Effects of Lips and Hands on Auditory Learning of Second-Language Speech Sounds

Yukari Hirata
Spencer D. Kelly
Colgate University, Hamilton, NY

Purpose: Previous research has found that auditory training helps native English speakers to perceive phonemic vowel length contrasts in Japanese, but their performance did not reach native levels after training. Given that multimodal information, such as lip movement and hand gesture, influences many aspects of native language processing, the authors examined whether multimodal input helps to improve native English speakers' ability to perceive Japanese vowel length contrasts.
Method: Sixty native English speakers participated in 1 of 4 types of training: (a) audio-only; (b) audio-mouth; (c) audio-hands; and (d) audio-mouth-hands. Before and after training, participants were given phoneme perception tests that measured their ability to identify short and long vowels in Japanese (e.g., /kato/ vs. /kato:/).
Results: Although all 4 groups improved from pre- to posttest (replicating previous research), the participants in the audio-mouth condition improved more than those in the audio-only condition, whereas the 2 conditions involving hand gestures did not.
Conclusions: Seeing lip movements during training significantly helps learners to perceive difficult second-language phonemic contrasts, but seeing hand gestures does not. The authors discuss possible benefits and limitations of using multimodal information in second-language phoneme learning.

KEY WORDS: lips, mouth, hand gesture, auditory learning, phonemic contrasts, second language

Foreign languages pose many challenges for learners. But learning to hear distinctions among novel speech sounds stands out as a particularly difficult obstacle. In the present study, we investigated whether this challenge can be overcome through various types of auditory and visual instruction.

Auditory Learning of Second-Language Phonemic Contrasts

Research in phonetic science and second-language acquisition has made progress over the past several decades investigating the limitations of adults in perceiving nonnative phonemic contrasts. Numerous studies have shown that even though adults are imperfect in learning to perceive certain phonemes of a second language, their perceptual inability can be improved by intensive auditory training in a laboratory (Bradlow, Pisoni, Yamada, & Tohkura, 1997; Logan, Lively, & Pisoni, 1991; Morosan & Jamieson, 1989; Pisoni & Lively, 1995; Yamada, Yamada, & Strange, 1995). This laboratory training typically involves auditorily presenting a member of a pair of words, such as *light-right* and *cloud-crowd* to Japanese

speakers, asking them to identify whether they have heard *l* or *r,* and providing immediate feedback on their responses. Although this auditory training for difficult second-language phonemes is proven to improve adults' perception, it is still difficult for them to reach native levels.

Another type of contrast that nonnative adults have difficulty perceiving is Japanese vowel length contrast (Hirata, 2004a; Hirata, Whitehurst, & Cullings, 2007; Landahl & Ziolkowski, 1995; Tajima, Kato, Rothwell, Akahane-Yamada, & Munhall, 2008; Tajima, Rothwell, & Munhall, 2002), which is the focus of the present study. Japanese has five short vowels (/i e a o u/) that contrast phonemically with the corresponding long vowels (/iː eː aː oː uː/)—for example, /i/ "stomach" versus /iː/ "good." The major acoustic correlate and perceptual cue for this length distinction is duration (Fujisaki, Nakamura, & Imoto, 1975), with few differences in their formant frequencies (Hirata & Tsukada, 2009; Tsukada, 1999; Ueyama, 2000). Long vowels are 2.2–3.2 times longer in duration than short vowels (Hirata, 2004b; Tsukada, 1999), but the difference between the short and long vowels could be as small as 50 ms when vowels are spoken quickly in a sentence (Hirata, 2004b).

Native English speakers have difficulty perceiving this vowel length distinction (Hirata et al., 2007; Tajima et al., 2008) because the English vowel system (e.g., American and British varieties) does not use vowel duration as a sole perceptual cue to phonemic distinction of vowels. For example, American English /i/ and /ɪ/, as in *heed* and *hid,* differ in both formant frequencies and duration, and the former is used as the primary cue (Hillenbrand, Clark, & Houde, 2000). Whether one utters *heed* with the vowel of 200 or 70 ms does not change the meaning of the word. Native English speakers' auditory ability to perceive Japanese vowel length contrasts can improve after hours of training, but it does not reach the native level (Hirata, 2004a; Hirata et al., 2007; Tajima et al., 2008), similar to the case of the *l* and *r* distinction mentioned earlier.

Effects of Visual Input—Lipreading

Spoken communication typically occurs in a rich multimodal context. In natural face-to-face interactions, people produce important information through such channels as facial expression, hand gestures, and tone of voice. Theories of communication claim that this multimodal information combines with speech to help people better comprehend language (Clark, 1996; Goldin-Meadow, 2003; McNeill, 1992). In addition to the auditory modality, we focused on two types of visual information in the present study: mouth movements and hand gestures.

Abundant research has focused on auditory and visual (AV) sensory integration of speech and lip (mouth) movements, showing that our perception of information in one modality is tightly connected to perception of information in the other (Green, Kuhl, Meltzoff, & Stevens, 1991; Massaro & Cohen, 1983; McDonald, Tedder-Sälejärvi, & Hillyard, 2000; Munhall, Gribble, Sacco, & Ward, 1996; Reisberg, McLean, & Goldfield, 1987; Sekiyama, 1997; Skipper, Goldin-Meadow, Nusbaum, & Small, 2009; Skipper, van Wassenhove, Nusbaum, & Small, 2007; Teder-Sälejärvi, McDonald, Di Russo, & Hillyard, 2002). Observing lip movements activates the auditory cortex, even in the absence of speech sounds (Calvert et al., 1997), suggesting that "seen speech" influences "heard speech" (known as the McGurk effect from McGurk & MacDonald, 1976) at very early stages of language processing. This bimodal integration helps perception and comprehension of speech for hearing-impaired listeners (Grant & Seitz, 1998), cochlear implant users (Desai, Stickney, & Zeng, 2008), and nonnative speakers (Hardison, 2003; Wang, Behne, & Jiang, 2008), as well as for normal hearing listeners (Arnold & Hill, 2001; Sumby & Pollack, 1954).

Most relevant to the present study is the work of Hardison (2003, 2005), who found that AV training produced better learning than auditory-only training for native Japanese and Korean speakers' perception of English /ɹ/ and /l/. One way that the present study differed from Hardison (2003, 2005) was that we investigated the distinct role that AV input played in the auditory learning of nonnative length or *quantity* contrasts, instead of contrasts that differ in *quality.* Hardison (2005) and other researchers (e.g., Wang et al., 2008) have focused on the benefits of *qualitative*ly different AV input, exploiting the visual differences in the production of phonemes differing in place or manner of articulation, such as /b/-/v/, /v/-/ð/-/z/, and /ɹ/-/l/. However, it is simply not known whether lip movements conveying the length of short and long vowels (in which the difference is only that of duration of the mouth opening) would be as visually salient and informative during training. Kubozono (2002) showed that nonnative speakers of Japanese depended more heavily on visual information than native speakers in the AV presentation of Japanese disyllables such as /sado/ "(place name)" and /sadoː/ "way of flower arrangement." Because Kubozono's study did not involve training, it is an empirical question whether the AV training is actually beneficial to nonnative speakers when they learn to hear important quantity distinctions. Thus, examining the role that mouth movements play in nonnative speakers' auditory learning of Japanese vowel length contrast was one of the major questions of the present study.

The degree to which this AV training is advantageous depends on many factors, such as the visual distinctiveness of the target speech sounds, phonetic contexts, learners' first language, and nonnative speakers' degree

of exposure to the target language (Hardison, 1999, 2003; Hazen, Sennema, Iba, & Faulkner, 2005; Wang et al., 2008). One interesting factor, which is relevant to the present study but is not well understood, is whether AV training with sentence stimuli, as opposed to that with isolated words, can be beneficial to nonnative speakers. Many studies (Hardison, 2003; Wang et al., 2008) have shown AV training benefits with isolated or excised words for learning to make nonnative phonemic distinctions. However, much less is known about AV benefits within sentences. One might think that sentence training could be distracting because it provides too much information, but the specific benefit of sentence training cannot be overlooked. Hirata (2004a) found that auditory-only *sentence* training assisted auditory learning in both isolated word and sentence contexts, whereas auditory-only *word* training was not as effective for the auditory learning in the sentence context. We extended this previous research in the present study and examined whether lip movements of sentences have a significant effect on the learning of nonnative phonemic distinctions.

Effects of Visual Input—Hand Gesture

Hand gestures that co-occur with speech are another prevalent aspect of face-to-face communication. These spontaneous hand movements are produced unconsciously and often convey information that reinforces and complements the speech they accompany. Cospeech gestures are so pervasive that McNeill (1992) theorizes that together with speech, they are part and parcel of language and integrated at a deep conceptual level. Among the many types of cospeech gestures, the most well studied are the iconic gestures, that is, gestures that convey imagistic information about object attributes, spatial relationships, and movements (e.g., a gesture representing the action of drinking). Researchers have demonstrated that these iconic gestures play a significant role in language comprehension (Beattie & Shovelton, 1999; Cassell, McNeill, & McCullough, 1999; Goldin-Meadow, Wein, & Chang, 1992; Kelly, Barr, Church, & Lynch, 1999; Kelly & Church, 1998; see Kelly, Manning, & Rodak, 2008, for a review). Iconic gesture and speech are also linked at the neural level during language comprehension (Holle & Gunter, 2007; Kelly, Kravitz, & Hopkins, 2004; Özyürek, Willems, Kita, & Hagoort, 2007; Wu & Coulson, 2007). Moreover, iconic gestures also play a role in processing and learning a foreign language (Church, Ayman-Nolley, & Mahootian, 2004; Kelly, McDevitt, & Esch, 2008; Quinn-Allen, 1995; Roberge, Kimura, & Kawaguchi, 1996; Sueyoshi & Hardison, 2005). For example, Sueyoshi and Hardison (2005) showed that Korean and Japanese learners of English (especially those with lower proficiency) understood English lectures best when lips and iconic hand gestures accompanied those lectures compared with audio alone.

In the present study, we investigated the communicative function of a different type of hand gesture—beat gestures. Rather than containing semantic content as iconic gestures do, beats convey information about the prosody and rhythm of speech (McNeill, 1992). Researchers have only begun to seriously consider the role that the beat gestures play in language perception and production. For example, Krahmer and Swerts (2007) demonstrated that beats (e.g., quick flicks of the hand) not only enhance the acoustic prominence of prosodic stress and pitch patterns during speech *production* but also draw attention to those prosodic elements during speech *perception*. Corroborating this finding, recent neuroimaging research using functional magnetic resonance imaging has shown that low-level auditory brain areas, such as the planum temporale in the superior temporal gyrus, are more active during language comprehension when beat gestures accompany speech than when speech is presented alone (Hubbard, Wilson, Callan, & Dapretto, 2008). The investigators concluded that beat gestures function to focus listeners' (viewers') attention to speech prosody, and this ultimately makes the phonological elements of speech clearer and more intelligible.

Much less clear is the role that gesture plays in learning phonemic contrasts of a second language. In one of the few studies on the topic, Roberge et al. (1996) taught native English speakers to make beat gestures of differing lengths to differentiate short and long vowels in Japanese. Learners were instructed to make pivoting motions of the hands below the elbows with flat palms extending from the center to the sides of the body, and this instruction was successful at helping English speakers to produce Japanese long vowels clearly. The authors speculated that the long stroke of hand movements might have helped physically sustain muscle tension of the vocal articulators necessary to produce long vowels. We built on this work and attempted to extend these findings from the production to the *perception* of gestures in the present study. Specifically, we investigated whether beat gestures that convey temporal information about the length of Japanese vowels help native English speakers to *perceive* the phonemic vowel length contrasts successfully.

Goals of the Present Study

As mentioned earlier, the auditory ability to distinguish difficult nonnative phonemic contrasts improves through intensive auditory training, but thus far, learners have not reached the native level (Hirata et al., 2007; Tajima et al., 2008). In the present study, we explored whether this limited auditory ability can improve even further by the use of two types of visual information:

mouth movements and hand gestures. Exploration of this possibility, in turn, provides insights into theories of gesture and multimodal communication, that is, whether the contributions of gestures for comprehending the semantics of a language can extend to auditory learning of novel phonemes. The questions were: How does multimodal information conveyed through speech sounds, mouth movements, and hand gestures facilitate the auditory/perceptual learning of difficult second-language phonemes? And how does this multimodal training method compare with the traditional audio-only training method?

In the present study, we investigated the effects of the following four types of training differing in the input modalities: (a) audio-only, (b) audio-mouth, (c) audio-hands, and (d) audio-mouth-hands. All four groups of participants completed a pretest, four sessions of one of the above training types, and a posttest over the course of a 2-week period. The pretest and posttest included only audio stimuli without mouth movements or hand gestures. The purpose of this format was to examine how the use of visual information, mouth movements and hand gestures, would ultimately improve participants' auditory ability to distinguish Japanese short and long vowels (and not how well participants ultimately learn to use multimodal information).

Drawing from native English speakers' significant perceptual improvement after auditory training in Hirata et al. (2007), we expected the present audio-only condition to show a moderate but significant improvement from the pretest to the posttest. Moreover, if there is a distinct role of seeing lip movements, then we would expect the auditory improvement to be greater for the audio-mouth than the audio-only condition. Similarly, if beat gestures help not only for understanding of one's native language (Hubbard et al., 2008; Krahmer & Swerts, 2007) but also for learning to perceive subtle phonemic differences in a second language, we would expect the auditory improvement to be greater for the audio-hands than the audio-only condition. And if there are combined effects of mouth and hand movements synchronized with auditory stimuli, then we would predict the auditory improvement to be highest for the audio-mouth-hands condition.

Method

Participants

Sixty students between the ages of 19 and 23 were recruited from a college in the northeastern United States and were paid for their participation. All were monolingual native speakers of American English and had not been exposed to spoken Japanese (with limited exposure to Japanese anime or films), nor had studied Japanese prior to this study. Most had studied a foreign language, but none had achieved native-level fluency. None of the participants had more than 9 years of musical training, and no one reported any hearing problems. Participants were randomly assigned to four experimental groups: audio-only (15 participants), audio-mouth (15 participants), audio-hands (16 participants), and audio-mouth-hands (14 participants).

Test Materials

All participants were given a pretest and then a posttest 2 weeks later. These tests were identical to those used in Hirata et al. (2007). In both the pretest and the posttest, there were 180 test items that comprised a carrier sentence and a target word. The target items were five pairs of Japanese words: /rubi/-/rubi:/, meaning *agate-ruby*, /ise/-/ise:/, (name of a place)-*opposite gender,* /rika/-/rika:/, *science-liquor,* /kato/-/kato:/, *transition*-(surname), and /saju/-/saju:/, *hot water–left and right*. The difference within each target pair occurred in the final vowel, with one word ending in a short vowel and the other ending in a long vowel. In the pretest, six carrier sentences were combined with every target word, so that 10 unique stimuli were formed from each carrier sentence. The target word was always inserted in the middle of the sentence, for example, /sore ga ___ da to omoimasu/ ("I think that is ___"). One male speaker recorded stimuli with three carrier sentences, and one female speaker recorded the same target words with three different carrier sentences. To create further diversity in the test items, the stimuli included three speaking rates: slow, normal, and fast. In total, there were 180 stimuli in the pretest (i.e., 5 vowels × 2 vowel lengths × 2 speakers × 3 sentences × 3 speaking rates). For the purposes of the present experiment, these different types of test stimuli were not treated as independent variables (but see Hirata et al., 2007, for more on these variables). For the posttest, stimuli were created with the identical target words with six different carrier sentences. As in the pretest, the same male and female speakers each recorded stimuli with three different carrier sentences.

Test Procedure

All participants took the pretest and the posttest in a quiet lab space while wearing Grado Labs SR125 headphones. The 180 test stimuli in the pre- and posttest were presented in a random order across six separate blocks, with each block using a different carrier sentence. The carrier sentences—but not target words—were written on the screen simultaneously with the presentation of audio stimuli. The participants were asked to identify whether the second vowel of the target words (e.g., /ise/ or /ise:/) was short or long (a two-alternative forced-choice identification task presented on the computer screen). The correct responses to this question served

as the dependent variable in our experiment. After each response, no feedback was given, and participants were asked to click a "play" button to hear the next audio stimulus. The pre- and posttest took approximately 30 min each.

Training Materials

In between the pre- and posttests, there were four training sessions that instructed participants on how to distinguish long and short vowels in Japanese. The audio stimuli were adapted from the "slow" training used in previous research by Hirata et al. (2007). Specifically, there was a spoken (and written) Japanese carrier sentence that was the same for every target word /soko wa ___ to kaite arimasu/ ("___ is written there"), and there were 10 different Japanese nonsense words (containing 5 short and 5 long vowels) that were spoken (but not written) within each sentence (/mimi/-/mimi:/, /meme/-/meme:/, /mama/-/mama:/, /momo/-/momo:/, and /mumu/-/mumu:/). Each training session had a total of 160 trials (i.e., 5 vowels × 2 lengths × 2 repetitions × 8 blocks). Materials in the four training sessions were identical but recorded by four speakers (two men and two women) who were different from the speakers in the tests.

The above audio stimuli from Hirata et al. (2007) were combined with video clips created by filming two male and two female native Japanese speakers. These four speakers were not the speakers originally recorded for the audio stimuli described above, but they were instructed to "lip synch" the training sentences. The audio from Hirata et al. (2007) was later dubbed into these videos so that it appeared that the new speakers were the original ones who had spoken the training sentences. The purpose of this dubbing procedure (rather than creating entirely new training items) was to maintain consistency in the auditory channel between the original study by Hirata et al. (2007) and the present study so that the results would be as comparable as possible. Prior to the experiment, naïve viewers (who were not the participants of the experiment) watched these training materials and verified that they believed the voice and the visual materials were synchronized and came from the same speaker.[1]

There were four different audiovisual training conditions. In the audio-mouth-hands condition (n = 14), the audio track was accompanied by the bodies of the Japanese speakers that conveyed information about the length of the vowels through their mouth movements and hand gestures. That is, the opening of the mouth corresponded to either a short or a long vowel (mouth opening was 2.2–3.2 times longer for long vowels than short vowels; Hirata, 2004b). In addition, hand gesture was added to convey information about the length of the vowels in the target words. For the words with two short vowels (e.g., /mama/), the speaker produced two quick hand flicks, and for the words with one short and one long vowel (e.g., /mama:/), the speaker produced a quick hand flick and a prolonged hand sweep extended horizontally. These hand movements represent the technique used by Roberge et al. (1996). The idea was that the mouth movements and hand gestures would convey information about the temporal properties of the vowels contained in the target words—temporal information that is very difficult to perceive auditorily for English speakers. The remaining conditions systematically removed the number of visual modalities available to the participants to determine the relative contribution of the different types of visual input. In the audio-mouth condition (n = 15), the face of the speaker was fully visible, and the mouth was clearly synchronized with the spoken sentences, but the body did not produce any hand gestures—a still frame of the body was superimposed to achieve this effect. In contrast, the audio-hands condition (n = 16) obscured the face of the speaker by using a digital pixelization technique, but the body produced hand gestures that corresponded with the length of vowels in the target words. Finally, the audio-only condition (n = 15) presented a complete still frame of the speaker (there was no information from the mouth or hands regarding the length of the target vowels) and played the audio track over that still image. A still frame rather than a blank screen was chosen because we wanted the audio-only condition to be as similar as possible to the other three conditions. Refer to Figure 1 for still frames at the end of the long and short vowel for each of the four training conditions. Note, in particular, that the gestures in the last two conditions finish at different locations, with gestures for short vowels finishing closer to the center than gestures for long vowels.

Training Procedure

Participants in all but the audio-only conditions were instructed to pay attention to the mouth and hand movements of the native Japanese speakers who appeared on the computer screen when they listened to the audio stimuli. In the two gesture conditions, they were told that the gestures would convey relevant information regarding the length of the vowels, with short vertical and long

[1]Even though undetected by naïve viewers, it is possible that the lip-synching technique resulted in some incongruencies between the visual and auditory channels. However, note that if anything, this would decrease the chances of us finding differences among our groups. Also note that the audio-mouth condition is more vulnerable than the audio-hands condition because there was more visual information to synchronize in the former condition (every word in the sentence) compared with the latter one (only the target word). However, as we show later in the Results and Discussion sections, benefits of the audio-mouth training are significantly greater than the other types of training. The risk of a Type II error was preferable to creating all new stimuli that would make it difficult to compare the original training data in Hirata et al. (2007).

Figure 1. Examples of video clips for four training conditions. Audio-only: no image was moving while audio stimuli were presented; audio-mouth: the speaker's body below the neck was not moving at all while the mouth was clearly moving and synchronized with the audio stimuli; audio-hands: the speaker's face was obscured while the hand movements clearly synchronized with audio stimuli; audio-mouth-hands: both the mouth movements and hand gestures were synchronized with audio stimuli. Note that, for audio-hands and audio-mouth-hands, hand gestures for short versus long vowels at the end of the target words clearly show the difference in the figure, with the right hand located in the middle for the short vowel, and to the right of the speaker for the long vowel.

horizontal strokes corresponding to short and long vowels, respectively. The participants in all but the audio-only conditions were reminded that, despite this visual information, they still would need to pay close attention to the spoken vowel sounds because they would hear audio stimuli only (and would see no visual material) in the posttest.[2]

[2]In a postexperiment interview, we learned that 4 participants (2 in each of the gesture conditions) occasionally relied on gesture alone to successfully navigate the instruction sessions. However, even when these participants were removed from the analysis, the results were not different from those presented in this article. After all, even if participants did not explicitly attend to the speech in the instructions sessions, they still had the experience of hearing it and at least covertly associating it with the gestures (see Lim & Holt, 2009, for implicit learning).

During each training session, participants were asked to identify whether the second vowel in each target word (e.g., /meme/) was short or long by clicking the appropriate button on the computer screen. If participants responded correctly, the word *Correct* appeared on the screen, and they received the next sentence. If they responded incorrectly, the message "Sorry, you are incorrect" appeared on the screen, and they were required to click a button labeled *Play again,* and the sentence was played three more times. Before the first and fifth blocks, participants were given examples of sentences and their correct responses.

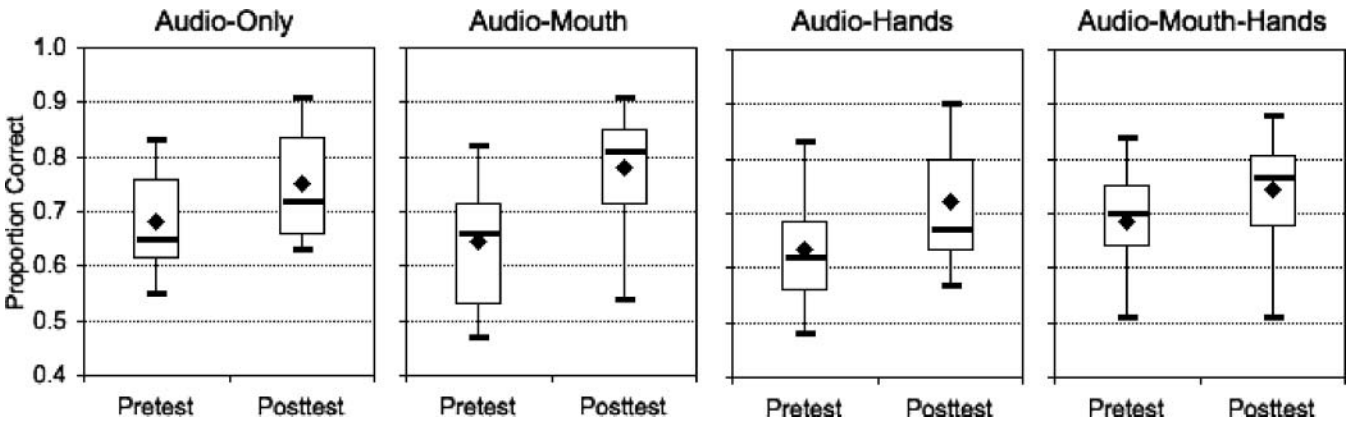
The four training sessions (approximately 30 min each) spanned a 2-week period, and any two sessions were separated by at least 1 day and no more than 4 days. Each training session involved a different "teacher" (each of the four different Japanese speakers). It is important to note that the training session included items that were not included in the pre- and posttest (i.e., different target words, carrier sentences, and speakers). In this way, any beneficial effects of training are the result of generalized learning rather than rote memorization. Data obtained during training was unfortunately lost due to an unexpected computer error, and thus the present article focused on results of pretest and posttest scores.

Results

We performed a mixed design analysis of variance (ANOVA), with test (pre- and post-) as the within-subjects factor and multimodal condition (audio-only, audio-mouth, audio-hands, audio-mouth-hands) as the between-subjects factor. The dependent measure was the pre- and posttest proportions of correct responses with arcsine transformations. Bonferroni t tests compared all orthogonal contrasts in the within-subjects condition (test), and Dunn's multiple t tests (tD, with adjusted p values) compared the pre- to posttest difference scores of the audio-only condition with the difference scores from the three multimodal conditions. The figures present raw percentages scores rather than arcsine-transformed scores.

The 2 (test time) × 4 (multimodal condition) ANOVA did not uncover a significant main effect of condition, $F(3, 56) = 0.57$, *ns,* but there was a significant main effect of test, $F(1, 56) = 71.00$, $p < .001$. This indicates that participants made significant improvement from the pretest ($M = 0.66$, $SD = 0.1$) to posttest ($M = 0.75$, $SD = 0.11$) across all training conditions (see Figure 2). Participants in the audio-only condition improved from 0.68 ($SD = 0.09$) to 0.75 ($SD = 0.10$), $t(14) = 5.28$, $p < .001$; the audio-mouth condition improved from 0.64 ($SD = 0.12$) to 0.78 ($SD = 0.10$), $t(14) = 5.88$, $p < .001$; the audio-hands condition improved from 0.64 ($SD = 0.11$) to 0.72 ($SD = 0.11$), $t(15) = 4.89$, $p < .001$; and the audio-mouth-hands condition

Figure 2. Pretest and posttest scores (in proportions) for the four conditions. All four groups made significant improvement from the pre- to posttest. However, there was an interaction, with the audio-mouth condition improving significantly more than the other three conditions.



improved from 0.69 (SD = 0.9) to 0.74 (SD = 0.10), $t(13)$ = 2.18, p = .028.

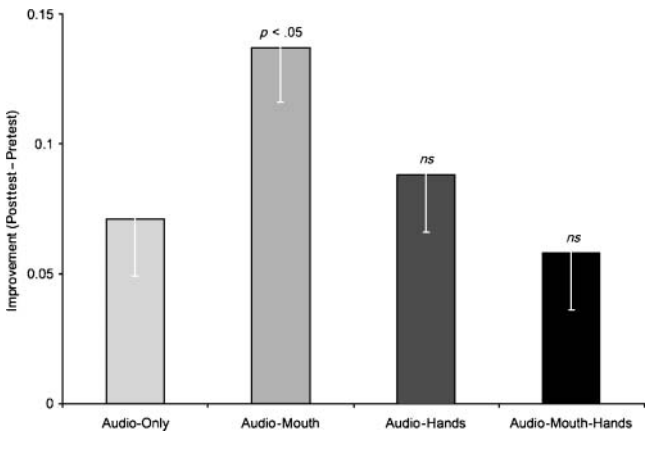
In addition, there was a significant Test × Condition interaction, $F(3, 56) = 2.70$, $p < .05$, suggesting that improvement in at least one condition was larger than improvement in at least one other condition. To explore this interaction further, we calculated difference scores by subtracting pretest from posttest proportions for each condition and then compared the three multimodal conditions with the audio-only condition. Only the audio-mouth condition (M = 0.14, SD = 0.10) produced a larger increase from pre- to posttest than the audio-only condition (M = 0.07, SD = 0.05), $tD(3, 28) = 2.29$, $p < .05$, whereas neither the audio-hands (M = 0.09, SD = 0.07), $tD(3, 28) = 0.47$, *ns,* nor the audio-mouth-hands (M = 0.05, SD = 0.10), $tD(3, 28) = 0.51$, *ns,* were different from the audio-only condition. Figure 3 presents the improvement (measured by subtracting the pretest scores from the posttest scores) for each of the multimodal conditions.

To address the possibility that the above interaction was driven by participants in different groups coming into the pretest at different competency levels, we further ran a one-way ANOVA across the four conditions for just the pretest. Importantly, there were no significant differences among conditions before instruction, $F(3, 56) = 0.94$, *ns,* and interestingly, there were no differences across the conditions after instruction either, $F(3, 56) = 0.87$, *ns*. However, when running an analysis of covariance (which used the pretest scores as a covariate) on the posttest scores, there is a marginally significant difference among the four groups, $F(3, 55) = 2.33$, p = .084. When the "adjusted" posttest scores of the audio-only condition are contrasted with the three multimodal posttests scores, only the audio-mouth condition produced significantly higher scores (p = .053). This, together with the significant interaction effect, suggests that it was the improvement—and not the absolute starting or finishing point—of the audio-mouth condition that separated it from the other multimodal training conditions.

Discussion

All groups in the present study made significant improvement (9 percentage points on average) from the pretest to the posttest. This improvement is similar to that in Hirata et al. (2007; 8–9 percentage points improvement by participants with three types of training). More specifically, the improvement of the audio-only condition in the present study (pre: 68%; post: 75%) was almost identical to the improvement in Hirata et al.'s (2007) slow-rate condition (pre: 67%; post: 75%), which used identical

Figure 3. Difference scores (in proportions) of the posttest minus pretest. Contrasting the three multimodal conditions with the baseline audio-only condition, only the audio-mouth group's improvement was significantly greater than that of the audio-only condition.



audio stimuli. Thus, the present audio-only results replicated Hirata et al. in that the amount of improvement was quite limited, and in both studies, strictly auditory training did not bring the participants' performance to the ceiling. Similar findings were reported in Tajima et al. (2008), who used different sets of phonemic length contrasts as auditory stimuli, and used more extensive training than Hirata et al. (2007). These results suggest that although nonnative speakers' perception of difficult phoneme contrasts does improve, there is room for further improvement.

Another general result seen for all four conditions is a high level of variability in their auditory abilities, as indicated by both the pre- and posttest scores in Figure 2. This high variability of scores across participants might explain, or have contributed to, the lack of significant differences among the nonadjusted posttest scores of the four groups. This variability could be the result of the age of our participants (Flege, 1995). Participants in the present study were college students with the age range of 19–23, a period known to yield greater performance variability than the younger age in second-language acquisition (Johnson & Newport, 1989; Oyama, 1976). In this sense, our results on native English adults learning Japanese vowel length contrast are similar to those in many other studies with different first- and second-language combinations (e.g., native Italian speakers learning English in Oyama, 1976).

Effects of Visual Input—Lipreading

One of the major goals of the present study was to identify the specific role that multimodal information plays—and does not play—in enhancing phonological processing abilities in a second language. Although all groups improved from pre- to posttest, the improvement was greatest when mouth movements, but not hand gestures, accompanied the auditory training. This finding is important because it suggests that the modest auditory/perception improvement in previous research using intensive auditory training alone (e.g., Hirata et al., 2007) could further be enhanced.

There is a long-established line of work demonstrating that visual information from the lips and mouth significantly impacts speech perception in one's native language (Green et al., 1991; Massaro & Cohen, 1983; McDonald et al., 2000; McGurk & MacDonald, 1976; Munhall et al., 1996; Skipper et al., 2007; Teder-Sälejärvi et al., 2002). And more recently, researchers have uncovered neural mechanisms for these behavioral effects, showing that information from the lips and mouth actually enhances neural activity to speech sounds in auditory brain regions (Calvert et al., 1997). AV input helps learners better perceive speech in a second language as well (Hardison, 1999, 2003; Wang et al., 2008). One explanation for these previous findings, as well as for the present audio-mouth results, is that the mouth conveys meaningful visual information that correlates with the sounds that it simultaneously accompanies. This natural coupling may create stronger perceptual traces of the phonemes (Calvert et al., 1997), which may make the speech sounds more salient and clear for later processing even in the absence of the visual information.

Although there are similarities with previous research, the present study extends this past work in an important way. The phonemic vowel length contrasts we studied entailed vowels that were qualitatively the same—they sounded and looked the same—but had different durations. With differences just in the length of the vowels, it was uncertain how well mouth movements would actually help participants learn these quantitative differences. However, results clearly indicated that even with this small difference in duration (approximately 100–200 ms), visual input from the mouth helped. This finding is especially notable given that this subtle temporal difference was embedded in a sentence, whereas many previous studies have used isolated syllables (Wang et al., 2008). In short, people seem to be sensitive to not only qualitative but also quantitative multimodal information conveyed through speech and mouth movements. The effects of seeing lips were small, however, and it would be interesting to investigate in future research whether more extensive training or training with more natural (i.e., non "lip-synched") multimodal stimuli might enhance these effects, and whether the learning would be sustained in the long term.

Effects of Visual Input—Hand Gesture

It is interesting that although mouth movements facilitated phoneme learning in the present study, hand movements did not. Of course, there may be methodological reasons for this null result. For example, even though we modeled our gestures after ones successfully used in previous research on language production (Roberge et al., 1996), we may have chosen the wrong type of gesture to distinguish long and short vowels for language perception. Another possibility is that participants in the two gesture conditions paid more attention to the visual distinctions between the hands than the auditory distinctions between the phonemes. However, even if participants were not explicitly paying attention to the speech, they were still implicitly processing it with the gesture—and if gesture was indeed beneficial, it should have still boosted learning (see Lim & Holt, 2009, for more on implicit learning). Alternatively, it is possible that beat gestures simply do not help in learning novel phonological contrasts. This possibility is interesting in light of research demonstrating that hand gestures do play an important role in phonological processing in one's native language

(Hubbard et al., 2008; Krahmer & Swerts, 2007). However, this previous research focused on the role that beat gestures play in the *suprasegmental* processing of speech. In contrast, the type of phonological processing in the present study focused on *segmental* processing—that is, whether beat gestures (i.e., long vs. short) could help to auditorily differentiate two small segmental units or vowels that differed only in duration. If gestures do not play a role in segmental processing, then it makes the results of the audio-mouth condition all the more interesting: It is possible that beat gestures are simply not as well suited as mouth movements to make such small and local temporal distinctions in language processing.

The fact that beat gestures did not help participants make phonemic distinctions is also interesting in light of other research that has uncovered beneficial effects of gesture on much higher linguistic levels in second-language learning (Kelly, McDevitt, & Esch, 2008; Quinn-Allen, 1995; Sueyoshi & Hardison, 2005). For example, Kelly, McDevitt, and Esch demonstrated that iconic gestures help people learn new word meanings (semantic level) in a second language. In the context of this previous work, one interpretation of the present results is that although gesture plays a clear semantic role in second-language learning, perhaps the phonological benefits are limited (see Skipper et al., 2009, for neural support for such a claim). This is provocative because it suggests that standard theories of gesture–speech integration—which focus primarily on the semantic level of analysis (Clark, 1996; Goldin-Meadow, 2003; Kendon, 2004, McNeill, 1992)—need to account for the possibility that gesture and speech may be integrated to different extents on different levels of language.

Effects of Lips and Hand

Another goal of the present study was to determine the relative contributions of different types of multimodal input—mouth movements and hand gestures—in phonological learning in a second language. The results of the audio-mouth-hands condition are particularly interesting in light of this goal. Even though there were three channels of information to help participants learn the novel phoneme contrasts, the audio-mouth-hands training was not as effective as when there were just two channels (audio-mouth) of information. This suggests that the third channel, the hand gesture channel, may have actually detracted from the benefits of the mouth movements. One possible explanation for this intriguing finding is that participants were "overloaded" with visual input, and this ultimately distracted them from reaping the benefits from the mouth and lips. That is, participants had to pay attention to auditory information along with two different pieces of visual information, and this integration may have overloaded working memory such that people could not properly encode the sounds into long-term memory (more on this memory hypothesis below). Indeed, although the improvement was significant, it is interesting to note that the audio-mouth-hands condition produced the smallest increment in learning (5 percentage points). Another possibility is that participants' visual attention was more heavily drawn to the gestures because they were bigger and more salient, and this caused them to not look at the more important and useful information conveyed through the mouth. In this way, viewing gestures may be distracting when learning phonemic elements of a second language. Both of these possibilities are interesting because they run contrary to almost all of the previous literature on the benefits of hand gesture in processing phonological, semantic, and pragmatic information in one's native language (for a review, see Kelly, Manning, & Rodak, 2008). Apparently, this is one case in which gestures are actually unhelpful, or worse yet, a liability.

The above points about cognitive overload and visual distraction have practical implications for second-language pedagogy. With the increasing availability of technology and other resources, there is a growing trend to use movie clips as language teaching materials, replacing the traditional audio-only "listen and repeat" method. We now know that it benefits learners to see how people articulate speech with respect to phonological learning and processing. However, this benefit might be cancelled out when too much visual information (e.g., hand gestures and body movements) accompanies language in the movie clips. Natural conversations with hand gestures and other body language might be too distracting for learners to reap the phonological benefits. Pedagogically, then, movie clips might need to be controlled (e.g., by way of editing, shortening, or simplifying) to a great extent, at least at relatively early stages of learning, to assure that learners can focus on relevant phonological information.

Along these lines, the findings have implications for special populations. For example, it has been long advocated that a "total communication" (TC) approach is an effective means to teach verbal aspects of language to hearing-impaired individuals (Matkin & Matkin, 1985). This approach couples auditory input with a wide range of accompanying visual information (lip movements, conventionalized sign, spontaneous cospeech gesture, finger spelling, pictures, etc.) to facilitate teaching different aspects of spoken and written language. The benefits of this approach, however, have recently been challenged (Geers & Brenner, 2004; Geers, Brenner, & Davidson, 2003). For example, Geers and Brenner (2004) studied 8- and 9-year-old children with cochlear implants (received before the age of 5) and found that strictly oral methods of instruction were superior to a TC approach in multiple outcomes, including speech perception, speech production, and reading. One possible interpretation of this finding is that TC

involves so much visual information that it distracts or overwhelms learners, making it difficult to attend to and benefit from the auditory input. Our results suggest that a middle ground—with oral instruction accompanied by only visually salient mouth movement—might be the optimal approach for teaching novel speech sounds to hearing and hearing-impaired individuals alike.

A Theoretical Perspective

As a final point, we would like to situate our findings into a relevant theory concerning phonological memory (Baddeley, 1986). Baddeley and colleagues have proposed a model in which the phonological loop—a component of working memory—is a mechanism for how people remember new linguistic items during language learning (Baddeley, Gathercole, & Papagno, 1998). According to the model, the phonological loop is dedicated to holding verbal information in working memory for short periods of time while that information can be rehearsed and ultimately transferred into long-term memory. This model has obvious implications for how people learn new speech sounds: To learn distinctions between novel speech sounds, there must be some mechanism for encoding and transferring those sounds into more permanent memory.

In fact, the role of the phonological loop in learning language is so important that there are claims that it is *specifically dedicated* to the job of learning a new language, either as a child acquiring a first language or as an adult learning a second (or third or fourth) language (Baddeley et al., 1998). According to the theory, difficulties in learning a new language arise when the phonological loop is taxed with novel speech sounds, and this disrupts the encoding of those novel sounds into permanent memories for new words. For example, when adults are asked to remember words in a language that contains phonologically familiar sounds (e.g., from one's native tongue), they do not have to rely heavily on the phonological loop when encoding the words into long-term memory (after all, they already have these sound representations in their repertoire), and performance is strong. However, when asked to remember words that do have novel speech sounds, the phonological loop is heavily taxed (i.e., people need to create new memories for these sounds), and performance is weak (Papagno, Valentine, & Baddeley, 1991; Papagno & Vallar, 1992). Thus, it appears that the phonological loop plays a special and dedicated role in learning new speech sounds in a second language.

Couched in this theory, the results from the present study make sense. The phonological loop of English speakers may be taxed by the novel vowel length distinctions in Japanese, and this may make it difficult to transfer those distinctions from working memory to more permanent long-term memories. However, with repetition and practice (i.e., instruction), people do eventually encode these sounds into long-term memory. Our study adds to this by demonstrating that visual input can enhance this process: We found that certain types of visual information—for example, lip movements—appear to strengthen this memory. Future research will be necessary to examine more directly how visual information interacts with verbal processing in the phonological loop to facilitate the transfer of novel speech sounds into long-term memory when learning a second language.[3]

Conclusion

In the fields of phonetics and second-language acquisition, most research to date has focused on the question of how one might maximize nonnative speakers' learning of difficult second-language phonemes by using different types of auditory training (e.g., a variety of voices and diverse phonetic contexts; Bradlow et al., 1997; Hirata, 2004a; Logan et al., 1991; Pisoni & Lively, 1995). The results from the present study provide insights into the added benefits of other modalities—above and beyond the auditory modality—in helping learners to further improve their ability to perceive and learn the distinction of difficult phonemic contrasts. We found that multimodal input from lips and speech combine to facilitate greater learning than just auditory input alone. This suggests that the limits of training learners with auditory input alone (e.g., Hirata et al., 2007) can be exceeded.

Finally, the present results suggest that although mouth movements are beneficial, hand gesture may not help auditory/perceptual learning of difficult phonemic contrasts. These results contrast with previous research in which multimodal input from speech and hand gesture does facilitate second-language learning on the semantic level, such as learning new vocabulary. Given the richness of natural face-to-face communication—which includes speech, mouth movements, hand gestures, and a whole host of other nonverbal behaviors—it will be important for future research to determine the optimal multimodal conditions for teaching and learning the many different aspects of a second language.

Acknowledgments

This study was supported by Harvey M. Picker Institute for Interdisciplinary Studies in the Sciences and Mathematics at Colgate University. We thank Emily Cullings, Jason Demakakos, Jackie Burch, Jen Simester, and Grace Baik for

[3]In Baddeley's original model of working memory, there was a visual component referred to as the visuospatial sketchpad (Baddeley & Hitch, 1974). The role of this component in language learning is not well understood, and for this reason, we do not speculate about it here. Nevertheless, it appears that visual input does play an important role in phoneme learning, and this opens the door to research on just *how* this information interacts with the phonological loop to help encode novel speech sounds into permanent memory.

their involvement at various stages this project. Portions of this study were presented at the conferences Acoustics '08 in Paris and WorldCall 2008 in Fukuoka as well as invited colloquiums at Advanced Telecommunications Research Institute International in Kyoto in 2008 and the Max Planck Institute for Psycholinguistics in Nijmegen (the Netherlands) in 2009.

References

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

Baddeley, A. D. (1986). *Working memory.* Oxford, England: Oxford University Press.

Baddeley, A. D., Gathercole, S. E., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105,* 158–173.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. A. Bower (Ed.), *Recent advances in learning and motivation* (Vol. 8, pp. 47–90). New York: Academic Press.

Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica, 123,* 1–30.

Bradlow, A. R., Pisoni, D. B., Yamada, R. A., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/-/l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America, 101,* 2299–2310.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., et al. (1997, April 25). Activation of auditory cortex during silent lipreading. *Science, 276,* 593–596.

Cassell, J., McNeill, D., & McCullough, K. E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics and Cognition, 7,* 1–34.

Church, R. B., Ayman-Nolley, S., & Mahootian, S. (2004). The role of gesture in bilingual education: Does gesture enhance learning? *International Journal of Bilingual Education and Bilingualism, 7,* 303–319.

Clark, H. H. (1996). *Using language*. Cambridge, England: Cambridge University Press.

Desai, S., Stickney, G., & Zeng, F. G. (2008). Auditory-visual speech perception in normal-hearing and cochlear-implant listeners. *Journal of the Acoustical Society of America, 123,* 428–440.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language speech research* (pp. 233–277). Timonium, MD: York Press.

Fujisaki, H., Nakamura, K., & Imoto, T. (1975). Auditory perception of duration of speech and non-speech stimuli. In G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 197–219). London: Academic Press.

Geers, A. E., & Brenner, C. A. (2004). Educational intervention and outcomes of early cochlear implantation. *International Congress Series, 1273,* 405–408.

Geers, A., Brenner, C., & Davidson, L. (2003). Factors associated with development of speech perception skills in children implanted by age five. *Ear and Hearing, 24,* 24–35.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge, MA: Belknap Press.

Goldin-Meadow, S., Wein, D., & Chang, C. (1992). Assessing knowledge through gesture: Using children's hands to read their minds. *Cognition and Instruction, 9,* 201–219.

Grant, K. W., & Seitz, P. F. (1998). Measures of AV integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America, 104,* 2438–2450.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics, 50,* 524–536.

Hardison, D. M. (1999). Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect. *Language Learning, 49,* 213–283.

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics, 24,* 495–522.

Hardison, D. M. (2005). Second-language spoken word identification: Effects of perceptual training, visual cues, and phonetic environment. *Applied Psycholinguistics, 26,* 579–596.

Hazen, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication, 47,* 360–378.

Hillenbrand, J., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *Journal of the Acoustical Society of America, 108,* 3013–3022.

Hirata, Y. (2004a). Effects of speaking rate on the vowel length distinction in Japanese. *Journal of Phonetics, 32,* 565–589.

Hirata, Y. (2004b). Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts. *Journal of the Acoustical Society of America, 116,* 2384–2394.

Hirata, Y., & Tsukada, K. (2009). Effects of speaking rate and vowel length on formant frequency displacement in Japanese. *Phonetica, 66,* 129–149.

Hirata, Y., Whitehurst, E., & Cullings, E. (2007). Training native English speakers to identify Japanese vowel length with sentences at varied speaking rates. *Journal of the Acoustical Society of America, 121,* 3837–3845.

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience, 19,* 1175–1192.

Hubbard, A. L., Wilson, S. M., Callan, D. E., & Dapretto, M. (2008). Giving speech a hand: Gesture modulates activity in auditory cortex during speech perception. *Human Brain Mapping, 30,* 1028–1037.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21,* 60–99.

Kelly, S. D., Barr, D., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of

speech and gesture in comprehension and memory. *Journal of Memory and Language, 40,* 577–592.

Kelly, S. D., & Church, R. B. (1998). A comparison between children's and adults' ability to detect conceptual information conveyed through representational gestures. *Child Development, 69,* 85–93.

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language, 89,* 253–260.

Kelly, S. D., Manning, S., & Rodak, S. (2008). Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education. *Language and Linguistics Compass, 2,* 1–20.

Kelly, S. D., McDevitt, T., & Esch, M. (2008). Neural correlates of learning Japanese words with and without iconic gestures. *Language and Cognitive Processes, 24,* 313–324.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge, England: Cambridge University Press.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language, 57,* 396–414.

Kubozono, H. (2002). Temporal neutralization in Japanese. In C. Gussenhoven & N. Warner (Eds.), *Papers in laboratory phonology VII* (pp. 171–201). New York: Mouton de Gruyter.

Landahl, K., & Ziolkowski, M. (1995). Discovering phonetic units: Is a picture worth a thousand words? *Papers from the 31st Regional Meeting of the Chicago Linguistic Society, 1,* 294–316.

Lim, S. J., & Holt, L. L. (2009). Investigating non-native category learning using a video-game-based training paradigm. *Journal of the Acoustical Society of America, 125,* 2768.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America, 89,* 874–886.

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 753–771.

Matkin, A. M., & Matkin, N. D. (1985). Benefits of total communication as perceived by parents of hearing-impaired children. *Language, Speech, and Hearing Services in the Schools, 16,* 64–74.

McDonald, J. J., Teder-Sälejärvi, W. A., & Hillyard, S. A. (2000, October 19). Involuntary orienting to sound improves visual perception. *Nature, 407,* 906–908.

McGurk, H., & MacDonald, J. (1976, December 23). Hearing lips and seeing voices. *Nature, 264,* 746–748.

McNeill, D. (1992). *Hand and mind: What gesture reveals about thought*. Chicago: University of Chicago Press.

Morosan, D., & Jamieson, D. G. (1989). Evaluation of a technique for training new speech contrasts: Generalization across voices, but not word-position or task. *Journal of Speech and Hearing Research, 32,* 501–511.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58,* 351–362.

Oyama, S. (1976). A sensitive period for the acquisition of a nonnative phonological system. *Journal of Psycholinguistic Research, 5,* 261–283.

Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience, 19,* 605–616.

Papagno, C., Valentine, T., & Baddeley, A. D. (1991). Phonological short-term memory and foreign language vocabulary learning. *Journal of Memory and Language, 30,* 331–347.

Papagno, C., & Vallar, G. (1992). Phonological short-term memory and the learning of novel words: The effect of phonological similarity and item length. *Quarterly Journal of Experimental Psychology, 44,* 47–67.

Pisoni, D. B., & Lively, S. E. (1995). Variability and invariance in speech perception: A new look at some old problems in perceptual learning. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language speech research* (pp. 433–459). Timonium, MD: York Press.

Quinn-Allen, L. (1995). The effects of emblematic gestures on the development and access of mental representations of French expressions. *The Modern Language Journal, 79,* 521–529.

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. 97–113). Hillsdale, NJ: Erlbaum.

Roberge, C., Kimura, M., & Kawaguchi, Y. (1996). *Nihongo no Hatsuon Shidoo: VT-hoo no Riron to Jissai* [Pronunciation training for Japanese: Theory and practice of the VT method]. Tokyo: Bonjinsha.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics, 59,* 73–80.

Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., & Small, S. L. (2009). Gesture orchestrates brain networks for language understanding. *Current Biology, 19,* 661–667.

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,* 2387–2399.

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

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

Tajima, K., Kato, H., Rothwell, A., Akahane-Yamada, R., & Munhall, K. (2008). Training English listeners to perceive phonemic length contrasts in Japanese. *Journal of the Acoustical Society of America, 123,* 397–413.

Tajima, K., Rothwell, A., & Munhall, K. G. (2002). Native and non-native perception of phonemic length contrasts in Japanese: Effect of identification training and exposure. *Journal of the Acoustical Society of America, 112,* 2387.

Teder-Sälejärvi, W. A., McDonald, J. J., Di Russo, F., & Hillyard, S. A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Cognitive Brain Research, 14,* 106–114.

Tsukada, K. (1999). *An acoustic phonetic analysis of Japanese-accented English*. Unpublished doctoral dissertation, Macquarie University, Sydney, Australia.

Ueyama, M. (2000). *Prosodic transfer: An acoustic study of L2 English vs. L2 Japanese*. Unpublished doctoral dissertation, University of California, Los Angeles.

Wang, Y., Behne, D., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *Journal of the Acoustical Society of America, 124,* 1716–1726.

Wu, Y. C., & Coulson, S. (2007). How iconic gestures enhance communication: An ERP study. *Brain and Language, 101,* 234–245.

Yamada, T., Yamada, R. A., & Strange, W. (1995). Perceptual learning of Japanese mora syllables by native speakers of American English: Effects of training stimulus sets and initial states. *Proceedings of the 14th International Congress of Phonetic Sciences, 1,* 322–325.

Received November 18, 2008

Accepted August 24, 2009

DOI: 10.1044/1092-4388(2009/08-0243)

Contact author: Yukari Hirata, Department of East Asian Languages and Literatures, 9B Lawrence Hall, Colgate University. E-mail: yhirata@mail.colgate.edu.