Vol.2011-SLP-87 No.7
2011/7/22

A Preliminary Perceptual Analysis on the Relationship of Phoneme Duration and Speaking Rate

Greg Short,[†1] Keikichi Hirose[†1]
and Nobuaki Minematsu[†1]

In Japanese phonemes can be of two perceptual durations: short and long. This distinction can be difficult for students of Japanese to acquire. We perform a preliminary perceptual experiment to look into the effects of speaking rate with the future purpose of developing CALL software for learning the tokushuhaku. Japanese words and non-sense words were recorded at three different speaking rates. Then, the duration of one of the vowels was extended in order to produce what would be perceived as a long vowel. From these results, there seemed to be a small shift in the duration at which the transition from short to long vowel occurs. Also, there were other factors besides only speaking rate that caused the transition duration from long to short to shift such as whether it was the final syllable or not.

1. Introduction

Recent years have seen great increase in the number of international students studying in Japan[1)]. For these people, Japanese language education is of high importance. However, in their classes, teachers often have little time for pronunciation training as class-time is highly limited. Therefore, building pronunciation systems to aid the teacher and supplement classroom instruction can be a great benefit[2)].

One area that can present trouble for learners of Japanese is phoneme duration[3)]. Japanese can have long phonemes that have two beats called tokushuhaku. The existence of tokushuhaku in Japanese gives Japanese a rhythm that is not found in a lot of languages. Tokushuhaku are morae that make up part of a long syllable and include lengthened phonemes and nasal finals. Rhythmically, Japanese is often considered to be a mora timed language, which differs from syllable-timed languages such as Chinese and stress-timed languages such as English. Syllables can be heavy (two morae/with tokushuhaku) or light (one mora/without tokushuhaku)[4)] and the duration of many phonemes can be short or long. This can often be a source of struggle for those wishing to acquire good pronunciation skills, though, as many languages do not differentiate sounds by their length. For example, English, Portuguese, and Chinese all do not have this distinction. Native speakers of these languages, though, comprise a large portion of the Japanese language learning population. Consequently, it is important for a pronunciation learning system of Japanese to integrate error detection of tokushuhaku.

Also, in a system that detects errors in pitch accent, tokushuhaku recognition is necessary prior to accent recognition. This is because the Japanese accent is dependent on the number of morae so with deletion or insertions of the number of mora have to be handled. In[5)] a method was developed to detect duration related (tokushuhaku) errors. However, this system did not take into account the speaking rate. By increasing or decreasing speech rate, the length of the sounds will increase, which, in turn, may cause the length at which a sound will be perceived as short or long to change. This leads to the above methods possibly producing inaccurate results. Hence, in[6)] and[7)], the effect of speaking rate was taken into consideration when a sound was classified as being a tokushuhaku or not. These approaches, though, are limited in their applicability so there is still a need for a system that can better discriminate tokushuhaku.

As a result of the above considerations, we are working toward the development of a method to automatically detect tokushuhaku errors taking speaking rate into account. Since there are a variety of issues to be looked at in order to develop this kind of system, we have conducted a perceptual experiment which will be discussed in this paper. In this experiment speech samples were recorded at three different speaking rates and were resynthesized to have longer vowels or consonants. This was in order to better understand the relationship between phoneme duration (tokushuhaku) and speaking rate. Specifically the issue of short and long vowels will be discussed and consderations will be made for future experiments based on these results.

2. Problem Overview

2.1 Tokushuhaku

Tokushuhaku are found in heavy syllables in Japanese. There are mainly considered to be three kinds of tokushuhaku. Those are geminate obstruents, long vowels, and nasal finals. As these sounds do not adhere to the typical (C)V, construction they are termed tokushuhaku.

The nasal sound is phonemically realized as being one sound, (ん) /N/, but acoustically it can be realized as a variety of different sounds. Geminate obstruents are called choked sounds and they are represented with the kana (ッ). In the case of plosives, this will be realized with a longer occlusion and in the case of fricatives, longer friction. For all of the tokushuhaku, the spectrogram is rather similar to the spectrograms of the short sounds so these sounds are almost completely differentiated from their short sound counterparts almost completely by duration[6)].

†1 京都大学
Tokyo University

2.2 Mora Timing

Mainly, there are considered to be three types of timing in languages. Japanese is said to be a mora timed language, which means that each mora is roughly of the same length. This means that a syllable with a tokushuhaku (a two mora syllable) will be roughly twice as long as one without it. In syllable-timed languages such as Chinese, each syllable is roughly the same length and in stress-timed languages such as English, the length between each stressed syllable is said to be roughly the same length[4].

If one 's native language timing is different from that of Japanese, there will be many difficulties in correctly pronouncing Japanese due to interference from one 's first language. To give an example, in a native English speaker of Japanese will lengthen syllables that should be one mora and shorten syllables that should be two morae due to the rules found in stress timing. It can be difficult to properly acquire the difference between long and short phones for a language in which the main timing scheme is based on stress or the syllable.[3].

2.3 Speaking Rate

Speaking rate is defined as being the frequency at which the unit of timing is repeated. In Japanese, as the speaking rate is dependent on morae, it is often calculated by dividing the number of morae by the duration of the utterance in Japanese. With increasing speaking rate, the absolute length of sounds will get longer, the length of one mora phones and two mora phones will increase accordingly. In[8], they explored whether relational acoustic invariance exists in Japanese vowel length distinction. In[9], speaking rate and geminate duration was examined. In those they found that although absolute measures of the duration of sounds was largely affected by speaking rate, relational measures such as long-to-short vowel ratio and vowel-to-word ratios were little affected by speaking rate. These results seem to support the concept of their being relational invariance in Japanese. Due to the effects of speaking rate on phoneme length, it is necessary to account for this in developing a system for learning tokushuhaku.

3. Previous Research

Systems for learning tokushuhaku have been proposed in[5],[7], and[6]. The approach in[5] was based on a perceptual experiment using a variety of minimal pairs. Phoneme sequences with tokushuhaku minimal pairs of the same pitch pattern were chosen. The phoneme that formed the tokushuhaku minimal pair was then lengthened in increments to produce both words that form the minimal pair. Almost all of the pairs chosen for this experiment had a minimal pair difference on the first syllable. In most cases, the transition from being perceived as a tokushuhaku to not a tokushuhaku was between 80 ms and 140 ms.

In[7], a method that accounts for speaking rate was proposed. This was also based on a perceptual experiment. The target sound length was linearly normalized based on the duration of the word and the perceptual experiment. To derive the equations for this method, they used minimal pair pronounced with a tokushuhaku and without such as /itto:/ and /ito:/ lengthening the /t/ phoneme in order to produce both variations. Then they applied Baye 's theorem in order to .

In[6], the relationship between speaking rate and phone length was investigated. Minimal pairs were utilized. The speaking rate for the utterance was then estimated and from this the ISR (inverse speaking rate) was calculated. Tokushuhaku minimal pairs were spoken at three speaking rates in carrier sentences. Equations were then derived based on linear regression for the part of the word that either possessed or did not possess a tokushuhaku in these minimal pairs to determine whether a phone was long or short. Using this linear regression equation words were classified.

While both[7] and[6] took speaking rate into account, these two methods were still inadequate. It is not apparent that normalizing by the average mora length or by the length of the word is appropriate. In[10], it was said that the length of the vowel in the preceding mora is important in determining whether or not a sound is tokushu or not. If this is the case, then normalizing by the word length would be improper. This would make sense in that it is more likely a length distinction between two sounds should not be dependent on the length of the word. Also, this would make it much easier to determine whether a sound is as there would not be any dependencies on word length. Therefore, it would be possible to make the tokushuhaku determination even in a situation where speech recognition is required and the boundaries of the word are unkown. Also, while normalization was performed to account for speaking rate, this was done under the assumption that the change in length of the tokushuhaku would be the same as the change in the length of an entire word with increasing speaking rate. To create different speaking rates, the length of the entire word was varied. Doing this would result in all phonemes being varied equally, which will likely not occur in natural speech.

Also, for[6] and[7] there methods have not been tested enough on a wide variety of word lengths. In that research, it was only tested on words with minimal pairs and in[7] it was only applicable for contrasts of specific words of specific lengths. However, it is important not to have to rely on word length. Using word length in tokushuhaku discrimination means that it is necessary to conduct experiments for a variety of lengths for words. Also, word length is dependent on the phonemes in the word as well. For instance, the phoneme sequence /aku/ will be shorter in most cases than /shaku/. Thus, it is necessary to work toward making a system that does not rely on word length measures.

To come up with a method that is not dependent on calculations based on word length or calculations of speaking rate, it will be necessary to look more in depth at the relationship between speaking rate and tokushuhaku perception. It was suggested in[6] suggested to carry out such an experiment.

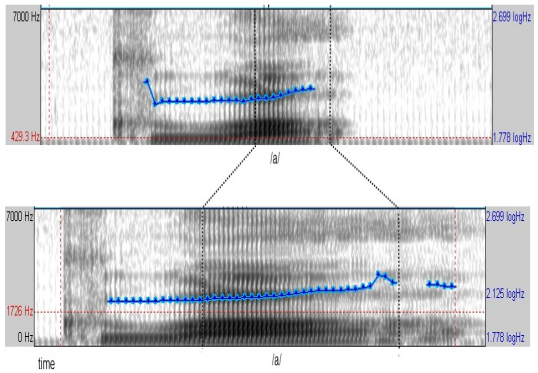

Fig. 1 Example of manipulation of /kuwa/ with the /a/ lengthened

4. Experiment

For this preliminary experiment, we mainly look at the relationship of speaking rate and phoneme duration perception in various situations. Also, we will look at other areas such as what kinds of effects there are from devoicing and position in the word. From there, we plan to construct a plan for a future perceptual experiment to get equations to automatically classify a phone as being long or short.

4.1 Description of Experiment Method

In this experiment we plan to look at how speaking rate affects how a consonant or vowel needs to be in order to be long, the affect of shortening the consonant before a vowel. We recorded a Japanese native speaker reading combinations of kana (wa,o,ka,shi) at three different speaking rates, slow, medium, and fast and also monomoraic words once. All of the recorded samples did not include tokushuhaku, but were made to have tokushuhaku using the duration manipulation utility in Praat. Some of the combinations ended up having meaning, and some of them were meaningless. Three different lengths were used: one, two, and three syllables. Then, using Praat one vowel in the utterance was lengthened incrementally. Fig. 1 shows an example of a lengthened /a/ in /kuwa/. Sixteen native Japanese speakers participated in the experiment, which was conducted online. The experiment was done using an internet-based Java program which displays two phoneme sequence written in katakana, one with a tokushuhaku and one without. A sound file is played and the subject selects which of the two words he or she heard. Table 1 shows the conditions for this experiment.

Table 1 *Conditions for the Perceptual Experiment*

Speakers	1 Japanese Male
Speaking Rates	3 (Slow, Medium, Fast)
Utterances	Isolated words and Non-sense words using the syllables o, ku, wa, shi
Subjects	16 Native Japanese Speakers
Test Method	Internet-based
Displayed Characters	All Katakana

We broke this experiment into six different sets. For each of these the vowel was lengthened to several durations and resynthesized.

For the first set, only the monosyllabic samples containing different phonemes were chosen to use for analysis in determining what kind of difference there was amongst the perception of the main phone combinations we are using.

In the second set, two syllable words were used and the length of the vowel on the second syllable was lengthened to be several different durations.

For the third set, we also shortened the length of the consonant of the second mora for a few of samples to see what the whether the length of the consonant is related to whether the vowel will be perceived as long or short. If there is no effect on the selection of the tokushuhaku compared to set 2 then it would seem that normalizing by the length of the word would be incorrect.

For the fourth set, two syllable words were used as well. However, in this set the vowel in the first mora was lengthened to see whether there would be a difference compared to the first syllable. This means was to see whether the syllable the lengthening is being done on is important in judging whether or not it is a long vowel or short vowel.

In the fifth set, three syllable words were used with subsets having the either the second, or third syllable vowel manipulated. This was to see whether there was a difference between perception of tokushuhaku in three syllable words was different form that in two syllable words.

For the sixth set, three syllable words with the vowel in the middle mora devoiced. This was to see whether there is any difference in selection rate if the preceding vowel is devoiced.

4.2 Results

In all of the graphs below, the x-axis indicates the duration in seconds of the target phone and the y-axis shows the rate of selection for the version of the phoneme sequence without the tokushuhaku for the target phone. Hereafter, S will be used to mean slow, M will be used to mean medium, and F will be used

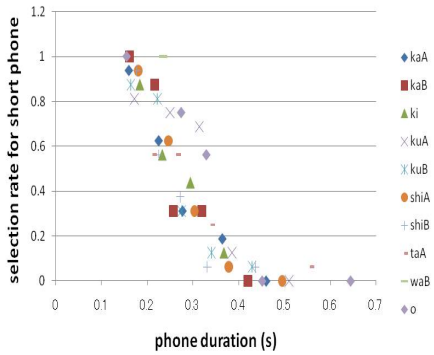

Fig. 2 Selection Results for Set 1: Monomoraic Syllables

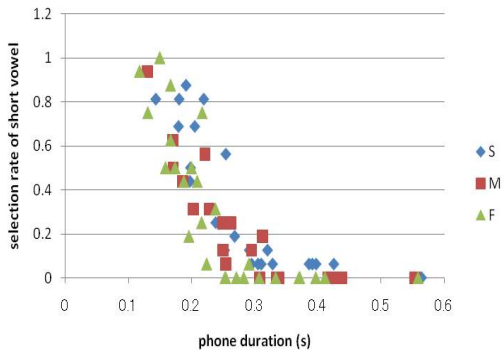

Fig. 3 Selection Results for Set 2: Manipulated Vowel is in Second Syllable

to mean fast speaking rate.

The results for the lengthening of the monomoraic samples are shown in Fig. 2. Two different subsets were created for some of the samples. The ‘ A ’ appended to the end of the sample name indicates that for this sample the entire vowel duration was lengthened and a ‘ B ’ indicates that only the steady state of the vowel was lengthened. That is to say that the area at the end of the vowel lengthening where the energy drops was not lengthened. This graph shows that there was not a large difference among the selection rate of different phones. The transition from being perceived as short from being perceived as long occurs for almost all samples between 200 ms to 350 ms. In some cases, the ‘A’ version resides to the left of the ‘ B ’, ‘ ku ’ and ‘ ka ’, and in the case of ‘ shi ’, the ‘ B ’ version is to the left of the ‘A’ version. As a result, there does not appear to be much of an effect from where the energy drops. Also, tokushuhaku discrimination does not appear to be largely affected by phoneme.

The results for the second set are shown in Fig. 3. This set includes bisyllabic phoneme sequences with the target for lengthening being the second mora for each sample. Three groups are graphed: S (slow speaking rate), M (medium speaking rate), and F (fast speaking rate). As there did not appear to be a large difference in the transitions for different phonemes in the above, all of the samples were grouped together based on the speaking rate. However, the selection rate for S trends slightly toward the right of the selection rate for F. If speaking rate has an effect on the discrimination of long and short vowels, these results would be expected, as an increase in duration should mean the duration for the transition should shift to the right.

Fig. 4 shows the results for the case when the consonant before the second syllable is shortened. The ‘1’ and ‘2’ appended to the beginning of the speaking rate indicates whether the consonant was shortened or not, with ‘1’ meaning it was not shortened and ‘2’ meaning it was shortened by half. In this graph, there does not appear to be much of a difference between the selection rates of the ones with shortened consonants and without, despite the length of the word dropping considerably. This could mean that the preceding consonant is not important for distinguishing short and long vowels. This was to determine whether or not a vowel is long or short.

The results for the fourth set show the two syllable samples for the case when the vowel of the first syllable was manipulated. Comparing it to set 2, Fig 5, overall the selection rate tends to reach 0% at a shorter duration. The transition from short to long also occurs roughly in the same region that they did for[5] as well. On this graph, there are a few samples that generally are to the right of the other samples as well. These all come from the samples for /oku/, the other two being /shishi/ and /washi/. It is possible that the difference is because /oku/ is a V syllable rather than a CV syllable like the other two.

Fig 6, is a graph of the results for set 5. From this graph the overall transition from long to short appears to be close to the same as it does in the set 2. There does appear to be some divergence in the samples for both F and S. This is probably because in this set one of the samples has the second syllable targeted for manipulation. Based on the results from set 4, it would be expected that the transition for the second syllable would be at a shorter duration than for the

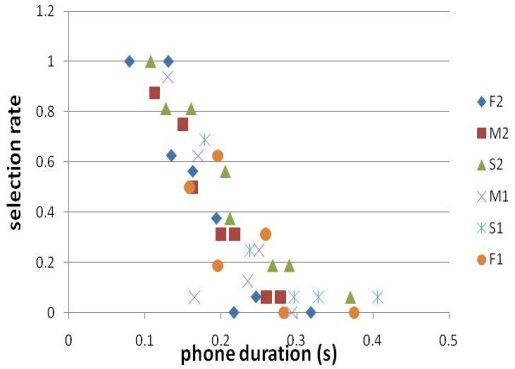

Fig. 4 Selection Results for Set 3: Manipulated Vowel is in Second Syllable

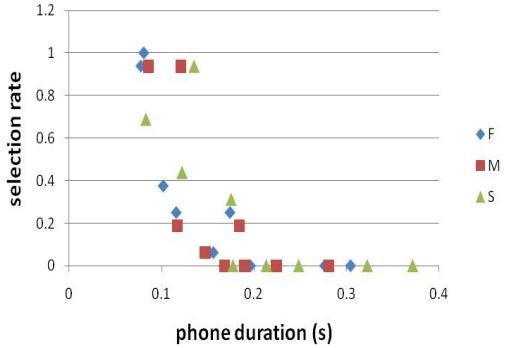

Fig. 5 Selection Results for Set 4: First Syllable Vowel Manipulated

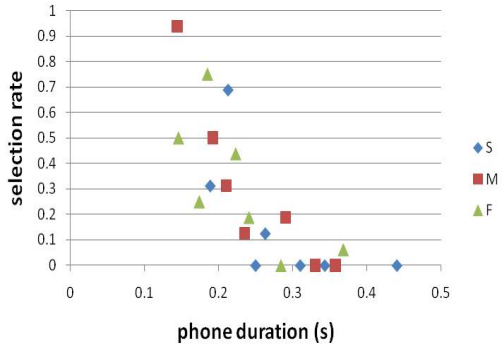

Fig. 6 Selection Results for Set 5: Vowel of Three Syllable Manipulated

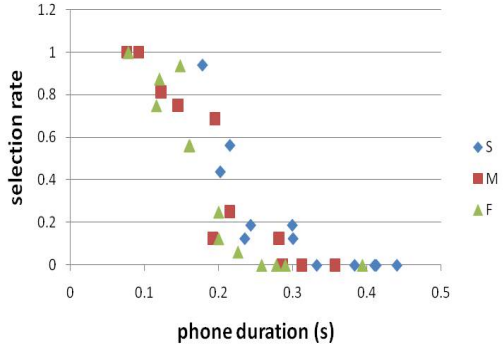

Fig. 7 Selection Results for Set 6

third syllable. It is possible that a decrease by half in the length of the consonant was insufficient.

The results for the sixth set show two syllable words with an unvoiced vowel in the middle mora such as /okushi/ and the vowel of the final syllable lengthened are in Fig. 7. In these samples, also, there seems to be a slight shift to the right for the slow speaking rate compared to the fast speaking rate, though it is not very large and there is still not enough data overall, the graph looks similar to the graph in set 2. It seems as though the durations at which the selection rate approaches 100% and 0% are slightly to the left of the two mora sample.

4.3 Discussion

As the amount of data was not sufficient and there was some overlap among the different speaking rates, even F and S, it is difficult to determine how to normalize the speaking rate. In most cases analyzed above, however, the distinction between non-tokushuhaku and tokushuhaku was slightly shifted toward longer durations for slower speaking rates. This shift was not very large in the case of the last mora and there was still some overlap between S and F, but if this shift stays the

same even with more data, it could easily result in misclassifications. It will be necessary to analyze this further in the future. Also, this shift might be larger for the first syllable of the word than it is for the final syllable as well. This shift due to speaking rate, though, did not appear as important as other factors such as position in the word or type of syllable, CV or V. Even with the difference in S and F, there was still only a shift of 30 or so ms. It will be necessary to tackle the other aspects in conjunction with speaking rate. There appeared to be other factors affecting the duration at which that transition occurred such as whether a syllable was V or CV and the position in the word. For the samples in which the vowel of the first mora was lengthened the transition from short to long occurred at a shorter duration than for the samples in which the second mora length was manipulated. The reason that the transition in selection of the short version of the vowel was further to the left than in the previous samples is likely due to reasons discussed in[11]. In that study the relationship between acceptability of durational changes and position in the word was investigated. It was found the further the mora is from the start of the word, the more acceptable durational changes are. In the case, in which the /o/ of /oku/ was extended, the duration at which it was perceived to be /o:/ was further to the right than in other cases. It is uncertain whether or not this is a common trend or not, but this will need to be analyzed more in depth.

5. Conclusion

In this paper, a preliminary experiment looking at the relationship between speaking rate and the perception of vowel duration. The ultimate goal of this is to correctly detect whether a learner of Japanese is producing a phone with the correct duration. This is an important issue as making mistakes in phone duration can degrade the intelligibility of speech[12]. In previous systems, they used metrics like word length and average mora duration in order to normalize the duration in the discrimination of long and short phones. These methods have not yielded satisfactory results, though.

The duration of a specific sound in a word was manipulated and the subject selected whether that sound was the short version or the long version. The difference in perception at the fast rate compared to the slow rate was not great, but there did appear to be some difference. The results of this experiment showed that using those metrics for normalization probably is not correct. They showed a tendency for longer durations being needed to change from a short to a long sound. Since the amount of data is still limited, it is difficult to come up with equations at this point so in the future it will be necessary to increase the amount of data to determine and difference in duration between the three different speaking rates and their influence on discriminating short and long phones. Also, since the variation was rather large for the different speaking rates, prior to manipulating the duration of the target sound, we will manipulate the durations of the phones for each speaking rate to be approximately uniform. Also, it appears to be better to use four speaking rates than only three, since in this experiment there was a lot of overlap between S(low) and M(edium). It may be important, also, to separate V and VC into two different sets. Other necessary things to do are use syllables with the nasal final and have the mora other than the target mora be long and use continuous speech and not solely isolated words. First, though, it is necessary to carry out an experiment that specifically looks at what causes a vowel to shift from being perceived as long or short without changing the duration of that vowel. Upon conducting another perceptual experiment, we will determine equations that factor in speaking rate and any other aspect we find to be important for tokushuhaku discrimination.

References

1) Ministry of Education, “300,000 International Student Plan,” *Ministry of Education*, pp.1–3, 2003.
2) G. Kawai, C.T. Ishi, “A System for Learning the Pronunciation of the Japanese Pitch Accent,” *Proc Eurospeech ’99*, 1999.
3) Motohashi-Saigo, M. et al, “Acquisition of L2 Japanese Geminates: Training with Waveform Displays,” *Language Learning And Technology*, vol. Vol. 13, no. No. 2, pp.29–47, 2009.
4) Auer, P, “Some ways to count morae: Prokosch’s Law, Streitberg’s Law, Pfalz’s Law, and other rhythmic regularities,” *The Acoustical Society of Japan*, vol. 27, pp.1071–1102, 1989.
5) Kawai, G. et al., “A Call System Using Speech Recognition to Teach the Pronunciation of Japanese Tokushuhaku,” *STiLL*, pp.73–76, 1998.
6) Ishi, Carlos et al, “Identification of Japanese “tokushuhaku” regarding the influence of speaking rate.,” *Technical Report of IEICE*, vol. Vol. 100, no. No. 97, pp.17–24, 2000.
7) Yamamoto, M. et al, “Computer Assisted Learning System for Japanese Special Mora and Its Evaluation,” *The Acoustical Society of Japan*, pp.1–8, 2000.
8) Hirata, Y., “Effects of speaking rate on the vowel length distinction in Japanese ,” *Journal of Phonetics*, vol.32, pp.565–589, 2004.
9) Hirata, Y., “Duration of Japanese singleton and geminate obstruents,” *Acoustics*, pp.2351–2356, 2008.
10) Toda, T. et al, “Perceptual Categorizationn of the Durational Contrasts by Japanese Learners,” *Tsukuba University Linguistics Repository*, vol.Vol. 33, pp.65–82, 1998.
11) Muto, M. et al, “Effect of speaking rate on the acceptability of change in segment duration,” *Speech Communication*, vol.Vol. 47, no.No. 3, pp.277–289, 2005.
12) Tsurutani, C., “Foreign accent matters most when timing is wrong,” *In INTERSPEECH*, pp.1854–1857, 2010.