subjects should imitate. However, the subjects in the Gentilucci and Bernardis study were instructed to repeat the utterances that they saw, possibly biasing them toward imitation. Although this may be a more minor concern, testing visual alignment with subjects who are instructed to simply say the words out loud could provide a more rigorous test of inadvertent (unconscious) alignment and would be more consistent with the existing alignment research.

A final question is whether visual speech alignment occurs in a perceptually relevant way. Gentilucci and Bernardis's (2007) evaluation of alignment involved measuring movement kinematics and voice spectra. In contrast, auditory speech researchers most often evaluate alignment using the aforementioned naive rater matching task. By having naive perceivers judge the relative similarity between utterances, researchers use this method to determine whether shadowed speech alignment occurs in a perceptually relevant manner (e.g., Goldinger, 1998; Goldinger & Azuma, 2004; Namy et al., 2002; Pardo, 2006; Shockley et al., 2004) . Recall that one proposed function of speech alignment is to facilitate communicative and social interactions. If this assertion is true, speech should align in a way that is perceptible. Although the use of rater matching tasks has established this to be true for auditory speech alignment, it has not yet been determined for visual speech.

To address whether visual speech alignment can occur in a way comparable to that in which auditory speech alignment does—in a way that is lexical, gender-relevant, unconscious, and perceptible—visual speech was tested using the alignment methods of the auditory speech research. In Experiment 1, we borrowed the shadowing methodology and AXB rating measure used by Goldinger (1998) and others. We tested both male and female subjects on an auditory and visual speech alignment task. The auditory task was borrowed directly from the method of Goldinger and involved shadowing of a word list adopted from Shockley et al. (2004). The visual speech task adapted these methods for lipreading. On each visual speech trial, subjects were asked to say out loud a word that they had just lipread from a model. The model was of the same gender as the subjects (Shockley et al., 2004). In order to make the lipreading task easier, each trial first included a presentation of two text words, one of which was the same as the word that they were to lipread. The subjects' utterances were recorded and presented to raters along with the model's auditory words and baseline (read) words spoken by the subjects before the shadowing task. If visual speech can induce the type of alignment induced by auditory speech, the raters should find that the subjects' shadowed utterances sounded more like the model's utterances than did their baseline utterances.

EXPERIMENT 1

Method

Participants

Two graduate students (1 male, 1 female) acted as models in the experiment and produced the original word list to be shadowed (e.g., Shockley et al., 2004). These models had no noticeable accents or speech impediments. Sixteen undergraduates (8 male, 8 female) acted as subjects who were asked to shadow the models' words. Thirty-two undergraduates acted as raters in an AXB matching task. All of the models, subjects, and raters were native speakers of American English with normal hearing and normal or corrected vision. The graduate student models were paid for their participation. The undergraduate subjects and raters participated in order to partially fulfill a course requirement.

Materials and Apparatus

A list of 74 bisyllabic, low-frequency English words were used as stimuli (see the Appendix). These words were derived from the list used by Shockley et al. (2004). The words had frequencies of less than 75 occurrences per million (Kučera & Francis, 1967), and they all began with the voiceless stop consonants (/p/, /t/, or /k/). This allowed us to ensure that our subjects were shadowing to a degree comparable to those of Shockley et al. (2004). In addition, low-frequency words were selected because it has been shown that they generally induce greater alignment in shadowers (e.g., Goldinger, 1998). In that this experiment constituted a first attempt to induce alignment with visual speech, it was thought that using low-frequency words would provide the best chance of doing so. However, it must be acknowledged that using low-frequency words does limit the scope of the study.

All of the stimuli were presented using PsyScope software. Text (baseline) and visual speech stimuli were presented on a 20-in. video monitor positioned 3 ft in front of the subjects. Auditory stimuli were presented through Sony MDR-V6 headphones. A Sony DSR-11 camcorder was used to videotape the models. The models and subjects responded verbally into a Shure SM57 microphone and were audio recorded at 44 kHz (16 bits) using Amadeus II software.

Procedure

The experiment took place in three phases. For all three phases, the individuals sat in a sound-attenuating chamber.

Phase 1. In Phase 1, two models (1 male, 1 female) were videotaped producing the 74 bisyllabic words. The word list was presented to the models as text on a video monitor. The words were randomly presented at a rate of one word per second. The models were asked to speak the words quickly but clearly into the microphone. These utterances were filmed using the camcorder, and these recordings were edited on a computer to produce tokens for later presentation to the subjects. The audiovisual recordings were digitized and edited using FinalCut Pro software into 74 audio and 74 silent video tokens. The silent video showed the entire head and a portion of the models' shoulders.

Phase 2. Phase 2 of the experiment consisted of the 16 subjects (8 male, 8 female) participating in three tasks: baseline word production (text reading), audio shadowing, and silent video shadowing (lipreading). Each task was presented in its own block (e.g., Goldinger, 1998; Shockley et al., 2004), and all of the subjects performed the baseline word production first. The order of the remaining two tasks was counterbalanced across subjects.

For the baseline word task, the subjects were audio recorded producing the original word list, which they read from a video monitor. The words were presented individually at 1-sec intervals. The subjects were asked to say the words that they saw quickly but clearly into the microphone. These utterances were later edited on a computer to create 74 baseline tokens for the ratings in Phase 3.

For the audio shadowing task, the subjects were audio recorded shadowing 1 of the model's 74 audio words, which they heard over headphones. The male subjects shadowed the male model, and the female subjects shadowed the female model (e.g., Shockley et al., 2004). The shadowing task required the subjects to say each word that they heard quickly but clearly into the microphone (e.g., Shockley et al., 2004). The subjects were never asked to imitate, or even repeat, the model. All shadowed utterances were recorded and later edited to create 74 audio shadowed tokens for comparison purposes in Phase 3.

For the silent video shadowing task, the subjects were again audio recorded shadowing a model's 74 words. However, in this condition, the subjects were asked to lipread the words from the model. Because of the difficulty that some individuals have with lipreading, a low-uncertainty forced choice task was used. The subjects were first presented with two text words—a target and distractor—shown side by side on the video monitor (e.g., *cabbage*, *camel*). These words were presented for 2 sec. Immediately afterward, the subjects would see the face of the model silently saying 1 of the words (e.g., *cabbage*). The subjects' task was to produce out loud into the microphone quickly but clearly the word that they had lipread. Again, they were never asked to imitate the model.

Each distractor word was chosen to be similar in initial segments to its paired target word (e.g., *cabbage*, *camel*). Pilot tests showed that this forced the subjects to pay attention to the articulated target words but allowed the subjects to correctly lipread the target words a majority of the time.

All shadowed utterances were audio recorded and later edited to create video shadowed tokens for comparison purposes in Phase 3.

Phase 3. In Phase 3, naive raters were asked to judge the similarity between the models' words and the subjects' shadowed words relative to that between the models' words and the subjects' baseline words. For these purposes, we used an AXB matching task, which is commonly used in speech alignment experiments (Goldinger, 1998; Namy et al., 2002; Pardo, 2006; Shockley et al., 2004). Rating methods were chosen over acoustical analysis for a number of reasons. First, rating methods provide a perceptually valid way of establishing similarities across stimuli and, thus, alignment across utterances (Goldinger, 1998; Namy et al., 2002; Pardo, 2006; Shockley et al., 2004). In addition, the method avoids the difficulty in determining to which of the many possible acoustical dimensions subjects are aligning (Goldinger, 1998). Finally, the method has been used to evaluate alignment in a majority of the studies in which the phenomenon was investigated (e.g., Goldinger, 1998; Namy et al., 2002; Pardo, 2006; Shockley et al., 2004).[1]

Thirty-two naive raters (23 female) judged whether a subject's shadowed token was more similar to the model's token than was the subject's baseline token. Two raters were assigned to judge the words produced by a given shadowing subject (16 shadowing subjects × 2 raters each = 32 raters) (e.g., Shockley et al., 2004).

A separate AXB triad was created for each word that the subjects produced. Each triad included presentations of the same word (e.g., *cabbage*) produced once by the model and twice by the subject. The model's spoken utterance always appeared as the middle (X) token. The subjects' shadowed tokens appeared either in the A (first) or B (third) position and the subjects' baseline (the read token) appeared in the remaining A or B position. Each subject's word, from each shadowing block (audio and visual) actually appeared in two triads: once when presented in the A position and once when presented in the B position. This means that, in principle, raters would judge a total of 296 separate triads: 74 words × 2 shadowing modalities (audio, video) × 2 triad orderings (once with the shadow word in the A position, once with it in the B position).

However, the number of triads derived from each shadowing subject's responses actually ranged between 256 and 284 ($M = 268$). The reason for this was as follows. Although the subjects were generally quite accurate at lipreading the model in the two-alternative forced choice task, all of the subjects lipread words wrong a few times during the session. The percentage correct on the lipreading task ranged from 86% to 96% for the 16 shadowing subjects, with a mean of 90% correct. Because only correctly lipread utterances could be used in the matching task performed by the raters, each subject's incorrectly lipread words were not including in the AXB sets for that subject. Furthermore, to ensure that comparisons across shadowing of audio and visual presentations were fair, the words incorrectly lipread by a subject were also removed from that subject's audio shadowed lists. Thus, if the word *cabbage* had been incorrectly lipread by a subject, *cabbage* would also be removed from that subject's audio shadowed list, baseline list, and model's list so that *cabbage* would not be part of the AXB stimuli for that subject. This accounts for the differential number of triads judged by the raters.

The triads based on the auditorily and visually derived utterances of a shadower were completely randomized together for presentation to the raters. The raters listened to the triads through headphones and were asked to choose which of the words—the first or third—sounded more similar to the second. The raters were instructed to press the key labeled "1" on the keyboard if the first word sounded more similar to the second or to press the key labeled "3" on the keyboard if the third word sounded more similar to the second.

Results and Discussion

Means were calculated for the number of shadowed utterances chosen as sounding more like those of the model for each rater and each subject. These individual means for male and female subjects, for both the audio and video shadow responses (averaged across words), are presented graphically in Figure 1. The overall mean proportion of the subjects' shadowed tokens considered better imitations of the models' tokens (than were the baseline read tokens) was .573 ($SE = .017$) for audio shadowing and .564 ($SE = .015$) for visual (lipread) shadowing. These proportions were compared with chance (.50) using *t* tests, which revealed that the subjects' shadowed tokens were judged to be better imitations of the models' tokens than were the baseline tokens for both the audio shadowed words [$t(31) = 4.892$, $p < .0001$, Cohen's d effect size = .87] and the visually shadowed words [$t(31) = 3.704$, $p = .0008$, Cohen's $d = .66$] (Thalheimer & Cook, 2002). A paired samples *t* test revealed that there was no significant difference in rater matching between the audio and visual shadowing tasks [$t(31) = 0.604$, $p = .5500$].

The effects of gender (between subjects; male vs. female) and modality (within subjects; audio vs. video) were evaluated (on the basis of values averaged across words and raters) using a factorial ANOVA. The results indicate a marginal main effect of gender [$F(1,30) = 3.524$, $p = .07$], with the female subjects aligning more than the male subjects. Still, *t* tests conducted for the male and female subjects revealed that the utterances for both gender groups were matched to their respective models at better-than-chance levels for both the audio and the video shadowing conditions ($p < .05$). No main effect of modality was found [$F(1,30) = 0.357$, $p = .55$], and there was no significant gender × modality interaction [$F(1,30) = 0.328$, $p = .57$]. Finally, a paired *t* test of the effect of presentation block was conducted and revealed that on the basis of the AXB ratings, more alignment occurred during the second block than during the first ($M = .586$, $SE = .017$, and $M = .551$, $SE = .015$, respectively) [$t(31) = 2.47$, $p = .019$]. This is not surprising, because the same 74 words were used in the two blocks. Past research has

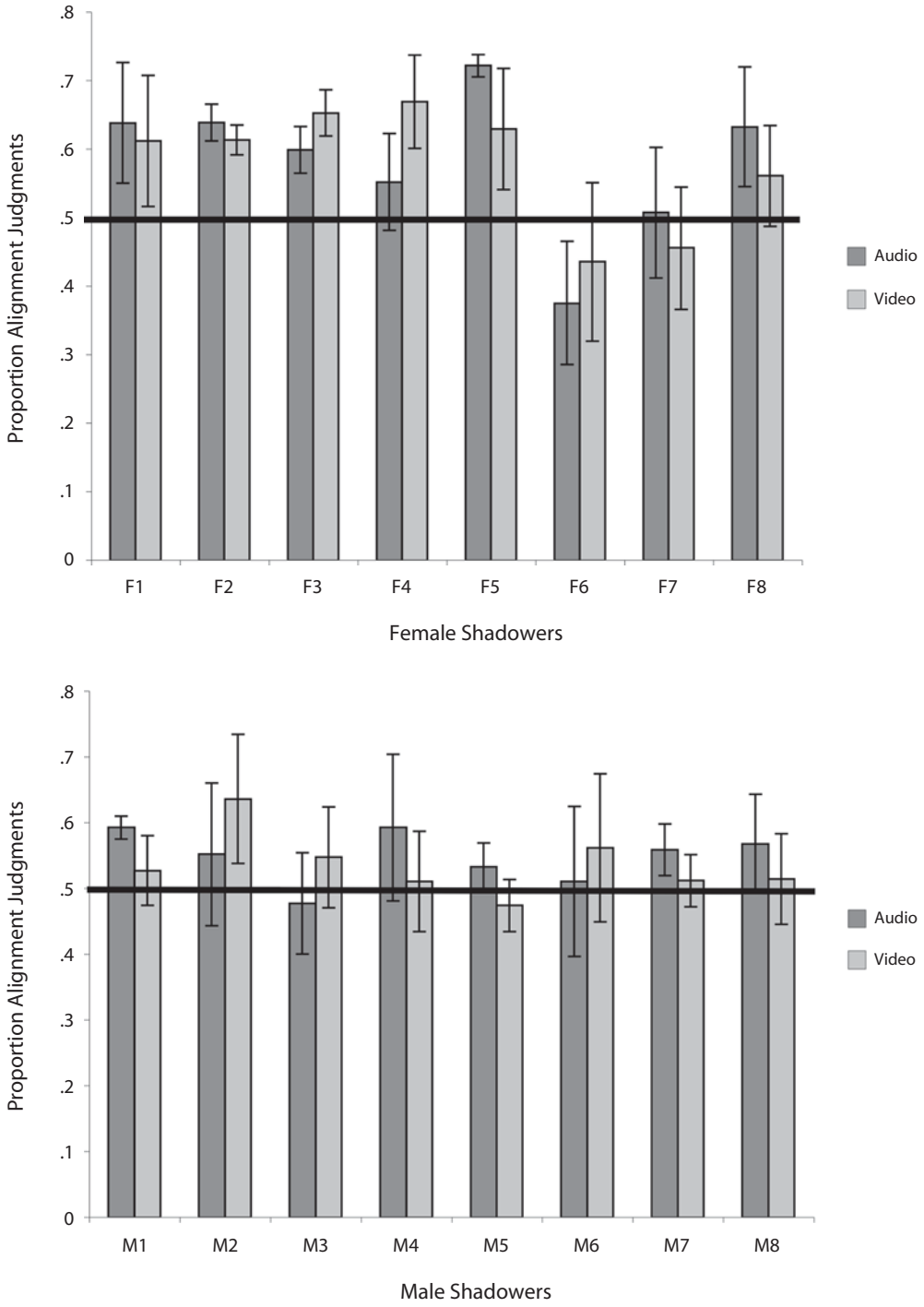

Figure 1. Mean proportion of model's words sounding more similar to subjects' shadowed words than did baseline words for audio and visual shadowing conditions for female subjects (top panel) and male subjects (bottom panel).

shown that the degree of alignment to a talker increases with the number of repetitions of a word spoken by that talker (Goldinger, 1998).

These results reveal that on the basis of the auditory judgments of naive raters, the subjects did align to the words that they both heard and saw the models say. In fact, the values were statistically equivalent for the auditory and visual shadowed conditions. This suggests that the two modalities provided a comparable amount of information to drive speech alignment.

Although the results portrayed in Figure 1 suggest that some subjects aligned more than others, the range of these values is similar to those of other alignment studies (e.g., Goldinger, 1998; Namy et al., 2002; Pardo, 2006; Shockley

et al., 2004). Also, the effect sizes for both the audio and the visual conditions were in the high-medium-to-large range (Thalheimer & Cook, 2002). Thus, although the results show that the alignment for both the audio and the visual conditions was often subtle, it was statistically sound and comparable to that of other alignment research.

Although recent evidence has shown that visual speech provides indexical information (Kaufmann & Schweinberger, 2005; Schweinberger & Soukup, 1998; Sheffert & Fowler, 1995; Yakel, Rosenblum, & Fortier, 2000; and see also Sheffert & Olson, 2004), it was unknown whether this visually specified information could unconsciously alter speech production responses. Prior research (e.g., Goldinger, 1998; Goldinger & Azuma, 2004; Namy et al., 2002; Pardo, 2006; Shockley et al., 2004) has shown that auditory speech has this potency in both conversational and shadowing contexts. In showing that lipread shadowed words are rated as auditorily similar to those of the model, the present results provide evidence that visually specified indexical talker information can modulate speech production responses.

These results go beyond those reported by Gentilucci and Bernardis (2007) by showing that visual speech alignment can occur with spoken words. Furthermore, the present results show that both female and male subjects align to visible speech and do so even when they are simply instructed to say out loud, rather than to repeat, the utterances that they perceive. Finally, Gentilucci and Bernardis evaluated alignment using acoustic and kinematic measurements of shadowed responses; the present results add evidence that visually induced alignment is robust enough to be perceived by naive raters in a matching task.

The results also revealed a marginal main effect of gender. Research findings on the impact of gender on speech shadowing have been inconsistent (Namy et al., 2002; Pardo, 2006). Using a shadowing paradigm, Namy et al. compared the alignment of male and female subjects shadowing models of the same or of a different gender. The researchers found that female shadowers tended to align more than male shadowers, although the shadowers, in general, tended to align more to the male models. This difference in alignment was attributed to gender differences in perceptual sensitivity. In other words, women may attend better to talker-specific properties than men. This interpretation is consistent with the results of Experiment 1.

However, Pardo (2006) found results that suggested that male talkers aligned more than female talkers. This difference may stem in part from differences in the experimental design. Rather than using shadowing to assess alignment, Pardo (2006) opted to use an interactive map task to induce alignment in the context of live conversation. Pardo (2006) attributes her observed gender effects in alignment to attentional differences with the task, rather than to differences in perceptual sensitivity, as such.

Future research can examine why women aligned marginally more than men in the present experiment. Because subject (shadower) gender was matched to model gender in the present experiment (following Shockley et al., 2004), the degree to which the subjects' versus the models' gender played a role in these effects is uncertain. Also, although gender may play an intricate role in alignment, it could also be that other factors distinguishing our two models (e.g., speech clarity, attractiveness, expression) drove these marginal effects.

EXPERIMENT 2

As was stated above, there is evidence in the literature that perceivers are sensitive to the articulatory-style information of a talker as it is reflected in both the auditory and the visual modalities (see Nygaard, 2005, and Rosenblum, 2005, for reviews). In fact, this information allows perceivers to match heard speech to lipread speech on the basis of talker identity (e.g., Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004a, 2004b, 2004c; Rosenblum et al., 2006). This suggests that speaking style can be perceived across modalities. Furthermore, the results of Experiment 1 show that raters are sensitive to the similarity in models' and shadowers' utterances whether the shadowing is based on audio or visual information of the model. This suggests that speaking style can, to some degree, be perceived across talkers.

If speaking style can be perceived across modalities and across talkers, an interesting prediction arises. Raters should be able to match aligned utterances across a model and shadower when each utterance is presented in a different modality. Put differently, if shadowers are taking on some of the articulatory style of the models, and articulatory style can be perceived across modalities, then observers should also be able to match a shadower's voice to the visible articulating face of the model that had been shadowed.

In Experiment 2, we tested this prediction using the audio and video recordings obtained in Experiment 1. Raters were asked to make cross-modal AXB matches. The raters were presented AXB trials on which a shadower's utterances (the A and B positions) were presented *auditorily*, whereas the model's words (X) were presented *visually* without sound. Thus, the raters were asked to match the similarity of utterances across two talkers (a model and a shadower) and two modalities (auditory and visual).

In this sense, the raters of Experiment 2 were actually the *subjects* whose perceptual sensitivity was tested. If the information for talker alignment can be conveyed across modalities, these subjects should be able to match a model's silent video token to the shadower's audio token (vs. baseline) at better-than-chance levels.

In addition, we incorporated the shadowed responses derived from both the audio and visual shadowing conditions of Experiment 1. In this sense, in Experiment 2, we tested a modified replication of Experiment 1 by examining whether matches (in this case cross-modal) can be made between a model's and a shadower's utterances when that shadow is based on lipread or auditory information.

Method

Participants

The graduate student models and undergraduate shadowers were the same as those used in Experiment 1. Thirty-two new undergraduates (23 female) acted as subjects in a modified AXB matching task.

These undergraduate subjects participated in order to partially fulfill a course requirement. None had participated in Experiment 1.

Materials and Apparatus

All materials and apparati were the same as those in Experiment 1. However, in this experiment, the models' silent video utterances recorded in Phase 1 of Experiment 1 were used for comparison with the shadowers' baseline and shadowed tokens recorded (auditorily) in Phase 2. All three types of shadowers' utterances (from Experiment 1) were used in Experiment 2: baseline productions, shadows of the model's audio tokens, and shadows of the model's video tokens. Shadowed modality (audio vs. video) from Experiment 1 was therefore considered a factor in Experiment 2.

Procedure

The subjects judged whether a shadower's shadowed tokens were more similar to the model's silent video tokens than were the shadower's baseline tokens. A separate AXB sequence was created for each shadower from Experiment 1 and was presented to 2 subjects (of Experiment 2). For each triad, a model's silent video token always appeared in the X position. The shadower's shadowed and baseline audio tokens appeared in the A and B positions, which were counterbalanced to create two orders for each triad. As was stated above, the shadowed tokens were taken from Phase 2 of Experiment 1, in which the subjects were asked to shadow both the audio (heard) and video (lipread) tokens of the model. This means that in principle, the full matching sequence would consist of 296 tokens: 74 words × 2 shadowing modalities (audio and video shadows from Experiment 1) × 2 AXB orderings. However, again, the total number of triads differed between sequences because of incorrect lipread responses for each shadower in Experiment 1 (see above).

The AXB triads were presented auditorily over Sony MDR-V6 headphones. The video tokens were presented on a 20-in. monitor 3 ft in front of the subjects. These tokens did not include sound. The subjects were asked to choose which of the utterances—the first or the third—was more similar to the video utterance presented as the second. The subjects were instructed to press the key labeled "1" on the keyboard if the first word was more similar to the second or to press the key labeled "3" on the keyboard if the third word was more similar to the second.

Results and Discussion

The individual means for the male and female subjects, for both the audio and the video shadowed responses, are presented graphically in Figure 2. The overall mean proportions of shadowers' shadowed tokens considered better imitations of the models' video tokens (than were the baseline read tokens) were .538 ($SE = .013$) for audio shadows and .559 ($SE = .016$) for visual (lipread) shadows. A comparison to chance (.50) revealed that the subjects judged the shadowers' shadowed tokens to be better imitations of the models' video tokens for audio shadowed words [$t(31) = 3.008$, $p < .01$, Cohen's $d = .535$] and for visually shadowed words [$t(31) = 3.658$, $p < .001$, Cohen's $d = .648$]. Again, a paired samples t test revealed that there was no significant difference in these judgments between words shadowed auditorily in Experiment 1 and those shadowed visually [$t(31) = -1.383$, $p = .177$].

An ANOVA on the factors of shadower gender and shadowed modality (on the basis of values averaged across words and raters) did not reveal a main effect of gender [$F(1,30) = 2.485$, $p > .05$] or modality [$F(1,30) = 2.229$, $p > .05$]. However, there was a significant gender × modality interaction [$F(1,30) = 6.143$, $p = .019$]. Pairwise comparisons revealed that the subjects of Experiment 2 matched the female shadowers' utterances to those of the model more often than they did the male shadowers' utterances when these shadowed utterances were based on the video stimuli. It is unclear why this interaction occurred, but the fact that the subjects in this experiment found that the female shadowers' shadowed utterances more often matched those of the model (when shadowing a video utterance) is consistent with the marginal gender effects reported in Experiment 1.

The results portrayed in Figure 2 suggest that, again, some subjects aligned more than others. Still, the range of these values is comparable to those of other alignment studies (e.g., Goldinger, 1998; Namy et al., 2002; Pardo, 2006; Shockley et al., 2004). The effect sizes for both the audio and the visual conditions were in the medium range (Thalheimer & Cook, 2002).

The results of Experiment 2 show that when subjects are asked to match shadowed utterances to a model's utterances, they can do so across modalities at better-than-chance levels. On each trial, the subjects in Experiment 2 were presented with an audio utterance from a shadower, a silent video utterance from a model, and then another audio utterance from the shadower. One of the shadower's utterances was produced when the shadower simply read the word (baseline), whereas the other was produced when the shadower shadowed the model. The subjects in Experiment 2 were able to determine, at better-than-chance levels, which of the shadower's utterances were produced when shadowing the model. In this sense, the subjects were able to detect speech alignment both across talkers and across modalities. This suggests that the indexical characteristics that are passed from one talker to another are perceptible across auditory and visual information. The implications of this finding will be addressed in the General Discussion section.

The results of Experiment 2 also showed that these matches could be made at better-than-chance levels when the shadowers of Experiment 1 shadowed either the visual or the auditory speech of the model. This finding is consistent with the results of Experiment 1 in again showing that speech alignment can be induced by either visual or auditory speech information.

GENERAL DISCUSSION

The experiments reveal that shadowers align to a model's spoken words whether those words are presented auditorily or visually. Although the results suggest that this alignment can be subtle, it seems comparable to that observed in previous alignment research (e.g., Namy et al., 2002; Pardo, 2006; Shockley et al., 2004). The present research also shows that this alignment between shadowers and models is perceivable across auditory tokens (Experiment 1, Phase 3), as well as across auditory and visual tokens (Experiment 2).

The finding of auditory alignment is consistent with past research showing alignment to auditory speech both during live conversation (Pardo, 2006) and when shadowing isolated tokens (Goldinger, 1998; Namy et al., 2002; Shockley et al., 2004). Indeed, our auditory results closely replicate

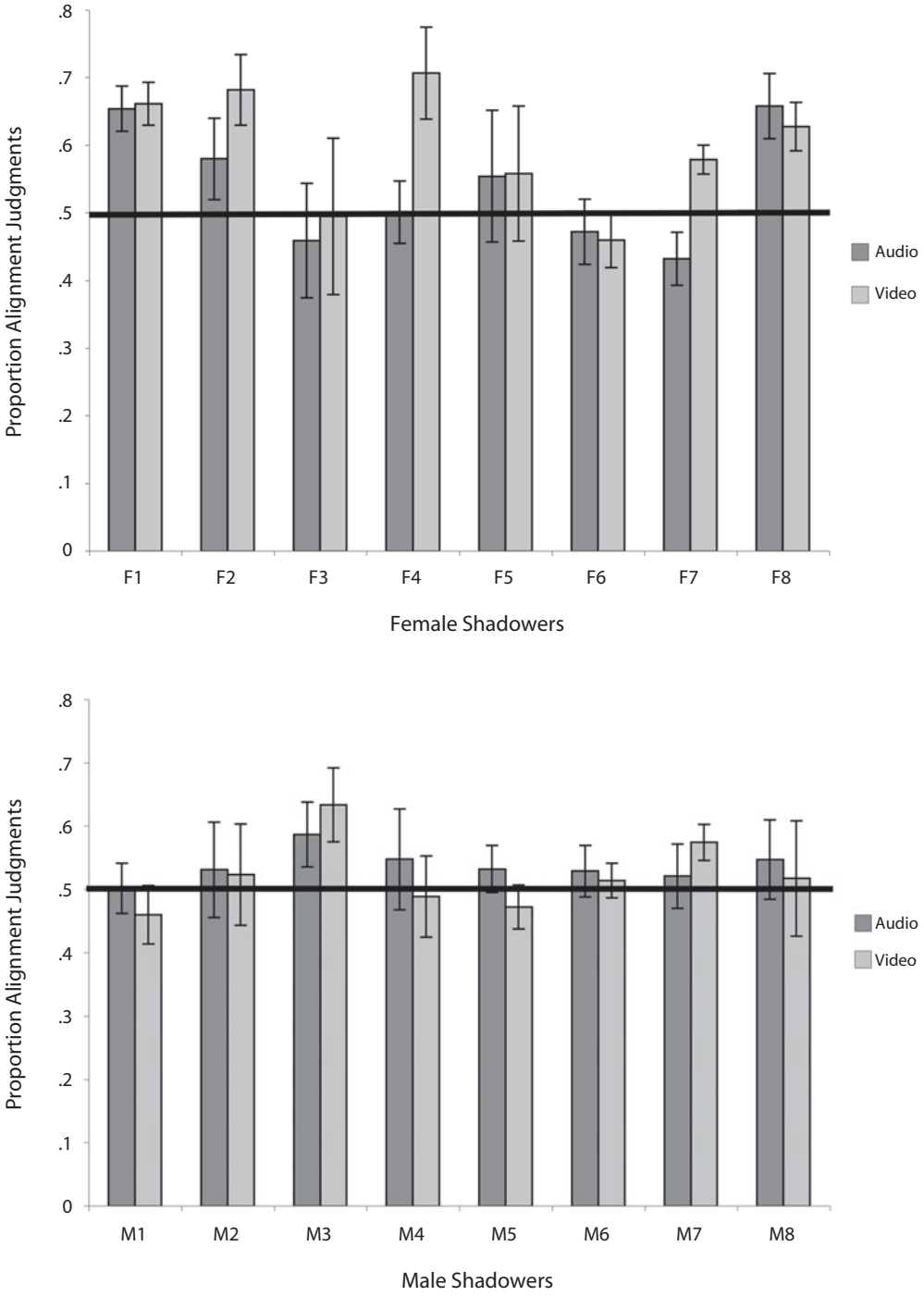

Figure 2. Mean proportion of model's visible words rated as more similar to subjects' shadowed words than were baseline words for audio and visual shadowing conditions for female subjects (top panel) and male subjects (bottom panel).

the findings of Shockley et al. (2004, Experiment 1), on which the method of the present study was partly based.

With regard to visual speech, our findings that shadowers align to visually presented stimuli are consistent with the initial report of Gentilucci and Bernardis (2007). As was noted above, however, our findings reach further than those researchers' work in showing that visual speech alignment occurs for multiple *word* stimuli, instead of for bisyllabic, nonsense stimuli, and occurs to the degree that it is perceivable by naive raters, not simply by measures of lip kinematics and acoustics. In this sense, the present findings show that alignment to visual speech can work in a methodological context comparable to that used for most auditory alignment demonstrations.

The results of Experiment 2 show that in a variation of the AXB task, matchers are perceptually sensitive to shadowers' aligned speech, despite this speech's being presented to them in different modalities. The results complement those of Experiment 1, wherein raters judged both auditory and visual speech alignment to occur when the tokens were compared auditorily. The ability to perceive alignment across modalities suggests that the indexical information carried across aligning talkers is available cross-modally. The conceptual implications of our findings will be discussed in the following sections.

Informational Basis of Alignment

Finding evidence for visual, as well as auditory, speech alignment poses interesting questions about the informational dimensions to which talkers align. Although the AXB matching method allows for confirmation of the perceptual salience of the aligned information between model and shadower, the method cannot easily determine the information to which talkers align. Still, speculation is warranted. Previous results in the auditory alignment literature have suggested that shadowers imitate models' produced acoustic dimensions, including intonational contour, acoustic vowel space, and VOT (e.g., Gentilucci & Bernardis, 2007; Goldinger, 1998; Pardo, 2004; Shockley et al., 2004).

With regard to visual speech alignment, it is unclear which talker-specific visible attributes shadowers might imitate. Gentilucci and Bernardis (2007) measured lip kinematics as their female subjects shadowed visible /aba/ bisyllables produced by male and female models. These researchers reported that the subjects produced greater lip excursions when shadowing male than when shadowing female model syllables, thereby aligning toward the extent of lip excursions produced by the models themselves (which were also measured). Although our own subjects may have also aligned to the model's lip excursions to some degree, it is unclear whether this dimension, considered by Gentilucci and Bernardis to distinguish talker gender, would be sufficient to induce the talker alignment observed in Experiment 1. It could very well be that other aspects of a model's visible articulatory movements also induced the alignment observed in Experiment 1.

Note that, unsurprisingly, the imitated dimensions for auditory alignment have been considered acoustic in nature and, for visual alignment, optic in nature. However, the results of Experiment 2 suggest an alternative formulation. In that subjects could match aligned utterances across modalities, the imitated information likely includes some dimensions that are instantiated in both the visible and the audible streams. In supporting cross-modal matches, this information might best be construed as amodal or modality neutral. In fact, the notion of amodal talker-specific information has been used to explain the cross-modal talker matching findings described earlier (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b, 2004c; Rosenblum et al., 2006). The authors of those reports suggest that cross-modal matching could be based on the extraction of common idiolectic information available across modalities. Idiolect is, after all, an amodal articulatory property that can potentially structure both the acoustic and the visual media. It could be that detection of amodal idiolectical properties also provided the basis for the cross-modal matching in our Experiment 2.

If shadowers align to properties of a talker's amodal idiolect, it still must be determined which of these properties are most salient. These properties could range in complexity from simple articulatory rate to more nuanced coarticulatory style. If the imitated dimensions are similar to those found salient for cross-modal matching, it is unlikely that the shadowers are imitating simple duration (e.g., Lachs & Pisoni, 2004b). Future research can be designed to determine which amodal and/or modality-specific properties shadowers imitate.

Indexical Influences on Word Perception

The present findings may also have implications for theories of word recognition. As was stated above, Goldinger (1998) interpreted his auditory alignment findings as supporting an episodic lexicon. Goldinger's theory proposes that episodic traces of heard words are present and accessible in lexical memory. Alignment is thought to emerge as a byproduct of responding to a particular talker whose indexical information contributes most recently to these episodes.

In finding evidence for alignment in shadowed responses to visible speech, the present results suggest a broadening of the form of episodic traces. Assuming that alignment phenomena reflect the nature of the episodic lexicon, the results indicate that episodic traces retain not only auditory, but also *visible* indexical dimensions. In fact, broadening the traces to include visual speech dimensions could allow Goldinger's (1998) theory to explain the talker-facilitation effects observed for visual speech perception. These effects show that visual familiarity with a speaking face can facilitate visible vowel identification (Kaufmann & Schweinberger, 2005; Schweinberger & Soukup, 1998), word lipreading (Lander & Davies, 2008), sentence lipreading from single- versus multiple-talker lists (Yakel et al., 2000), and—most germane to Goldinger's theory—memory for lipread words (Sheffert & Fowler, 1995; see also Sheffert & Olson, 2004).

However, following from the discussion in the preceding section, it could be that Goldinger's (1998) theory would be best served by considering the traces as composed of amodal idiolectic information. As was stated above, this could account for the results of the cross-modal and cross-talker alignment results observed in Experiment 2. Moreover, considering the retained talker information as amodal could explain results recently reported in our laboratory (Rosenblum, Miller, & Sanchez, 2007). We observed evidence that familiarity with a talker gained through one modality can carry over to facilitate speech in another modality. In this experiment, subjects were asked to lipread sentences from a single talker for 1 h. Afterward, they were asked to listen to auditory speech-in-noise sentences produced by a talker who was either the same as or different from the talker that they had just lipread. The subjects who listened to the same talker that they had just lipread were better able to recover the speech-in-noise sentences than were the subjects who lipread one talker and heard a

different talker. We interpret this finding as evidence that amodal idiolectic information is extracted from the visual speech signal of a talker and is then used to facilitate auditory speech recovery from that talker. If episodic traces also contain amodal idiolectic information, rather than simply auditory details, it could explain these cross-modal talker facilitation results, as well as the results presented in the present report. In fact, Goldinger himself entertains the possibility that the episodes might take a gestural, rather than simply auditory, form (Goldinger, 1998).

Perceptual Regulation of Speech Production Responses

Speech alignment effects have also been interpreted as demonstrating perceptual regulation of produced speech based on the input from an interlocutor (Fowler, 2004; Pardo, 2006; Pardo & Remez, 2006). As was stated above, alignment research has shown that auditory speech information can influence a talker's rate, accent, and intonational contour (e.g., Giles et al., 1991; Gregory, 1990; Natale, 1975; Sancier & Fowler, 1997), as well as phonetic dimensions (Shockley et al., 2004). These alignment phenomena show that the perceptual effects on the self-regulation of produced speech can be impressively fast. In fact, there is evidence that speech production responses to perceived speech can be disproportionately faster than speech responses to nonspeech stimuli (Fowler, Brown, Sabadini, & Weihing, 2003; Kozhevnikov & Chistovich, 1965; Porter & Castellanos, 1980; Porter & Lubker, 1980). Similarly, the reaction time differences between simple single and choice response types are especially small for shadowed speech (Fowler et al., 2003; Porter & Castellanos, 1980; Porter & Lubker, 1980).

These reaction time findings have been interpreted as evidence for an exceptionally close connection between the speech production and perception functions and that a common currency is shared between the functions (Fowler, 2004; Fowler et al., 2003; Sancier & Fowler, 1997; Shockley et al., 2004). Fowler and her colleagues argued that this common currency takes the form of articulatory gestures that are both perceived and produced, and they cited alignment phenomena as supporting this claim (e.g., Fowler, 2004). These researchers further suggested that the common currency thesis is consistent with neurophysiological evidence for mirror neuron-type functions in human speech perception (e.g., Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Sundara, Namasivayam, & Chen, 2001).

The present visual speech alignment results seem consistent with the common currency thesis (see also Kerzel & Bekkering, 2000). For the currency to be truly common between production and perception, the primitives for perception would need to be gestural and not auditory in nature. In showing that these gestures, visually conveyed, can modulate a production response, the present results demonstrate that the primitives need not be auditory. In fact, there is evidence that, as for auditory speech, visual speech stimuli can induce mirror-type activity in articulatory musculature (Sundara et al., 2001).

Furthermore, the results of Experiment 2 show evidence for the presence of cross-modal alignment information. As was argued above, this finding, along with findings reported on cross-modal talker recognition, call for a consideration of the relevant indexical information as amodal. To the degree that this amodal information takes a gestural form, these results are also consistent with the common currency proposal.

Before we conclude, an important caveat must be acknowledged. The auditory alignment literature has revealed that although these phenomena can appear rapid, unconscious, and inadvertent, it would be wrong to consider alignment as a reflexive, direct, or automatic phenomenon (see Pardo & Remez, 2006, for a review). Intervening variables such as interlocutor gender and role, as well as the lexical frequency and presentation repetitions of word stimuli, strongly affect auditory alignment phenomena (Goldinger, 1998; Pardo, 2004, 2006; Pardo & Remez, 2006). It is likely that these same factors will bear on visual speech alignment, as was hinted by the marginal gender effects found in Experiment 1, as well as the intersubject variability observed in both experiments (see Figures 1 and 2). In fact, it is easy to imagine that the visual information for an interlocutor could bear even more strongly on the social aspects of alignment. Future research can be designed to test for this possibility, as well as to further examine the claims of the amodal and common currency theses.

AUTHOR NOTE

This research was supported by NIDCD Grant 1R01DC008957-01. The authors thank three anonymous reviewers and editor Lynne Nygaard for helpful comments on an earlier version of this article. Correspondence concerning this article should be addressed to L. D. Rosenblum, Department of Psychology, University of California, 900 University Ave., Riverside, CA 92521 (e-mail: rosenblu@citrus.ucr.edu).

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.

Calvert, G. A., Bullmore, E., Brammer, M. J., Campbell, R., Iversen, S. D., Woodruff, P., et al. (1997). Silent lipreading activates the auditory cortex. *Science*, **276**, 593-596.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality & Social Psychology*, **76**, 893-910.

Davis, C., & Kim, J. (2001). Repeating and remembering foreign language words: Implications for language teaching systems. *Artificial Intelligence Review*, **16**, 37-47.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience*, **15**, 399-402.

Fowler, C. A. (2004). Speech as a supermodal or amodal phenomenon. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processing* (pp. 189-201). Cambridge, MA: MIT Press.

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory & Language*, **49**, 396-413.

Gentilucci, M., & Bernardis, P. (2007). Imitation during phoneme production. *Neuropsychologia*, **45**, 608-615.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequences. In H. Giles, N. Coupland, & J. Coupland (Eds.), *Contexts of accommodation: Developments in applied sociolinguistics* (pp. 1-68). Cambridge: Cambridge University Press.

GOLDINGER, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, **105**, 251-279.

GOLDINGER, S. D., & AZUMA, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin*, **11**, 716-722.

GRANT, K. W., & SEITZ, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, **108**, 1197-1208.

GREGORY, S. W. (1990). Analysis of fundamental frequency reveals covariation in interview partners' speech. *Journal of Nonverbal Behavior*, **14**, 237-251.

KAMACHI, M., HILL, H., LANDER, K., & VATIKIOTIS-BATESON, E. (2003). "Putting the face to the voice": Matching identity across modality. *Current Biology*, **13**, 1709-1714.

KAUFMANN, J. M., & SCHWEINBERGER, S. R. (2005). Speaker variations influence speechreading speed for dynamic faces. *Perception*, **34**, 595-610.

KERZEL, D., & BEKKERING, H. (2000). Motor activation from visible speech: Evidence from stimulus–response compatibility. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 634-647.

KOZHEVNIKOV, V., & CHISTOVICH, L. (1965). *Speech: Articulation and perception* (JPRS Publication 50, 543). Washington, DC: Joint Publications Research Service.

KUČERA, H., & FRANCIS, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

LACHS, L., & PISONI, D. B. (2004a). Crossmodal source identification in speech perception. *Ecological Psychology*, **16**, 159-187.

LACHS, L., & PISONI, D. B. (2004b). Cross-modal source information and spoken word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **30**, 378-396.

LACHS, L., & PISONI, D. B. (2004c). Specification of cross-modal source information in isolated kinematic displays of speech. *Journal of the Acoustical Society of America*, **116**, 507-518.

LANDER, K., & DAVIES, R. (2008). Does face familiarity influence speechreadability? *Quarterly Journal of Experimental Psychology*, **61**, 961-967.

MACSWEENEY, M., AMARO, E., CALVERT, G. A., CAMPBELL, R., DAVID, A. S., MCGUIRE, P., ET AL. (2000). Silent speechreading in the absence of scanner noise: An event-related fMRI study. *NeuroReport*, **11**, 1729-1733.

MACSWEENEY, M., CALVERT, G. A., CAMPBELL, R., MCGUIRE, P. K., DAVID, A. S., WILLIAMS, S. C. R., ET AL. (2002). Speechreading circuits in people born deaf. *Neuropsychologia*, **40**, 801-807.

MCGURK, H., & MACDONALD, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.

MELTZOFF, A. N., & MOORE, M. K. (1997). Explaining facial imitation: A theoretical model. *Early Development & Parenting*, **6**, 179-192.

MILLS, A. E. (1987). The development of phonology in the blind child. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 145-162). Hillsdale, NJ: Erlbaum.

NAKAMURA, M., IWANO, K., & FURUI, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their of effects on recognition performance. *Computer Speech & Language*, **22**, 171-184.

NAMY, L. L., NYGAARD, L. C., & SAUERTEIG, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language & 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 & Social Psychology*, **32**, 790-804.

NAVARRA, J., & SOTO-FARACO, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of L2 sounds. *Psychological Research*, **71**, 4-12.

NYGAARD, L. C. (2005). The integration of linguistic and non-linguistic properties of speech. In D. Pisoni & R. Remez (Eds.), *Handbook of speech perception* (pp. 390-414). Malden, MA: Blackwell.

PARDO, J. S. (2004). Acoustic–phonetic convergence among interacting talkers. *Journal of the Acoustical Society of America*, **115**, 2608.

PARDO, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, **119**, 2382-2393.

PARDO, J. S., & REMEZ, R. E. (2006). The perception of speech. In M. Traxler & M. A. Gernsbacher (Eds.), *The handbook of psycholinguistics* (2nd ed., pp. 201-248). New York: Academic Press.

PORTER, R. J., JR., & CASTELLANOS, F. X. (1980). Speech production measures of speech perception: Rapid shadowing of VCV syllables. *Journal of the Acoustical Society of America*, **67**, 1349-1356.

PORTER, R. J., JR., & LUBKER, J. F. (1980). Rapid reproduction of vowel–vowel sequences: Evidence for a fast and direct acoustic–motoric linkage. *Journal of Speech & Hearing Research*, **23**, 593-602.

REISBERG, D., MCLEAN, J., & GOLDFIELD, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97-114). Hillsdale, NJ: Erlbaum.

ROSENBLUM, L. D. (2005). The primacy of multimodal speech perception. In D. Pisoni & R. Remez (Eds.), *Handbook of speech perception* (pp. 51-78). Malden, MA: Blackwell.

ROSENBLUM, L. D., MILLER, R. M., & SANCHEZ, K. (2007). Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects. *Psychological Science*, **18**, 392-396.

ROSENBLUM, L. D., NIEHUS, R. P., & SMITH, N. M. (2007). Look who's talking: Recognizing friends from visible articulation. *Perception*, **36**, 157-159.

ROSENBLUM, L. D., SMITH, N. M., NICHOLS, S. M., HALE, S., & LEE, J. (2006). Hearing a face: Cross-modal speaker matching using isolated visible speech. *Perception & Psychophysics*, **68**, 84-93.

ROSENBLUM, L. D., YAKEL, D. A., BASEER, N., PANCHAL, A., NORDARSE, B. C., & NIEHUS, R. P. (2002). Visual speech information for face recognition. *Perception & Psychophysics*, **64**, 220-229.

SANCIER, M. L., & FOWLER, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, **25**, 421-436.

SCHWEINBERGER, S. R., & SOUKUP, G. R. (1998). Asymmetric relationships among perceptions of facial identity, emotion, and facial speech. *Journal of Experimental Psychology: Human Perception & Performance*, **24**, 1748-1765.

SHEFFERT, S. M., & FOWLER, C. A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Memory & Language*, **34**, 665-685.

SHEFFERT, S. M., & OLSON, E. (2004). Audiovisual speech facilitates voice learning. *Perception & Psychophysics*, **66**, 352-362.

SHOCKLEY, K., SABADINI, L., & FOWLER, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, **66**, 422-429.

SHOCKLEY, K., SANTANA, M. V., & FOWLER, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception & Performance*, **29**, 326-332.

SUMBY, W. H., & POLLACK, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**, 212-215.

SUNDARA, M., NAMASIVAYAM, A. K., & CHEN, R. (2001). Observation–execution matching system for speech: A magnetic stimulation study. *NeuroReport*, **12**, 1341-1344.

THALHEIMER, W., & COOK, S. (2002, August). *How to calculate effect sizes from published research articles: A simplified methodology*. Retrieved November 31, 2002, from http://work-learning.com/effect_sizes.htm.

YAKEL, D. A., ROSENBLUM, L. D., & FORTIER, M. A. (2000). Effects of talker variability on speechreading. *Perception & Psychophysics*, **62**, 1405-1412.

NOTE

1. In theory, there are some possible shortcomings of the AXB rating measure used here and in the extant speech alignment research. For example, in each AXB triad, two of the utterances are based on *read* speech (the model's token and subject's baseline), whereas the third is based on *shadowed* speech (the subject's shadowed token). It is known that there are audible differences between read and shadowed speech (Nakamura, Iwano, & Furui, 2008), and these differences could influence the raters' matches. Note, however, if the raters made matches on the basis of the similarity of the two read tokens, they would more often match the model's tokens to the subjects' *baseline* tokens, since both are read. This is an outcome *opposite* to that hypothesized and typically observed in the alignment literature.

APPENDIX
English Bisyllable Words

/k/	Frequency (per million)	/p/	Frequency (per million)	/t/	Frequency (per million)
cabbage	4	*package*	20	*tailor*	2
cable	7	*panther*	1	*tamper*	1
camel	1	*pardon*	8	*target*	45
campus	33	*parrot*	1	*taxi*	16
canyon	12	*partner*	32	*teaspoon*	4
capture	17	*passion*	28	*temper*	12
carpet	13	*patience*	22	*temple*	38
cartridge	6	*payment*	53	*tender*	11
castle	7	*pedal*	4	*tennis*	15
cocoa	2	*pencil*	34	*terrace*	9
combat	27	*penny*	25	*ticket*	16
comet	2	*perfect*	58	*tidy*	1
compass	13	*pester*	1	*tiger*	7
concert	39	*pigeon*	3	*timber*	19
contact	63	*pillow*	8	*timing*	11
contest	26	*pizza*	3	*token*	10
copper	13	*poison*	10	*tonic*	1
cottage	19	*poker*	6	*topic*	9
courage	32	*poodle*	2	*towel*	6
culture	58	*poster*	4	*tuba*	1
curtain	13	*posture*	13	*tulip*	4
cushion	8	*punish*	3	*tumble*	3
custom	14	*puppy*	2	*tunnel*	10
kennel	3	*puzzle*	10	*turkey*	9
kitten	5			*turtle*	8
M	17.5		14.6		10.7

Note—Adapted from "Imitation in Shadowing Words," by K. Shockley, L. Sabadini, and C. A. Fowler, 2004, *Perception & Psychophysics*, *66*, p. 428. Copyright 2004 by the Psychonomic Society, Inc. Word frequencies based on Kučera and Francis (1967).

(Manuscript received October 28, 2008;
revision accepted for publication April 4, 2010.)