JSLHR

Research Article

Engaging the Articulators Enhances Perception of Concordant Visible Speech Movements

Matthew Masapollo[a] and Frank H. Guenther[a,b]

Purpose: This study aimed to test whether (and how) somatosensory feedback signals from the vocal tract affect concurrent unimodal visual speech perception.
Method: Participants discriminated pairs of silent visual utterances of vowels under 3 experimental conditions: (a) normal (baseline) and while holding either (b) a bite block or (c) a lip tube in their mouths. To test the specificity of somatosensory–visual interactions during perception, we assessed discrimination of vowel contrasts optically distinguished based on their mandibular (English /ɛ/–/æ/) or labial (English /u/–French /u/) postures. In addition, we assessed perception of each contrast using dynamically articulating videos and static (single-frame) images of each gesture (at vowel midpoint).
Results: Engaging the jaw selectively facilitated perception of the dynamic gestures optically distinct in terms of jaw height, whereas engaging the lips selectively facilitated perception of the dynamic gestures optically distinct in terms of their degree of lip compression and protrusion. Thus, participants perceived visible speech movements in relation to the configuration and shape of their own vocal tract (and possibly their ability to produce covert vowel production–like movements). In contrast, engaging the articulators had no effect when the speaking faces did not move, suggesting that the somatosensory inputs affected perception of time-varying kinematic information rather than changes in target (movement end point) mouth shapes.
Conclusions: These findings suggest that orofacial somatosensory inputs associated with speech production prime premotor and somatosensory brain regions involved in the sensorimotor control of speech, thereby facilitating perception of concordant visible speech movements.
Supplemental Material: https://doi.org/10.23641/asha.9911846

The central goal of research on sensorimotor integration for speech processing is to explicate the mechanisms of perception, how perception influences articulatory and phonatory movements, and how those movements, in turn, affect perception (Hickok, Houde, & Rong, 2011). Work to date has firmly established that sensory feedback signals play a critically important role in guiding and coordinating speech movements (see Guenther, 2016, Chapters 5 and 6, for a thorough review). Over the years, numerous studies have consistently demonstrated that speakers (both adult and child) automatically adjust their movements of unimpeded articulators to compensate for unexpected perturbations in auditory (e.g., Abur et al., 2018; Cai et al., 2012; Golfinopoulos, Tourville, Bohland, Ghosh, & Guenther, 2011; Houde & Jordan, 1998; MacDonald, Johnson, Forsythe, Plante, & Munhall, 2012; Mollaei, Shiller, Baum, & Gracco, 2016; Stuart, Kalinowski, Rastatter, & Lynch, 2002; Villacorta, Perkell, & Guenther, 2007) or somatosensory feedback (e.g., Nasir & Ostry, 2006; Tremblay, Shiller, & Ostry, 2003). Thus, the evidence indicates that speakers effectively monitor their own self-generated auditory and somatosensory feedback online to guide, correct, and fine-tune vocal production parameters.

Moreover, and of particular relevance to the present research, there is now a growing body of evidence that orofacial somatosensory inputs associated with speech production can feed back to influence concurrent speech perception (Bruderer, Danielson, Kandhadai, & Werker, 2015; Ito, Tiede, & Ostry, 2009; Sams, Möttönen, & Sihvonen, 2005; Sato, Troille, Ménard, Cathiard, & Gracco, 2013; Yeung & Werker, 2013). To take one well-cited example, Ito et al. (2009) created a "head"–"had" auditory series, in which the "head" (/hɛ̲d/) versus "had" (/hæ̲d/) distinction was specified by small, incremental changes in the first and

[a]Department of Speech, Language and Hearing Sciences, Boston University, MA
[b]Department of Biomedical Engineering, Boston University, MA
Correspondence to Matthew Masapollo: mmasapol@bu.edu

Editor-in-Chief: Bharath Chandrasekaran
Editor: Kate Bunton
Received April 17, 2019
Revision received June 27, 2019
Accepted June 28, 2019
https://doi.org/10.1044/2019_JSLHR-S-19-0167

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

second formant frequencies of the vocalic portion of the signal. They then presented a group of listeners with randomized sequences of the series and asked them to identify each member as an exemplar of the word *head* or *had*. While the subjects listened to the members of the series, the researchers used a robotic device to stretch their facial skin on each side of the mouth in either an upward or downward direction. During a baseline (control) condition, listeners consistently identified stimuli on one side of the continuum as /hɛd/ and those one on the other side of the continuum as /hæd/. However, the location of the "head"–"had" boundary systematically shifted in the experimental conditions depending on the direction of the skin stretch perturbation. Specifically, when subjects' skin was stretched and perturbed upward, as is normally evoked during the production of vowel /ɛ/, they were more likely to report "hearing" the spectrally ambiguous members in the middle of the series as the word *head* than *had*. Conversely, when their skin was stretched downward, as is normally evoked during the production of the vowel /æ/, subjects were more likely to report hearing the spectrally ambiguous members of the series as the word *had* than *head*. In other words, the somatosensory feedback they received from the facial skin deformation biased perceptual judgments toward the concordant articulations (i.e., an *assimilation* effect).

Other evidence in favor of action's effect on speech perception comes from studies showing that silently articulating syllables improves concurrent perception of concordant syllables but not discordant syllables (Sams et al., 2005; Sato et al., 2013). For example, in one experiment, Sams et al. (2005) instructed subjects to silently mouth either the syllable "pa" or "ka" in synchrony with an acoustic stimulus specifying a consonant–vowel syllable, and then identify the syllable they heard. The results showed that there was an effect of the uttered syllable on perception. Subjects were more accurate at identifying the acoustic stimulus when the uttered syllable was phonetically congruent with the heard syllable. Conversely, subjects were less accurate when the uttered syllable was phonetically incongruent with the heard syllable (see Sato et al., 2013, for similar results). Sams et al. (2005) invoked the analysis by synthesis (A × S) theory of speech perception (Kuhl, Ramirez, Bosseler, Lin, & Imada, 2014; Skipper, van Wassenhove, Nusbaum, & Small, 2007; Stevens, 1960, 2002) and suggested that the subjects were internally simulating the sensory consequences of the produced vocal tract maneuvers, which, in turn, biased perception toward the concordant syllable in the form of an efference copy from the motor system (i.e., a *facilitation* effect).

There are other reports, however, in the developmental literature that somatosensory inputs from the vocal tract can inhibit, rather than facilitate, concurrent speech perception (Bruderer et al., 2015; Yeung & Werker, 2013). For example, Bruderer et al. (2015) reported that preventing articulatory maneuvers consistent with what speech sounds were being perceived interfered with phonetic discrimination in preverbal infants. Specifically, 6-month-old English-learning infants were found to discriminate a nonnative Hindi dental-retroflex stop contrast (/d̪/-/ɖ/) while sucking on a "gum-teether" pacifier that allowed free tongue motion but failed when sucking on a "flat-tongue" pacifier that prevented tongue tip motion. (Note that these researchers used ultrasound methodology to confirm the hypothesized teether effects on tongue position and motion.) This result was especially surprising because infants this age have not yet acquired the fine-grained articulatory control necessary to produce such segments.[1] One possible explanation of these findings is that the simultaneous activation of motor and perceptual representations for speech leads to the inhibition of those representations during concurrent perception (Galantucci, Fowler, & Goldstein, 2009; cf. Sams et al., 2005; Sato et al., 2013; Yeung & Werker, 2013). However, Hickok and Poeppel (2016) argued for an alternative explanation, namely, that the flat-tongue teether may have drawn more attention away from the task than the gum-teether, which, in turn, led to the failure to discriminate. Whatever the reason for the inhibition effect reported by Bruderer et al. may turn out to be, the aforementioned findings raise the intriguing possibility that speech is perceived in relation to the shape and configuration of one's own vocal tract and ability to act, even prior to the acquisition of well-specified speech production targets.

Whereas these and other psychophysical experiments have demonstrated complex somatosensory–auditory interactions during phonetic perception at a behavioral level, neuroimaging studies indicate that visual speech cues in talking faces influence blood oxygen level–dependent (BOLD) responses in premotor cortex, primary motor cortex, and somatosensory cortex above and beyond acoustic speech cues alone (Matchin, Groulx, & Hickok, 2014). Specifically, Matchin et al. (2014) used functional magnetic resonance imaging methods to examine BOLD activity patterns while subjects passively listened to or lipread a speaker silently talk (with no overt motor task). These researchers found inferior frontal gyrus (pars opercularis), dorsal motor cortex, and inferior parietal lobe to be more active during the lipreading task than the listening task. Skipper, Nusbaum, and Small (2005) also reported that activity in premotor and primary motor cortical regions during bimodal (auditory–visual) speech perception was modulated by the visual salience of speech stimuli. Furthermore, Sundara, Namasivayam, and Chen (2001) demonstrated using transcranial magnetic stimulation that perception of (silent) visual speech, but not acoustic speech, elicits enhanced motor-evoked potentials in the vocal tract muscles recruited to articulate speech. Thus, understanding the contribution of potential somatosensory–visual interactions during speech processing may yield additional key insights into action effects on phonetic perception.

The purpose of the present research was to investigate whether, and if so, how, the somatosensory system is

[1]A complementary result was obtained by Yeung and Werker (2013), who found that 4- to 5-month-old infants' ability to cross-modally match audiovisual vowels was also disrupted by teething toys that constrained the shape and movements of the lips.

involved in the perceptual processing of unimodal visual speech. We addressed this question by examining whether engaging the articulators influences concurrent discrimination of visual speech using either dynamically articulating videos (Experiment 1) or still pictures (Experiment 2) of a speaker. Previous studies have shown that perceivers have some ability to lipread from photographs of faces (Rosenblum, 2005), and that the processing of dynamic and static visual speech cues is carried out by similar neural substrates (Calvert & Campbell, 2003). A finding that facial somatosensory inputs modulate perception of dynamic, but not static, facial displays would indicate that the somatosensory system is especially involved in tracking time-varying characteristics of seen speech. In contrast, a finding that proprioceptive inputs modulate perception of both dynamic and static visual speech would suggest that the somatosensory system is involved in extracting configural information about the filter state of the vocal tract.

To test the specificity of somatosensory–visual interactions during phonetic perception, we also experimentally manipulated the position of subjects' lips or jaw and tested whether this affected their discrimination of two optically distinct vowel contrasts (see details below) that involve dramatic movements of either the lips or the jaw in their production. Subjects discriminated both contrasts under one of three experimental conditions: (a) normal (baseline, i.e., no oral–motor manipulation) and while holding either (b) a bite block between the upper and lower teeth or (c) a tube between the lips. If there are somatosensory–visual interactions during visual speech perception, then engaging the jaw should selectively influence (i.e., facilitate or inhibit) discrimination of gestures optically distinguished by their mandibular postures, whereas engaging the lips should selectively influence discrimination of gestures optically distinguished by their labial postures. In addition, if engaging the articulators selectively affects how perceivers track orofacial speech movements, rather than changes in target mouth shapes, then there should only be an effect of condition during discrimination of the dynamic facial displays (Experiment 1), but not the static facial displays (Experiment 2). Alternatively, if engaging the articulators influences perception of both configural and time-varying phonetic information, then there should be an effect of condition across both experiments, regardless of the type of facial displays used. Yet another possibility is that simultaneously engaging the articulators during concurrent perception may simply increase attentional processing load, which, in turn, will lead to a decline in overall discrimination performance, regardless of which specific articulator is activated.

Experiment 1

Materials and Method

Subjects

Forty-eight participants (11 males; age range = 18–32 years, M = 21.2 years, SD = 3.1) from Boston University completed this experiment for pay. All were native, monolingual American English speakers who reported normal hearing, normal (or corrected-to-normal) vision, and no history of speech, language, or other neurological disorder.

Stimuli

As already mentioned, two vocalic contrasts were selected to provide maximal opportunity for observing somatosensory–visual interactions during phonetic perception. Specifically, we used close back rounded English /u/ versus French /u/, and open-mid front English /ɛ/ versus near-open front English /æ/. Previous cross-language vowel production studies (e.g., MacLeod, Stoel-Gammon, & Wassink, 2009; Noiray, Cathiard, Ménard, & Abry, 2011) have demonstrated that French speakers produce more extreme /u/ gestures, with a greater degree lip rounding (lip compression and protrusion) and tongue backness, compared to English speakers. Consequently, English /u/ and French /u/ are optically distinct in terms of their lip postures. As for the other contrast, English /ɛ/ and /æ/ gestures are optically distinct in terms of their mandibular position; the production of /æ/ involves a greater degree of jaw lowering than /ɛ/.

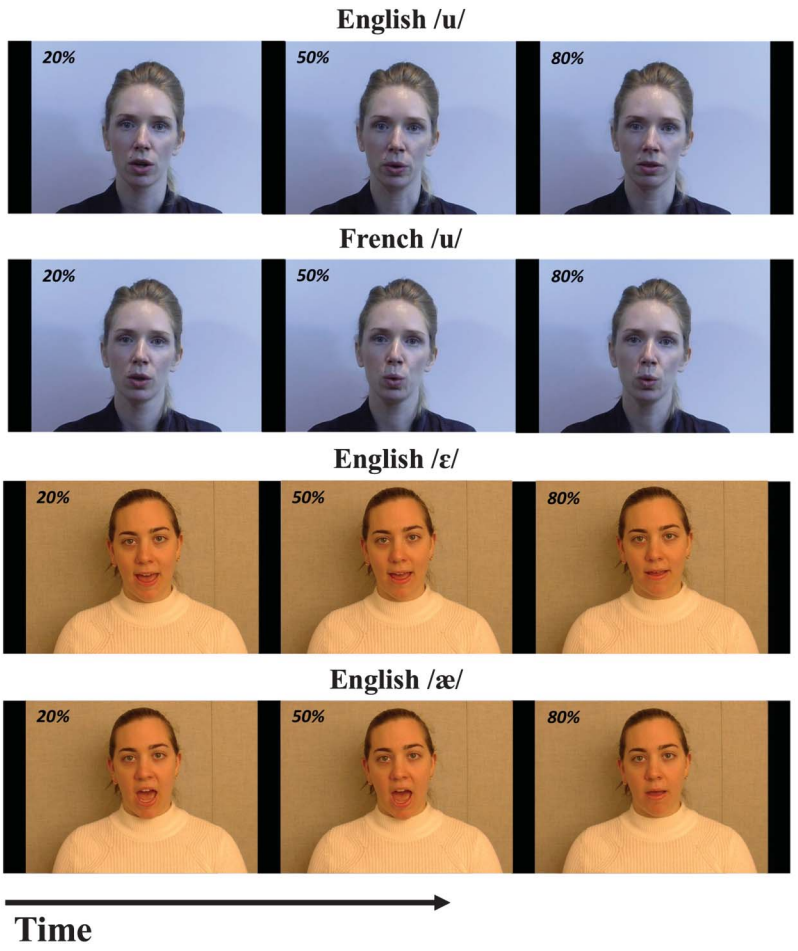
The stimulus vowels were produced by a native female speaker of the source languages (a simultaneous English–French bilingual speaker from Montréal, Québec, and an English monolingual speaker from Austin, TX, respectively). We recorded both model speakers producing stop-initial consonant–vowel (/gV) syllables instead of isolated vowels to facilitate cross-stimulus splicing for other bimodal (auditory–visual) vowel perception experiments. The speakers were instructed to produce clear and distinct vowels embedded at the end of the carrier phrase, "I'm going to tell you about ____." The speakers produced multiple runs of each vowel; each run consisted of 10 repetitions of the target vowel in the carrier phrase. The productions were audiovisually recorded using a digital camcorder (Panasonic AG-DVX100B; 29.97 frames/s and 1,400 × 1,000 pixels; audio at 44.1 kHz) from a straight, face-on view in a sound-treated booth.

The duration, fundamental frequency, and first and second formant frequencies of the vowel in each of the recorded syllables were measured using Praat (Boersma & Weenink, 2019). Five different video tokens of each vowel were selected as stimuli based on their visual similarity in head position and facial expression. In addition, we selected tokens that were roughly matched in the duration of their vocalic portions. Figure 1 shows example video frames of the visible vocal tract configuration of the model speakers during the production of each vowel type (at 20%, 50%, and 80% of the acoustic vowel duration). Using Adobe Premiere (San Jose, CA), the video-only stimuli were created by removing the audio track from the audiovisual video recordings of the model speakers' productions.

Procedure and Design

Subjects completed a categorical same–different (AX) discrimination test for each of the two vowel contrasts in a

Figure 1. Sample images of the model speakers' visible vocal tract configuration during the production of each vocalic gesture at 20%, 50%, and 80% of vowel duration. Note that, in Experiment 2, the stimuli only consisted of static (single-frame) images of the talking faces taken at 50% of vowel duration (as shown in the center panels). As the images show, French /u/ is executed with a greater degree of visible lip compression and protrusion, and English /æ/ is implemented with a lower mandibular position than English /ɛ/.



unimodal visual-only mode (see, e.g., Masapollo, Polka, & Ménard, 2017, Experiment 2). The order of the two contrasts was counterbalanced to counteract potential fatigue effects. Prior to the start of each AX test, subjects were informed that they would see a speaker articulating two different types of vowels, and that their task was to try to differentiate between these two different types of vowels. They were also told that each sequence contained either two different instances of the same vowel type (same pairs) or instances of two different vowel types (different pairs). On each trial, subjects watched a sequence of two unimodal viseme tokens, and then judged whether they were the "same" or "different" by pressing one of two buttons on a response pad. For each same trial, different tokens of the same vowel type were paired (e.g., two different English /ɛ/ tokens or two different English /æ/ tokens were paired). For each different trial, tokens from the two different vowel types were paired (e.g., an English /ɛ/ token was paired with an English /æ/ token). Thus, subjects had to indicate whether pairs of physically different stimuli were members of the same vowel set or members of the two different vowel sets.

Each of the two AX tests contained 180 trials organized into two blocks. Subjects saw every possible type of pairing of the 10 tokens per stimulus set, separated by an interstimulus interval (ISI) of 1,500 ms in both presentation orders. Each block had 90 trials, which consisted of each possible pairing (i.e., 50 different-type trials and 40 same-type trials). Because these within-category pairs did not consist of physically identical pairings, subjects had to generalize across small optical differences to perceptually group (or categorize) the stimuli. Several practice trials were included at the start of the experiment to confirm that subjects understood the instructions and were able to perform the task. Subjects took a short break after they completed the first AX test. No feedback was provided.

To test the specificity of a somatosensory influence on visual vowel discrimination, subjects were randomly

assigned to one of three experimental conditions (16 in. each): baseline without any oral–motor manipulations, with a tube inserted between the lips, or with a bite block inserted between the upper and lower teeth. The tube was intended to selectively restrict lip movements, whereas the bite block was intended to selectively restrict mandibular movements. The subjects in the lip tube condition were instructed to hold a PVC pipe (2.7 mm in diameter; 4.2 mm in length) between their lips while keeping their lips in a fixed, rounded (i.e., compressed and protruded) position. The subjects in the bite block condition were instructed to hold an athletic mouth guard (Under Armour, Baltimore, MD) between their upper and lower teeth while keeping their jaw in a fixed, closed position. Given these articulatory configurations, we hypothesized that, if there is a visual–somatosensory interaction during concurrent speech perception, then engaging the lips would influence perception of the English /u/–French /u/ contrast (relative to baseline), whereas engaging the jaw would influence perception of the English /ɛ/–/æ/ contrast (relative to baseline). Subjects in the baseline group were given no explicit instruction to impede oral-facial movements, and thus were free to move their articulators in whatever spontaneous manner they chose.[2]

Subjects were tested individually in a quiet laboratory room that was dimly lit. The experiment was programmed using the SuperLab 5.0 software package (Cedrus Corporation, San Pedro, CA), which controlled the presentation of the stimuli, and collected subjects' responses. The stimuli were presented on a 22-in. flat screen monitor about 0.58 m (23 in.) in front of the subject.

Data Analysis

We employed a signal detection theory analysis to assess perceptual sensitivity; the dependent measure was *A*-prime (Grier, 1971). A' is an unbiased index of discrimination performance that ranges from .50 (chance) to 1.0 (perfect discrimination). The following formula (from Grier, 1971) was used to compute each score: $A' = 0.5 + (H - FA)(1 + H - FA)/[4H(1 - FA)]$, where H = proportion of hits (i.e., the proportion of trials in which subjects correctly responded to a category difference between two vowel stimuli) and FA = proportion of false alarms (i.e., the proportion of trials in which subjects incorrectly responded to a category difference between two vowel stimuli). The data were used to calculate separate A' scores for each subject for each vowel contrast (English /u/–French /u/ vs. English /ɛ/–/æ/) in each experimental condition (baseline vs. lip tube vs. bite block).

Results

Subjects' mean A' scores as a function of experimental condition (baseline vs. bite block vs. lip tube) and vowel contrast (English /u/–French /u/ vs. English /ɛ/–/æ/) are presented in Figure 2. To examine whether there was a visual–somatosensory interaction during discrimination, these scores were submitted to a 3 × 2 mixed analysis of variance (ANOVA) with experimental condition (baseline vs. lip tube vs. bite block) as a between-subjects factor, and vowel contrast (English /u/–French /u/ vs. English /ɛ/–/æ/) as a within-subject factor. A significant main effect of vowel contrast was observed, $F(1, 45) = 40.439$, $p < .001$, $\eta^2 = .473$, such that discrimination was better for the English /ɛ/–/æ/ contrast ($M = .89$, $SD = .06$) compared to the English /u/–French /u/ contrast ($M = .82$, $SD = .08$), likely indicating that it is easier to discriminate a cross-category viseme contrast than a within-category viseme contrast. The effect of condition did not reach statistical significance, $F(2, 40) = 2.744$, $p = .075$, $\eta^2 = .109$. Critically, however, there was a highly significant interaction between condition and vowel contrast, $F(2, 45) = 6.193$, $p = .004$, $\eta^2 = .216$. Post hoc *t* tests revealed that the mean A' scores for the English /u/–French /u/ contrast were significantly higher in the lip tube condition ($M = .87$, $SD = .05$; $t(30) = 2.905$, $p = .007$, Cohen's $d = 1.09$), but not in the bite block condition ($M = .80$, $SD = .08$; $t(30) = 0.285$, $p = .777$), when compared to baseline ($M = .79$, $SD = .09$). In contrast, the mean A' scores for the English /ɛ/–/æ/ contrast were marginally higher in the bite block condition ($M = .92$, $SD = .05$; $t(30) = 1.921$, $p = .064$, Cohen's $d = 0.74$), but not in the lip tube condition ($M = .89$, $SD = .05$; $t(30) = 0.835$, $p = .410$), when compared to baseline ($M = .87$, $SD = .08$).[3]

Discussion

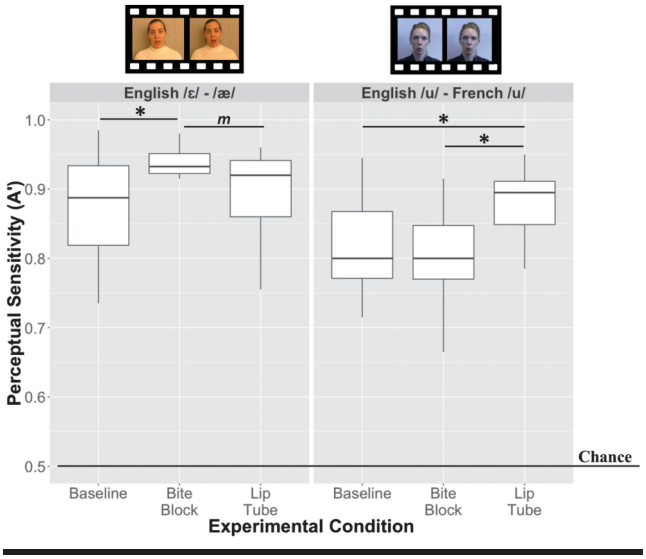
The results of Experiment 1 revealed an interaction between vowel contrast and experimental condition. Subjects who held a tube in their mouths showed enhanced discrimination of the vocalic gestures that were visually distinguished by their degree of lip movements when compared to baseline. In contrast, subjects who held a bite block in their mouths showed enhanced discrimination of the vocalic gestures that were visually distinguished by their degree of jaw movements when compared to baseline. Thus, engaging the lips or jaw led to heightened discrimination of visible speech movements involving the concordant articulator during perception.

This facilitation effect is broadly consistent with other findings described earlier that silently articulating syllables enhances identification of concordant syllables specified acoustically (Sams et al., 2005; Sato et al., 2013). Although our subjects were not instructed to produce phonetic gestures during concurrent perception, they still had to internally control the shape, position, and motion of their articulators in response to the present manipulations while watching the two model speakers talk. Furthermore, because they were perceiving speech under impoverished conditions

[2]Note that we did not video-record the subjects while they performed the discrimination task and, therefore, did not measure any covert vowel production–like movements that they might have produced while viewing the videos.

[3]See Supplemental Material S1 for further details regarding analyses of stimulus order effects (see, e.g., Masapollo et al., 2018).

Figure 2. Perceptual sensitivity (mean *A′* scores) to dynamically articulating visemes (Experiment 1) as a function of experimental condition (baseline vs. bite block vs. lip tube) and vowel contrast. Chance performance (*A′* = 0.5) is shown by the black horizontal line.



(i.e., they had to decode the speech signal in the absence of acoustic cues), our subjects may have been mentally simulating the sensory consequences of moving their lips or jaw to help discriminate the visual stimuli. Such an account would be consistent with the A × S perspective on speech perception, which posits that the generation of an internal model improves perceptual processing of a concordant speech stimulus (i.e., a facilitatory priming effect), especially when the speech signal is ambiguous or degraded (Sato et al., 2013; Skipper et al., 2007).

However, a much simpler interpretation of these results can be offered: The present manipulations may have biased attention toward properties of the environment that were congruent with the actions being performed. In other words, activating the lip muscles with the lip tube may have selectively biased attention toward perceived visible lip actions, whereas activating the jaw muscles with the bite block may have selectively biased attention toward perceived visible jaw actions. Moreover, it could be that it was not perception of the kinematic information itself that was affected by engaging the articulators. Perhaps these manipulations were instead affecting perception of static facial features (i.e., mouth shapes) rather than dynamic facial motion. Experiment 2 was conducted to address these competing accounts.

Experiment 2

Experiment 2 was designed to achieve two goals. The first goal was to test whether the effect of engaging the articulators enhanced perceivers' ability to track visually perceived speech movements or changes in visible vocal tract position and configuration, independent of a kinematic form. Toward this end, we investigated whether identical manipulations would influence perception of stilled speech. Our logic was as follows: If engaging the articulators facilitated perception of subtle changes in vocal tract posture, then we should observe effects comparable to those observed in Experiment 1 when the dynamically articulating visemes are replaced with stilled speech face image sequences. Specifically, the two oral–motor manipulations should each lead to better discrimination of images depicting static vocalic gestures produced with the concordant articulator compared with images of gestures produced with a different articulator. If, on the other hand, the manipulations facilitated perception of concordant orofacial speech movements, then we should fail to elicit such effects during the discrimination of resting face images showing the same differences in target (i.e., movement end point) lip and jaw position.

The second goal of Experiment 2 was to rule out the possibility that articulator activation simply biases attention toward that articulator during concurrent face perception. If this is the case, then the manipulations used in Experiment 1 should lead to the same overall pattern of results, because controlling the posture of an articulator should bias attention toward that articulator while viewing another person's face, regardless of whether it has a motion path.

Materials and Method

Subjects

Forty-seven participants (15 males, 32 females; age range = 18–31 years, *M* = 21.6 years, *SD* = 3.0) from Boston University completed this experiment. All were native, monolingual American English speakers who reported normal hearing and normal (or corrected-to-normal) vision. The experiment took approximately 1 hr, and subjects were paid for their participation.

Stimuli

The stimuli consisted of static single-frame images of the model speakers' visible vocal tract configuration during the production of the two vowel contrasts. The stimuli were created by taking a screenshot of the visual vowel tokens at vowel midpoint (see Figure 1, center panels). The images were presented for an equal amount of time as the corresponding video tokens in Experiment 1. Thus, any differences in task performance could not be attributed to an effect of shorter stimulus presentation.

Procedure and Design

The experimental protocol for Experiment 2 matched the procedures used in Experiment 1, except that subjects were instructed to discriminate static images depicting vocalic gestures, as opposed to dynamically articulating videos of the model speakers producing the two different vowel contrasts.

Data Analysis

As in Experiment 1, the dependent variable was the mean A' score (see calculation details above) averaged across subjects for each vowel contrast (English /u/–French /u/ vs. English /ɛ/–/æ/) in each experimental condition (baseline vs. lip tube vs. bite block).

Results

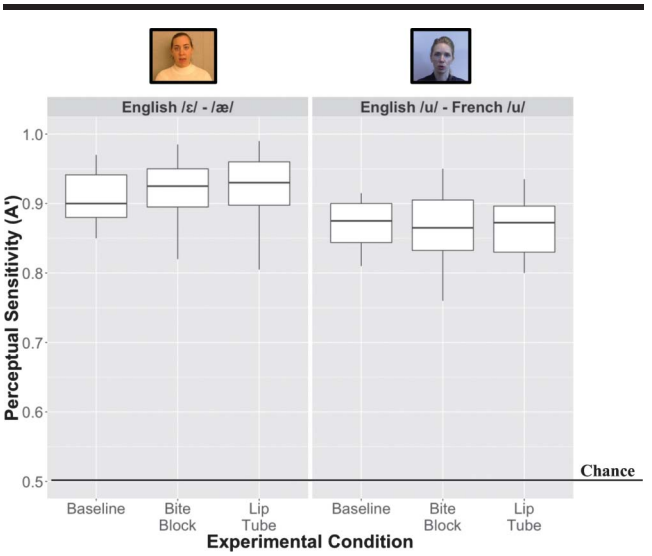
The critical question in Experiment 2 was whether engaging the lips or jaw would boost concurrent visual discrimination of stilled phonetic gestures produced with the concordant articulator compared to those produced with the discordant articulator. To address this question, we examined subjects' mean A' scores for each experimental condition as a function of vowel contrast, which are displayed in Figure 3. These scores were submitted to an ANOVA with experimental condition (baseline vs. lip tube vs. bite block) as a between-subjects factor, and vowel contrast (English /u/–French /u/ vs. English /ɛ/–/æ/) as a within-subject factor. A significant main effect of vowel contrast was observed, $F(1, 44) = 35.899$, $p < .001$, $\eta^2 = .449$, such that discrimination was again better for the English /ɛ/–/æ/ contrast ($M = .92$, $SD = .04$) compared to the English /u/–French /u/ contrast ($M = .86$, $SD = .06$). Critically, however, there was no significant main effect of condition, $F(2, 44) = 0.012$, $p = .998$, $\eta^2 = .001$, or interaction between condition and vowel contrast, $F(2, 44) = 1.309$, $p = .280$, $\eta^2 = .056$.[4]

In a second analysis, task performance was directly compared across Experiments 1 and 2. Mean A' scores were submitted to a three-way ANOVA with Experiments (1 vs. 2) and condition (baseline vs. lip tube vs. bite block) as between-subjects factors, and vowel contrast (English /u/–French /u/ vs. English /ɛ/–/æ/) as a within-subject factor. A significant main effect of experiment was observed, $F(1, 88) = 7.860$, $p = .006$, $\eta^2 = .082$, such that overall task performance was better in Experiment 2 ($M = .89$, $SD = .04$) compared to Experiment 1 ($M = .86$, $SD = .05$). This is perhaps unsurprising given that direct visual comparisons of the visemes may have been made easier by the removal of the oral-facial kinematic cues (see also, Masapollo et al., 2018). As in Experiment 1, a highly significant main effect of vowel contrast was also observed, $F(1, 88) = 55.525$, $p < .001$, $\eta^2 = .387$, such that discrimination was better for the English /ɛ/–/æ/ contrast ($M = .91$, $SD = .05$) compared to the English /u/–French /u/ contrast ($M = .84$, $SD = .07$). There was one significant interaction, the three-way Experiment × Condition × Vowel Contrast interaction, $F(2, 88) = 4.175$, $p = .019$, $\eta^2 = .087$. There were no other reliable main effects or interactions ($p > .135$, in all instances).

Discussion

The results of Experiment 2 suggest that the oral–motor manipulations in Experiment 1 affected subjects' perception of time-varying kinematic information in talking faces, rather than visible changes in target mouth shapes. When dynamic facial cues were not present, these manipulations did not facilitate perception of visemes produced with the concordant articulator compared to those produced with the discordant articulator.[5] Thus, the present manipulations appear to affect perceptual mechanisms that operate on dynamic visual–facial motion information. Finally, these findings are also inconsistent with the hypothesis that articulator activation generally biased visual attention toward that articulator during concurrent face perception.

Figure 3. Perceptual sensitivity (mean A' scores) to stilled facial speech images (Experiment 2) as a function of experimental condition (baseline vs. bite block vs. lip tube) and vowel contrast. Chance performance ($A' = 0.5$) is shown by the black horizontal line.



General Discussion

In the present research, we investigated whether and how somatosensory inputs from the vocal tract influence perception of visual speech. Recent studies (Bruderer et al., 2015; Ito et al., 2009) have provided evidence that manipulating the configuration and/or motion of the articulators shifts or constrains perception of some acoustic properties of speech. The present experiments extend this work by providing the first evidence, to our knowledge, that engaging the articulators also influences perception of dynamic

[4]See Supplemental Material S1 for further details regarding analyses of stimulus order effects (see, e.g., Masapollo et al., 2018).

[5]The finding that the present oral–motor manipulations did not influence the perception of static visual speech may be indirectly related to other findings showing that the acoustic and visual information for speech perception includes dynamic (time-varying) information (such as formant transitions and oral-facial kinematic patterns) and static target information (e.g., Masapollo et al., 2018; Masapollo, Zhao, Franklin, & Morgan, 2019; Strange, 1989; Viswanathan, Magnuson, & Fowler, 2014).

visual information in talking faces. Specifically, we found in Experiment 1 that engaging the lip muscles facilitated perception of concordant vocalic gestures optically distinct in terms of their degree of lip rounding (English /u/–French /u/), whereas engaging the jaw muscles facilitated perception of concordant vocalic gestures optically distinct in terms of their degree of jaw lowering (English /ɛ/–/æ/). By comparison, in Experiment 2, when the dynamically articulating visemes were shown under static conditions, those same oral–motor manipulations had no effect on discrimination performance. Thus, when the configuration and motion of the vocal tract is constrained, the perception of time-varying concordant visual speech movements is systematically affected, rather than perception of target, movement end point mouth shapes.

Taken together, these findings are inconsistent with the hypothesis that increased attentional load (or other task-related processing load) associated with the present oral–motor manipulations would lead to a decline in overall discrimination performance, regardless of which specific articulator was being engaged. Rather, these findings serve to further bolster previous claims that, during speech processing, perceivers analyze segmentally relevant information in relations to one's own vocal tract and ability to act (Bruderer et al., 2015; Ito et al., 2009; Sams et al., 2005; Sato et al., 2013). Manipulating the configuration and motion of the articulators had a consequence on how our subjects processed dynamic visual articulatory information. Specifically, subjects showed heightened discrimination of vocalic visemes that were congruent with the intrinsic motor properties of the articulator being engaged. Consistent with the A × S perspective on speech perception (Kuhl et al., 2014; Skipper et al., 2007; Stevens, 1960, 2002), these results may be interpreted as evidence that somatosensory feedback from the articulators may prime premotor, primary motor, and somatosensory brain regions involved in the sensorimotor control of speech, thereby facilitating perception of concurrent speech movements via an efference copy from the motor system.

The present findings may also be related to the finding that, when transcranial magnetic stimulation is applied to parts of the primary motor cortex controlling the lips or tongue during concurrent auditory speech perception, it facilitates identification of labial /b/, /p/, or dental /d/, /t/ stop consonants (D'Ausilio et al., 2009). That is, the priming of a motor representation for a given phoneme seems to bias perceptual judgments toward the congruent articulation. D'Ausilio et al. argued that such findings are compatible with the motor theory of speech perception (Galantucci, Fowler, & Turvey, 2006), which posits that the motor system is recruited during speech perception and that the object of perception is articulatory or gestural in nature. However, we would like to suggest an alternative interpretation of these findings from the perspective of the "directions into velocities of articulators" model of speech production (Guenther, 2016). According to this model, during the planning and execution of speech movements, neurons in the ventral motor cortex send inputs to neurons in the auditory cortex (i.e., Heschl's gyrus, posterior superior temporal gyrus), which encode the time-varying sensory expectations associated with those movements. Such inputs would allow speakers to effectively monitor their own self-generated auditory feedback for production errors. Consistent with this idea, Wise, Greene, Buchel, and Skott (1999) found, using positron emission tomography methods, reduced superior temporal gyrus activations during speech production compared to a passive listening task. This account would seem to reinforce our interpretation of the present findings as well as those reported by D'Ausilio et al. (2009). That is, speech perception might be selectively affected by the concurrent activation of the motor system because, by activating the motor programs for speech, the motor system affects the neural activity of the perceptual system (via an efference copy).

Although it is tempting to conclude on the basis of the present results that engaging the articulators enhanced perception of concordant visible speech movements, this conclusion may still need to be qualified. An alternative interpretation is that the present manipulations affected higher level cognitive processes (e.g., phonetic categorization) rather than low-level sensory discrimination per se. Theoretical accounts that focus on explicating the role of particular task demands on speech perception, such as those of Werker and Tees (1984), Macmillan, Goldberg, and Braida (1988), and more recently Strange (2011), propose that, when an experimental task places greater demands on verbal working memory, subjects often label stimuli in terms of their distance from salient (or "easy-to-remember") reference points within perceptual space. By this account, working memory fades quickly, and when the ISI is increased, perceivers must encode stimuli in terms of phonetic categories to complete the task. Consistent with this view, Pisoni (1973) found that sensitivity to auditory vowel stimuli within a phoneme category was higher at shorter ISIs compared to longer ISIs, purportedly because it was easier for listeners to compare acoustic details when the amount of time that each stimulus was stored in memory was shorter (cf. Werker & Logan, 1985). It is possible, then, that the subjects tested in the present experiments were interpreting the stimuli in terms of discrete labels in "face space" due to the memory demands imposed by the relatively long ISI (i.e., 1,500 ms). Thus, it is not entirely clear from the foregoing results whether the oral–motor manipulations affected perceptual processes or working memory and phonetic categorization processes.

Moreover, it is unknown whether the present manipulations influence speech processing across sensory modalities. It could be that the facilitation/priming effect observed in Experiment 1 is limited to the visual domain. Evidence substantiating this hypothesis comes from a functional magnetic resonance imaging study (Skipper et al., 2005) showing that the presence of visual speech movements modulated activity in brain regions associated with speech production and proprioception, presumably because the optical signal provides more direct information about the configuration and motion of some articulators. That visual speech cues

influence BOLD activity in motor and somatosensory cortices above and beyond auditory speech cues alone (Matchin et al., 2014) raises the possibility that the facilitation effect observed for visual vowel perception will not generalize to auditory vowel perception.

Alternatively, the present manipulations may enhance perceivers' ability to track and extract dynamic articulatory information reflected in both the acoustic and optical speech signals. Consistent with this view, other studies already discussed have reported that the shape and/or motion paths of the articulators influence concurrent auditory speech perception (Bruderer et al., 2015; Ito et al., 2009). A finding that the present manipulations also enhance perception of unimodal auditory-only vowels would be compatible with the direct realist perspective of speech perception, which posit that auditory and visual information jointly specify distal vocal tract gestures and that articulatory information is detected in each modality (e.g., Best, Goldstein, Nam, & Tyler, 2016; Fowler, 2004; Galantucci et al., 2006; Masapollo et al., 2017, 2018).

In sum, the present findings provide evidence that manipulating the configuration and motion of the articulators influences concurrent perception of visible articulatory movements. Somatosensory inputs from the vocal tract selectively enhanced perception of concordant phonetic gestures. These findings increase our understanding of perception–action linkages for speech by showing that perception involves processes that relate visual articulatory information to the perceiver's current vocal tract posture and potential for action. Such findings have important implications for theories of speech production and perception, which must explicate the nature of the complex interplay between the articulatory motor and speech perception systems.

Acknowledgments

Research reported in this publication was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health Grant R01DC002852, awarded to Frank H. Guenther. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We are grateful to Barbara Holland and Farwa Faheem for assistance with subject recruitment, data collection, and analysis. This work benefited from helpful discussions with, or comments from, Julia Irwin, David Ostry, Linda Polka, Henny Yeung, Lauren Franklin, Christina Zhao, Janet Werker, and members of the audience at the 10th Annual Meeting of the Society for the Neurobiology of Language.

References

Abur, D., Lester-Smith, R. A., Daliri, A., Lupiani, A. A., Guenther, F. H., & Stepp, C. E. (2018). Sensorimotor adaptation of voice fundamental frequency in Parkinson's disease. *PLOS ONE, 13*(1), e0191839. https://doi.org/10.1371/journal.pone.0191839

Best, C. T., Goldstein, L. M., Nam, H., & Tyler, M. D. (2016). Articulating what infants attune to in native speech. *Ecological Psychology, 28,* 216–261.

Boersma, P., & Weenink, D. (2019). *Praat: Doing Phonetics by Computer* (Version 6.1.03) [Computer program]. Retrieved from http://www.praat.org/

Bruderer, A. G., Danielson, D. K., Kandhadai, P., & Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America, 112*(44), 13531–13536.

Cai, S., Beal, D. S., Ghosh, S. S., Tiede, M. K., Guenther, F. H., & Perkell, J. S. (2012). Weak responses to auditory feedback perturbation during articulation in persons who stutter: Evidence for abnormal auditory-motor transformation. *PLOS ONE, 7*(7), e41830. https://doi.org/10.1371/journal.pone.0041830

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience, 15,* 57–70.

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The motor somatotopy of speech perception. *Current Biology, 19*(5), 381–385.

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

Galantucci, B., Fowler, C. A., & Goldstein, L. (2009). Perceptuomotor compatibility effects in speech. *Attention, Perception & Psychophysics, 71*(5), 1138–1149.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13*(3), 361–377.

Golfinopoulos, E., Tourville, J. A., Bohland, J. W., Ghosh, S. S., & Guenther, F. H. (2011). fMRI of unexpected somatosensory feedback perturbation during speech. *NeuroImage, 55,* 1324–1338.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75,* 424–429.

Guenther, F. H. (2016). *Neural control of speech*. Cambridge, MA: MIT Press.

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: Computational basis and neural organization. *Neuron, 69*(3), 407–422.

Hickok, G., & Poeppel, D. (2016). How do chinchillas, pigeons, and infants perceive speech? Another comment on Skipper et al. *Talking Brains*. Retrieved from http://www.talkingbrains.org/2016/11/how-do-chinchillas-pigeons-and-infants.html

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279,* 1213–1216.

Ito, T., Tiede, M., & Ostry, D. J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences of the United States of America, 106*(4), 1245–1248.

Kuhl, P. K., Ramirez, R. R., Bosseler, A., Lin, J.-F. L., & Imada, T. (2014). Infants' brain responses to speech suggest analysis by synthesis. *Proceedings of the National Academy of Sciences of the United States of America, 111,* 11238–11245.

MacDonald, E. N., Johnson, E. K., Forsythe, J., Plante, P., & Munhall, K. G. (2012). Children's development of self-regulation in speech production. *Current Biology, 22,* 113–117.

MacLeod, A., Stoel-Gammon, C., & Wassink, A. B. (2009). Production of high vowels in Canadian English and Canadian French: A comparison of early bilingual and monolingual speakers. *Journal of Phonetics, 37*(4), 374–387.

Macmillan, N. A., Goldberg, R. F., & Braida, L. D. (1988). Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua. *The Journal of the Acoustical Society of America, 84,* 1262–1280.

Masapollo, M., Polka, L., & Ménard, L. (2017). A universal bias in vowel perception–By ear or by eye. *Cognition, 166,* 358–370.

Masapollo, M., Polka, L., Ménard, L., Franklin, L., Tiede, M., & Morgan, J. (2018). Asymmetries in unimodal visual vowel perception: The roles of oral-facial kinematics, orientation, and configuration. *Journal of Experimental Psychology: Human Perception and Performance, 44*(7), 1103–1118.

Masapollo, M., Zhao, T. C., Franklin, L., & Morgan, J. L. (2019). Asymmetric discrimination of nonspeech tonal analogues of vowels. *Journal of Experimental Psychology: Human Perception and Performance, 45*(2), 285–300.

Matchin, W., Groulx, K., & Hickok, G. (2014). Audiovisual speech integration does not rely on the motor system: Evidence from articulatory suppression, the McGurk effect, and fMRI. *Journal of Cognitive Neuroscience, 26*(3), 606–620.

Mollaei, F., Shiller, D. M., Baum, S. R., & Gracco, V. L. (2016). Sensorimotor control of vocal pitch and formant frequencies in Parkinson's disease. *Brain Research, 1646,* 269–277.

Nasir, S. M., & Ostry, D. J. (2006). Somatosensory precision in speech production. *Current Biology, 16,* 1918–1923.

Noiray, A., Cathiard, M.-A., Ménard, L., & Abry, C. (2015). Test of the movement expansion model: Anticipatory vowel lip protrusion and constriction in French and English speakers. *The Journal of the Acoustical Society of America, 129*(1), 340–349.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics, 13*(2), 253–260.

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.

Sams, M., Möttönen, R., & Sihvonen, T. (2005). Seeing and hearing others and oneself talk. *Cognitive Brain Research, 23,* 429–435.

Sato, M., Troille, E., Ménard, L., Cathiard, M.-A., & Gracco, V. L. (2013). Silent articulation modulates auditory and audiovisual speech perception. *Experimental Brain Research, 227*(2), 275–288.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *NeuroImage, 25*(1), 76–89.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex, 17*(10), 2387–2399.

Stevens, K. N. (1960). Toward a model for speech recognition. *The Journal of the Acoustical Society of America, 32,* 47–55.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America, 111*(4), 1872–1891.

Strange, W. (1989). Evolving theories of vowel perception. *The Journal of the Acoustical Society of America, 85*(5), 2081–2087.

Strange, W. (2011). Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics, 39,* 456–466.

Stuart, A., Kalinowski, J., Rastatter, M. P., & Lynch, K. (2002). Effect of delayed auditory feedback on normal speakers at two speech rates. *The Journal of the Acoustical Society of America, 111,* 2237–2241.

Sundara, M., Namasivayam, A. K., & Chen, R. (2001). Observation-execution matching system for speech: A magnetic stimulation study. *NeuroReport, 12*(7), 1341–1344.

Tremblay, S., Shiller, D. M., & Ostry, D. J. (2003). Somatosensory basis of speech production. *Nature, 423,* 866–869.

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbation of vowel acoustic and its relation to perception. *The Journal of the Acoustical Society of America, 122*(4), 2306–2319.

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2014). Information for coarticulation: Static signal properties or formant dynamics? *Journal of Experimental Psychology: Human Perception and Performance, 40*(3), 1228–1236.

Werker, J. F., & Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics, 37,* 35–44.

Werker, J. F., & Tees, R. C. (1984). Phonemic and phonetic factors in adult cross-language speech perception. *The Journal of the Acoustical Society of America, 75*(6), 1866–1878.

Wise, R. J., Greene, J., Buchel, C., & Skott, S. K. (1999). Brain regions involved in articulation. *Lancet, 353,* 1057–1061.

Yeung, H. H., & Werker, J. F. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science, 24*(5), 603–612.