Identification of Japanese Double-Mora Phonemes Considering Speaking Rate for the Use in CALL Systems

Carlos Toshinori Ishi, Keikichi Hirose and Nobuaki Minematsu
University of Tokyo, Department of Information and Communication Engineering
[c_t_ishi, hirose, mine]@gavo.t.u-tokyo.ac.jp

ABSTRACT

Effect of speaking rate on phone duration was investigated to improve a Computer-Aided Language Learning (CALL) system that teaches Japanese double-mora phoneme pronunciation [1]. The influence of the speaking rate was formulated as a linear regression line for each phone, based on the analysis results of a male speaker’s non-sense word utterances. The formulations were used to develop a method to estimate the speaking rate of the utterance over two hypothesis (presence and absence of double-mora phoneme), and decide which of the hypothesis is the more suitable. Experiments indicated that good identification between single- and double-mora phonemes was possible irrespective of the speaking rate.

1. Introduction

The Japanese is a language whose rhythm is based on “mora” units. The mora has a basic (C)V structure, but there are a set of special mora that we will call double-mora phonemes.

There are three types of Japanese double-mora phonemes: long vowels /N:/, a mora nasal /N/ and a mora obstruct /Q/. Although not identical to their short (regular) phoneme counterparts, they have very similar spectral features, being duration the main distinctive feature between long and short phonemes. However, duration changes according to speaking rates, and its normalization is necessary for the reliable distinction.

For non-native speakers of Japanese, correct duration control of double-mora phonemes is rather difficult. Therefore, a CALL (Computer-Aided Language Learning) system is desired for the training of double-mora phoneme pronunciation. Although pronunciation evaluation is possible by measuring the phoneme duration, the effect of speaking rate should be taken into account, since learners may utter in different speaking rates when using a CALL system.

Figure 1 shows an example of two utterances “Sorewa oQtodesu” and “Sorewa otodesu” of the same speaker, uttered in different speaking rates. The first utterance contains a long phone /Qt/, but its duration is shorter than the short phone /t/ in the second utterance, showing that the influence of the speaking rate has to be taken into account for duration analysis. Possible solutions could be normalize the duration relative to the previous phone [3], but it could be not convenient since phone durations may be affected by various factors, such as phone type, neighboring phones, accented/stressed or not, function of word, besides speaking rate.

Although influence of these factors on segmental duration was investigated by multiple regression analyses [2], research works on speaking rate were rather rare, especially related to the CALL systems. From this respect, an investigation has been conducted to formulate the influence of speaking rate on phone duration. This paper reports the results of the investigation, and the application of these results on the single- and double-mora phoneme identification.

Fig.1. Two utterances “Sorewa oQtodesu” (fast speech) and “Sorewa otodesu” (slow speech) of a same speaker. The duration of the long phone /Qt/ in the upper utterance is shorter than the short phone /t/ in the lower utterance, showing the danger of using absolute values.
2. Formulation of the effects of speaking rate on phone duration

2.1. Framework for the analysis

Although analysis of all possible phoneme combinations, i.e. all tri-phones, is desirable, it requires processing of a huge amount of data. So, to decrease the necessary size of data, instead of analyzing all the combinations, influences of the preceding and following phones over 5 Japanese vowels were analyzed separately for non-sense Japanese-like words, with the form “arV, XV, ra”, where

\[ V_1, V_2 = \{ a, i, u, e, o \} \text{ and} \]
\[ X = \{ k, t, p, kj, ch, ts, s, sh, h, hj, f, g, gj, d, dh, b, z, zh, r, \]
\[ rj, y, w, n, nj, m, ng, ngj \}. \]

Here, symbols with “j” represent the palatalized phones. They were treated separately, because differences in the articulation points may cause differences in segmental duration from their non-palatalized phones.

By analyzing the tri-phone parts of r-V-X sequence and those of X-V-r sequence for utterances of these non-sense words, influences of preceding and following phones can be extracted separately.

A list of these nonsense words were prepared and read by a male speaker in 5 speaking rates. Each speaking rate was realized by asking the speaker to try to synchronize his utterance to metronome beats. By using a metronome, a stable control of speaking rate is possible, though decreased naturalness of utterances may happen as a penalty. Discussions on this trade-off issue are left for the future study. Use of a metronome has another advantage; speaking rate can be determined as the metronome beat as we did for the formulation. Concretely, in order to increase the naturalness of the utterances, control of 2 morae per a beat was asked to the speaker taking the Japanese rhythm feature into account.

The utterances were manually segmented into phonemes referring to their spectrograms for further duration analysis.

2.2. Formulation for phone duration considering the speaking rate

For each r-V-X or X-V-r tri-phone sequence, phone duration was separately plotted with respect to the speaking rate. As shown in Fig. 2, a good approximation of the relationship between phone duration and speaking rate was obtained for each sequence by a linear regression line. The abscissa of the figure is ISR (Inverse Speaking Rate), which is calculated as the metronome beat period in ms/mora and is assumed to be the average mora duration of an utterance.

2.3. Extension for the formulation of phone duration

Since we investigated separately the influences of the preceding and following phones on the current phone, we need some hypotheses to extend the approximations to all X1-V-X2 sequences. Assuming the influence of the preceding phone on the current phone duration is independent to that of the following phone, we can formulate as;

\[ \text{dur}_{X1-V-X2} = \text{dur}_{X1-V} \pm \text{dur}_{V-X2} = \text{dur}_{r-V-r} \text{.} \quad (1) \]

Then, we can calculate duration of each X1-V-X2 sequence as follows as a function of speaking rate;

\[ \text{dur}_{X1-V-X2}(\text{ISR}) \equiv \]
\[ \text{dur}_{X1-V}(\text{ISR}) + \text{dur}_{V-X2}(\text{ISR}) - \text{dur}_{r-V-r}(\text{ISR}). \quad (2) \]

Since all the items of the right side of the equation are given from the analysis of utterances, formulation of vowel duration can be obtained in any contexts.

3. Estimation of speaking rate

3.1. Method 1

The simplest way to estimate speaking rate of an utterance (accent phrase in our experiment) is to divide its duration with its number of morae;

\[ \text{ISR} = \frac{\text{dur}_{\text{utterance}}}{N_{\text{mora}}} . \quad (3) \]

Even with this formulation, a stable estimation of the speaking rate may be possible irrespectively to the mora contents for long utterances (say, utterances with more than 5 morae). However, for short utterances, the duration may vary a lot depending on the mora combinations. For example, in “soto” and “oto”, the duration of “soto” is remarkably greater than “oto”. Furthermore, when unvoiced stops come in the initial position of an utterance, like “k” in “koto”, they lack from the closure periods, resulting in under estimations of ISR.

3.2. Method 2

Therefore, we proposed a new method, which could estimate speaking rate taking the phone types into account. This method is based on the formulations of the tri-phone duration in Section 2. Let’s represent an utterance as a sequence of tri-phones with X1-V-X2 sequences. The first step of the speaking rate estimation is to obtain the
formulation of duration for each constituting tri-phone of the utterance using equation (2). Then, we estimate the utterance duration for any speaking rates as the sum of tri-phone duration as follows:

\[
dur_{\text{utterance}} = dur_{\text{triphone1}}(\text{ISR}) + dur_{\text{triphone2}}(\text{ISR}) + \ldots
\]

(4)

Here, \(dur_{\text{utterance}}\) is the measured duration of the utterance, and \(dur_{\text{triphone}}(\text{ISR})\) is the formulation of each tri-phone. Solving this equation, we can obtain an estimation for ISR.

4. Identification of Japanese double-mora phonemes

4.1. Method of Japanese double-mora phoneme identification

The developed method of double-mora phoneme identification first estimates the speaking rate of the utterance, containing the target phoneme, under two hypotheses: the target phoneme being a double-mora phoneme or a single-mora phoneme. The estimation was conducted by both of methods 1 and 2 in Section 3 for comparison. Since the presence or absence of double-mora phoneme causes a difference in the number of morae of the utterance, the estimated speaking rate should be different for each hypotheses. Let’s denote the speaking rate under the first hypothesis as \(\text{ISR}_{\text{short}}\), and that under the second hypothesis as \(\text{ISR}_{\text{long}}\).

Then, for each estimated speaking rate (\(\text{ISR}_{\text{short}}\) or \(\text{ISR}_{\text{long}}\)), the target phone duration (\(\text{dur}_{\text{short}}\) or \(\text{dur}_{\text{long}}\)) was predicted. From triphone \(\text{dur}_{\text{short}}\) or \(\text{dur}_{\text{long}}\), the long and short versions of the predicted duration were calculated:

\[
dist_{\text{short}} = \text{abs}(\text{dur}_x - \text{dur}_{\text{short}}(\text{ISR}_{\text{short}}))
\]

(5)

\[
dist_{\text{long}} = \text{abs}(\text{dur}_x - \text{dur}_{\text{long}}(\text{ISR}_{\text{long}}))
\]

(6)

When \(\text{dist}_{\text{short}}\) was smaller/larger than \(\text{dist}_{\text{long}}\), the utterance was decided to include no/a double-mora phoneme.

Nevertheless, there is a problem in comparing directly durations predicted in different ISR, because the curves of \(\text{dur}_{\text{short}}(\text{ISR})\) and \(\text{dur}_{\text{long}}(\text{ISR})\) become distant as ISR increase. Thus, we proposed a normalized distance measure relative to the distance between \(\text{dur}_{\text{short}}(\text{ISR})\) and \(\text{dur}_{\text{long}}(\text{ISR})\):

\[
\text{normdist}_{\text{short}} = \frac{\text{abs}(\text{dur}_x - \text{dur}_{\text{short}}(\text{ISR}_{\text{short}}))}{\text{abs}(\text{dur}_{\text{short}}(\text{ISR}_{\text{short}}) - \text{dur}_{\text{short}}(\text{ISR}_{\text{long}}))}
\]

(7)

\[
\text{normdist}_{\text{long}} = \frac{\text{abs}(\text{dur}_x - \text{dur}_{\text{long}}(\text{ISR}_{\text{long}}))}{\text{abs}(\text{dur}_{\text{long}}(\text{ISR}_{\text{long}}) - \text{dur}_{\text{short}}(\text{ISR}_{\text{short}}))}
\]

(8)

4.2. Experiments of the identification of Japanese double-mora phonemes

Twenty four minimal word pairs (“oto / oQto”, “itsu: / iQtsu:”, “ichi / iQchi”, “kata / kaQta”, “kako / kaQko”, “jiQkaN / jiQkaN”, “hakeN / haQkeN”, “bushi / buQshi”, “iseki / iQseki”, “kado / kado”, “kado / kado”, “chizu / chizu”, “biru / biru”, “tsuchi / tsuchi”, “sekii / seki”,”kaite / kaite”, “heya / heya”, “iQsho / iQsho”, “kuro / kuro”, “toru / toru”, “hamoN / haNmoN”, “sama / saNma”, “tanin / taNnin”, “kona / koNna”), one of each pair including a Japanese double-mora phoneme and the other not, were inserted in a carrier phrase “Sorewa ... desu” (“This is ...”). These sentences were read by the same male speaker in 3 speaking rates (without metronome beats).

4.3. Results of the experiments

For each minimal pair, two panels as these in Fig. 2 were obtained. In these panels, 2 points (\(\Delta\) and \(\times\)) are plotted for each utterance of the minimal pair. Symbol \(\Delta\) depicts \(\text{dur}_{\text{short}}(\text{ISR}_{\text{short}})\) and symbol \(\times\) depicts \(\text{dur}_{\text{long}}(\text{ISR}_{\text{long}})\). The upper line is the formulation for phone duration of double-mora phoneme, and the lower line is that of single-mora phoneme.

The left panel of Fig. 2 shows the results for utterances

![Fig. 2. Results of Japanese double-mora phoneme identification for minimal pair “oto” / “oQto”.](attachment:image.png)
without double-mora phoneme, and the right one shows those for utterances with double-mora phonemes. Smaller \( \text{dist}_{\text{short}} \) in the left panel and smaller \( \text{dist}_{\text{long}} \) in the right panel indicate the good discrimination between double-mora phoneme and single-mora phoneme is possible with the proposed method (method 2). Actually, no discrimination error was obtained also for other minimal pairs. When the speaking rate was estimated as the average mora duration of an utterance (method 1), some identification errors arise. These errors are showed in the table 1. Furthermore, even in the cases where correct identification was done using method 1, we can note that the difference between short and long distance measurements is greater in method 2 than in method 1.

There was no significant difference between non-normalized and normalized distance results. However, it's expected that in extreme cases, the normalized distance should be better. Additional experiments are being conducted to verify this hypothesis.

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
& \text{dist} & & & \text{normdist} & & \\
& \text{fast} & \text{normal} & \text{slow} & \text{fast} & \text{normal} & \text{slow} \\
\hline
\text{"seki"} & \text{short} & 33.2045 & 22.9173 & 15.52 & 0.222518 & 0.10742 & 0.426818 \\
& \text{long} & 28.197 & 46.73078 & 72.98667 & 1.255938 & 2.167063 & 0.615615 \\
\hline
\text{"kado"} & \text{short} & 43.50175 & 60.86651 & 73.05075 & 0.312966 & 0.354563 & 0.332882 \\
& \text{long} & 35.33217 & 39.57999 & 53.2995 & 0.409247 & 0.342856 & 0.376142 \\
\hline
\end{array}
\]

\(a. \) Results using method 1.

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
& \text{dist} & & & \text{normdist} & & \\
& \text{fast} & \text{normal} & \text{slow} & \text{fast} & \text{normal} & \text{slow} \\
\hline
\text{"seki"} & \text{short} & 20.7581 & 8.799296 & 2.42076 & 0.191038 & 0.063763 & 0.01281 \\
& \text{long} & 42.49579 & 62.94999 & 93.59753 & 1.082396 & 2.113333 & 3.859714 \\
\hline
\text{"kado"} & \text{short} & 56.16495 & 77.00138 & 93.34657 & 0.448944 & 0.500127 & 0.473405 \\
& \text{long} & 26.51585 & 28.34665 & 39.16924 & 0.280286 & 0.225015 & 0.252783 \\
\hline
\end{array}
\]

\(b. \) Results using method 2.

Table 1. Results for non-normalized and normalized distance measurements using method 1 and method 2 for speaking rate estimation. These were the two only cases where identification error was detected using method 1. (No errors for method 2). The distance measurements for Japanese double-mora identification errors are underlined.

5. Conclusion

Through the duration analyses of Japanese nonsense words, a formulation was proposed to relate speaking rate and phone duration. The relations were represented as linear regression lines. We applied these relations for the Japanese double-mora phoneme identification. Experimental results showed a good identification was possible irrespective of the speaking rate.

The final goal of this research is to integrate the developed method of double-mora phoneme identification into CALL systems that teach the pronunciation of Japanese double-mora phonemes and Japanese rhythm. Further investigations are still necessary to achieve this goal, such as those on accurate automatic segmentation, influence of the pitch accent over long-short decision, and perceptual tolerance of duration errors in pronunciation.

Acknowledgements

We would like to thank Katsuhiro Fujimoto and Akira Ishida of the University of Yamanashi for contributions in data processing.

References