Psychon Bull Rev (2015) 22:1048–1053
DOI 10.3758/s13423-014-0774-3

BRIEF REPORT

# The self-advantage in visual speech processing enhances audiovisual speech recognition in noise

**Nancy Tye-Murray · Brent P. Spehar · Joel Myerson · Sandra Hale · Mitchell S. Sommers**

Published online: 25 November 2014


**Abstract** Individuals lip read themselves more accurately than they lip read others when only the visual speech signal is available (Tye-Murray et al., Psychonomic Bulletin & Review, 20, 115–119, 2013). This self-advantage for vision-only speech recognition is consistent with the common-coding hypothesis (Prinz, European Journal of Cognitive Psychology, 9, 129–154, 1997), which posits (1) that observing an action activates the same motor plan representation as actually performing that action and (2) that observing one's own actions activates motor plan representations more than the others' actions because of greater congruity between percepts and corresponding motor plans. The present study extends this line of research to audiovisual speech recognition by examining whether there is a self-advantage when the visual signal is added to the auditory signal under poor listening conditions. Participants were assigned to sub-groups for round-robin testing in which each participant was paired with every member of their subgroup, including themselves, serving as both talker and listener/observer. On average, the benefit participants obtained from the visual signal when they were the talker was greater than when the talker was someone else and also was greater than the benefit others obtained from observing as well as listening to them. Moreover, the self-advantage in audiovisual speech recognition was significant after statistically controlling for individual differences in both participants' ability to benefit from a visual speech signal and the extent to which their own visual speech signal benefited others. These findings are consistent with our previous finding of a self-advantage in lip reading and with the hypothesis of a common code for action perception and motor plan representation.

**Keywords** Audiovisual speech recognition · Lip reading · Visual speech benefit · Self-advantage · Common coding hypothesis

N. Tye-Murray (✉) · B. P. Spehar
Department of Otolaryngology, Washington University School of Medicine, Campus Box 8115, 660 South Euclid Avenue, St. Louis, MO 63124, USA
e-mail: murrayn@ent.wustl.edu

J. Myerson · S. Hale · M. S. Sommers
Department of Psychology, Washington University in St. Louis, St. Louis, MO, USA

## Introduction

Perhaps surprisingly, people can lip read themselves more accurately than they can lip read other talkers even though they have rarely seen themselves talk (Tye-Murray, Spehar, Myerson, Hale, & Sommers, 2013). This "self-advantage" in vision-only speech recognition is consistent with the common-coding hypothesis, which posits that individuals' perceptions of observed actions and their motor plans share a common code, and that because of this, perceiving an action activates the same representations of motor plans that would be activated if one were actually performing or planning the action (Hommel, Musseler, Aschersleben, & Prinz, 2001; Prinz, 1997). Moreover, because individuals differ in the ways that they perform the same action, one's motor plans are hypothesized to be more strongly activated by perceiving one's own actions than by perceiving the actions of others.

Consistent with this hypothesis, individuals viewing a video clip can recognize previous movements as being self- rather than other-generated, even when little or no identifying information is available other than idiosyncratic aspects of the movements themselves (Repp & Knoblich, 2004). Importantly, the evidence for common coding is not restricted to its role in identifying agency. For example, participants can better predict the outcome of an action from videos of themselves than from videos of others, leading in part to the

suggestion that people's ability to map perceived actions onto their own action repertoire enables direct understanding of such actions (Knoblich & Sebanz, 2006).

Some researchers have argued that the perception of action, and speech perception in particular, entails activation of the very motor representations in the brain necessary to produce the action (Rizzolatti & Craighero, 2004). Visual (Turner, Fridriksson, Baker, Eoute, Bonilha, & Rorden, 2009) as well as auditory (Wilson, Saygin, Sereno, & Iacoboni, 2004) speech stimuli produce such activation, suggesting that similar processes may be involved in both visual and auditory speech recognition. Because individuals have unique motor signatures for speech gestures, just as they do for other actions, there should be greater congruity between visually perceived speech gestures and the corresponding motor plans and associated kinesthetic experiences when the talker and the observer are the same person. This, in turn, should lead to speech motor plans being more activated when individuals see themselves speaking than when they see others speak, and this greater activation may be what underlies people's enhanced ability to lip read themselves (Tye-Murray et al., 2013). We predict that the greater activation of appropriate speech motor plans also should affect people's ability to use visual speech information under audiovisual conditions when listening is difficult. Accordingly, the purpose of the present study was to test our common coding account by asking whether what we term the *visual speech benefit* would be larger (i.e., whether participants would benefit more from adding the visual signal to the auditory signal) when participants were presented with recordings of their own visual and auditory speech signals than when the talker was someone else.

In one of the few studies of how audiovisual speech perception is affected by the talker's identity, Aruffo and Shore (2012) examined the classic McGurk illusion (McGurk & McDonald, 1976), in which presentation of the auditory signal corresponding to one phoneme together with the visual signal corresponding to another phoneme leads to auditory perception of a third phoneme. For example, an auditory /aba/ paired with a visual /aga/ may lead to perception of /ada/ or /ala/. Aruffo and Shore reported participants were less likely to experience the McGurk illusion when the stimuli were recordings of their own auditory and visual speech signals than when the talker was someone else, a result that might be interpreted as being consistent with a self-advantage in audiovisual speech recognition.

Aruffo and Shore (2012) also examined two mismatched conditions in which participants either heard themselves but saw someone else or vice versa. Notably, the McGurk illusion was reduced in the condition in which they heard their own voice but not in the condition where they saw their own face. These results suggest that the identity of the talker may differentially affect processing of visual and auditory speech information, and raises the possibility that participants might not show a greater visual benefit when perceiver and talker are the source of both auditory and visual signals if a difficult listening environment makes listener/observers more reliant on visual speech information.

The present study directly tested the prediction of the common coding hypothesis regarding the benefits of adding visual speech information when auditory speech information is already available. Participants were asked to recognize previously recorded sentences spoken by themselves and by others. There were two experimental conditions: an auditory-only (A-only) condition with speech embedded in noise and an auditory-plus-visual (AV) condition in which the visual speech signal was added. At issue was whether adding the visual signal to the auditory signal would result in greater improvements in participants' performance if they themselves were the talkers than if the talkers were other participants in the study.

## Method

### Participants

Two groups of young adults volunteered to participate. Group 1 (mean age = 23.6 years, SD = 5.3) consisted of ten participants who had previously participated in the Tye-Murray et al. (2013) lip-reading study. Group 2 (mean age = 25.5 years, SD = 1.7) consisted of ten newly recruited participants. Both groups were recruited through the Volunteers for Health program at Washington University School of Medicine in St. Louis. All participants were screened for normal hearing (20 dB HL or better) at octave frequencies between 250 and 8000 Hz using a calibrated audiometer as well as for corrected visual acuity better than 20/30 and for normal visual contrast sensitivity. Participants received $10 per hour for their time and effort.

### Stimuli

The study was conducted in a sound-treated booth with participants positioned approximately 0.5 meters from an ELO Touchsystems monitor on which visual stimuli were presented (e.g., the talker's head and shoulders in the audiovisual condition). Auditory stimuli were presented over two loudspeakers positioned at±45 degrees azimuth.

All test stimuli were based on audiovisual recordings of the participants speaking items from the *Build-A-Sentence* (BAS) test (Tye-Murray, Sommers, & Spehar, 2007). The BAS material consisted of lists of sentences created based on random combinations of the same 36 high-frequency keywords (mean log frequency = 9.8, SD = 1.4; Balota et al., 2007). Each list consisted of three two-, six three-, and three four-keyword sentences (e.g. The *bird* watched the *boys*; The *dog* and the

*girls* watched the *snail*; The *boys* and the *cow* watched the *saint* and the *bird*; for more details regarding the BAS, see the Appendix of Tye-Murray et al., 2013). For testing, the BAS sentences were embedded in four-talker babble presented at 60 dB SPL, and the audio for the sentences themselves was set individually for each participant to a level determined during the pre-testing phase of the study (see Procedure).

For Group 1 (the returning participants), the lists of sentences used as stimuli in the present study were derived from the audiovisual recordings of lists of BAS sentences that they had read for the Tye-Murray et al. (2013) study more than 18 months previously. To prevent participants from remembering those sentences, we originally had them record over 30 lists (more than 360 sentences), on average, of which only two were used as stimuli in that study. From this previously recorded material, we selected audio clips that the participants had never heard before and audiovisual clips that they had never heard or seen before for use as stimuli.

For Group 2 (the new participants), stimuli were created in a two-hour recording session that took place at least four weeks before their test session. The recording set-up, teleprompter, and video processing were the same as described by Tye-Murray et al. (2013). To control for variability in speaking levels, the audio track for each sentence was first extracted from the original audiovisual recording and then all sentences were leveled for amplitude based on RMS using Adobe Audition. The audio tracks were then re-combined with the video tracks using Adobe Premiere Elements. All participants in Group 2 recorded 48 lists (576 sentences), of which only two were used as test stimuli. As in the previous study, the bulk of the recordings were not used during testing but were only made in order to reduce the possibility that participants would remember specific sentences that they had recorded.

## Procedure

For purposes of round-robin testing, Groups 1 and 2 were further divided into two subgroups each, and each participant in a subgroup was paired with every member of their subgroup (including themselves) twice: once as a talker and again as a listener/observer. None of the participants in any of the subgroups reported knowing each other, and therefore testing of all participants began with presentation of audio and audiovisual recordings of sample sentences uttered by all of those in their subgroup (including themselves). Group 1 was divided into one subgroup of six participants (two females), all of whom had been members of the same group in the Tye-Murray et al. (2013) study, and another subgroup of four (two females), who had been members of the other group in that study. Group 2 was divided into subgroups of five each (two females in one and three in the other).

The two-hour test session consisted of two phases: pre-testing and data collection. In both phases, sentences were presented in a closed-set format with the 36 keywords appearing on the screen after each sentence was presented, and participants were instructed to respond by repeating the sentence aloud and to make guesses from the list of words on the screen whenever they were not sure. No cue regarding the length of the sentence was given. The experimenter, who was outside the booth and could not see the test screen, entered their responses into the computer before the next sentence was presented. Scores were based on the percentage of total keywords correctly identified. Credit was given for a word even if it was recalled in the wrong position within the sentence.

The purpose of the pre-testing phase was to establish the signal-to-babble ratios (SBRs) needed so that each participant could perform at approximately 30 % accuracy in the A-only condition of the subsequent data collection phase. This level of accuracy was chosen to ensure that performance would be above floor in the A-only condition of the data collection phase, and that when the visual speech signal was added in the AV condition, performance would be below ceiling. This was necessary because the presence of either floor or ceiling effects would preclude accurate estimation of the visual speech benefit (AV accuracy minus A-only accuracy).

To determine the SBRs needed, we used a two-down, one-up staircase procedure programmed in LabView in which the level for each talker in the participant's subgroup was varied while the level of background babble was held constant. After three reversals for a specific talker, the program continued to present that talker's recordings at the determined level so that the experimenter could determine whether the participant was achieving approximately 30 % accuracy, and if not, adjust the levels further.[1] Before testing began, participants were given practice with each talker, including themselves, in each condition in order to familiarize them with the other talkers and to reduce the novelty of seeing themselves talking. In the data collection phase, participants were presented with A-only and AV stimuli from each of the talkers in their group, including themselves, in a quasi-random order with the constraint that two clips of the same talker were not presented in a row. The

[1] For Group 1, the adjusting procedures used in the pre-testing phase to determine the SBRs for the data collection phase calculated accuracy as the percentage of sentences for which all of the keywords were correctly identified, whereas the program used in the data collection phase calculated accuracy as the percentage of keywords correctly identified. This mismatch resulted in accuracy levels that were too close to ceiling once visual speech information was added (i.e., in the audiovisual condition). Therefore, participants in Group 1 received further testing with different sentences at SBRs that were 4 dB harder than those determined based on their performance in the pre-test phase. Four participants were recalled for a 30-minute session to complete testing at the harder level. This problem was identified and the mismatch in the program was fixed prior to the testing sessions for Group 2, and no extra testing was necessary for them. Despite the differences in procedure, however, both groups showed the same pattern of results (i.e., greater visual benefit when the talker and the participant were the same person; see Fig. 1).

same individually determined SBRs were used in both the A-only and AV conditions.

## Results

Preliminary analyses revealed that the data met the prerequisites for accurate estimation of the visual speech benefit. On average, participants in Groups 1 and 2 scored 27.1 % words correct (SD = 7.0) in the A-only condition of the testing phase when they were the talker and 32.9 % (SD = 5.4) when another participant was the talker, values that were close to the target value of 30 %. Interestingly, the SNRs established in the pre-testing phase were significantly greater when the talker was the participant him- or herself (−11.1 dB) than when the talker was another participant (−9.9 dB) ($t$ (19) = 4.07, $p < .001$), which may have contributed to the difference in A-only scores ($t$ (19) = 3.19, $p < .01$). In the AV condition, participants recognized 77.1 % words correct (SD = 12.0) when they were the talker and 74.8 % correct (SD = 7.3) when another participant was the talker, values indicating that AV performance was below ceiling, a difference which was in the predicted direction but was itself not significant ($t$ (19) = 1.05, p > .05).

The goal of the present study was to determine whether there was a self-advantage in the visual speech benefit (AV accuracy minus A-only accuracy), as would be predicted by the common currency hypothesis and our previous finding of a self-advantage in lip reading (Tye-Murray et al., 2013). In this regard, it is important to note that at the individual level, the visual speech benefit depends on at least two kinds of individual differences: differences in the quality of the visual speech signal produced by the talker and differences in the lip-reading ability of the listener/observer (Tye-Murray et al., 2013), which add measurement error unless they are controlled, either experimentally or statistically. Therefore, we compared the visual speech benefit that individual participants obtained when they were both the talker and the listener/observer with (1) the benefit they obtained when they were the listener/observer but the talker was someone else (see the left panel of Fig. 1), and (2) the benefit they obtained when they were the talker but the listener/observer was someone else (see the right panel of Fig. 1). Both kinds of comparisons revealed that the visual speech benefit is significantly greater when the talker and the listener/observer are the same person: On average, the benefit a participant obtained when he or she was the talker was greater than the benefit he or she obtained when others were the talker, $t$(19) = 2.74, $p$ = .013, as well as being greater than the benefit others obtained when the participant was the talker, $t$(19) = 2.78, $p$ = .012.

In addition, we developed a novel analytic approach that uses multiple regression to simultaneously control for both kinds of individual differences that are involved in the visual speech benefit. In the current context, an individual's lip-reading ability may be inferred from the visual speech benefit he or she obtained when the talker was someone else, and the quality of an individual's own visual speech signal may be inferred from the benefit others obtain from seeing the individual's face as well as hearing the person's voice. In the following equation, these abilities are represented by *OS* and *SO*, respectively, and the predicted visual speech benefit when the talker and the listener/observer are the same individual is represented by *SS*, where the first letter (*S* for Self or *O* for Other) indicates the

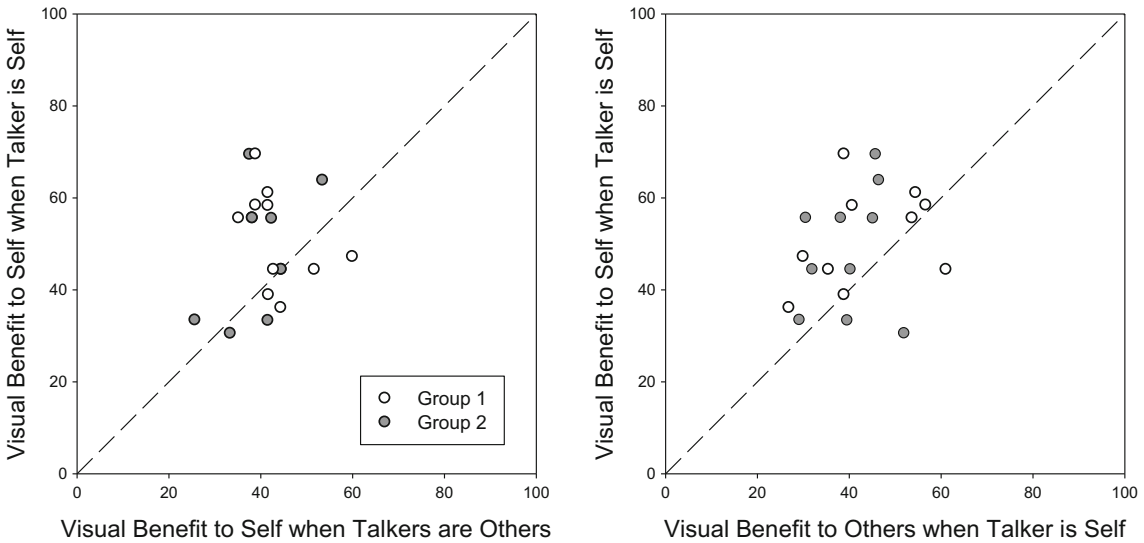


**Fig. 1** Scatterplots comparing participants' visual speech benefit when they were the talker with the visual benefit when other participants were the talker (left panel) and also with other participants' visual benefit when they were the talker. Results are shown for members of Group 1 and 2 separately, and as may be seen, despite differences in procedure and the makeup of the groups, similar patterns of results were observed

identity of the talker and the second letter indicates the identity of the listener/observer:

$$SS = a_{*}OS + b_{*}SO + c,$$

where $a + b = 1$.

The null hypothesis is that the visual speech benefit that individuals obtain when they are both the talker and the listener/observer depends only on these two abilities (i.e., c = 0); the hypothesis to be tested is that when individuals are both the talker and the listener/observer, they will benefit more than would be predicted based on their lip-reading ability and the quality of their visual speech signal alone (i.e., c > 0). The constraint ($a + b = 1$) reflects the fact that under the null hypothesis, the predicted benefit is simply the weighted average of the observed benefits used to infer the abilities involved.

To test our hypothesis, we first estimated the visual speech benefit (*SS*) predicted under the null hypothesis ($SS = 0.484_{*}OS + 0.516_{*}SO$) and then calculated the difference between the observed and predicted values. In 16 out of the 20 cases, the benefit a participant received when he or she was both the talker and the listener/observer was greater than the benefit predicted taking into account both the benefit he or she received when the talker was someone else and the benefit others received when the participant was the talker, $t(19) = 3.19$, $p = .005$.

## Discussion

Taken together, the present findings indicate that adding the visual speech signal to the auditory signal results in greater improvements in individuals' speech recognition performance (i.e., the visual speech benefit is larger) if the talker and the listener/observer are the same person, as predicted by the common-coding hypothesis (Prinz, 1997). Three analyses revealed this self-advantage in audiovisual speech recognition. First, on average, participants obtained greater visual speech benefits when they were the talkers than when the talker was someone else. Second, participants received greater visual speech benefits when they were the talkers than others obtained from observing and listening to them. Third, a final analysis revealed a self-advantage in audiovisual speech recognition after statistically controlling for individual differences in both participants' ability to benefit from a visual speech signal and the extent to which their own visual speech signal afforded benefit to others, factors that logically must both play a role in determining the size of individual visual speech benefits.

These findings suggest that a strong linkage exists between motor and sensory speech representations, and that this linkage is important for speech recognition. Although some researchers (Hickok, Houde, & Rong, 2011) believe that the primary purpose of this linkage is to inform action rather than perception, as in providing online auditory feedback during ongoing speech production, they acknowledge that activated motor representations may inform speech perception under difficult perceptual conditions like those studied here.

One possibility is that activation of sensory representations spreads to activation of motor representations, consistent with the common-coding hypothesis. A growing body of studies using a variety of experimental approaches provide converging evidence for this view. Neuroimaging studies, for example, have revealed that cortical motor areas are not only activated during speech production but also when listening to speech (Wilson et al., 2004). Not only does watching silent videos of oral speech movements produce such activation, it also produces more activation than when watching non-speech oral movements (Turner et al., 2009), suggesting that these regions are not part of a generalized mirror neuron system, but are specialized for the production and perception, both auditory and visual, of speech stimuli.

Further neurophysiological evidence of linkage between the systems underlying speech production and perception, both auditory and visual, comes from the finding that the size of motor potentials in the lips (Watkins, Strafella, & Paus, 2003) and tongue (Fadiga, Craighero, Buccino, & Rizzolatti, 2002) elicited by transcranial magnetic stimulation of the corresponding regions of the motor cortex are increased both while listening to speech and while viewing oral speech movements. Thus, it seems likely that the speech stimuli in the present study, both visual and auditory, also elicited simultaneous activation of sensory and motor representations, and that consistent with the common coding hypothesis, participants benefited more from their own visual speech signal because it, and perhaps the auditory signal as well, corresponded better with the representations of their own motor patterns than did the visual and auditory speech signals of others.

The present results suggest a number of research questions that could shed further light on the self-advantage for speech recognition. For example, what role does a person's intentions play? In the present study, participants were prepared to respond by saying what they had just seen and heard. It is possible that this intention facilitated activation of participants' speech motor plans by visual and auditory stimuli and that the self-advantage effect was heightened as a result. If so, then little or no self-advantage might be observed if participants had to respond by touching or clicking on the appropriate items from among a list displayed on a screen rather than by saying them aloud. Another important issue concerns how long a speech sample must be for a self-advantage to be observed (e.g., will a single nonsense syllable suffice or do results like those reported here emerge only with multi-syllabic or multi-word stimuli)? The answer to this question is important because knowing the time course of this effect could help in linking phenomena at the behavioral and neural

levels and might inform theories of speech production regarding the units of motor planning.

In addition, however, the present findings may have implications for our understanding of the effects of talker variability on speech recognition that go well beyond the self-advantage per se. This is because the difficulty of recognizing a talker's utterances likely depends in part on where the talker's utterances fall along the dimension of similarity to the utterances of the listener/observer, rather than simply on categorical perception of the talker's identity (self vs. non-self). Talker similarity has been shown to be important for spoken language processing (e.g., Goldinger, 1996), and previous research has examined the dimensions underlying auditory aspects of talker similarity (e.g., Baumann & Belin, 2010; Walden, Montgomery, Gibeily, Prosek, & Schwartz, 1978). So far as we know, there have been no tests of the prediction of the common coding hypothesis of a self-advantage in auditory speech recognition, perhaps because of the difficulty of ruling out familiarity as a cause were such an advantage to be observed. The situation is different with respect to visual speech recognition, of course, because most individuals are relatively unfamiliar with the sight of themselves talking.

The question of whether there is a self-advantage in auditory speech recognition is an important, albeit difficult, one that deserves researchers' attention. The present results can shed little light on this issue, however, because in addition to the auditory familiarity confound, the design is not well suited for comparing different A-only performances in that comparisons of different talkers would typically involve different SNRs. Nevertheless, the present design is appropriate for the purpose for which it was intended, looking at the visual speech benefit, because for each participant, the SNR for each talker was the same in both A-only and AV conditions so that the benefit was always calculated based on comparisons across the same SNRs.

To the best of our knowledge, there have been no previous studies of visual similarity among talkers. Such research is needed because determining which aspects of a talker's visual speech gestures are the most important determinants of such similarity may shed light on the processes underlying "resonance," the activation of motor plans in response to observing the actions of others (Hommel et al., 2001), and thus on why an individual listener/observer may find some talkers' utterances easier to recognize than others' utterances. Pair-specific resonance could also play a role in the perceptual costs of switching between talkers, a well-established phenomenon with respect to auditory speech recognition and one that recent research suggests may be associated with even larger effects on audiovisual speech recognition (Heald & Nusbaum, 2014). In short, variability in the intelligibility of speech signals is widespread, and the finding of a self-advantage in the recognition of such signals suggests an approach, based on the common coding hypothesis, to understanding the role of individual differences in such variation.

**Acknowledgments** This research was supported by NIH grant AG018029. We thank Kristin Van Engen for her helpful comments, and Elizabeth Mauze and Shannon Sides for their assistance in preparing the stimuli and testing the participants.

## References

Aruffo, C., & Shore, D. I. (2012). Can you McGurk yourself? Self-face and self-voice in audiovisual speech. *Psychonomic Bulletin & Review, 19,* 66–72.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B.,…Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*, 445–459.

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research, 74,* 110–120.

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.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1166–1183.

Heald, S., & Nusbaum, H. C. (2014). Talker variability in audiovisual speech perception. *Frontiers in Psychology, 5,* 698.

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: Computational basis and neural organization. *Neuron, 69,* 407–422.

Hommel, B., Musseler, J., Aschersleben, G., & Prinz, W. (2001). The Theory of Event Coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24,* 849–937.

Knoblich, G., & Sebanz, N. (2006). The social nature of perception and action. *Current Directions in Psychological Science, 15,* 99–104.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology, 9,* 129–154.

Repp, B. H., & Knoblich, G. (2004). Perceiving action identity: How pianists recognize their own performances. *Psychological Science, 15,* 604–609.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27,* 169–192.

Turner, T. H., Fridriksson, J., Baker, J., Eoute, D., Bonilha, L., & Rorden, C. (2009). Obligatory Broca's area modulation associated with passive speech perception. *Neuroreport, 20,* 492–496.

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Audiovisual integration and Lipreading abilities of older adults with normal and impaired hearing. *Ear and Hearing, 28,* 656–668.

Tye-Murray, N., Spehar, B. P., Myerson, J., Hale, S., & Sommers, M. S. (2013). Reading your own lips: Common-coding theory and visual speech perception. *Psychonomic Bulletin & Review, 20,* 115–119.

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia, 41,* 989–994.

Walden, B. E., Montgomery, A. A., Gibeily, G. J., Prosek, R. A., & Schwartz, D. M. (1978). Correlates of psychological dimensions in talker similarity. *Journal of Speech, Language, and Hearing Research, 21,* 265–275.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7,* 701–702.