JSLHR

Research Article

Effects of Hand Gestures on Auditory Learning of Second-Language Vowel Length Contrasts

Yukari Hirata,[a] Spencer D. Kelly,[a] Jessica Huang,[a] and Michael Manansala[a]

Purpose: Research has shown that hand gestures affect comprehension and production of speech at semantic, syntactic, and pragmatic levels for both native language and second language (L2). This study investigated a relatively less explored question: Do hand gestures influence auditory learning of an L2 at the segmental phonology level?
Method: To examine auditory learning of phonemic vowel length contrasts in Japanese, 88 native English-speaking participants took an auditory test before and after one of the following 4 types of training in which they (a) observed an instructor in a video speaking Japanese words while she made syllabic-rhythm hand gesture, (b) produced this gesture with the instructor, (c) observed the instructor speaking those words and her moraic-rhythm hand gesture, or (d) produced the moraic-rhythm gesture with the instructor.
Results: All of the training types yielded similar auditory improvement in identifying vowel length contrast. However, observing the syllabic-rhythm hand gesture yielded the most balanced improvement between word-initial and word-final vowels and between slow and fast speaking rates.
Conclusions: The overall effect of hand gesture on learning of segmental phonology is limited. Implications for theories of hand gesture are discussed in terms of the role it plays at different linguistic levels.

Research in phonetic science and second language (L2) acquisition has progressed over the past several decades, investigating how and why adults are limited in learning to perceive and produce an L2 (Piske, MacKay, & Flege, 2001; Strange, 1995). Even though adults plateau in learning to perceive certain L2 phonemes when taking classes or living in an L2-speaking country, their auditory inability can be helped by intensive auditory training in a laboratory (Bradlow, Pisoni, Yamada, & Tohkura, 1997; Logan, Lively, & Pisoni, 1991; Pisoni & Lively, 1995). One of the well-studied problems in this field is nonnative adults' inability to perceive phonemic vowel length contrasts in Japanese (for a review, see Hirata, in press). The length of vowels, whether short or long, is phonemic in Japanese (e.g., /dʒo/ with a short vowel means "introduction," but /dʒoː/ with a long vowel means "emotion"). The only difference between short and long vowels is that of duration. Long vowels are 2.2–3.2 times longer in duration than short vowels (Tsukada, 1999), but the difference between them could be as small as 50 ms when vowels are spoken quickly in a sentence (Hirata, 2004a). Because there is no such phonemic distinction in English, native English speakers have difficulty perceiving this distinction, although auditory training does improve their perception (Hirata, 2004b).

Auditory Learning of Difficult Japanese Speech Contrasts

Two major factors are known to affect L2 learners' auditory difficulty in the distinction of short and long vowels: speaking rate and position of the contrasting vowels within words. With regard to the speaking rate, training in one rate does not generalize to one's ability to distinguish the same vowel length contrasts spoken at a faster rate (Hirata, Whitehurst, & Cullings, 2007; Tajima, Kato, Rothwell, Akahane-Yamada, & Munhall, 2008). However, an L2 training method with higher variability in speaking rate

[a]Center for Language and Brain, Colgate University, Hamilton, NY
Correspondence to Yukari Hirata: yhirata@colgate.edu

Editor: Jody Kreiman
Associate Editor: Ewa Jacewicz
Received February 13, 2014
Revision received June 1, 2014
Accepted July 19, 2014
DOI: 10.1044/2014_JSLHR-S-14-0049

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

enables more rate-general auditory learning[1] (Hirata et al., 2007), consistent with Pisoni and Lively's (1995) high phonetic variability hypothesis. Regarding the position of short and long vowels within a word as the second major factor, L2 learners have more difficulty in accurately identifying the vowel length when the vowels are in the word-final position (e.g., /joko/ "side" vs. /jokoː/ "rehearsal") than in the word-initial position (e.g., /soka/ "simple refreshments" vs. /soːka/ "flower arrangement"; Minagawa, Maekawa, & Kiritani, 2002).

Despite the fact that training does help adult L2 speakers to overcome some of these challenges, there is currently no available training method that brings them to the native level in perceiving these difficult L2 phonemic contrasts. The present study was an attempt to search for a more effective training method that enhances such potential perceptual learning. We investigated the extent to which visual input conveyed through hand gestures helps native English speakers to auditorily perceive Japanese vowel length contrasts and learn new words. Given the variety of roles that hand gestures play in language comprehension and learning (as reviewed below in the following three subsections), the first question we addressed was whether auditory learning takes place with the proposed training methods using hand gestures regarding L2 vowel contrasts that vary in speaking rate and in word-internal position, particularly in the difficult contexts, such as word-final position spoken at a fast rate. The present training method was modeled after that in Hirata et al. (2007) using stimuli of two speaking rates. To enable generalized auditory learning, we also trained participants with vowel contrasts both in the word-initial and the word-final positions (see Table 1). We examined whether the patterns of auditory difficulty found in previous studies are replicated, and whether there are different amounts of auditory improvement on word-initial and word-final vowel length contrasts spoken at a slow rate and a fast rate.

This study was designed and developed on the basis of a large body of previous literature on L2 acquisition by learners at various levels from naive monolingual speakers to advanced learners. The present investigation was limited to examining effects of the proposed training method for monolingual native speakers of American English with no knowledge of Japanese, but we hope that this line of investigations will continue to include L2 learners of Japanese with more experience with the language.

Roles of Hand Gesture in Language Comprehension and Learning

Hand gestures that accompany speech, or *cospeech gestures,* are a pervasive part of spoken communication. Researchers have theorized that speech and gesture together are a fundamentally integrated system of communication (Kendon, 2004; McNeill, 1992, 2005). These theories originated in the realm of language production, but researchers have recently argued that this integrated relationship extends to the comprehension domain as well (see Kelly, Özyürek, & Maris, 2010). This is not limited to one's native language, as gestures also facilitate L2 learning in adults (Kelly, McDevitt, & Esch, 2009; Quinn-Allen, 1995; Sueyoshi & Hardison, 2005; Tellier, 2008). For example, Kelly et al. (2009) found that iconic hand gestures (visually depicting object motions, attributes, and spatial relationships) that accompanied spoken words (e.g., saying "*nomu* means 'drink'" with drinking gesture) helped English speakers learn Japanese words.

Although the studies above focused on the role that iconic gestures play on the semantic and pragmatic domains, much less is known about gestures that are associated with phonological aspects of speech. In one of the few studies, Krahmer and Swerts (2007) demonstrated that listening/watching words in sentences with beat gestures—quick flicks of the hand—increased native speakers' perception of the acoustic prominence of those words. Corroborating this finding, a recent functional magnetic resonance imaging study has shown that low-level auditory brain areas, such as the planum temporale, are more active during comprehension of a native language when beat gestures accompany speech than when speech is presented alone (Hubbard, Wilson, Callan, & Dapretto, 2008). In fact, research using event-related potentials (ERPs) suggests that beat gestures influence the brain's processing of phonemes as early as 100 ms of hearing a spoken word (Biau & Soto-Faraco, 2013).

One of the few empirical studies on the role of hand gestures in L2 phonological learning is Hirata and Kelly (2010), in which native English speakers saw videos of speakers producing Japanese short and long vowels with and without short and long beat gestures (shown in the syllable gestures in Figure 1). These gestures were referred to as *beats* because they conveyed temporal information about the accompanying syllables, although one could also describe those hand movements as *metaphorics* because the length of the visual gesture metaphorically mapped onto the length of the corresponding spoken vowel. Hirata and Kelly found that participants did not learn to perceive the short/long vowel contrasts in the speech–gesture condition any better than the speech–alone condition. One interpretation of this result is that hand gestures might not play a role in the segmental processing of speech, suggesting a potential lower limit of the integration of gesture and speech in language comprehension. However, the authors also

[1] In this article, we use the terms *auditory learning* and *auditory ability* as equivalent to *perceptual learning* or *learning or ability to identify* nonnative phonemic contrasts in phonetics and speech science literature. This was intended simply to distinguish these terms from visual perception and visual learning, or from integration of visual and auditory stimuli as involved in our study. We are aware that what is really meant by the "auditory" ability is extremely complex and multidimensional in speech, psychoacoustic, and perception research (e.g., Kidd, Watson, & Gygi, 2007). Even within perception of speech, the tasks of identification versus discrimination measure different aspects of people's perception system (e.g., Tsukada, 2011). In the present study, we avoid using the term *perceptual learning* to simply prevent confusion from visual perception and learning during training that we conducted.

Table 1. Training stimuli.

Word-initial vowel contrasts				Word-final vowel contrasts			
Word	**Length**	**Pitch accent**	**Meaning**	**Word**	**Length**	**Pitch accent**	**Meaning**
seki	SS	HL	seat	çaɾe	SS	LH	joke
seːki	LS	HLL	century	çaɾeː	SL	LHH	honorarium
kedo	SS	HL	but	goke	SS	LH	widow
keːdo	LS	HLL	slight degree	gokeː	SL	LHH	word form
toço	SS	HL	book	joko	SS	LH	side
toːço	LS	HLL	at the beginning	jokoː	SL	LHH	rehearsal
koɟʑi	SS	HL	orphan	iso	SS	LH	seashore
koːɟʑi	LS	HLL	construction	isoː	SL	LHH	transport
kuɾo	SS	HL	black	ɟʑiçu	SS	LH	to turn yourself in
kuːɾo	LS	HLL	air path	ɟʑiçuː	SL	LHH	self-study

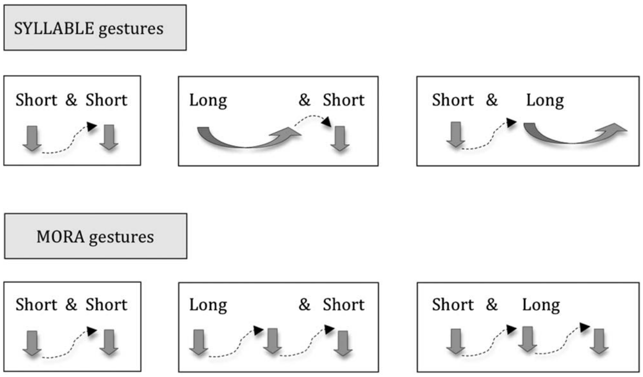
Note. In the "Length" column, S = short; L = long. In the "Pitch accent" column, H = high; L = low.

pointed out a possibility that the study might not have used the most effective type of gesture. This led to the present study that considers another type of gesture (shown in the mora gestures in Figure 1).

Syllable Versus Mora Gestures

The basic rhythmic unit of Japanese is the *mora* (Ladefoged, 1975), which is similar to the notion of *syllable* except that the mora is duration sensitive. For example, a Japanese short vowel (by itself or with a preceding consonant, e.g., /dʒo/ "introduction") counts as one mora, and a long vowel (by itself or with a preceding consonant, e.g., /dʒoː/ "emotion") counts as two moras, although they are both just one syllable. Native Japanese speakers' rhythm consists of equal beats of moras (Vance, 1987), and thus, /dʒo/ has one beat, and /dʒoː/ has two equal beats. Abundant research has shown that the duration of words corresponds to the number of moras that they contain (although a durational increment for an addition of a mora in a word is smaller for faster speech; Han, 1994, Hirata & Whiton, 2005; Port, Dalby, & O'Dell, 1987). Rhythmicity of moras including long vowels is not perfectly manifested in spoken utterances —for example, two- and three-mora word pairs such as /ise/ (name of a place) and /iseː/ "opposite gender" had mean ratios ranging from 2 to 2.70–2.95 in Hirata (2004a), not quite reaching a perfect 2–3. However, when a mora is shorter than what it should be, neighboring moras compensate to balance the overall mora timing (Brady & Port, 2007). According to one proposal (Brady & Port, 2007; Port, Cummings, & Gasser, 1995), native Japanese speakers' *adaptive oscillator,* that is, a rhythm-detecting neural mechanism, makes them perceive imperfect rhythms of speech signals as more rhythmic so as to perceive regular timing of moras.

Figure 1. Hand movements used in the syllable and mora conditions in training.



Given native English speakers' syllable system, it is intuitive for them to perceive Japanese words according to syllables (Hirata, 2004b). Referring to the syllable gestures in Figure 1 (e.g., for the word /seːki/, a long vowel followed by a short vowel), this translates visually into a gesture with one long dip (Roberge, Kimura, & Kawaguchi, 1996) followed by a short downward chopping movement. In contrast, it is rhythmically intuitive for native Japanese speakers to see the same word, /seːki/ (with three moras), as having three chopping downward movements coinciding with their speech, as shown in the mora gestures in Figure 1. This idea of three mora beats is intuitive for native Japanese instructors to actually use in their teaching and is well supported by the adaptive oscillator model by Brady and Port (2007) and Port et al. (1995). It may be counterintuitive for native English speakers to perceive the first long vowel as having two beats, but this apparent incongruence may help them recognize the rhythmic beats that are different from those of English. Thus, the second question we addressed in the present study was whether the unfamiliar and incongruent mora gestures (see Figure 1) help learners to auditorily perceive the vowel length distinction better than the intuitive and congruent gestures of syllables (see Figure 1). There is a reason to believe that this counterintuitive gesture may promote learning by highlighting new and useful strategies that help the learner, according to Goldin-Meadow's (2003, 2010) mismatching gesture hypothesis. Goldin-Meadow's work focused mainly on children learning mathematics and other conceptual problems, but if it extends to learning of L2 speech rhythm, the mora condition should yield more effective learning than the syllable condition.

Observing Versus Producing Hand Gestures

As the third question in the proposed study, we asked whether producing gestures oneself, in addition to observing gestures of other people, facilitates even greater auditory learning. There are good reasons to believe that observing and producing gestures may be different from just observing them in the context of learning. Indeed, imitation is a powerful learning tool, perhaps because of the many neural mechanisms that link others actions with one's own actions (Iacoboni, 2005). With specific regard to gesture, neuroimaging work has demonstrated that imitating gestures activates a more distributed network of neural regions than simply observing gestures (Montgomery, Isenberg, & Haxby, 2007). Finally, producing gestures has been shown to help children learn challenging mathematical concepts (Cook, Mitchell, & Goldin-Meadow, 2008), and producing gestures is more effective than observing gestures in learning lists of sentences (the *enactment effect*; Engelkamp & Dehne, 2000).

Research has also demonstrated how gesture production and speech interact during L2 learning (for reviews, see Gullberg, 2006; Gullberg, de Bot, & Volterra, 2008). In the field of L2 pedagogy, there have been some suggestions as to how learning can be assisted by the use of physical actions or gestures associated with auditory speech sounds, for example, Asher's (1969) total physical response technique. Some studies showed positive effects of producing iconic gestures—such as for running, writing, and eating—on the learners' improvement in oral comprehension of word meaning (e.g., Gary, 1978). More recently, Macedonia, Müller, and Friederici (2011) showed that imitating iconic cospeech gestures helps adults to remember the meaning of words in an invented language more than imitating unrelated hand movements. In an observational account of the role of gesture production in L2 instruction, Roberge et al. (1996) taught nonnative speakers to produce beats of differing lengths to differentiate short and long vowels in Japanese. They observed that these hand gestures helped nonnative speakers make significant progress in their short and long vowel production in Japanese. This suggests that producing hand gestures may be a powerful way of learning novel phoneme distinctions in an L2. As mentioned earlier in the Roles of Hand Gesture in Language Comprehension and Learning section, Hirata and Kelly (2010) showed no effect of observing (syllable) gestures (see Figure 1) on L2 learners' auditory learning (compared with listening to audio only), and it is possible that the hand gestures are helpful only when learners produce, instead of just observe, them.

In summary, despite the potential differences between gesture observation and production, no study, to our knowledge, has directly compared whether *observing* versus *observing and producing* (*producing* henceforth) gestures play different roles in perceiving and learning L2 speech, and thus this is one focus of our study. This question is important for theories of gesture and more generally for theories on the embodiment of language and learning, which claim that high-level cognitive processes are deeply rooted in the body (Barsalou, 1999; Decety & Grezes, 2006; Fischer & Zwaan, 2008; Glenberg & Kaschak, 2002; Rizzolatti & Craighero, 2004).

Overview of the Experiment

The present experiment compared effects of four types of training—Syllable-Observe (SO), Syllable-Produce (SP), Mora-Observe (MO), and Mora-Produce (MP)—on auditory learning of Japanese vowel length contrasts in the word-initial and the word-final positions spoken at slow and fast rates. We note that the present SO condition is the same training as one used in Hirata and Kelly (2010), except that we used different word pairs of phonemic vowel length contrasts. Another difference is that the present training involved asking participants to remember English translations of Japanese words (i.e., vocabulary learning), but this component will be reported in a separate article.

Method

Participants

Eighty-eight right-handed monolingual native speakers of English (men and women) with no knowledge of Japanese language (18–23 years of age) were recruited from undergraduate students at a liberal arts college in the northeastern United States. A questionnaire also screened participants so that no participants grew up in bilingual family environments nor had extensive auditory input of Japanese prior to the experiment. Participants' experience with the formal study of foreign languages included less than 6 years of French, Spanish, German, Italian, Russian, Mandarin Chinese, Arabic, Hebrew, Latin, or Greek. None of the participants had more than 6 years of continual music training.

These participants were assigned randomly to one of four conditions (n = 22 in each condition): SO, SP, MO, and MP.

Stimuli and Procedure

The overall structure of the experiment for all participants was as follows: an auditory pretest for Day 1, four sessions of training for Days 2 and 3, a vocabulary test and an ERP test for Day 4 (results of which are not reported in the present article), and an auditory posttest for Day 5.

Training stimuli. Ten pairs of Japanese words contrasting in length of vowels /e o u/ were used as training stimuli (see Table 1). Five pairs had a contrasting vowel in the first syllable (e.g., /seki/ "seat" vs. /seːki/ "century"), and the other five pairs had a contrasting vowel in the second syllable (e.g., /joko/ "side" vs. /jokoː/ "rehearsal"). The first five word pairs, in which the vowel length contrasts were in the first syllable, had the pitch accent patterns of HL (H = high; L = low) and HLL, and thus the contrasting vowels received H versus HL pitch patterns. The pitch accent patterns of the other five pairs, in which the vowel length contrasts were in the second syllable, were LH and LHH, and thus the contrasting vowels received H versus HH pitch patterns. We were aware that the pitch accent covaried with the word-internal position of contrasting vowels, and this was due to constraints of our experimental design. Minagawa et al. (2002) found that learners of Japanese had more perceptual errors when a long vowel was in the word-final position in the LL accent pattern (e.g., /joto:/ HLL "a ruling government party") than in the HL or HH patterns. If vowels in the word-final position cause auditory difficulty, regardless of their accent patterns as in Minagawa et al.'s study, we would expect the present word-final HH long vowels to be still more difficult than the word-initial long vowels.

For auditory stimuli, two right-handed female native speakers of Japanese spoke these words in isolation twice, each with slow and fast speaking rates. The definition of *slow* speaking rate was "slower than one's normal rate, clearly enunciating," and *fast* speaking rate was "faster than one's normal rate but still comfortable and accurate." After the speakers received this definition and practiced, it was up to each speaker to determine the actual speaking rates. A total of 80 digital audio files (10 word pairs × 2 lengths × 2 repetitions × 2 speakers) were made. These audio files were used in the auditory portion of training (see Step 1 in Figure 2).

For visual stimuli, the same two speakers above were videotaped, and 40 video clips (10 word pairs × 2 lengths × 2 speakers) were made. These video clips were used in the auditory–visual portion of training (see Steps 4 and 6 in Figure 2). For each video clip for short vowel words (e.g., /seki/ or /joko/), the speaker said a word with the hand gesture of two small downward chopping movements. For words with long vowels (e.g., /seːki/ or /jokoː/), the speakers made two clips, one for the syllable condition and the other for the mora condition. For the syllable condition, a long vowel was represented by the speaker's hand making one horizontal dip as in Figure 1 (as in Roberge et al., 1996), followed or preceded by a short vowel represented in a small downward chopping movement. For the mora condition, a long vowel was represented by the speaker's hand making two small downward chopping movements as he or she spoke, and the total number of vertical movements in long-vowel words (e.g., /seːki/ or /jokoː/) was three, corresponding with the number of moras in them. The video clips showed the upper half of the body, including the speaker's face, speaking the word. (Lip movements play a significant role in auditory learning as in Hirata & Kelly, 2010, but the present study focused only on gesture by having the mouth visible in all conditions.)

Prior to videotaping, the two native Japanese speakers received precise instructions as to how their hands should move and how their speaking and gesturing rates should be. The definition of the speaking rate was the same as when their audio recordings were made. The speakers used their right hand, and the videos were digitally flipped so that it appeared to be the left hand (see Figure 2). This was intended to make it easier for participants to mirror the gestures they saw with their own right hand.

After auditory and video stimuli were created separately, we deleted the audio in the original video, and we dubbed the auditory stimuli that were made solely for the audio recording onto the video stimuli. This was done because the acoustic properties of speech are affected by cospeech hand gesture (Krahmer & Swerts, 2007), and we wanted to ensure that the four conditions contained identical auditory information so that any differences across conditions could be attributed to gesture and not actual differences in the acoustic signal.

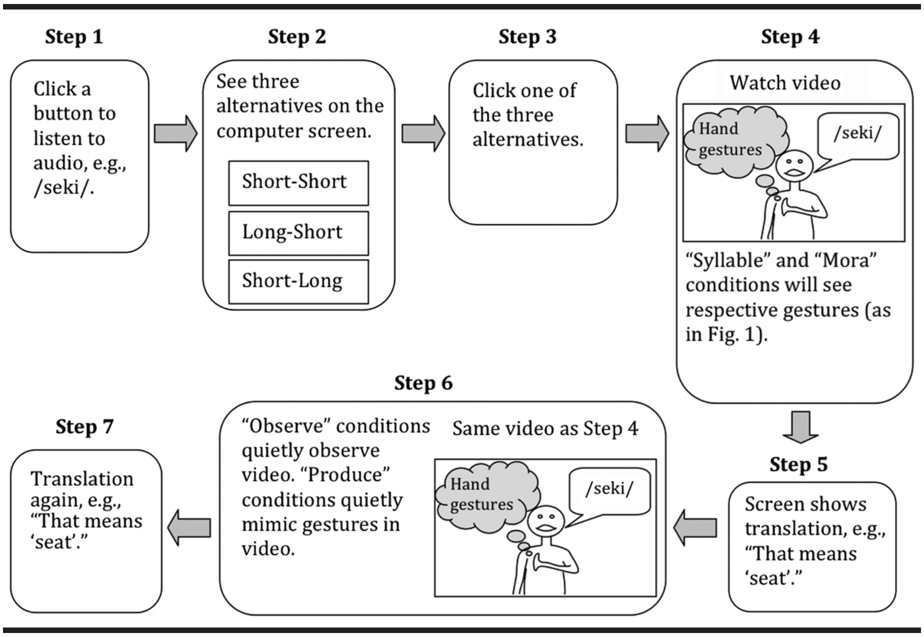
Training procedure. Participants completed four training sessions in 2 days. The 2 training days were separated by at least 1 day and no more than 3 days (as in Hirata & Kelly, 2010). Each training session contained 80 trials, including all of the 20 words spoken by the two speakers repeated twice, presented in a randomized order. Session 1 contained only slow rate stimuli, Session 2 had only fast rate stimuli, and Sessions 3 and 4 had both slow and fast rate stimuli mixed.

Seven steps were involved for each trial of the 20 words, summarized in Figure 2. In Steps 1–3, all participants listened to the audio file of the word (produced by one of the two Japanese speakers), saw three choices on the computer screen and the corresponding keys on the keyboard, that is, "short–short" (as in /seki/ or /joko/), "long–short" (as in /seːki/), and "short–long" (as in /jokoː/), and chose one of them that corresponded to what they heard. In Step 4, participants watched a video clip in which the speaker said the word along with the accompanying hand gestures (in this way, participants received indirect feedback through the gestures as to whether they were correct in Step 3). Participants in the syllable and mora conditions saw the Japanese speakers gesturing syllables and moras, respectively, as in Figure 1. In Step 5, all participants saw an English translation of the word written on the screen. In Step 6, participants saw a countdown, "3," "2," and "1" on the screen in 3 s, and they saw the same videos as in Step 4. At this time, those in SO and MO groups quietly observed the respective videos, whereas those in SP and MP groups mimicked the respective gestures in the videos. In Step 7, the translation of the word was visually presented to all participants again as in Step 5. Participants then pressed the space bar to listen to the next word (Step 1). For all of the four conditions, participants were silent the whole time. Steps 1 and 3 (pressing the keys on the keyboard to play the audio and to choose one of the three alternatives) were self-paced, but the rest of the steps were automated by a computer program.

During training, participants were monitored through a live video camera by the experimenters to ensure adherence to their expected tasks in the four conditions. To motivate participants, the participants were told at the beginning of the first training session that the person who improved most from a pretest to a posttest (for each condition) would receive a prize.

Auditory test stimuli. Auditory tests used target words and sentence contexts that were different from those used in training, and they were produced by a novel female

Figure 2. Seven steps taken for each trial in training.


Step 1
Click a button to listen to audio, e.g., /seki/.
Step 2
See three alternatives on the computer screen.
Short-Short
Long-Short
Short-Long
Step 3
Click one of the three alternatives.
Step 4
Watch video
Hand gestures
/seki/
"Syllable" and "Mora" conditions will see respective gestures (as in Fig. 1).
Step 5
Screen shows translation, e.g., "That means 'seat'."
Step 6
"Observe" conditions quietly observe video. "Produce" conditions quietly mimic gestures in video.
Same video as Step 4
Hand gestures
/seki/
Step 7
Translation again, e.g., "That means 'seat'."

speaker of Japanese who did not appear in training. This was because our aim was to determine participants' ability to generalize to new stimuli and not the ability to respond to trained stimuli well. There were a total of 120 stimuli each for a pretest and a posttest. Of a total of 10 word pairs used for each test (see Table 2), five pairs had the vowel length contrast in the first syllable (e.g., /eki/ "station" vs. /eːki/ "energetic spirit"), and the other five pairs had the length contrast in the second syllable (e.g., /mizo/ "ditch" vs. /mizoː/ "unprecedented").

Both the pretest and the posttest used the same 20 words but were spoken in different carrier sentences. The target words were placed in a sentence medial position of four carrier sentences (e.g., /sore wa ___ da to omou/ "I think that is ___"), of which two were used for the pretest, and the other two were used for the posttest. The speaker spoke each sentence at slow and fast speaking rates. The definition of these speaking rates given to the speakers was the same as that in the training stimuli, and the actual rates of speech were determined by each speaker.

To eliminate any bias, we needed to match the number of the following three types of words: "short + short" (e.g., /eki/ and /mizo/), "long + short" (e.g., /eːki/), and "short + long" (e.g., /mizoː/). The breakdown of the 120 stimuli in each pre- or posttest was as follows:

- "short + short" (10 words × 2 rates × 2 carrier sentences × 1 repetition) = words
- "long + short" (5 words × 2 rates × 2 carrier sentences × 2 repetitions) = 40 words
- "short + long" (5 words × 2 rates × 2 carrier sentences × 2 repetitions) = 40 words

Within each condition of SO, SP, MO, and MP, half of participants heard Carrier Sentences 1 and 2 at the pretest and Carrier Sentences 3 and 4 at the posttest, and this order was switched for the other half of participants.

Auditory test procedure. Within each of the pretest and the posttest, stimuli were randomly presented among word pairs, carrier sentences, and speaking rates (i.e., the random sequence was different for the pretest and posttest). The participants' task was the same three-alternative forced choice identification as in training: "short + short," "long + short," and "short + long." For each trial, a carrier sentence, such as "sore wa ___ da to omou," was written on the computer screen. The participants' task was to listen to varying words inserted in the underlined location and to choose one of three buttons that matched the vowel length pattern. There were six blocks in each of the pre- and posttests, and a short break was given between blocks. No feedback on the tests was given to participants at any time. Each test took about 20–30 min to complete. Participants were required to take the auditory posttest within 3 days after completing their final training session.

Other tests. A vocabulary test and an ERP test were conducted after the training and the above auditory posttest were completed. Details of these tests and results are not reported in this article.

Analyses and Predictions

A 2 × 2 × 2 × 4 analysis of variance (ANOVA) was conducted with auditory percentage-correct test scores. Test (pretest, posttest), rate (slow, fast), and position (initial, final) were within-subjects factors, and condition (SO, SP, MO, MP) was a between-subjects factor. An *improvement*

Table 2. Test stimuli.

Word-initial vowel contrasts				Word-final vowel contrasts			
Word	**Length**	**Pitch accent**	**Meaning**	**Word**	**Length**	**Pitch accent**	**Meaning**
eki	SS	HL	station	jome	SS	LH	bride
eːki	LS	HLL	energetic spirit	jomeː	SL	LHH	one's remaining years
doki	SS	HL	earthenware	oɾe	SS	LH	male use of "I"
doːki	LS	HLL	same period	oɾeː	SL	LHH	gratitude
toɾo	SS	HL	fatty tuna	kaze	SS	LH	wind
toːɾo	LS	HLL	the authorities	kazeː	SL	LHH	taxation
soka	SS	HL	simple refreshments	mizo	SS	LH	ditch
soːka	LS	HLL	flower arrangements	mizoː	SL	LHH	unprecedented
humi	SS	HL	(woman's name)	sotsu	SS	LH	pointless action
huːmi	LS	HLL	flavor	sotsuː	SL	LHH	communication

score was also calculated as the posttest minus the pretest score. The embodied cognition stance predicts a significant Test × Condition interaction: The improvement made from the pretest to the posttest would be greater for the two "produce" conditions than the two "observe" conditions. In addition, the mismatch gesture hypothesis (Goldin-Meadow, 2010) would predict that the pre-/posttest improvement would be greater for the two mora conditions than the syllable conditions.

Results

Overall Test Scores

All four groups significantly improved their auditory test scores on average (70.3% for the pretest and 79.4% for the posttest), as indicated by the significant main effect of test, $F(1, 84) = 66.4$, $p < .001$, $\eta^2 = .44$. Contrary to the predictions made by the embodied cognition stance or the mismatch gesture hypothesis, neither the Test × Condition interaction, $F(3, 84) = 0.013$, $p = .998$, $\eta^2 = .000$, nor the main effect of condition, $F(1, 3) = 0.023$, $p = .995$, $\eta^2 = .001$, was significant. This suggests that all four groups improved equally from the pretest to the posttest. The amount of improvement (posttest minus pretest) was very similar for all four groups (MO: 9%; MP: 8.9%; SO: 8.9%; SP: 9.5%). Refer to Figure 3.

Test Item Factors

With regard to the test item factors of rate and position, their main effects were significant: rate, $F(1, 84) = 78.3$, $p < .001$, $\eta^2 = .48$; position, $F(1, 84) = 73.5$, $p < .001$, $\eta^2 = .47$. As expected from previous research (e.g., Hirata et al., 2007; Minagawa et al., 2002), the test scores were higher for the slow rate than the fast rate (79.5% vs. 70.2%) and were higher for the word-initial (e.g., /eki/-/eːki/) than the word-final (e.g., /mizo/–/mizoː/) vowels (78.7% vs. 71.0%). In addition, a Text × Rate interaction was significant, $F(1, 84) = 4.53$, $p = .036$, $\eta^2 = .051$, and a three-way Test × Rate × Position interaction was also significant, $F(1, 84) = 14.1$, $p < .001$, $\eta^2 = .14$, indicating that the amount of improvement from the pretest to the posttest depended on speaking rate and word-internal position. Figure 4 shows the nature of this interaction. Post hoc *t* tests with Bonferroni corrections were conducted on the improvement scores (i.e., posttest minus pretest scores). The largest amount of improvement of 13.5% was made for the word-final vowels at the slow rate. This improvement was significantly greater than that in the other three item categories—improvement for slow initial: 7.9%, $t(87) = 3.7$, $p < .001$; improvement for fast initial: 8.1%, $t(87) = 2.7$, $p = .008$; improvement for fast final: 6.9%, $t(87) = 3.5$, $p = .001$. The amount of improvement among the initial slow, initial fast, and final fast categories did not significantly differ from each other (t = 0.096, 0.575, and 0.805, respectively; $p > .1$). Note that the improvement for all of these four types of test stimuli, however small, was significant, that is, the posttest scores were significantly higher than the pretest scores for all of these categories ($p < .001$).

Finally, there was a significant Test × Rate × Position × Condition interaction, $F(3, 84) = 3.3$, $p < .025$, $\eta^2 = .10$, and no other significant interaction was found. The significant four-way interaction indicates that the ways in which the training groups improved in their test scores differed for different rates and word-internal positions. Figure 5

Figure 3. Auditory test scores at the pretest and the posttest for each of the four conditions. Error bars represent 1 standard error of the mean. The asterisks indicate significant differences between the pretest and the posttest.

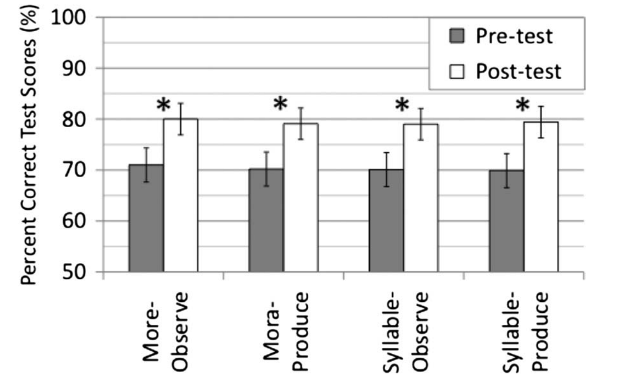
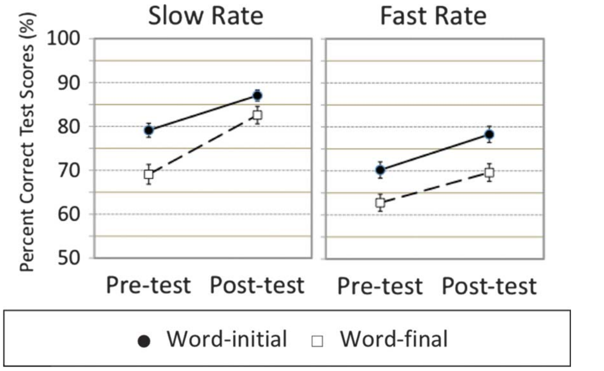


Figure 4. Auditory test scores on vowel contrasts in the word-initial and word-final positions spoken at two rates, pooled across all four groups. Error bars represent 1 *SEM*.

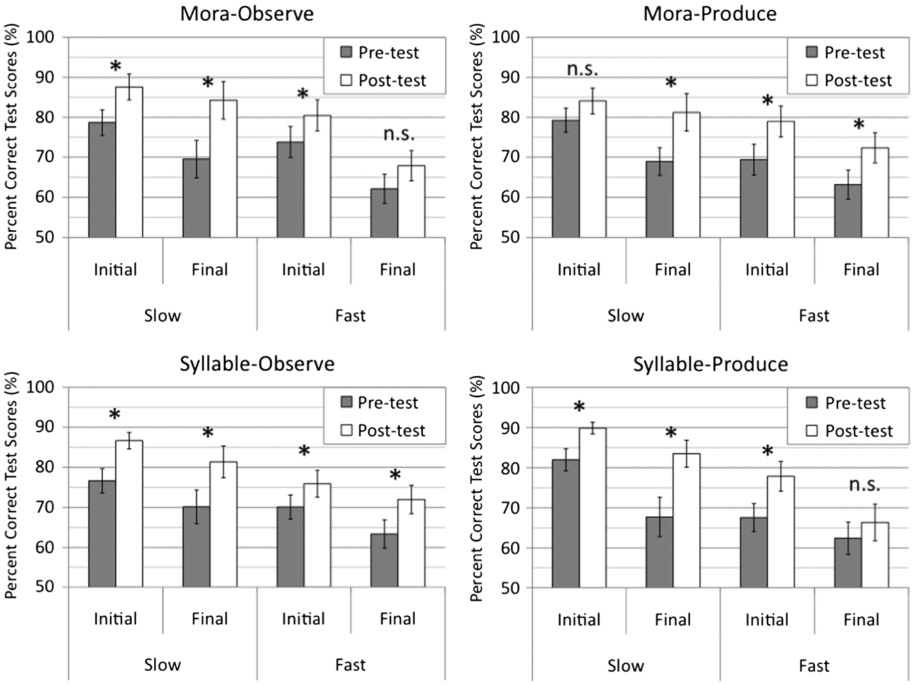


shows the nature of the four-way interaction, plotting the pretest and posttest scores of the word-initial versus word-final vowels at the two rates, separately for each of the four groups. In this figure, the asterisks indicate a significant difference, and "n.s." indicates a nonsignificant difference as a result of a paired-sample t test for the orthogonal comparison between the pretest and the posttest within each item category. What was common among the four groups was that they all made the largest improvement in the word-final position at the slow rate, as shown in the Test × Rate × Position interaction reported earlier (see Figure 4). What was different across the four groups was that SO was the only group that made significant improvement ($p < .05$) in all of the four item categories. The other groups, MO, MP, and SP, each showed one item category in which their improvement was not significant ($p > .05$). The item categories in which this happened differed among these three groups: the final vowels at the fast rate for MO and SP, and the initial vowels at the slow rate for MP. Taken together, among the four groups, the SO training yielded the most balanced auditory improvement across all of the item categories of the word-initial and word-final vowels at the slow and fast rates.

Discussion and Conclusions

This study examined whether different types of audio–visual training with hand gestures would yield different amounts of auditory learning. The first question addressed in the present study was whether auditory learning of L2 vowel contrasts takes place with the present training methods using hand gestures, especially when stimuli varied in speaking rate and in word-internal position. The results provide insights into the learners' initial auditory ability and learning ability with regard to different stimulus types. We found that speaking rate and the word-internal position (where accent types covaried) were two factors that affected auditory learning. Throughout the experiment, accuracy was higher in the word-initial vowels whose pitch patterns were H versus HL, for example, in /seki/ (HL) versus /se:ki/ (HLL), than in the

Figure 5. Auditory test scores on vowel contrasts in the word-initial and word-final positions spoken at two rates for each of the four groups. Error bars represent 1 *SEM*. The asterisks indicate significant differences between the pretest and the posttest, and "n.s." indicates nonsignificant differences.



word-final vowels whose pitch patterns were H versus HH, for example, in /ɕaɾe/ (LH) versus /ɕaɾeː/ (LHH). This finding fits well with work by Minagawa et al. (2002), who found that L2 learners of Japanese were more likely to make errors in the word-final than the word-initial position. In the present study, the long vowels in LHH (such as /ɕaɾeː/), which are supposed to be easier to identify than those in HLL pitch pattern (Minagawa et al., 2002), were still more difficult to identify than the first vowels (such as /seːki/). This suggests that it is generally more difficult to perceive vowel length in the word-final position than in the word-initial position. Despite this difficulty, however, participants on average showed the greatest improvement of 13.5 percentage points in this word-final position when the words were spoken slowly (see Figure 4).

With regard to speaking rate as a factor, vowel length distinction was found to be generally easier at a slower than a faster rate even though the test items were identical, consistent with findings in previous studies (e.g., Hirata et al., 2007). In terms of auditory learning, the present training paradigm that included both slow and fast stimuli yielded auditory improvement on the posttest in both slow and fast items. This finding is consistent with Hirata et al. (2007), and together, the two studies suggest that combining slow and fast items during training is a good way to help people improve at perceiving novel L2 phonemes spoken at either rate.

The second and third questions of this study were whether training yielded greater amount of improvement when participants were given the mora-like gestures rather than syllable-like gestures, and when they produced rather than observed hand gestures. Results show that all of the four training groups—MO, MP, SO, and SP—improved their overall auditory test scores by about 9 percentage points (see Figure 3), but the amount of improvement did not differ among the four groups. Thus, when collapsing across different speaking rates and different positions of vowels within words, we found no unique advantage of producing versus perceiving hand gestures in learning to hear phonemic vowel length distinction. There also seems to be no unique advantage of using the mora gestures as opposed to syllable gestures (see Figure 1).

How do the present results compare with Hirata and Kelly (2010) as discussed in the introduction? Their training utilized hand gestures along with Japanese vowel length pairs of words, and their "audio-mouth-hands" condition was basically the same method as the present SO training. Hirata and Kelly's audio-mouth-hands training was no more effective (5 percentage points improvement) than audio-only training (7 percentage points improvement) or audio-hands training (in which hand gesture, but not mouth movements, was provided; 9 percentage points improvement). In contrast, the audio-mouth training in which an instructor's mouth (but no hand gestures) was presented as he or she spoke words yielded a significantly greater improvement (14 percentage points improvement) than the audio-mouth-hands training. Taken together, the best training method in terms of yielding the largest amount of improvement seems to be Hirata and Kelly's audio-mouth (with no gesture) training. We conclude that hand gestures representing phonemic vowel length distinction do not have a unique effect in enabling auditory learning.

What are the implications for theories of gesture–speech integration (Kendon, 2004; McNeill, 1992)? It is important to understand that these theories focus primarily on the conceptual relationship between speech and gesture. For example, McNeill (1992) has argued that during language production, the core of a thought—or the growth point, to use his term—is fundamentally composed of linguistic meaning and imagistic meaning, and this unified meaning manifests in different ways through speech and gesture. Thus, speech and gesture, according to theory, are fundamentally linked at the level of meaning, or in linguistic terms, at the semantic and pragmatic levels of language.

The present study carves out an area in which gestures do not seem to integrate with speech, namely, the phonological level of language. In the context of L2 learning, Kelly et al. (2009) demonstrated that iconic gestures help learning of new L2 word meanings, and research has accumulated to reveal beneficial effects of gesture on the semantic and other higher linguistic levels in L2 learning (Macedonia et al., 2011; Quinn-Allen, 1995; Sueyoshi & Hardison, 2005). In contrast, we conclude from the present study that such beneficial effects do not seem to exist for segmental phonology at least at the very beginning stage of L2 learning. This is interesting in light of research showing that hand gestures do play an important role in phonological processing in one's native language (Biau & Soto-Faraco, 2013; Hubbard et al., 2008; Krahmer & Swerts, 2007). However, these studies examined the role that beat gestures play in the *suprasegmental* processing of speech, that is, a speech property that goes beyond an individual segment, such as sentential intonation and a focus within a sentence. Note that the ultimate goal of gestures in this case is to accentuate what information is semantically most relevant in the context. In contrast, the function of gestures in the present study is *segmental*—to accentuate the phonemic contrast of vowel length in which the difference is local and minimal. Thus, these gestures conveying segmental information highlight its auditory properties as an end in and of itself. We conclude that, at this segmental phonology level, such gestures representing long versus short vowels are not integrated into speech in the way that iconic gestures are for representing the semantic content of speech, or even in the way that beat gestures are for pragmatically focusing attention on the most relevant aspects of a spoken utterance. This conclusion comes from our effort in exhaustively comparing multiple methods using different types of gestures (mora vs. syllable) and different modalities (imitating vs. observing). To solidify our conclusion, of course, it would be necessary to compare the results of the present study with roles of gestures in segmental phonology in other languages, such as Chinese lexical tones. However, some preliminary results from our laboratory suggest that hand movements visually representing the four Mandarin tones do not uniquely help L2 learners to produce those tones more accurately than speech alone (Zheng, Cho, Kelly, & Hirata, 2013).

Despite the similarity of overall auditory improvement made across all of the four groups, we did find a small effect that distinguished SO training from the other three training groups. The SO group made significant improvement on all four item categories (the word-initial and word-final vowels at the slow and fast rates), whereas the other groups improved in only three of the four categories. It is useful here to be reminded that participants in the present study were also asked to memorize meaning of words while they were learning to hear the vowel length distinction. Training required a heavy cognitive load for participants—the task involved watching or producing hand gestures, as well as paying attention to the instructor speaking the words in the video, while focusing on subtle differences in vowel duration and trying to memorize the meaning of the words. That is a lot to do in one task. One possible explanation for the advantage of SO condition was that this training is least demanding in terms of the cognitive load it imposes on learners. Compared with the other training conditions, this group may have been able to more clearly focus on the novel speech sounds without being distracted by having to mimic gestures or observe gestures that were foreign to them (recall that the mora gesture is much less familiar to native English speakers than the more intuitive syllable gesture). This more manageable cognitive load might have allowed the SO participants to generalize their learning to different types of item categories. Indeed, recent research has suggested that when learning higher level aspects of an L2 (semantics and syntax), observing (Kelly & Lee, 2012) and producing (Post, Van Gog, Paas, & Zwaan, 2013) gestures help only when cognitive demands are low. In fact, these studies suggest that high cognitive loads may turn gestures into a distraction during L2 learning.

One final note that is worth mentioning is that although most previous training studies (e.g., Hirata & Kelly, 2010; Hirata et al., 2007) focused on auditory learning only, the learners in the present training paradigm were able to improve their auditory ability while also trying to memorize new vocabulary items. This additional task makes it even more impressive that there was such robust improvement on participants' auditory ability in the present study. This is good news when considering practical applications of our research: Learners are capable of multitasking even when the main goal of training is to enable perceptual learning of difficult L2 phonemic contrasts. The success of this more naturalistic way of learning foreign speech sounds—by embedding them within vocabulary learning—suggests that L2 auditory training does not always have to be as dry and tedious as just attending to subtle differences among speech sounds detached from all meaning.

The present attempt to include vocabulary learning also has theoretical implications as well. According to the model put forth by Baddeley, Gathercole, and Papagno (1998), difficulties in learning a new language arise when the phonological loop is taxed with novel speech sounds, and this disrupts the encoding of those new sounds into permanent memories for new words. The majority of previous research studying how people learn novel L2 phoneme contrasts focuses on phoneme learning as isolated from vocabulary learning. The vocabulary results of the present experiment are to be reported elsewhere, but Baddeley et al.'s model makes it clear that future research should explore the relationship between phonological and vocabulary learning in tandem during L2 acquisition.

In summary, the present research attempted to contribute to a relatively new area of study for the fields of phonetics and L2 acquisition by examining an understudied behavior of observing and imitating different types of hand movements. This attempt was a step forward from existing research in multimodal speech processing, which has focused predominantly on the role that visible mouth movements play with speech processing. Although hand gestures have been theorized to tightly integrate with semantic and pragmatic levels of speech processing (Kendon, 2004; McNeill, 1992), the results of the present study suggest that effects of gestures on segmental phonology acquisition are limited.

Of course, it is important to add a caveat to our conclusions that the auditory learning that we captured was of a fairly short term, spanning only within a week, and this study only examined naive learners who had the very first exposure to spoken Japanese through this training experiment. L2 learning can be a long process in one's lifetime, and different kinds of input are likely to have differential effects on learning at different stages. Future studies, therefore, should examine whether this training has differential effects on learners of Japanese who are more experienced and advanced in their study of Japanese. In addition, although it would be labor intensive, future studies should expose learners to these kinds of training with hand gesture in a more extended period of time, such as several months, and should examine their long-term effects. Finally, given Roberge et al.'s (1996) original pedagogical insight about use of hand gesture in production of Japanese vowel length contrasts, it would be interesting to test effects of the present training on learners' production in a controlled experimental setting.

Acknowledgments

Portions of this study have been presented at the Fifth Conference of the International Society for Gesture Studies, Lund, Sweden, July 2012; the 35th Annual Meeting of the Cognitive Science Society, Berlin, Germany, August 2013; and the 167th Annual Meeting of the Acoustical Society of America, Providence, Rhode Island, May 2014. This study was supported by National Science Foundation Grant 1052765 awarded to Yukari Hirata and Spencer D. Kelly. We thank Carmen Lin, Zach Zhao, April Bailey, and Kristen Weiner for being an integral part of designing stimuli and collecting data.

References

Asher, J. J. (1969). The total physical response approach to second language learning. *Modern Language Journal, 53,* 3–17.

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

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22,* 577–660.

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain & Language, 124,* 143–152.

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. *The Journal of the Acoustical Society of America, 101,* 2299–2310.

Brady, M. C., & Port, R. F. (2007). Quantifying vowel onset periodicity in Japanese. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 337–342). Saarbrücken, Germany: International Congress of Phonetic Sciences.

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesture makes learning last. *Cognition, 106,* 1047–1058.

Decety, J., & Grezes, J. (2006). The power of simulation: Imagining one's own and other's behaviour. *Cognitive Brain Research, 1079,* 4–14.

Engelkamp, J., & Dehne, D. M. (2000). Item and order information in subject-performed tasks and experimenter-performed tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 671–682.

Fischer, M. H., & Zwaan, R. A. (2008). Embodied language: A review of the role of the motor system in language comprehension. *The Quarterly Journal of Experimental Psychology, 61,* 825–850.

Gary, J. O. (1978). Why speak if you don't need to? The case for a listening approach to beginning foreign language learning. In W. C. Ritchie (Ed.), *Second language acquisition research—Issues and implications* (pp. 185–199). New York, NY: Academic Press.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review, 9,* 558–565.

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

Goldin-Meadow, S. (2010). When gesture does and does not promote learning. *Language and Cognition, 2,* 1–19.

Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition. *International Review of Applied Linguistics in Language Teaching (IRAL), 44,* 103–124.

Gullberg, M., de Bot, K., & Volterra, V. (2008). Gestures and some key issues in the study of language development. *Gesture, 8,* 149–179.

Han, M. (1994). Acoustic manifestations of mora timing in Japanese. *The Journal of the Acoustical Society of America, 96,* 73–82.

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. *The Journal of the Acoustical Society of America, 116,* 2384–2394.

Hirata, Y. (in press). L2 phonetics and phonology. In H. Kubozono (Ed.), *Handbook of Japanese phonetics and phonology*. Berlin, Germany: De Gruyter Mouton.

Hirata, Y., & Kelly, S. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research, 53,* 298–310.

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

Hirata, Y., & Whiton, J. (2005). Effects of speaking rate on the single/geminate stop distinction in Japanese. *The Journal of the Acoustical Society of America, 118,* 1647–1660.

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.

Iacoboni, M. (2005). Neural mechanisms of imitation. *Current Opinion in Neurobiology, 15,* 632–637.

Kelly, S. D., & Lee, A. (2012). When actions speak too much louder than words: Gesture disrupts word learning when phonetic demands are high. *Language and Cognitive Processes, 27,* 793–807.

Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes, 24,* 313–334.

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science, 21,* 260–267.

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

Kidd, G. R., Watson, C. S., & Gygi, B. (2007). Individual differences in auditory abilities. *The Journal of the Acoustical Society of America, 122,* 418–435.

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.

Ladefoged, P. (1975). *A course in phonetics*. Orlando, FL: Harcourt Brace Jovanovich.

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

Macedonia, M., Müller, K., & Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping, 32,* 982–998.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL: University of Chicago Press.

McNeill, D. (2005). *Gesture and thought*. Chicago, IL: University of Chicago Press.

Minagawa, Y., Maekawa, K., & Kiritani, S. (2002). Effects of pitch accent and syllable position in identifying Japanese long and short vowels: Comparison of English and Korean speakers. *Journal of the Phonetic Society of Japan, 6,* 88–97.

Montgomery, K. J., Isenberg, N., & Haxby, J. V. (2007). Communicative hand gestures and object-directed hand movements activated the mirror neuron system. *Social Cognitive Affective Neuroscience, 2,* 114–122.

Piske, T., MacKay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29,* 191–215.

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.

Port, R. F., Cummings, F., & Gasser, M. (1995). A dynamic approach to rhythm in language: Toward a temporal phonology. In B. Luka & B. Need (Eds.), *Proceedings of the Chicago Linguistic Society* (pp. 375–397). Chicago, IL: University of Chicago, Department of Linguistics.

Port, R. F., Dalby, J., & O'Dell, M. (1987). Evidence for mora timing in Japanese. *The Journal of the Acoustical Society of America, 81,* 1574–1585.

Post, L. S., Van Gog, T., Paas, F., & Zwaan, R. A. (2013). Effects of simultaneously observing and making gestures while studying grammar animations on cognitive load and learning. *Computers in Human Behavior, 29,* 1450–1455.

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.

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

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, Japan: Bonjinsha.

Strange, W. (Ed.). (1995). *Speech perception and linguistic experience: Issues in cross-language speech research*. Timonium, MD: York Press.

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

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

Tellier, M. (2008). The effect of gestures on second language memorization by young children. *Gesture, 8,* 219–235.

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

Tsukada, K. (2011). The perception of Arabic and Japanese short and long vowels by native speakers of Arabic, Japanese, and Persian. *The Journal of the Acoustical Society of America, 129,* 989–998.

Vance, T. J. (1987). *An introduction to Japanese phonology*. Albany, NY: State University of New York Press.

Zheng, A., Cho, Y., Kelly, S. D., & Hirata, Y. (2013, May). *Hand gestures and head nods do not aid L2 Mandarin tonal production.* Paper presented at the 25th Annual Convention of the Association for Psychological Science, Washington, DC.