Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Effects of hand gesture and lip movements on auditory learning of second language speech sounds

Spencer Kelly[a], Yukari Hirata[b], Jen Simester[a], Jackie Burch[c], Emily Cullings[b] and Jason Demakakos[a]

[a]Colgate University, 13 Oak Drive, Department of Psychology, Hamilton, NY 13346, USA
[b]Colgate University, 13 Oak Drive, Department of East Asian Languages and Literatures, Hamilton, NY 13346, USA
[c]University of Rochester, 585 Elmwood Ave., Box 645, Rochester, NY 14642, USA
yhirata@mail.colgate.edu

Previous research has found that auditory training helps native English speakers to perceive phonemic vowel length distinctions in Japanese, but that their performance has never reached native levels. Given that multimodal information, such as hand gestures and lip movements, influences semantic aspects of language processing and development, we examined whether multimodal information helps to improve native English speaker's ability to perceive Japanese vowel length distinctions. Sixty native English speakers participated in one of four types of training: (1) Audio-Only; (2) Audio-Mouth; (3) Audio-Hands; and (4) Audio-Mouth-Hands. Before and after training, participants were given phoneme perception tests that measured their ability to distinguish between short and long vowels in Japanese, e.g., /kato/ versus /katoː/. Our original prediction was that more modalities of training would result in greater learning. Although all four groups improved from pre- to post-test, the participants in the Audio-Mouth condition improved the most, whereas the other two conditions involving hand gestures were no different from the Audio-Only condition. We discuss possible benefits and limitations of using multimodal information in second language phoneme learning.

1 Introduction

Adults face difficulty in learning to perceive and produce a second language, and it takes extraordinary effort and time to become like native speakers. Non-native speakers' perception and production abilities do not easily reach the level of native speakers, often resulting in a detectable foreign accent. Research in phonetic science and second language acquisition has progressed over the past several decades, investigating why and how adults are limited in learning a second language [1]. One of the most well-studied problems in these fields is the inability of native Japanese speakers to perceive English phonemic contrast /ɹ/ versus /l/ [2, 3, 4]. Because the Japanese language does not have /ɹ/ and /l/ as phonemes, native speakers have difficulty perceiving the distinction when they learn English.

Phonetic Science and Second Language Acquisition

Numerous studies have shown that, even though adults are limited in learning to perceive certain phonemes of a second language, their perceptual inability can be remedied by intensive auditory training in a laboratory [2, 3, 4, 5, 6, 7]. This laboratory training typically involves providing pairs of words auditorily, such as 'light'-'right' and 'cloud'-'crowd' to Japanese speakers, asking them to identify whether they have heard 'l' or 'r', and providing immediate feedback on their responses. Although this auditory training for difficult second language phonemes is proven to improve adults' perception, their perceptual abilities still do not reach the native level even after many hours of training. There is currently no available training method that brings adults to the native level in perceiving these difficult second language phonemes.

In the field of second language 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 (e.g., 'Total Physical Response Technique' [8]). Some studies examined effects of the use of physical gestures, such as for running, writing, and eating, on the learners' improvement in oral comprehension of word meaning [9]. However, their goals have been more practical than theoretical, and few studies have investigated the association of physical actions or gestures with speech sounds at the cognitive level, using scientifically valid methods.

Multimodal Communication and Language

Spoken communication occurs in a rich multimodal context. In natural face-to-face interactions, people produce important information through such non-linguistic channels as, facial expression, hand gesture and tone of voice. Theories of communication claim that this multimodal information combines with speech to help people better comprehend language [10, 11]. The present study focuses on two types of multimodal information: mouth movements and hand gestures.

Mouth movements are an inherent by-product of spoken language. Researchers have long noted that these visual movements correlate with particular speech sounds in a language [12]. Not surprisingly, people use lip movements to better comprehend speech sounds [13, 14]. Moreover, neuroscience research further supports the link between lip movements and speech sounds: Calvert and colleagues used fMRI to show that observing lip movements activates the auditory cortex, even in the absence of speech sounds, suggesting that "seen speech" influences "heard speech" at very early stages of language processing [15].

Hand gestures are another prevalent aspect of face-to-face communication. Iconic gestures convey visual information about object attributes, spatial relationships and movements. McNeill theorizes that these gestures, together with speech, are part and parcel of language and are integrated a deep conceptual level [11]. Behavioral research has shown that these gestures significantly impact language comprehension [16, 17]. For example, Kelly et al. has shown that a spoken sentence and gesture mutually disambiguate meaning of one another (semantics) during language comprehension—that is, gesture not only disambiguates the meaning of speech, but speech itself disambiguates the meaning of gesture [16]. Moreover, recent functional imaging research using event-related potentials (ERPs) has demonstrated that hand gestures influence semantic stages of the neural comprehension of words [18]. These studies have clearly demonstrated that hand gestures influence how people comprehend the semantics, or meaning, of a language. However, few studies have explored effects of gesture in the learning of second language speech sounds, or phonemes, which are the sensory foundation for understanding the semantics of spoken language.

Goals of the Proposed Project

In the above two sections, we have laid out the limitations of current research in the two major fields, one in phonetic science of second language acquisition and the other in studies on multimodal communication. The limitation of the phonetic science findings has been that, even though auditory training is found to improve second language learners' perception of difficult phonemes, their performance has never reached the native level. The limitation of studies on multimodal communication is that although previous research has investigated the relative and combined contributions of mouth movements and hand gestures to understanding of the semantics of a language, no study has investigated this question with regard to how people learn novel phonemes.

The present study addresses the following questions: how does multimodal information conveyed through speech sounds, mouth movements and hand gestures facilitate the perceptual learning of difficult second language phonemes? How does this multimodal training method compare with the traditional audio-only training method? Is the limitation of learners' perceptual improvement from currently available auditory training attributable to the cognitive limits of adults learning a second language? Or, is this limitation a methodological one, as researchers have mostly focused so far on training with audio stimuli in the absence of natural multimodal cues?

We will address these questions in the context of native English adults learning Japanese. Japanese has five pairs of short (/i e a o u/) and long vowels (/iː eː aː oː uː/). The length of a vowel, whether it is short or long, is phonemic in Japanese, i.e., it distinguishes meaning of two words. For example, /i/ with a short vowel means 'stomach,' but /iː/ with a long vowel means 'good.' The only difference between the short and long vowels is that of duration. Long vowels are 2.2-3.2 times longer in duration than short vowels [19], but the difference between the short and long vowels could be as small as 50 milliseconds when vowels are spoken quickly in a sentence [20]. Since there is no such phonemic distinction in English, native English adults have difficulty perceiving this Japanese vowel length distinction [7, 21, 22].

The present study investigates the effects of the following four types of training:

1. Audio-Only: Participants hear only audio stimuli during training.

2. Audio-Mouth-Hands: Participants are exposed to all three modalities during training. That is, the speaker's mouth movements and hand gestures are shown simultaneously with the auditory presentation of the target words.

3. Audio-Mouth: Participants hear the auditory stimuli and see mouth movements, but the speaker's hand gestures are blocked during training.

4. Audio-Hands: Participants hear the auditory stimuli and see hand gestures, but the speaker's mouth movements are blocked during training.

All four groups of participants completed a pretest, four sessions of one of the above training types, and a post-test over the course of two-week period. The pretest and post-test included only audio stimuli without mouth movements or hand gestures. The purpose of this format was to examine how the use of visual information, mouth movements and hand gestures, would ultimately improve participants' auditory ability to distinguish Japanese short and long vowels.

Hypothesis

Given the findings of robust effects of multimodal information in the semantic processing of speech, we hypothesize that the visual information (mouth and hand movements) synchronized with auditory stimuli will ultimately help participants to hear the distinction between short and long vowels in Japanese. Drawing from Hirata et al.'s results [23], we expect the Audio-Only condition to show a moderate but significant improvement from the pretest to the post-test. However, we predict the improvement in the test scores to be distinctly higher for the Audio-Mouth-Hands condition than the Audio-Only condition. The results of the Audio-Mouth and Audio-Hands conditions will help disambiguate the unique and relative contributions of mouth and hand movements.

2 Methods

2.1 Participants

Sixty students were recruited from a college in the Northeast United States and were paid for their participation. All were monolingual native speakers of American English and had not been exposed to Japanese phoneme training nor had studied Japanese prior to this study.

2.2 Test Materials

All participants were given a pre-test and then a post-test two weeks later. These tests are identical to those used in [23]. Each test was composed of 180 stimuli, each composed of a carrier sentence and a target word. The target words were five pairs of real Japanese words: /rubi/-/rubiː/, /ise/-/iseː/, /rika/-/rikaː/, /kato/-/katoː/, and /saju/-/sajuː/. The difference within each pair occurs in the final vowel, with one word ending in a short vowel and the other ending in a long vowel. Each carrier sentence was combined with every target word, so that ten unique stimuli were formed from each carrier sentence. The tests were broken into 6 blocks of 30 trials each, and each block has a different carrier sentence. During the presentation of each spoken sentence, the carrier sentence appeared in written form on the computer screen, with a blank taking the place of the target word. These stimuli were obtained by fully crossing the following five factors: 2 speakers x 3 speaking rates x 3 sentences x 5 vowels x 2 vowel lengths. The 2 native Japanese speakers in the tests, who were different from those who recorded training stimuli, spoke each sentence at slow, medium, and fast rates.

2.3 Test Procedure

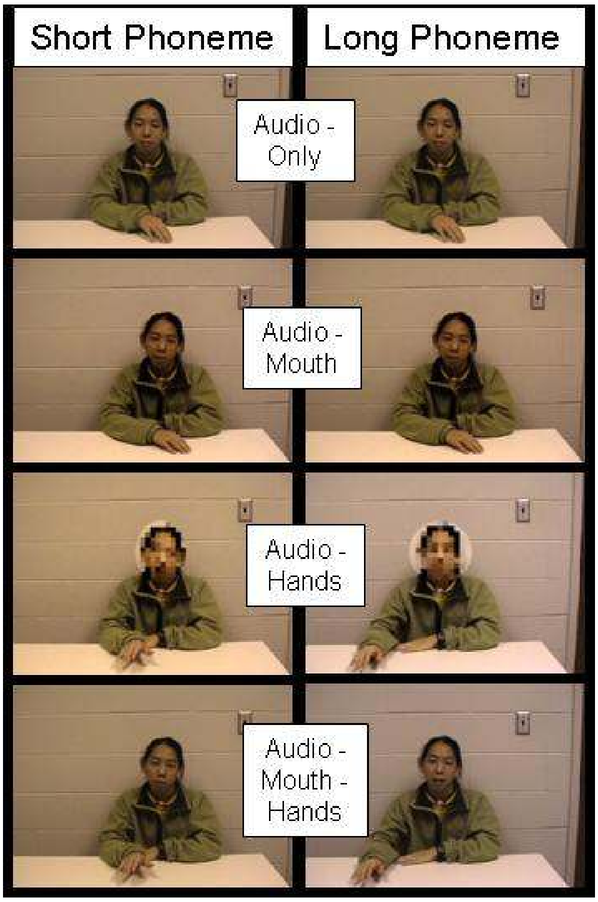
All participants took the pretest and the post-test in a quiet lab space using Grado Labs SR125 headphones. The 180 test stimuli (audio-only) in each test were presented in a random order across rates in 6 blocks. The participants were asked to identify whether the second vowel of the target words, e.g., /ise/ or /iseː/, in a carrier sentence was short or long on the computer screen (a two-alternative forced-choice identification task). The carrier sentences were written on the screen simultaneously with the presentation of audio stimuli. After each response, no feedback was given, and participants were asked to click a "play" button to hear the next audio stimulus. No word, sentence, or speaker in the tests appeared in training. This was to examine the participants' genuine ability to perceive the length of Japanese vowels in words that they had not practiced with, and not to test how many words or sentences they remembered from the training materials. The pre- and post-test took approximately 30 minutes each.

2.4 Training

Stimuli in each of four training sessions consisted of a total of 160 audiovisual clips presented on a computer, divided into 8 blocks. Each of the four sessions contained a different native Japanese speaker. These audiovisual stimuli were a combination of the audio clips used in [23] and video clips recorded for the present study. This blend of old and new clips were used instead of creating entirely new audiovisual stimuli in order to compare the results obtained in the two studies. The same Japanese carrier sentence was used for each audio clip. Each of the 10 sentences also contained a unique target word, which were nonsense Japanese word pairs including short and long vowels: /mimi/-/mimi:/, /meme/-/meme:/, /mama/-/mama:/, /momo/-/momo:/, and /mumu/-/mumu:/.

The video clips were created by videotaping 4 native Japanese speakers, who were not the speakers originally recorded in the audio clips described above. The speakers were recorded speaking the 10 sentences in concert with the audio clips of the same sentences. The speakers made a hand gesture during each target word representing the length of each vowel, moving their hand for two quick beats for short vowel words such /mama/, and for one quick beat and one long beat for long vowel words such as /mama:/. Their visual clips were combined with the audio clips from [23], so that it appeared as though the voices in the audio clips belonged to the videotaped speakers. For the Audio-Only condition, there was a static picture of the speaker not producing any lip movements or hand gestures. For the Audio-Mouth condition, the speakers' hand movements digitally removed by inserting a still frame of the body over the gestures, and for the Audio-Hands condition, the speakers' mouth movements were blocked by a pixel scrambling technique. Refer to Figure 1.

There were four training sessions (30 minutes each) over a two week period. During each training session, participants were asked to identify whether the second vowel in each target word, e.g., /meme/, was long or short by clicking the appropriate button on the computer screen. When they clicked the "play" button, they heard the audio and saw the video clip of the assigned condition. If participants responded correctly, the word "Correct" appeared on the screen, and they received the next sentence. If they responded incorrectly, the word "Sorry…" appeared on the screen, and they were required to click a button labeled "Play again," and the sentence was played three more times. Before the first and fifth blocks, participants were given examples of sentences and their correct responses.



.Fig.1 Training Conditions and Short/Long Phonemes.

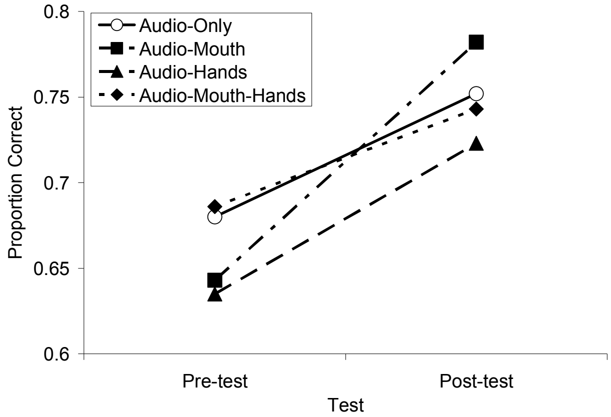
2.5 Design and Analysis

This is a mixed design, with test session as the within subjects factor and training condition as the between subjects factor. A 2 (Pre, Post) by 4 (Audio-Only, Audio-Mouth, Audio-Hands, Audio-Mouth-Hands) Analysis of Variance was performed on the pre- and post-test scores, and Dunn's multiple contrasts compared difference scores from pre- to post-test across the four training conditions.

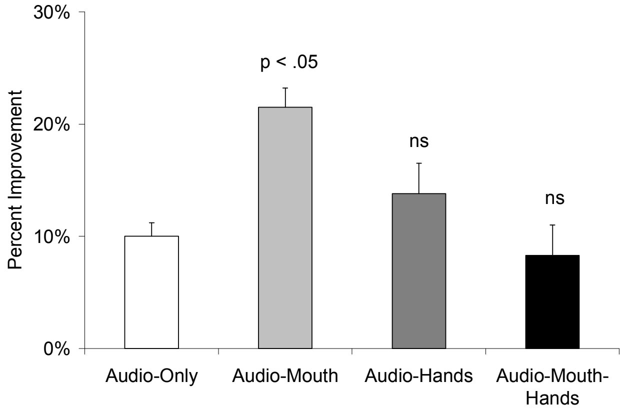
3 Results

The 2 (Test Time) by 4 (Training Condition) ANOVA did not uncover a significant main effect of Training, $F(3, 56) = 0.56$, ns, but there was a significant main effect of Test $F(1, 56) = 69.51$, $p < .001$, with participants improving from the pre- to post-test across all training conditions. One-way ANOVAs within the pre- and post-tests indicated that there were no significant differences among training condition before instruction, $F(3, 56) = 0.93$, ns, or after instruction, $F(3, 56) = 0.87$, ns.

However, there was a significant interaction between Test and Training, $F(3, 56) = 2.81$ $p > .05$. Refer to Figure 2. This effect was driven by the Audi-Mouth condition producing more learning than the Audio-Only condition, $tD(3, 28)$ 2.37, $p < .05$. However, neither the Audio-Hands, $tD(3, 28) = 0.77$, ns, nor the Audio-Mouth-Hands, $tD(3, 28) = 0.48$, ns, were different from the Audio-Only condition. Figure 3 presents the data in terms of percentage improvement from pre- to post-test for each of the training conditions.



.Fig.2 Improvement across the Four Training Conditions.



.Fig.3 Percentage Improvement from Pre- to Post-Test.

P values show significant differences of the three multi-modal training conditions from the Audio-Only baseline

4 Discussion

The results of the present study confirmed one part of our main hypothesis. Although all training groups improved from pre- to post-test, only the Audio-Mouth training was greater than the Audio-Only baseline condition. In contrast, the two training conditions with gesture did not improve performance beyond the baseline.

The finding that mouth movements helped English speakers best learn the vowel length distinctions in Japanese is consistent with previous research using other languages [24, 25]. For example, Hardison demonstrated that Japanese and Korean speakers improved their ability to distinguish between English /ɹ/ and /l/ to a greater extent after seeing mouth movements that were congruent with the phonemes, compared to hearing the contrasts without seeing the mouth. One explanation for this finding, and for the findings from the present study, is that the mouth conveys meaningful non-arbitrary visual information that correlates with the sounds that it simultaneously accompanies. This natural coupling may create stronger perceptual traces of the phonemes [15], which may make them more salient for future processing.

Interestingly, although lip and mouth movements facilitated phoneme learning in the present study, hand movements did not. In fact, when gestures accompanied lip and mouth movements (Audio-Mouth-Hands condition), the benefits of the face were lost. One possible explanation for this intriguing finding is that participants were “overloaded” with visual input, and this distracted them from reaping the benefits from the mouth and lips.

But why would one type of multimodal input—mouth and lip movements—facilitate phoneme learning, but another type—hand gestures—not? This question is particularly interesting because we know from previous research that hand gestures do help English-speakers learn the *semantics* of Japanese words [26]. For example, Kelly and colleagues showed that English speakers learn the meaning of Japanese words better when iconic hand gestures accompany versus do not accompany new Japanese verbs (e.g., learning that “nomu” means “drink” while seeing a drinking gesture accompany the two words). One possible explanation for this finding is that hand gestures—although very well suited for conveying higher-level meaning—are not particularly well suited for conveying lower-level acoustic information, such as phoneme contrasts.

So combining what we have learned so far, it is possible that lip and mouth movements play a significant role during the auditory encoding of the speech stream, but only when correctly encoded, hand gestures step in and help people understand the meaning of words in that speech stream. In this way, the benefits of multimodal input may vary according to different stages of linguistic processing, with lip and mouth input playing a more important role during phonological stages, and hand gestures playing a more important role during semantic stages. This interpretation has implications for theories about how gesture and speech are cognitively related. According to McNeill [11], speech and gesture form an integrated system during language production and comprehension. The results from the present study clarify this theory, and suggest that this integration most likely occurs more at the semantic level, and less at the phonological one.

The findings also have implications for language learning and instruction. In the fields of phonetics and second language acquisition, researchers have so far focused on the question of how one might maximize non-native speakers' learning of difficult second language phonemes using auditory only stimuli, e.g., by way of using a variety of voice and different phonetic contexts, and examining length of training and methods of feedback [2, 3, 4, 21]. The results from the present study provide insights into what other modalities, besides the auditory modality, can help learners to hear the distinction of difficult phoneme pairs. Apparently, when teaching novel speech sounds, multimodal input from lips mouth and speech may combine to facilitate learning. And from what we know about previous research on hand gesture, when teaching new vocabulary, multimodal input from the hands and speech may result in the best learning.

5 Conclusions

The present study replicates previous research that auditory training improves the phonological discrimination of novel phonemes in a second language [23]. However, although we hypothesized that adding a multimodal dimension to this training would enhance this phonological learning, there was increased performance only when auditory training was coupled with visual information from the face, but not from the hands. This finding, together with previous research, suggests that information from the lips and mouth may help with phonological processing and learning in a second language, whereas information from the hands may help only with semantic processes involved with vocabulary learning.

Acknowledgments

This project was supported by Picker Institute for Interdisciplinary Studies at Colgate University. We would like to thank Ron Crans for providing technical support throughout the study; Satomi Nunokawa, Yoichi Aizawa, and Shouhei Yamagami for taking the time to be filmed in our training videos; and finally Grace Baik and Marla Pfenninger for helping during data collection.

References

[1] W. Strange, Ed., *Speech Perception and Linguistic Experience: Issues in Cross-Language Speech Research.* Timonium, MD: York Press. (1995)

[2] J. S. Logan, S. E. Lively, D. B. Pisoni, "Training Japanese listeners to identify English /r/ and /l/: A first report", *J. Acoust. Soc. Am.* 89, 874-886 (1991)

[3] D. B. Pisoni, S. E. Lively, "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.* Timonium, MD: York Press. pp. 433-459 (1995)

[4] A. R. Bradlow, D. B. Pisoni, R. A. Yamada, Y. Tohkura, "Training Japanese listeners to identify English /r/-/l/: IV. Some effects of perceptual learning on speech production", *J. Acoust. Soc. Am.* 101, 2299-2310 (1997)

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

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

[7] K. Tajima, A. Rothwell, K. G. Munhall, "Native and non-native perception of phonemic length contrasts in Japanese: Effect of identification training and exposure", *J. Acoust. Soc. Am.* 112, 2387 (2002)

[8] J. J. Asher, "The total physical response approach to second language learning", *Mod. Lang. J.* 53, 3-17 (1969)

[9] J. O. Gary, "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.* New York: Academic Press. pp. 185-199 (1978)

[10] H. H. Clark, *Using language.* Cambridge, GB: Cambridge University Press. (1996)

[11] D. McNeill, *Hand and mind: What gesture reveals about thoughts.* Chicago, IL: University of Chicago Press. (1992)

[12] P. Ladefoged, *A Course in Phonetics.* Orlando, FL: Harcourt Brace Jovanovich, Inc. (1975)

[13] W. H. Sumby, I. Pollack, "Visual contributions to speech intelligibility in noise", *J. Acoust. Soc. Am. 26*, 212–215 (1954)

[14] D. Reisberg, J. McLean, A. Goldfield, "Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli" in B. Dodd, R. Campbell, Eds., *Hearing by eye: The psychology of lip-reading.* Hillsdale, NJ: Erlbaum. pp. 97 –113 (1987)

[15] G. A. Calvert, et al., "Activation of auditory cortex during silent lip-reading", *Science* 276, 593-596 (1997)

[16] S. D. Kelly, D. Barr, R. B. Church, K. Lynch, "Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory", *J. of Memory and Lang.* 40, 577-592 (1999)

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

[18] S. D. Kelly, C. Kravitz, M. Hopkins, "Neural correlates of bimodal speech and gesture comprehension", *Brain and Lang.* 89, 253-260 (2004)

[19] K. Tsukada, "An acoustic phonetic analysis of Japanese-accented English", Doctoral dissertation: Macquarie University. (1999)

[20] Y. Hirata, "Effects of speaking rate on the vowel length distinction in Japanese", *J. of Phonetics* 32, 565-589 (2004a)

[21] Y. Hirata, "Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts", *J. Acoust. Soc. Am.* 116, 2384-2394 (2004b)

[22] K. Landahl, M. Ziolkowski, "Discovering phonetic units: Is a picture worth a thousand words?", *Papers from the 31st Regional Meeting of the Chicago Ling. Soc.* 1, 294-316 (1995)

[23] Y. Hirata, E. Whitehurst, E. Cullings, "Training native English speakers to identify Japanese vowel length with sentences at varied speaking rates", *J. Acoust. Soc. Am.* 121, 3837-3845 (2007)

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

[25] K. Sekiyama, "Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects", *Perception and Psychophysics* 59, 73-80 (1997)

[26] S. D. Kelly, M. Esch, T. McDevitt, "Neural correlates of learning Japanese words with and without iconic gestures", *Cognitive Neuroscience Abstracts* 1, 84 (2007)