Factors Affecting the Perception of Japanese Moraic Rhythm by Second Language Learners

Yukari Hirata*

日本語学習者のモーラリズム知覚に影響を与える要因

平 田 由香里*

要旨：本論文では、日本語らしいモーラリズムの重要な要素である長短音節対立に焦点をあて、日本語学習者の知覚に影響を与える三つの要因（提示文脈・発話速度・複数モダリティインプットの学習効果）に関する研究を紹介する。日本語母語話者が文脈の単語の長短対立を知覚する際、単語内部分のリズムや速度も有用な手段として用いるが、初級学習者は自動的にそのような知覚ができない。また、知覚学習中の提示文脈・発話速度は、学習過程の重要な要因であることがわかった。さらに、視覚インプットとしての話者の口の動きは長短対立の聴覚学習に貢献するが、長短対立を手の動きで表すビデオジェスチャーは、聴覚学習に貢献しないことがわかった。

Key words: Japanese length contrast, moraic rhythm, non-native Japanese speakers, word and sentence contexts, speaking rate, multimodal input, lip movement, beat gesture

1. Introduction

Japanese is known as a mora-timed language, in which moras are the basic unit of timing, creating equal rhythmic beats (Vance 1987, Ladefoged 1993). Abundant research has been conducted to show the isochrony and phonetic reality of Japanese moras manifested in the acoustic dimension of duration in careful, read speech (Han 1962, Homma 1981, Sagisaka and Tohkura 1984, Port et al. 1987, Han 1994). For example, there is a tendency that duration of words corresponds roughly to the number of moras they contain, /ite/ (2 moras) and /ite/ (3 moras) as having a 2:3 ratio and /supai/ (3 moras) and /supai/ (4 moras) as having a 3:4 ratio (Han 1994). These precise ratios change with different speaking rates: the durational increment of an additional mora in one- to six-mora words is roughly equal within each speaking rate, but the increment is smaller for faster speaking rates (Port et al. 1987, Hirata and Whitton 2005).

Length of vowels, obstruents, and nasals is phonemic in Japanese, and the length distinction in pairs of words is an important aspect of the moraic rhythm. Duration is the major manifestation of vowel and consonant length contrasts in Japanese, e.g., single and geminate stop distinction (Han 1962, Fujisaki et al. 1975, Homma 1981, Han 1992, Han 1994, Hirata and Whiton 2005). The present paper focuses on this basic characteristic of Japanese moraic rhythm through the acoustic dimension of duration, and assumes that the local, smallest rhythmic units of vowel and consonant length contrasts are a crucial element as part of the global rhythm in streams of speech.

The acoustic dimension of duration is used to classify rhythmic patterns of different languages, such as stressed-timed, syllable-timed, and mora-timed languages (Ramus et al. 1999), but durational patterns do not perfectly conform to these three types of rhythm categories (Grabe and Low 2002). For Japanese, although evidence for the acoustic reality of the mora has accumulated as mentioned earlier, a debate continues as to whether the mora is an acoustically real unit (Warner and Arai 2001a). For example, Warner and Arai’s (2001b) durational analysis of spontaneous Japanese speech did not show as strong evidence for the mora as careful or read speech. Furthermore, other parameters, such as fundamental frequency and intensity, affect production and perception of rhythmic units in different ways for different languages (Lehiste 1970, Barry et al. 2009). This complexity of acoustic speech signals and their correspondence to phonological units is a theoretical pursuit itself, and this makes cross-language studies

* Associate Professor, Colgate University, Department of East Asian Languages and Literatures（コルゲート大学東アジア言語文学部准教授）
exponentially more complex. This paper limits its scope in native Japanese (NJ) speakers’ production and perception of the rhythmic units of vowel and consonant length, and uses their results as a means of evaluating non-native Japanese (NNJ) speakers’ perceptual abilities. Thus, this paper is not about providing evidence for or against the acoustic reality of Japanese moras.

NNJ speakers whose native languages are stress-timed (as in English), syllable-timed (as in French), or tonal (as in Chinese) are known to have difficulty acquiring Japanese length contrasts. Toda (2003) and Toda (2007) give a thorough review on this topic. These papers discuss various issues involved in non-native speakers’ perception and production of Japanese length contrasts, as compared with those by native speakers, and discuss cross-sectional and longitudinal methodologies for examining development of their abilities. The present paper does not intend to provide a comprehensive review on the above topic, but rather focuses on three important factors that affect non-native speakers’ auditory learning of Japanese moraic rhythm, namely, (1) speech contexts, (2) speaking rates, and (3) multi-modality of input and its perception. This paper also focuses mostly on NNJ speakers who are beginning learners, starting with no knowledge of Japanese, and compares their performance with that of the NJ speakers. Through this comparison, this paper aims at defining the distance between the NNJ speakers’ starting point and their ultimate goal, using useful acoustic and perceptual measures. The paper, thus, does not include discussion on advanced learners of Japanese, which should deserve full attention elsewhere.

2. Word versus sentence contexts

The first factor that is crucial in characterizing NNJ speakers’ perception and learning of Japanese moraic rhythm is the contexts in which target words are produced and perceived: words in isolation versus words in sentences. In reality, length of utterances varies in any number from one to several dozens of moras, from a one-word utterance, a word and a particle as a functionally complete utterance, to a long sentence with a syntactically complex structure. In this paper, however, for the sake of clarity in making an important point, an explicit dichotomy is made between words spoken and heard in isolation (word context) and words spoken and heard in a carrier sentence without pauses (sentence context). This paper discusses why it is important to explicitly compare these two contexts by reviewing data from native and non-native speakers.

2.1 Native Japanese (NJ) speakers

For NJ speakers’ production, durational structures of word-internal segments differ when identical words are spoken in isolation and in sentence-medial position. For example, a final vowel in a word spoken in isolation or a vowel in a prepausal or utterance-final position in a sentence lengthens significantly and systematically when the word has no accentual fall (Mori 2001). To take an example from Mori (2001), the vowel /e/ is significantly longer in /sake/ (with no pitch accent) ‘rice wine’ than in /sâke/ (with the first mora accented) ‘salmon’ when these words are spoken in isolation, but the duration of this vowel does not differ when the words are followed by a phrase, e.g., as in /sake no umami/. Durational variability is also high for vowels in prepausal and utterance-final positions (Mori 2004). This phenomenon of final vowel lengthening might be a position sensitive durational rule that overrides the general mora-timing rules in which segmental duration is compensated within and across moras found in Sagisaka and Tohkura (1984), Campbell (1999), Port et al. (1987), and Han (1994).

Another example that points to the importance of distinction between the word and sentence contexts is on NJ listeners’ perception of single and geminate consonants. There is a solid agreement in literature that the major perceptual cue to this distinction is consonant duration, e.g., duration of stop closure in /ita/ ‘stayed’ and /ita/ ‘went’ or frication of /s/ in /ise/ (‘place name’) and /ise/ ‘one’s entire life’ (Fujisaki et al. 1975, Otsubo 1981). However, the precise duration value that indicates a boundary between the paired words depends on duration of other units and speaking rates (Fujisaki et al. 1975), and there has been little consensus as to what unit is the most crucial one that plays an anchor role to the consonant duration. For example, this anchor was found to be the preceding vowel (e.g., /i/ in /it(a)/) in Watanabe and Hirato (1985) and Hirata (1990a), and the following vowel did not play a significant role in Hirato and Watanabe (1987) (i.e., the second /a/ in /ap(a)/) and in Idemaru (2005) (i.e., /a/ in /set(a)/) when experimental stimuli were disyllables in isolation. However, when stimuli were words in a carrier sentence, the segments following the contrasting stops played a significant role in Hirata (1990a) and Ofuka et al. (2005). The discrepancy in the use of different units as perceptual cues could be explained if we assume that different durational rules would apply in the word versus sentence contexts. The results that the vowel following the contrasting stop did not play a significant role in the perception of this distinction in isolated...
disyllables may be related to a decreased perceptual sensitivity in utterance/phrase-final positions (Kato et al. 2003). It might also relate to the high durational variability in the production of a final vowel in these positions, as mentioned earlier.

When NJ listeners perceive phonemic vowel length in disyllables spoken in a carrier sentence, they use not only the internal durational structure of words but also rhythmic and timing information outside those words as perceptual cues. In Hirata and Lambacher (2004), accuracy in identifying vowel length was near ceiling (98.1%) when word triples such as /mama/-/maːma/-/mama/ and /mumu/-/muːmu/-/mumu/ were heard in an original carrier sentence even though speaking rate varied from trial to trial. However, accuracy suffered significantly (69.5%) when these words of different speaking rates were excised out of the original carrier sentence context and presented in isolation (an ‘excised’ condition, see Fig. 1). Note that the nonsense disyllabic words themselves had two vowels whose duration clearly indicated the short versus long vowel distinction, as shown in Fig. 1. These two vowels alone were

![Fig. 1](image_url)

Fig. 1 Spectrograms and fundamental frequency (f0) contours of a sentence used in Hirata and Lambacher (2004). The panels show a formant frequency rage of 0–5000Hz and an f0 range of 0–500Hz in the y-axis. The time rage in the x-axis is 0–4.573s for the ‘slow intact’ panel, and 0–4.587s for the ‘fast intact’ panel. One can see that the absolute duration of the short vowel spoken slowly in the ‘slow-excised’ panel is about the same as, or longer than, the long vowel spoken fast in ‘fast-excised’ panel, but that the long vowel is relatively longer than the short vowel within the disyllable of each rate.
found to be insufficient for accurate identification of vowel length for NJ listeners. The decreased identification accuracy in the excised condition might partially be caused by the unnaturalness of stimuli due to the abrupt excision. This needs to be followed up with an experiment in which the non-durational aspects of excised stimuli are edited to sound more natural, e.g., in terms of pitch contours. However, given the position-specific durational rules such as final vowel lengthening as discussed earlier in utterance-final positions (Mori 2004), it is very likely that NJ listeners readily have different expectations about the durational structure of words in isolation versus in sentences. The result implies the importance of sentential contexts, i.e., global timing information, when perceiving local units such as phonemic vowel length.

2.2 Non-native Japanese (NNJ) speakers

How much do non-native speakers know (consciously and unconsciously) and perform on these durational rules that are different between the word and the sentence contexts? Are those rules for Japanese readily available to NNJ speakers who are learning Japanese? Hirata (2004a) aimed at answering these questions and conducted an experiment with beginning learners of Japanese. Native English speakers with no prior knowledge of Japanese were trained to perceive Japanese moras correctly in the word versus sentence contexts. After receiving instructions on Japanese moras, the participants were asked to count the number of moras in words of one- to six-mora length, including short and long vowels and consonants, e.g., /ja/ (1 mora), /kju:/ (2 moras), /otosu/ (3 moras), /odota/ (4 moras), /kaseka/ (5 moras), and /bosozoku/ (6 moras). Hirata (2004a) explicitly compared the participants’ performance in perceiving the number of moras of the words that are produced in isolation (word context) and in sentences (sentence context). For the sentence context, a carrier sentence was written on the computer screen with a blank, and participants were asked to count the number of moras of target words they had heard in the carrier sentence. The first group participated in training only in the word context, the second group only in the sentence context, and the third group did not participate in any training. All three groups took a pre-test and a post-test. The first important finding was that the overall performance of all three groups was consistently lower for the sentence than the word context. Second, the participants who were trained exclusively in the sentence context were able to generalize their learned ability to the word context. Third, those trained only in the word context also generalized their learned ability from the word context to the sentence context, but the amount of generalization was significantly less than that of the sentence training group.

One implication of these results is that NNJ speakers’ ability to perceive isolated words accurately does not guarantee the ability to perceive words in the sentence context. Beginning learners of Japanese might start from recognizing the correct timing of words in isolation, but they may lag behind in the abilities required for accurate recognition of the moras of words in sentences. One possible factor that is responsible for the sentence training advantage is the ability to segment a stream of speech into words and segments, which might have developed in the sentence-trained participants but not in those trained only with isolated words. Another factor responsible for the sentence training advantage is speaking rate of sentences, which is the second focus of the present paper and discussed in the next section. It is possible that what is lacking for beginning learners of Japanese is the ability to identify length of vowels and consonants in relation to the global timing and rate information available in the entire duration of sentences.

Another study, Hirata and Ueyama (2009), provides interesting insight into this topic further. In this study, following the same method as in Hirata (2004a) above, native Italian speakers’ perceptual ability was compared to that of native English speakers to identify the number of moras in Japanese words in the word and the sentence contexts. Italian has consonant length contrasts, similar to Japanese consonant length distinction, but does not have vowel length contrasts, while English has neither consonant nor vowel length contrasts. The question addressed in Hirata and Ueyama (2009) was whether there was an advantage for the native Italian group in accurately perceiving moras in Japanese words, particularly those including geminate consonants, e.g., identifying the correct number of moras for words such as /odota/ as 4 moras. First, the performance of the Italian group was similar to that of the English group in that their performance was consistently poorer for the sentence than the word context. Second, the Italian group did not perform any better than the English group for words with long vowels; they both tended to count the number of syllables such as /kju:/ (2 moras) as having 1 mora, overlooking the long vowel /u/ that contributes to one extra mora. This poor perception in both groups made sense since neither Italian nor English have short/long vowel contrasts in their respective native languages. Finally, what made
Factors Affecting the Perception of Japanese Moraic Rhythm by Second Language Learners

a difference between the two groups was that the Italian group performed significantly better (43.3%) than the English group (31.1%) in counting the moras of words with geminates in the sentence context, while their performances were not significantly different in the word context. The Italian speakers’ sensitivity to consonant length in Italian might have played a role in accurately detecting geminate consonants in Japanese sentences. If that is the case, the ability that the Italian group, but not the English group, possessed might be associated with getting the global timing of a long stretch of an utterance, just like native Japanese speakers showed for vowel length perception in the previous section.

3. Speaking rate

The above discussion on the word versus sentence contexts all points to the important role that a stream of speech plays for perception of more local rhythmic units such as short and long vowels or single and geminate consonants. But what are there in a stream of speech that carry useful information for accurate perception of rhythms? In this section we focus on speaking rate as such an element. Although there are many different ways in which speaking rate variations are manifested, such as differences in vowel formants (van Son and Pols 1990, Weismer and Berry 2003, Hirata and Tsukada 2009), in the activity level of the muscles and velocity of articulators (Gay et al. 1974, Gay 1981), the present paper focuses on ‘duration’ as a parameter that reflects the speaking rate.

3.1 NJ speakers

What is the effect of speaking rate on the duration of local rhythmic units such as short and long vowels and single and geminate consonants? Are long vowels and geminate consonants always longer than their short counterparts regardless of speaking rates? Hirata (2004b) examined these questions with NJ speakers’ short and long vowels spoken in a carrier sentence of varied speaking rates, slow, normal, and fast (see Port 1977, for definitions of these speaking rates). The results of duration analysis showed that long vowels were consistently longer than short vowels within each rate, but that there was a significant durational overlap between the two categories when all rates were combined (Fig. 2). In particular, short vowels spoken slowly and long vowels spoken quickly were completely overlapped in their duration. Similar findings were obtained for single and geminate stops across rates (Hirata and Whiton 2005).

If absolute duration of the short and the long segments overlap across rates, what might be a consistent indicator of the phonemic length of short and long vowels or single and geminate consonants in NJ speakers’ production? Both Hirata (2004b) and Hirata and Whiton (2005) found that ratios of these vowel or consonant duration to other parts of speech were more indicative of their phonological length distinctions. For example, in Hirata and Whiton (2005), the durational ratio of single stop closure to the preceding vowel was smaller, on average, than that of geminate stop closure to the preceding vowel, and these ratios overlapped between the two categories much less than the absolute duration did. This indicates that the phonological length of consonants and vowels was manifested in the relational form in the sense that the crucial portion of speech units (such as closure for single/geminate stops) had an anchor to which it was normalized (such as the preceding vowel). Among such possible anchors, in the case of single and geminate stop distinction, are the vowels preceding or following the contrasting stops (Hirata 2007), the entire words (Hirata and Whiton 2005), and the mora preceding the contrasting stops (Idemaru 2005, Yi 2007). For discussion in more details, see Hirata (2007).

While the above studies show NJ speakers’ production, how do NJ listeners perceive length contrasts in
the context of sentences spoken at different speaking rates? What is the precise role of speaking rate of a carrier sentence? How much information is contained in it that is useful for perception of local rhythmic units? Hirata and Lambacher (2004) addressed these questions with Japanese disyllables spoken originally in a carrier sentence of slow and fast rates. One of the conditions was a ‘mismatch’ condition, in which target disyllables that had originally been produced at the slow or fast rate was embedded in a carrier sentence that had been originally produced at the other rate, e.g., a slow target word combined with a fast carrier sentence. These stimuli with the mismatched rates were presented randomly to NJ listeners. Their accuracy in identifying the length of vowels in the target disyllables were 52.4%, significantly lower than the intact condition (98.1%) in which they heard original sentences as they were. Furthermore, this accuracy in the mismatch condition was significantly lower than that in the excited condition mentioned earlier (69.5%). There is a possibility that the unnatural-sounding stimuli, created by connecting fast and slow speech segments, might have contributed to the decreased accuracy. Another likely possibility, however, is that NJ listeners use speaking rate information available outside the target disyllables for identifying the local rhythmic units such as short and long vowels. Even though target disyllables have two vowels whose duration clearly indicates their relative length, it may not be enough for accurate perception of vowel length.

It is important to note that the use of word-external contexts in the perception of word-internal segments is not unique to Japanese or length contrasts. For example, English vowels were often misidentified when monosyllables were excised from original carrier sentences (Johnson and Strange 1982). Speaking rate of a sentence was shown to affect perception of /p/ and /b/ in English in Miller (1987) and Wayland et al. (1994). They found that longer voice onset time was needed for perception of /p/ when a carrier sentence was spoken at a slow than a fast rate. Similarly, Gottfried et al. (1990) found that native English listeners use duration of vowels or rate of sentences in identifying English /i/-/æ/ continua. It is also interesting to note that Miller (1987) claimed that the use of information in a sentence in perceiving local speech segments is obligatory, i.e., when rate information is available, listeners cannot make a phonetic judgment without taking the rate information into account.

3.2 NNJ speakers

As mentioned in NJ speakers’ production studies in the previous section, long vowels spoken fast had complete overlap in their absolute duration with short vowels spoken slowly. However, NJ listeners hardly make errors in perceiving the length of these vowels when the speaking rate information surrounding the target words is available. This is not the case for NNJ speakers. Studies have shown that this ability to take speaking rate of the surrounding context into account in perceiving length contrasts is not always present for NNJ speakers (Hirata 1990b), and that their perception is affected by absolute duration. For example, NNJ listeners were biased to perceive long vowels spoken in fast speech as short, and short vowels spoken slowly as long, because of their absolute duration (Hirata 2005, Tajima 2006).

We might ask, then, how NNJ speakers would learn to use speaking rate information in a sentence to accurately perceive length contrasts. What kind of input would help in their learning process? Hirata et al. (2007) examined these questions by training native English speakers with no knowledge of Japanese to identify Japanese vowel length in disyllables spoken in a carrier sentence of different speaking rates. Three groups of native English speakers each received training with different speaking rates of sentences: the first group with a slow rate only, the second with a fast rate only, and the third with slow and fast rates. The training materials and the amount of training were identical except for their speaking rates. If speaking rate is an ignorable factor in NNJ speakers’ learning of vowel length, i.e., they automatically adjust their perception to different speaking rates, the three groups’ improvement after training would not differ significantly from each other. However, if speaking rate is a crucial factor that needs to be taken into consideration for effectiveness of their learning, differential amount of improvement would be observed.

The results showed that those who received input of slow and fast rates improved their perception more than those who received input of only one rate, slow or fast. Furthermore, those who received only fast rate speech in their training did not differ significantly from a control group that did not receive any training. An important implication of these results is that, at their early stage of learning to perceive Japanese vowel length contrasts, NNJ speakers do not handle speaking rate variations effortlessly and automatically, which is consistent with results of Tajima et al. (2008). Hirata et al.’s (2007) results suggest that speech materials of varied rates are beneficial to their perceptual learning of length contrasts.
4. Multimodal input and auditory perception

The third factor that the present paper focuses on is different input modalities. How does visual input interact with auditory input in NNJ speakers’ learning of Japanese speech rhythm? Why is it important to consider this multimodal input factor?

One type of visual modality that has been known to affect auditory perception is visual information conveyed through speakers’ lips. Seeing lip movements as a speaker produces speech is strongly integrated into what one hears auditorily, known as the McGurk effect (McGurk and MacDonald 1976). This auditory and visual integration occurs for both native and non-native speakers (Sumby and Pollack 1954, Green et al. 1991, Arnold and Hill 2001, Hardison 2003, Wang et al. 2008). Visual information synchronized to auditory stimuli helps non-native speakers to perceive difficult rhythm and prosody of speech. For example, Krahmer et al. (2008), Hirata et al. (2008), and Hirata and Kelly (2010) showed that quick flicks of hands accompanying speech enhanced the acoustic realization of stress and pitch patterns, and that seeing these gestures as one heard speech also drew perceivers’ attention to those prosodic elements.

If these two types of visual modality, lips and hand gesture, are such an integral part of speech perception, do they play a role in NNJ speakers’ learning of Japanese rhythmic units such as short and long vowels? Would these two types of visual input help or distract auditory learning for learners of Japanese? Kelly, Hirata et al. (2008), Hirata et al. (2008), and Hirata and Kelly (2010) examined relative roles of these two types of visual input in improving NNJ speakers’ auditory perception of Japanese short and long vowels. They explicitly compared four types of training: audio-only (control), audio-lips, audio-hands, and audio-lips-hands. Training materials were audio stimuli of Japanese disyllables with short or long vowels spoken in a carrier sentence, and video clips in which NJ speakers spoke those sentences and produced beat gesture, with short vertical hand flicks and long horizontal movements for representing the short and long vowels of the target disyllables, respectively. For the audio-lips condition, participants heard speech and saw speakers’ lips producing it, while the view of hand gesture was blocked. For the audio-hands condition, they heard speech and saw hand gesture that synchronized well with short and long vowels of the target words, but the view of the speakers’ mouths was blocked. For the audio-lips-hands condition, they had all three pieces of information available.

The results showed that the audio-lips condition improved significantly more than the audio-only condition did, indicating a unique contribution of seeing lips in NNJ speakers’ auditory learning. This result supported previous studies showing effects of lip movements on auditory learning of a second language (Hardison 1999, Hardison 2003, Wang et al. 2008). An interesting finding was that the gesture conditions, the audio-hands and the audio-lips-hands, did not improve more than the audio-only condition. This result indicates that hand gestures not only had a null effect on the NNJ speakers’ auditory learning of Japanese vowel length, but also had a negative effect in the case of the audio-lips-hands, since the gesture cancelled out a positive effect of lips.

Research in multimodal input and auditory learning of non-native rhythms is still an uncharted area, and studies such as above motivate a number of follow-up studies. The above results on effects of hand gesture might sound counterintuitive, given the significant effects of hand gestures found in Krahmer and Swerts (2007). It is possible that hand gestures are more useful in perceiving and producing global rhythms and prosody over a stretch of an utterance than local, segmental units such as short and long vowel distinction. Another possibility is that hand gestures help NNJ speakers produce short and long vowels, as reported in Roberge et al. (1996), but may not help to perceive them as in Hirata and Kelly (2010). Alternatively, hand gestures might be most directly related to semantic, but not phonological or rhythm processing, and they might help mostly for higher level processing such as second language learners’ vocabulary learning (Kelly, McDevitt, and Esch 2008) or listening comprehension of a lecture (Sueyoshi and Hardison 2005). These possibilities have interesting theoretical and pedagogical implications, and should be pursued separately in the future.

5. Conclusions

The present paper focused on the three factors that
affect perception of Japanese moraic rhythm by NNJ speakers: word versus sentence contexts, speaking rate, and multimodal input. On the first factor, various studies in NJ speakers’ production suggested that durational structures of words are different between the word and sentence contexts. NJ speakers’ perception indicated that significant information is present in a stream of speech outside target words that helps NJ listeners to perceive length contrasts accurately. Global contexts across words that provide information on sentential rhythm may help the perception of small local rhythmic units of short and long vowels accurately. This is not to say that isolated words cannot be perceived accurately, as NJ listeners do perceive them with high accuracy when words are produced naturally in isolation (Tajima et al. 2008). But listeners may have different expectations about the word-internal durational structure in isolated word versus sentence contexts, such as word-final vowel lengthening (Mori 2001).

One of the many pieces of useful information present in the sentence context was speaking rate, which is the second factor focused in this paper. NJ speakers’ production studies indicated that speaking rate variations affect duration of short and long vowels or single and geminate consonants so as to yield substantial overlap in their duration between the two categories. However, the durational ratios of the target vowels or consonants to other parts of speech are more indicative of their phonological length distinctions. Even so, the word-internal duration information alone may not always be sufficient for NJ listeners’ accurate perception of vowel length. The speaking rate of word-external contexts is one of the useful information for identifying vowel length.

In contrast, NNJ speakers might not automatically possess the perceptual abilities to utilize useful information in the word-external contexts, and that might contribute to the imperfect perception of Japanese moraic rhythm. For NNJ speakers, perception of length contrasts in sentences might require a distinct skill that is different from the skill of simply telling a difference between duration of two segments in isolation. Training studies discussed in section 2.2 suggest that NNJ speakers might benefit from training that explicitly provide global contexts for the perception of local speech units as in words-in-sentence (as opposed to words-in-isolation) training. Furthermore, NNJ speakers benefit from receiving sentences spoken at different speaking rates more than those spoken at only one rate, as discussed in section 3.2.

What are theoretical implications of these findings? Models of second language speech learning, such as Flege’s (1995) Speech Learning Model and Best’s (1995) Perceptual Assimilation Model, have advanced the field with a strong focus on the segmental level of learning for non-native phonemic contrasts that are difficult for learners to acquire. The findings summarized in the present paper point to the importance of non-native speakers’ ability and inability to use global speech contexts that contain rich information on rhythm and prosody, which has not received much attention in extant models. Further research in this topic is necessary for our complete understanding of non-native speakers’ acquisition of fluent second language speech.

The final factor discussed in this paper was multimodal input that affects perception of Japanese moraic rhythm by NNJ speakers. Although it is still an uncharted area, studies suggest that seeing a speaker’s mouth producing a sentence helps NNJ speakers’ auditory learning of Japanese rhythmic units of short and long vowels. This result extends previous findings on effects of mouth movements producing isolated words to include those with sentences. Seeing hand gestures synchronizing with the spoken short and long vowels, on the other hand, did not help their auditory learning. This finding was counter to robust effects of hand gesture in many other areas of language processing and learning (see Kelly, Manning and Rodak 2008, for a summary), as well as some evidence for the neural basis of gesture-speech integration (Hubbard et al. 2008). Future research is necessary to narrow down the aspects of language processing and learning on which hand gestures have direct effects.

Finally, I point out some pedagogical implications of the factors discussed in this paper. In Japanese language classrooms, teachers are busy introducing many important points of grammar and vocabulary, and helping students improve all four skills of listening, speaking, reading, and writing. It is easy for teachers not to be mindful about these three factors discussed in this paper for the perception of Japanese moraic rhythm, which is a crucial aspect for native-like pronunciation. For example, oral and aural practice in minimal pairs of isolated words including vowel and consonant length distinctions should not be assumed to be sufficient, and learners will benefit from explicit training with sentences for these distinctions. Some teaching methods assume that it is best for teachers to always speak at a rapid rate as they normally talk with other native speakers. Model recordings that accompany textbooks rarely come with materials spoken at multiple speaking rates. Aiming at enabling learners to perceive Japanese vowel length distinction, the fast-only training in Hirata et al.
(2007) was found to be similar to having no training. Although this finding was based on beginning learners, it is possible that they might not have a chance to learn to make fine distinctions of length contrasts as they advance their studies of Japanese. Pedagogically, then, it might help learners if teachers consciously change speaking rates or if learning materials are recorded by multiple people who speak at different speaking rates. Finally, use of video clips and visual input for language teaching and learning has become more common with advancement of technology today, but we might need to be careful in ways to present them, depending on specific goals. We might need different types of video and audio clips for phonological benefits of acquiring speech rhythm versus for semantic benefits of understanding contents of passages.

Acknowledgments

For many of the studies introduced in this paper, I thank my collaborators, Jackie Burch, Emily Cullings, Jason Demakakos, Spencer Kelly, Stephen Lambacher, Jen Simester, Kimiko Tsukada, Motoko Ueyama, Elizabeth Whitehurst, and Jake Whiton. This research was supported by the Research Council of Colgate University.

References


Factors Affecting the Perception of Japanese Moraic Rhythm by Second Language Learners

*its Applications: Acquisition of Linguistic Organization and Human Audition*, 40.


(Received Nov. 18, 2009, Accepted Mar. 1, 2010)