UNIVERSITY of CALIFORNIA PRESS | Collabra

Kelly, S., et al. (2017). Metaphoric Gestures Facilitate Perception of Intonation More than Length in Auditory Judgments of Non-Native Phonemic Contrasts. *Collabra: Psychology*, 3(1): 7, DOI: https://doi.org/10.1525/collabra.76

ORIGINAL RESEARCH REPORT

Metaphoric Gestures Facilitate Perception of Intonation More than Length in Auditory Judgments of Non-Native Phonemic Contrasts

Spencer Kelly*, April Bailey† and Yukari Hirata‡

It is well established that hand gestures affect comprehension and learning of semantic aspects of a foreign language (FL). However, much less is known about the role of hand gestures in lower-level language processes, such as perception of phonemes. To address this gap, we explored the role that metaphoric gestures play in perceiving FL speech sounds that varied on two dimensions: length and intonation. English speaking adults listened to Japanese length contrasts and sentence-final intonational distinctions in the context of congruent, incongruent and no gestures. For intonational contrasts, identification was more accurate for congruent gestures and less accurate for incongruent gestures relative to the baseline no gesture condition. However, for the length contrasts, there was no such clear and consistent pattern, and in fact, congruent gestures made speech processing more effortful. We conclude that metaphoric gestures help with some—but not all—novel speech sounds in a FL, suggesting that gesture and speech are phonemically integrated to differing extents depending on the nature of the gesture and/or speech sound.

Keywords: foreign language; speech; hand gesture; phoneme; multimodal; length contrasts; intonational contrasts; perception

It is now well established that hand gestures play a significant role in native language production and comprehension, but much less is known about their role in the process of learning a new language. Traditionally, most of the research on this topic has focused on gesture's semantic function—its ability to add imagistic meaning to speech—but only recently have researchers begun to systematically explore how gestures function on lower levels of a language, such as learning to hear novel speech sounds. The present study explores the earliest stage of this learning process by asking what role metaphoric gestures play in English-speakers' ability to initially perceive different types of novel Japanese phoneme contrasts.

Most theories and models of gesture production place a central focus on the semantic and pragmatic functions of gesture (de Ruiter, in press; Hostetter & Alibali, 2008; Kendon, 2004; Kita & Özyürek, 2003; McNeill, 1992, 2005). McNeill's prominent "integrated system" theory claims that during language production, gesture and speech arise from the same semiotic starting point, and even though the two modalities ultimately represent semantic information in two different ways, they play equally important roles in capturing the meaning of an utterance. Iconic gestures—which visually represent object attributes, motions or spatial relationships—are particularly well suited to adding semantic richness to speech. For example, suppose a friend describes riding on a roller coaster by saying, ". . . and then we went upside down". Now imagine she said this while making either several horizontal tight corkscrew gestures vs. one large vertical looping gesture. Note how these two iconic gestures combine with speech to create two different pictures of your friend's topsy-turvy ride.

Building on McNeill's theory of gesture production, Kelly, Özyürek & Maris (2010) propose that iconic gesture and speech form a fundamentally integrated system during language comprehension as well. Focusing on two key elements of McNeill's theory—bi-directionality and obligatory integration—they presented evidence that gesture and speech mutually and automatically affect the semantic processing of one another during the comprehension process (for more on this relationship during comprehension, see Hostetter, 2011).

This tight semantic relationship appears to extend beyond one's native language and apply to second languages (L2) and foreign languages (FL) as well (Gullberg; 2006; Hardison, 2010; Kelly, McDevitt, & Esch, 2009; Macedonia, Müller, & Friederici, 2011; McCafferty & Stam,

* Department of Psychology, Center for Language and Brain, Colgate University, US

† Department of Psychology, Yale University, US

‡ Department of East Asian Languages and Literatures, Center for Language and Brain, Colgate University, US

Corresponding author: Spencer Kelly (skelly@colgate.edu)

2009; Quinn-Allen, 1995; Tellier, 2008). Focusing on FL comprehension, Kelly and colleagues (2009) showed that people learn and remember novel vocabulary items better when they are instructed with iconic hand gestures, and one neural mechanism responsible for this learning may be a strengthening of visual semantic representations. Further, Macedonia and colleagues (2011) showed that words learned with meaningful gestures were remembered better and produced more activation in the pre-motor cortex (a region involved in the planning and processing of motor movements) than words learned with meaningless gestures. Together, these studies show that iconic gestures help learners remember the meaning of new words by creating a rich network of semantic representations across multiple modalities in long-term memory.

Despite these semantic benefits of gesture, learning words is only just one of the many challenges in mastering a FL. A more fundamental and pressing challenge—at least for phonemically dissimilar languages—is learning to correctly perceive the set of novel speech sounds that combine to create words in a FL. Indeed, the meaning of a word cannot be committed to memory if the phonemes that comprise that word are not properly distinguished. Because phoneme perception accounts for many learners' difficulties mastering a foreign language (Broersma, & Cutler, 2011), it is important to understand the role hand gestures play in helping learners overcome this challenge.

Anecdotally, it is common practice for instructors to use hand gestures to teach novel phoneme contrasts in a FL. Taking Mandarin as an example, it is commonplace for teachers to use hand gestures to metaphorically represent the various "shapes" of the four Mandarin tones: A rising gesture may visually represent the acoustically rising nature of Tone 2, and a falling gesture may visually represent the acoustically falling nature of Tone 4. In this way, these sorts of gestures produced in FL classrooms serve a metaphoric function in that they abstractly map salient visual information onto the acoustic contours of the to-be-learned speech sounds (for a similar metaphoric function in the context of musical pitch perception, see Connell, Cai, & Holler, 2013).

Although there is plenty of anecdotal evidence—and even some preliminary observational support (for an example in Japanese instruction, see Roberge, Kimura, & Kawaguchi, 1996)—showing that metaphoric gestures may help with teaching novel phoneme contrasts in an actual FL classroom, surprisingly little research has experimentally tested whether these gestures help with phonemic learning. In one of the few lines of empirical research investigating this topic, Hirata and Kelly conducted multiple experiments to examine the effects of metaphoric gestures on Japanese vowel length distinctions among native English speakers (Hirata & Kelly, 2010; Hirata, Kelly, Huang, & Manansala, 2014; Kelly, Hirata, Manansala, & Huang, 2014). In Japanese, vowel length is phonemic such that shortening or lengthening spoken vowels changes word meanings, and research has shown that for many native speakers of English (e.g., American English), there is great difficulty in perceiving this length distinction (Hirata, 2004). Because auditory training alone produces only modest improvements in vowel length distinctions (Hirata, 2004), Hirata and Kelly combined the metaphoric length gestures described by Roberge et al. (1997) with audio training to determine whether this learning could be further enhanced. The results did not support Roberge and colleague's (1997) correlational findings: Training with Roberge's metaphoric length gestures—long sweeping gestures for long vowels and short sweeping gestures for short vowels—did not improve scores from a pre- to post-test more than an audio-only control, suggesting that hand gestures may not play a role in some aspects of phonemic processing (for more on the boundaries of gesture's influence on speech, see Kelly, in press).

However, recent research suggests that perhaps the type of FL phoneme matters. In a recent experiment by Hannah, Wang and colleagues (Hannah, Wang, Jongman, & Sereno, 2016), native-English speakers were exposed to Mandarin tones with congruent (rising hand and rising tone) and incongruent (rising hand and falling tone) metaphoric gestures.[1] Of relevance to the present study, when hand gestures metaphorically conveyed incongruent information to the speech, perceivers often heavily relied on that visual information to make judgments about what they heard, resulting in people actually "mishearing" the actual spoken tone. This suggests that at the earliest stages of FL exposure, learners may use metaphoric gestures as a cue to identifying certain phonemic intonational contours in a new language.

The current study builds on this research in two ways. First, it directly compared how metaphoric gestures affect perception of phonemic length (Hirata & Kelly, 2010) vs. intonational contrasts (Hannah et al., 2016). To accomplish this, we exposed English speakers to videos of a native Japanese speaker producing congruent and incongruent metaphoric gestures with Japanese sentences involving two types of acoustic distinctions, vowel length and intonational contour. Unlike American English, both of these distinctions are phonemic in Japanese, in that length and intonational distinctions *alone* differentiate lexical items (Vance, 1987).[2] In this way, it is possible to directly compare the role of metaphoric gestures in assisting with intonational vs. length perception during the earliest stages of FL learning.

Second, we build on previous work by using a more direct measure of gesture-speech integration than the paradigm used by Hannah et al. (2016). In that study, judgments were made about tones without specifying whether participants should attend to speech or gesture. These instructions made it difficult to determine if participants truly "integrated" gesture and speech, or whether they simply switched from relying on one modality to the other in performing the task. In the present study, we borrowed from Kelly et al. (2010) and asked participants explicitly to ignore gestures and pay attention *only* to speech. Using these stricter instructions, if gestures affect how participants hear the speech sounds, it would provide evidence that perception of speech sounds is actually affected by perception of gesture, which is a more direct test of whether the two modalities are fundamentally integrated at the lowest level of language comprehension.

Method

Participants

The study consisted of 57 (36 females and 21 males) English-speaking undergraduates aged 18 to 22 years. Participants were recruited as part of a *Research Methods* course and received no payment for volunteering. None had formal exposure to the Japanese language. Data collection was conducted under the approved ethical standards of the university's Internal Review Board.

Materials

The stimulus video presented novel phonemic distinctions—vowel length and sentence final intonational contrasts—in Japanese. Regarding vowel length distinctions, these contrasts are phonemic in Japanese (Fujisaki, Nakamura, & Imoto, 1975), which means that, unlike English, the length of a vowel *alone* can change the meaning of a word in Japanese. For example, the word, "kuro" ("black") has a short first vowel and the word, "kuuro" ("air path") has a long first vowel, and this length distinction alone determines the two different meanings of these words. With regard to intonation contrasts, some intonational distinctions at the end of a sentence are also phonemic (Vance, 1987). Unlike English declarative and question intonations, in Japanese the acoustic patterns throughout the sentence are identical and only the intonation of the final syllable changes . For example, when the sentence-ending particle, "ka↑", is used with a rising intonation, it means, "is it?", but when used with a falling intonation, "ka↓", it means, "it is". Note that the intonation of the earlier part of a sentence preceding "ka" is identical in Japanese, and the only difference between the two sentences is the rising or falling pitch of the phoneme, "ka", at the end of the sentence. Combining these two types of phonemic distinctions, the stimulus video contained four different sentences: "*Kuro desu ka↑*" with a rising intonation ("Is that black?"), "*Kuro desu ka↓*" with a falling intonation ("I see, that is black"), "*Kuuro desu ka↑*" with a rising intonation ("Is that an air path?"), and "*Kuuro desu ka↓*" with a falling intonation ("I see, that is an air path").

These stimuli were created in iMovie as short video clips. The audio component was comprised of a native female Japanese speaker producing one of four sentences in each clip. The actual audio component (approximately one second in length) was recorded separately without any gestures and dubbed onto the appropriate video segment. This was done to ensure that the audio would be identical across the three video conditions (see Krahmer & Swerts, 2007, for how producing gestures changes phonetic properties of speech).

The visual component consisted of a female native Japanese speaker shot from her shoulders to her waist. This was done to block access to mouth movements, which have been shown to influence phonemic perception in an L2 (Hirata & Kelly, 2010; Hardison, 2003, 2005). The gestures chosen for the video portion were based on standard gestures used in actual Japanese classrooms to teach length and intonational distinctions. There were three video conditions for each sentence. The first was the no gesture condition (baseline) and consisted of no actual gesturing (see **Figure 1a**). The second condition was the congruent condition, in which the gesture was congruent with the length and/or the intonation contrast. For example, suppose the speech was "Kuro desu ka↑". A congruent metaphoric length gesture would be one short, sliding palm movement perpendicular to the ground (corresponding to the short vowel, "ku").[3] See **Figure 1b**. Following the length gesture, a congruent intonational gesture—a gesture with the hand (starting where it left off for the length gesture) moving upward with the rising intonation—accompanied the final word in the sentence of the word, "ka↑". See **Figure 1d**. The third condition was the incongruent condition, in which the gesture was incongruent with the length contrast and/or the intonation contrast. Referring to the above example, "Kuro desu ka↑", an incongruent length gesture would be one *long*, sliding palm movement perpendicular to the ground (this movement corresponds to the long vowel, "kuu"). Following the length gesture, an incongruent intonational gesture would be the palm moving *downward* (this movement corresponds to the falling intonation of the word, "ka↓"). See **Figures 1c** and **1e** respectively. All gesture videos contained gestures for both length and intonation—that is, there were always two gestures per video—and the gesture onsets were timed with the onsets of the length and intonational phonemes.

In sum, there were a total of four different spoken sentences: 1) "*Kuro desu ka↓*", 2) "*Kuuro desu ka↓*", 3) "*Kuro desu ka↑*" and 4) "*Kuuro desu ka↑*". Each of the four sentences was presented in five different ways: 1) no gesture for both vowel length and final intonation, 2) gesture congruent with speech for both vowel length and final intonation, 3) gesture congruent with vowel length but incongruent for final intonation, 4) gesture incongruent with vowel length but congruent for final intonation, and 5) gesture incongruent with speech for both vowel length and intonation. These combinations yielded a total of 20 sentences (five combinations for four sentences), but to equate numbers of stimuli in each condition (eight congruent, eight incongruent, and eight no gesture for length and intonation contrasts each), we presented the no gesture videos twice. In the end, there were a total of 24 videos for each participant.

Procedure

Upon arrival, participants were first asked to fill out an informed consent form. Following this, they were brought to one of six testing cubicles and were seated in front of an iMac computer. To familiarize these native English speakers with the Japanese language, there was a short tutorial presented as a PowerPoint slide show. The first slide introduced the concept of long and short vowels and rising and falling intonation in the Japanese language. The second slide allowed participants to play audio clips of the four sentences used in the experiment to familiarize them with the length and intonational distinctions. Importantly, there was no video information for these familiarization trials. Because these phonemic distinctions are novel and challenging for native English-speakers, they could listen to these sentences as many times as they chose until the experiment started.

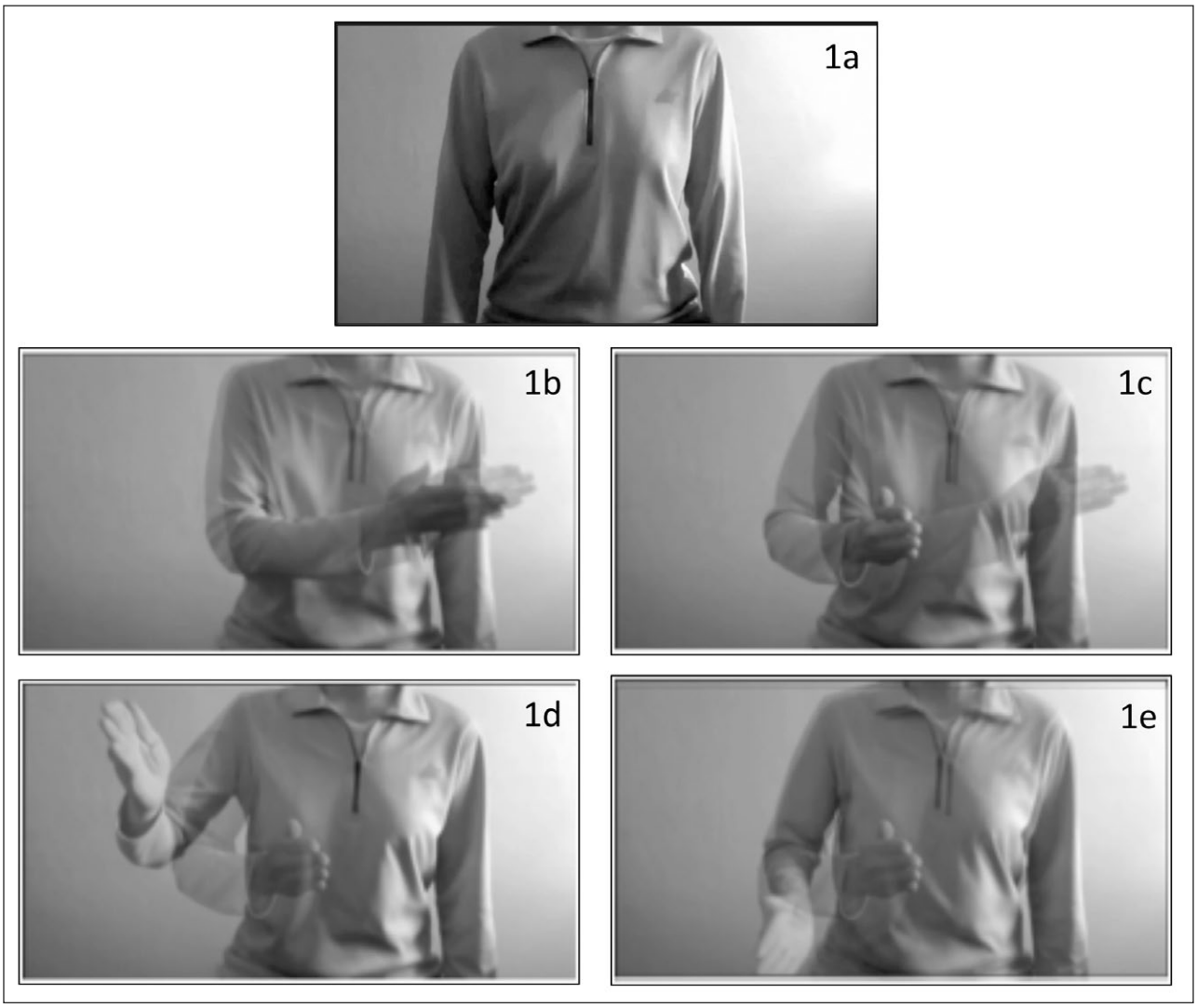

Figure 1: Still frames of the gesture relationships for length and intonation using the example speech, "Kuro desu ka↑." **(1a)**: No gesture (baseline) condition. **(1b)**: Short vowel gesture (congruent length condition). **(1c)**: Long vowel gesture (incongruent length condition). **(1d)**: Rising gesture (congruent intonation condition). **(1e)**: Falling gesture (incongruent intonation condition).

Following familiarization, participants were given a third slide presenting the instructions for the experiment. They were told that after viewing each sentence, there would be two consecutive forced-choice questions about the audio portion of the sentences. For the length question, participants were to press the "z" button if they heard a short vowel and the "/" if they heard a long vowel at the start of the sentence; for the intonation question, they were to press the "z" button if they heard a rising intonation and the "/" if they heard a falling intonation at the end of the sentence. Half the participants were questioned about vowel length first and vowel intonation second, and the other half received the reverse order. This was done to ensure that order of question would not be confounded with the order of presentation in the sentences, which was always held constant because phonemic intonational distinctions occurred only at the end of sentences.

If participants did not have any questions after these instructions, they were given five practice trials—without feedback—that consisted of the same sentences of the test trials. After addressing any remaining questions following the practice trials, the experimenter left the room and the participants engaged in the experiment.[4] Error rates and response times (RTs) were recorded to each of the questions. For the RT measure, outliers two standard deviations from the mean were removed (Ratcliff, 1993), and only responses that were correct were entered into the final analyses.

The order of the 24 experimental trials was randomized for each participant. After the experiment, participants were given a post-experiment questionnaire asking whether they kept their eyes on the screen during the experiment and whether they had difficulty ignoring the hand gestures (most did). Afterwards, they were debriefed by telling them the purpose of the experiment. The whole procedure lasted approximately 25 minutes.

Results

We performed a 2X3 repeated measures analysis of variance (ANOVA) on each dependent variable – error rates and RTs – with contrast type (length vs. intonation) and relationship (no gesture, congruent and incongruent) as

the two within-subjects factors. To follow up on significant interaction effects, we performed Dunn's corrected *t* tests to compare all possible pairings of no gesture, congruent, and incongruent conditions within each level of contrast type. Dunn's corrections adjust p value thresholds to make them more conservative when running multiple comparisons on the same data, so for each level of contrast type—length and intonation—three *t* tests were run (which shows up as the number "3" preceding the degrees of freedom in the denominator).

Error Rates

For error rates, there was a main effect of contrast type, $F(1, 54) = 4.42$, $p = .041$, $\eta_p^2 = .076$, with intonation producing more errors than length. In addition, there was a main effect of relationship, $F(2, 108)$, 14.64, $p < .001$, $\eta_p^2 = .213$, and a significant interaction of contrast type by relationship, $F(2, 108) = 9.23$, $p < .001$, $\eta_p^2 = .146$. This interaction was driven by there being less variability in the length than the intonation contrasts. For length, the incongruent condition produced significantly more errors than the no gesture condition, $tD(3, 54) = 2.88$, $p < .05$. However, there were no significant differences between the incongruent and congruent conditions, $tD(3, 54) = 2.17$, ns, or between the congruent and the no gesture conditions, $tD(3, 54) = 0.29$, ns. On the other hand, there was a clear and robust linear effect for intonation contrasts: The incongruent condition produced significantly more errors than the no gesture condition, $tD(3, 54) = 3.61$, $p < .005$, which produced significantly more errors than the congruent condition, $tD(3, 54) = 4.06$, $p < .001$. By extension, the incongruent condition produced more errors than congruent condition, $tD(3, 54) = 4.90$, $p < .001$. See **Figure 2a**.

Response Times

For RTs, there was a main effect of contrast type, $F(1, 54) = 4.01$, $p = .050$, $\eta_p^2 = .069$, with intonation producing slower RTs than length. In addition, there was a main effect of relationship, $F(2, 108) = 3.70$, $p = .047$, $\eta_p^2 = .064$, and a significant interaction of phoneme contrast by relationship, $F(2, 108) = 5.24$, $p = .018$, $\eta_p^2 = .089$. Unlike the error rates, this interaction was driven by there being more variability among the length than the intonation contrasts. For length, the congruent condition was slower than the no gesture condition, $tD(3, 54) = 3.42$, $p < .005$, and incongruent condition, $tD(3, 54) = 4.29$, $p < .001$. The no gesture condition was not significantly different the incongruent condition, $tD(3, 54) = 2.46$, ns. For the intonational contrasts, there were no significant differences in RTs: The congruent condition was no faster than the no gesture condition, $tD(3, 54) = 2.05$, ns, or incongruent condition, $tD(3, 54) = 0.18$, ns; nor was there a difference between the incongruent and no gesture conditions, $tD(3, 54) = 0.51$, ns. See **Figure 2b**.

Summary

Taken together, the error rate and RT data tell an interesting story. First, people were slower and made more errors on the intonational contrasts compared to length contrasts. Second, gesture had different effects within each type of phonemic contrast.

Intonational contrasts. For the intonational contrasts, the error rates clearly demonstrated that congruent gestures helped and incongruent gestures hurt intonation perception relative to the no gesture baseline. There were no significant RT differences, suggesting that congruent and incongruent metaphoric gestures affected precision, but not speed, of accurately identifying intonation.

Length contrasts. The error rates and RTs from the length contrasts were less straightforward. Although participants made more errors in the incongruent condition compared to the no gesture condition, suggesting that incongruent gestures compromised accuracy, they produced slower reaction times in the congruent condition compared to the no gesture and incongruent conditions, suggesting that congruent gestures made length processing more effortful.

Discussion

The results provide answers to our two research questions. Our first goal was to determine whether metaphoric gestures would have a similar or different influence on perception of FL length and intonational contrasts. The results suggest that the two are processed differently: Metaphoric gestures had a more systematic and straightforward influence on perception of intonation than perception of length. Our second goal was to explore whether metaphoric gestures influenced perception of length and intonation even when the task did not require attention to the visual input. We found that even when instructions focused on speech perception, participants were still influenced by intonational gestures, and to a less consistent extent, length gestures. Below, we discuss these finding in light of the previous literature on gesture's influence on FL perception and learning.

Length Contrasts

The results from the error rates and response times for the length contrasts complement and extend previous research on gesture's role in FL phoneme processing. Focusing first on response times, the findings are consistent with previous research by Hirata and Kelly (2010) demonstrating no positive influence of congruent metaphoric gestures when training native English speakers to hear long and short vowel distinctions in Japanese. Similar to that study, we showed that congruent metaphoric gestures do not play a helpful role above and beyond speech alone in distinguishing vowel length differences in Japanese. In fact, we found that length contrasts with congruent gestures produced slower RTs than baseline (and contrasts with incongruent gestures), suggesting that, if anything, congruent gestures actually disrupted processing of the length information.

One possible explanation for this finding is that the congruent gestures actually encouraged more elaborate processing of the length contrasts than the other conditions. Because phonemic length contrasts are so novel to many English speakers, adding congruent gestures to them may have caused participants to analyze them in a

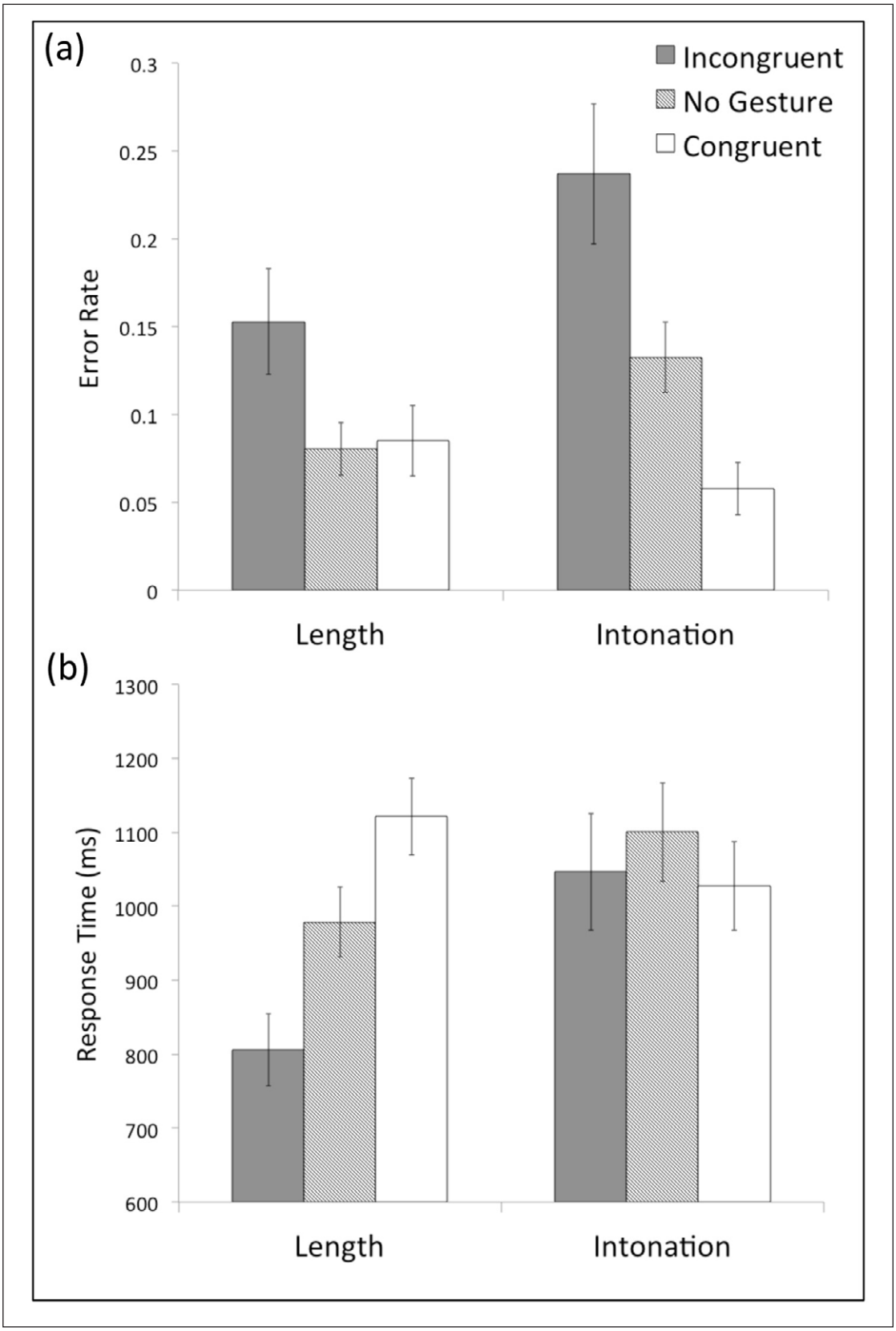

Figure 2: (a) Error rates (proportion incorrect) across the three relationship levels (no gesture, congruent and incongruent) for each of the two levels of phonemic contrasts (length and intonation). For the length contrast, the incongruent condition produced significantly more errors than the no gesture condition. For the intonation contrast, the incongruent condition produced significantly more errors than the no gesture condition, which in turn produced significantly more errors than the congruent condition. **(b)** Response times (in milliseconds) across the three relationship levels (no gesture, congruent and incongruent) for each of the two levels of phonemic contrasts (length and intonation). For the length contrast, the congruent condition produced significantly slower RTs than the no gesture condition and the incongruent condition. For the intonation contrast, there were no significant differences among the conditions.

more "linguistic" fashion than the other stimuli. That is, because participants had just learned about vowel length being phonemic in Japanese at the start of the experiment, perhaps they tried to match what they saw in gesture onto what they had heard in speech when length information was presented in a congruent fashion across modalities. In contrast, perhaps participants used a more "low level" approach when they heard length information in the absence of gestures or with incongruent gestures. There is some evidence for this possibility in other FL research on phoneme processing. For example, Werker & Logan (1985) found that native English speakers had more difficulty processing novel FL phonemes when those speech sounds were presented with larger than smaller intervals between stimuli. They argued that the longer time between phonemes encouraged a more high-level and abstract comparison between the novel phonemes and familiar phonemes in English, and this resulted in deeper and more laborious processing of the novel FL speech sounds. In contrast, presenting stimuli quickly forced a simpler, and ultimately more accurate, processing of the sounds that relied on a basic and low-level analysis of the acoustic properties of those sounds.

Similarly, participants in the present study may have processed the "*Kuro/Kuuro*" distinctions without gesture (and with incongruent gestures) quickly because they relied only on the low-level acoustic information that was available to them. In contrast, they may have processed the congruent items more slowly because they first processed the low-level acoustic information, and then after observing the temporal overlap between the speech and gesture, they did a higher-level reanalysis of the phonemes in the context of the congruent gestures.

The error rates show a different effect of gesture on length processing. Participants produced more errors in the incongruent condition compared to the speech only condition. Considering the error rates and response times together, one possible interpretation of this pattern is that there may have been a speed accuracy tradeoff (Winkelgren, 1977). Note in **Figure 2b** that there is a trend for the length contrasts with incongruent gestures to be processed faster than length contrasts without gesture. Although this trend was not significant (even though the difference was almost 200 milliseconds), it is possible that participants made more errors in the incongruent vs. no gesture condition because they were more hasty when incongruent gestures accompanied speech. Overall, the combined pattern of error rates and response times is not easy to interpret and suggests that gestures may play an inconsistent and idiosyncratic role with length contrasts in an FL.

Intonational Contrasts

The results from the intonational contrasts tell a more straightforward story. Although there were no differences in the RTs, the error rates for intonation clearly showed that congruent gestures facilitated, and incongruent gestures disrupted, the ability to accurately identify the correct intonation. There are at least three possible explanations for why the intonational results revealed a more coherent pattern than the length results. The first considers the distinct nature of the two types of phonetic contrasts; the second considers conceptual differences between the two types of metaphoric gestures; and the third considers both.

There is reason to believe that American English speakers may find intonational contrasts to be more familiar than length contrasts, which can be explained by the difference between segmental and supra-segmental linguistic processing of speech sounds. Durational distinctions, such as "kuro" versus "kuuro", operate only at the segmental level because the distinctions occur *within* a single syllable. In contrast, pitch distinctions can operate both within a single syllable, as in the Japanese "ka↑" versus "ka↓ distinction", or supra-segmentally *across* words in a sentence, as in finishing a sentence in English with a rising intonation to ask a question and a falling intonation to make a statement. Because American English speakers have familiarity with supra-segmental function of pitch and intonation, perhaps Japanese segmental intonational contrasts were easier to process, making it smoother to connect the sound to the gesture. Although plausible, the data do not support this account. Recall that there was a main effect of contrast type, with intonational contrasts producing *more* errors and *slower* RTs than the length contrasts. This suggests that even though intonation contrasts may be more familiar to American English learners, this familiarity did not make it easier to process in Japanese. Although not a central question of the experiment, it is interesting to speculate on why participants did worse on the intonational than length contrasts. One possibility is that although intonational contrasts may be more familiar than length contrasts for American English speakers, Flege (1987) has argued that it is paradoxically harder to process phonemic distinctions in an L2 that are familiar vs. unfamiliar. Another possibility is that the intonational contrasts were harder simply because they were always presented after the length contrasts in the videos. This might have caused participants to allocate more cognitive resources to the length contrasts, leaving fewer resources available to process the intonational contrasts. Either way, the inconsistency in the present findings—i.e., gestures produced more reliable and straightforward effects on intonational contrasts than length contrasts—cannot be explained by the former being inherently easier to process than the latter in Japanese.

The second possibility concerns conceptual differences in the metaphoric gestures used to represent information conveyed through vowel length versus intonation. The length gestures used in the present study were the same as the gestures used by Roberge et al. (1997) and Hirata and Kelly (2010), which metaphorically mapped the length of time it took to make a sweeping gesture to the length of time that the vowel sound was sustained. It is possible that these gestures were more difficult to process than intonational gestures because they are relatively more abstract. Although people often use their bodies to spatially represent temporal distinctions (Casasanto, 2008; Cooperrider, & Núñez, 2009), metaphoric gestures convey meaning about length only with respect to one another—that is,

it is difficult to represent absolute size with a gesture because size is inherently relative. In contrast, metaphoric gestures used to represent pitch information can capture intonation in a more *absolute* sense. For example, mapping the metaphorical incline of a hand movement to the rising intonation pattern of speech (or any sound, for that matter) conveys directional information in a more independent fashion. That is, unlike relative length of sounds, it is possible to understand the directional concept of "up" independently of "down", which may make the meaning of intonational gestures more easy to grasp.

Finally, a third possibility is a combination of differences between the types of phonemic contrasts and metaphoric gestures representing those distinctions. Given that the intonational contrasts were inherently more difficult to auditorily judge than length contrasts (as the error rates attest), perhaps participants were more likely to turn to gesture when they had difficulty hearing the intonational information. This make sense given research on native language processing showing that people rely on visual information more when they struggle to process auditory information (Sumby & Pollack, 1954; and specifically for hand gesture, Drijvers & Özyürek, 2016; Obermeier, Dolk, & Gunter, 2012). Although the present study cannot definitely determine whether the second or third account is correct, the results still make a novel contribution to the literature: In contrast to recent findings showing that metaphoric length gestures do not help with learning vowel length distinctions (Hirata & Kelly, 2010; Hirata et al., 2014), we have shown that metaphoric intonation (pitch) gestures do help non-native speakers process phonemic intonational information in FL speech (see also, Hannah et al., 2016). This finding is significant because it suggests that regardless of whether one type of FL speech contrast is inherently harder to process than the other, gesture plays different roles within each one, suggesting that some possible boundaries of gesture-speech integration at the phonemic level (more on this below).

Obligatory Integration

The present findings are interesting in light of the recent study by Hannah et al. (2016) also investigating the role of metaphoric hand gestures on English speakers processing of intonational contrasts (in Mandarin, not Japanese). Recall that in that study, participants were asked to attend to both speech and metaphoric gestures in order to judge what Mandarin tone they had just perceived. The main finding was that when gestures conveyed information that conflicted with the spoken tone, participants paid more attention to the gesture than the speech. These interesting results suggest that when given a choice to attend to contradictory pitch information, gesture trumps speech.

The present study adds to these findings by removing "choice" from the equation. As with Kelly et al. (2010), we instructed participants to respond *only to* the auditory modality when judging the sentences. Even with these strict instructions, participants' ability to accurately perceive the sentence final intonation was strongly influenced by the gestures that accompanied it: Participants misheard the intonation just under 6% of the time in the congruent condition, but when no gestures accompanied the speech, the error rate more than doubled. And when speech was accompanied by incongruent gestures, error rates quadrupled. These numbers are even more striking when one considers that the task did not require any response to the visual information. This fits nicely with the work by Connell and colleagues (2013) showing that even when people are asked to focus only on the pitch of musical notes during a song sequence, metaphoric gestures cannot be ignored and ultimately affect people's judgments about what they heard. Both sets of results fit within a neural framework that the planum temporale, located within the left hemisphere auditory cortex, is designed not only for auditory pitch processing, but visual-spatial processing as well (Griffiths & Warren, 2002).

Although the present study is a step in the direction towards making more definitive claims about the integration of gestured and spoken phonemic information in FL comprehension, additional research using more direct measures of "perception" would allow for even stronger conclusions. For example, it is impossible to determine whether the effects for intonational contrasts are the results of integration of gesture and speech during the actual online encoding of the two modalities, or whether they are the result of off-line memory or decision-making processes. That is, perhaps participants accurately encoded the speech in all conditions, but when explicitly forced to decide on the directionality of the intonation they heard, only then did gesture enter the picture and affect the decision-making process. Future studies using a more direct measure of online processing, such as taking RTs (Kelly et al., 2010) or ERPs (Biau & Soto-Faraco, 2013) during the actual encoding phase, would address this subtle, but important, distinction.

Implications and Conclusion

Before we discuss implications, some additional limitations of the present study should be addressed. One, the present study used real language stimuli instead of an invented language in order to connect to actual teaching classrooms and to better relate to previous L2 learning research on phoneme processing and learning. This decision made it hard to carefully control our variables (e.g., intonational contrasts may be more familiar to American English speakers than length contrasts), so future work might follow up with more carefully controlled stimuli to test gesture's role in processing different types of speech sounds. And two, the present study used a relatively narrow range of stimuli and low number of presentations. Although we had medium effect sizes for our main findings (e.g., the relationship by contrast type interaction for error rates was $\eta_p^2 = .146$), the results would be strengthened if they generalized to additional instances of length and intonational stimuli. We are optimistic of this generalization because the present findings are consistent with previous research on FL length (Hirata & Kelly, 2010) and intonation (Hannah et al., 2016) using different stimuli—and in the case of the latter study, an entirely different language.

Even with these important future lines of inquiry, the present findings have immediate implications for FL instruction. We already know from previous research that multimodal input is useful when teaching FL learners novel speech sounds (Hardison, 2003, 2005; Hirata & Kelly, 2010; Wang et al., 2008). These studies have convincingly shown that presenting congruent lip movements with auditory phoneme instruction helps people learn novel phoneme contrasts above and beyond auditory input alone. However, there is evidence that layering too much multimodal information onto FL speech instruction may over-load the system and actually produce decrements in learning (Hirata & Kelly, 2010; Kelly & Lee, 2012). For example, Hirata and Kelly showed that whereas seeing lip movements with speech helped English learners to distinguish Japanese long and short vowels better than speech alone, adding metaphoric gestures to lip and audio training actually removed the positive effects of the mouth.

The present findings add an interesting wrinkle to these studies. When learners have difficulty mapping the meaning of gestures onto novel speech sounds (as with hand gestures conveying information about length of phonemes), it may be wise to eliminate this form of multimodal input from the instruction process. In contrast, when learners can more easily connect gestures to speech sounds (as with hand gestures conveying information about intonation), it may be helpful to add this piece of multimodal input. Thus, it appears that more input is not always better in FL instruction. Although the present study focused only on hand gestures and speech, it will be important for future studies to investigate the synergistic—or disruptive, as with Hirata and Kelly (2010)—contributions of other channels of multimodal input, such as mouth movements, facial expressions and eye gaze.

More broadly, the present findings are interesting to consider in light of recent claims about the role of "iconicity" in language (Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015; Imai & Kita, 2014). For example, Dingemanse and colleagues make a compelling case that not all aspects of language are arbitrary, and indeed, some aspects of language, such as phonetics, iconically map onto to elements of meaning across languages. The study of hand gestures provides an interesting lens through which to consider these sorts of claims. For example, metaphoric gestures, such as the ones used in the present experiment, add a visual layer to the acoustic contours of speech, providing a possible iconic bridge to phoneme perception. The fact that perception was facilitated for only certain speech sounds—rising and falling intonation at the ends of sentences—but not others—varying lengths of syllables—suggests that perhaps phonemic contrasts across and within languages vary in their inherent iconicity. Returning to an earlier point, perhaps phonemic distinctions that are more absolute, such as up vs. down, lend themselves more easily to iconic representation than more relative phonemic distinctions, such as long vs. short.

In conclusion, although there is clear evidence that gestures are tightly integrated with speech during the semantic and pragmatic processing of a language—both native and foreign—the present study suggests that this integrated system is murky at lower levels of language processing, at least for a FL. It appears that for relatively familiar phonemic distinctions to American English speakers, such as intonational contrasts, the metaphoric content of gesture links up naturally with rising and falling intonational patterns in FL speech. But for more unfamiliar distinctions, such as phonemic length contrasts, metaphoric gestures may not connect smoothly with FL speech, and in fact, may actually disrupt it. Together, these contrasting FL perception findings add texture to previous work showing that metaphoric gestures do not help with learning of FL length contrasts (Hirata & Kelly, 2010; Hirata et al., 2014; Kelly & Lee, 2012; Kelly et al., 2014). We are now one step closer to identifying what aspects of FL phoneme processing are more open to the influence of gesture and what aspects are more closed. Identifying this boundary is not only useful for practical matters such as how—and when—to use gestures in teaching different aspects of a FL, but also for providing nuance to theoretical claims that gesture and speech form a fundamentally "integrated system" (Kelly, in press; McNeill, 1992, 2005).

Acknowledgements

We thank Tim Collett for programing the experiment, Dr. Bruce Hansen for help formatting the figures, and the students from Colgate's Research Methods course for assistance in collecting the data. Funding for the study was supplied by the Center for Language and Brain at Colgate University.

Competing Interests

The authors have no competing interests to declare.

Notes

[1] It is worth emphasizing that although rising and falling intonation gestures may seem more "iconic" in their use of space than short and long length gestures (at least intuitively), they both fall into the category of metaphoric gestures. After all, there is nothing inherently visual or spatial about the acoustical signal, so any manual gesture representing that signal visuospatially is by definition *metaphoric*. That said, it is possible that some metaphors are easier to grasp than others—we will return to this interesting possibility in the Discussion.

[2] Note that American English does make distinctions based length of vowels (e.g., "pick" vs. "peak") and intonation of words (rising vs. falling intonation at the ends of sentences distinguish questions from statements). However, these distinctions are not truly phonemic—that is, they do not operate purely at the segmental level—in the same way as they are in Japanese.

[3] Although there are different ways to metaphorically represent the "length" of a speech sound (e.g., using two flat hands with palms facing one another to show the length, as in "about this long"), we chose our way for two reasons: they were the gestures used in Roberge et al. (1996) and Hirata & Kelly (2010), and it is

common for Japanese language instructors to represent phonemic length distinctions in this fashion.

[4] After the first half of data collection (27 participants), it became clear in the post-experiment questionnaire that a small percentage of people adopted a strategy of closing their eyes during some of the trials. Two participants from this first session were eliminated from the experiment for this reason. To address this problem, the remaining participants were explicitly instructed that it was important that they keep their eyes on the screen during the experiment. There were no significant differences between the two sets of participants.

References

Biau, E., & **Soto-Faraco, S.** (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, *124*(2), 143–152. DOI: https://doi.org/10.1016/j.bandl.2012.10

Broersma, M., & **Cutler, A.** (2011). Competition dynamics of second-language listening. *The Quarterly Journal of Experimental Psychology, 64*(1), 74–95. DOI: https://doi.org/10.1080/17470218.2010.499174

Casasanto, D. (2008). Who's afraid of the big bad Whorf? Crosslinguistic differences in temporal language and thought. *Language Learning*, *58*, 63–79. DOI: https://doi.org/10.1111/j.1467-9922.2008.00462.x

Connell, L., Cai, Z. G., & **Holler, J.** (2013). Do you see what I'm singing? Visuospatial movement biases pitch perception. *Brain and Cognition*, *81*(1), 124–130. DOI: https://doi.org/10.1016/j.bandc.2012.09.005

Cooperrider, K., & **Núñez, R.** (2009). Across time, across the body Transversal temporal gestures. *Gesture*, *9*(2), 181–206. DOI: https://doi.org/10.1075/gest.9.2.02coo

De Ruiter, J. P. (in press). Gesture–speech unity — what it is, where it comes from. In Church, R. B., Alibali, M. W., & Kelly, S. D. (Eds.), *Why gesture? How the hands function in speaking, thinking and communicating.* John Benjamins Publishing: Amsterdam.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & **Monaghan, P.** (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences, 19*, 603–615. DOI: https://doi.org/10.1016/j.tics.2015.07.013

Drijvers, L., & **Özyürek, A.** (2016). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research*, 1–11. DOI: https://doi.org/10.1044/2016_JSLHR-H-16-0101

Flege, J. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics, 15*, 47–65.

Fujisaki, H., Nakamura, K., & **Imoto, T.** (1975). Auditory perception of duration of speech and non-speech stimuli. In Fant, G., & Tatham, M. A. A. (Eds.), *Auditory analysis and perception of speech*. London: Academic Press, pp. 197–219. DOI: https://doi.org/10.1016/B978-0-12-248550-3.50017-9

Griffiths, T. D., & **Warren, J. D.** (2002). The planum temporale as a computational hub. *Trends in Neurosciences, 25*, 348–353. DOI: https://doi.org/10.1016/S0166-2236(02)02191-4

Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). *IRAL-International Review of Applied Linguistics in Language Teaching*, *44*(2), 103–124. DOI: https://doi.org/10.1515/IRAL.2006.004

Hannah, B., Wang, Y., Jongman, A., & **Sereno, J. A.** (2016). Cross-modal association between auditory and visual-spatial information in Mandarin tone perception. *The Journal of the Acoustical Society of America, 140*(4), 3225–3225. DOI: https://doi.org/10.1121/1.4970187

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics, 24*, 495–522. DOI: https://doi.org/10.1017/S0142716403000250

Hardison, D. M. (2005). Second-language spoken word identification: Effects of perceptual training, visual cues, and phonetic environment. *Applied Psycholinguistics, 26*, 579–596. DOI: https://doi.org/10.1017/S0142716405050319

Hardison, D. M. (2010). Visual and auditory input in second-language speech processing. *Language Teaching*, *43*(01), 84–95. DOI: https://doi.org/10.1017/S0261444809990176

Hirata, Y. (2004). Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts. *Journal of the Acoustical Society of America, 116*, 2384–2394. DOI: https://doi.org/10.1121/1.1783351

Hirata, Y., & **Kelly, S. D.** (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research, 53*, 298–310. DOI: https://doi.org/10.1044/1092-4388(2009/08-0243)

Hirata, Y., Kelly, S. D., Huang, J., & **Manansala, M.** (2014). Effects of hand gestures on auditory learning of second-language vowel length contrasts. *Journal of Speech, Language, and Hearing Research, 57*(6), 2090–2101. DOI: https://doi.org/10.1044/2014_JSLHR-S-14-0049

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin, 137*, 297–315. DOI: https://doi.org/10.1037/a0022128

Hostetter, A. B., & **Alibali, M. W.** (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, *15*(3), 495–514. DOI: https://doi.org/10.3758/PBR.15.3.495

Imai, M., & **Kita, S.** (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B, 369*(1651), 20130298. DOI: https://doi.org/10.1098/rstb.2013.0298

Kelly, S., Hirata, Y., Manansala, M., & **Huang, J.** (2014). Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. *Frontiers in Psychology, 5*, 673. DOI: https://doi.org/10.3389/fpsyg.2014.00673

Kelly, S. D. (in press). Exploring the boundaries of gesture-speech integration during language comprehension.

In Church, R. B., Alibali, M. W. & Kelly, S. D. (Eds.), *Why gesture? How the hands function in speaking, thinking and communicating.* John Benjamins Publishing: Amsterdam.

Kelly, S. D., & **Lee, A. L.** (2012). When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high. *Language and Cognitive Processes*, *27*(6), 793–807. DOI: https://doi.org/10.1080/01690965.2011.581125

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. DOI: https://doi.org/10.1080/01690960802365567

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*(2), 260–267. DOI: https://doi.org/10.1177/0956797609357327

Kendon, A. (2004). *Gesture: Visible action as utterance.* Cambridge: Cambridge University Press. DOI: https://doi.org/10.1080/15427580701340790

Kita, S., & **Özyürek, A.** (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*(1), 16–32. DOI: https://doi.org/10.1016/S0749-596X(02)00505-3

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. DOI: https://doi.org/10.1016/j.jml.2007.06.005

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*, 982–98. DOI: https://doi.org/10.1002/hbm.21084

McCafferty, S. G., & **Stam, G.** (eds.). (2009). *Gesture: Second language acquisition and classroom research.* Routledge.

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

McNeill, D. (2005). *Gesture and thought.* University of Chicago Press. DOI: https://doi.org/10.7208/chicago/9780226514642.001.0001

Obermeier, C., Dolk, T., & **Gunter, T. C.** (2012). The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals. *Cortex, 48*(7), 857–870. DOI: https://doi.org/10.1016/j.cortex.2011.02.007

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. DOI: https://doi.org/10.1111/j.1540-4781.1995.tb05454.x

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*, 510–532. DOI: https://doi.org/10.1037/0033-2909.114.3.510

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

Sumby, W. H., & **Pollack, I.** (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*, 212–215. DOI: https://doi.org/10.1121/1.1907309

Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, *8*(2), 219–235. DOI: https://doi.org/10.1075/gest.8.2.06tel

Vance, T. J. (1987). *An introduction to Japanese phonology.* SUNY Press.

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. DOI: https://doi.org/10.1121/1.2956483

Werker, J. F., & **Logan, J. S.** (1985). Cross-language evidence for three factors in speech perception. *Perception and Psychophysics, 37*, 35–44. DOI: https://doi.org/10.3758/BF03207136

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41*(1), 67–85. DOI: https://doi.org/10.1016/j.visres.2011.09.007

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://doi.org/10.1525/collabra.76.pr

How to cite this article: Kelly, S., Bailey, A., & Hirata, Y. (2017). Metaphoric Gestures Facilitate Perception of Intonation More than Length in Auditory Judgments of Non-Native Phonemic Contrasts. *Collabra: Psychology*, 3(1): 7, DOI: https://doi.org/10.1525/collabra.76

Submitted: 03 January 2017 **Accepted:** 19 February 2017 **Published:** 14 March 2017