Considerations on Vowel Durations for Japanese CALL System

*Taro Mouri**, *Keikichi Hirose***, *Nobuaki Minematsu****

*Graduate School of Engineering, **Graduate School of Frontier Sciences,
***Graduate School of Information Science and Technology
University of Tokyo, Japan
{mouri,hirose,mine}@gavo.t.u-tokyo.ac.jp

Abstract

Due to various difficulties in pronunciation, utterances by non-native speakers may be lacking in fluency. The Japanese pronunciation is said to have mora-synchronism, and, therefore, we assume that the disfluency may cause larger variations in vowel durations. Analyses of vowel (and CV) durations were conducted for Japanese sentence utterances by 2 non-Japanese speakers and one Japanese speaker (all female speakers). Larger variations were clearly observed in non-Japanese utterances. Then, 10 Japanese speakers were asked to rate the non-Japanese utterances. Strong negative correlations were observed between durational variations and pronunciation ratings. Based on the result, a method was developed for automatic evaluation of non-Japanese utterances. The ratings by the method were shown to be close to those by native speakers. Also, in order to offer a corrective feedback in learner's voice, non-Japanese utterances were modified in their vowel durations by referring to native Japanese utterances. The modification was done using TD–PSOLA scheme. The result of listening test indicated some improvements in nativeness.

1. Introduction

Due to recent internationalization, a number of foreigners come to and reside in Japan. They need an extensive training of Japanese language. In the meantime, there is an increasing request of learning Japanese in learners' own countries. These situations require a large number of Japanese teachers, which will be hard to realize. The recent advancement of speech technologies and high performance of personal computers made automatic language training devices possible, and, therefore, many researchers are trying to develop such Computer Assisted Language Learning (CALL) systems.

In most CALL systems, the training of speaking skill is done in the following three steps: to evaluate learner's pronunciation using speech recognition technology, to show the difference in acoustic features of learner's voice from those of teacher's voice, and to repeat teacher's utterance as a corrective feedback. However, this training strategy includes two major problems. One problem is that the speech recognizer can only tell the distance of learner's utterance from a reference utterance (this can be an average of native utterances, teacher's utterance, and so forth), and cannot tell if the learner's utterance is acceptable by native speakers. The other problem is that most learners cannot get a good idea on how to correct their pronunciation only from the difference in acoustic features and the teacher's utterance. If learners could, they need not do language training from the first point.

We have been developing several CALL systems to solve these problems. For instance, in the system for *tokushuhaku* phoneme (long vowels, double consonants, nasal stops) pronunciation training [1], the learner's utterance was scored by the percentage of native speakers possible to recognize learner's speech. This scoring criterion was obtained through a listening test of synthetic speech by native speakers. We also developed a system for Japanese lexical accents, where, after recognizing learner's word accent types, his/her utterance was corrected in prosodic features and was served as a corrective audio feedback [2]. This corrective feedback was shown to be beneficial for the learner to correct his/her lexical accent pronunciation.

In the current paper, we try to develop an index to evaluate learner's pronunciation skill as a whole. It is widely said that mora-synchronism is observable in Japanese utterance rhythm, though arguments are still going on. Here, mora is a basic unit of Japanese utterance coinciding with a syllable, except when it includes a *tokushuhaku* phoneme. This mora-synchronism is often hard to be realized for non-Japanese learners. It could be distorted by the rhythm in learner's mother tongue and by other factors. In general, acquisition of the correct rhythm of foreign language usually requires a long period of training. Based on these considerations, we have attempted to evaluate learner's pronunciation level with the variations of syllable/mora durations. As the starting point, we investigated on how the durational variation of each vowel is different between non-Japanese speakers and Japanese speakers. We also copied teacher's vowel durations during the analysis-synthesis process of learner's speech, and checked if this process could have some positive effect on the learner's speech sounding as a native's utterance. in nativeness. The modified speech thus obtained can be served as a corrective feedback to the learner.

The rest of the paper is constructed as follows: In section 2, an analysis is conducted on the vowel durations for non-Japanese and Japanese utterances. In section 3, pronunciation ratings by Japanese teachers are checked if they have any correlations with the variations in vowel durations. Then, speech modification experiment is conducted in section 4. Section 5, have a brief discussion on the results, and in section 6, concludes the paper with the future work plan.

2. Analysis of vowel duration

Out of "Japanese Learner's Database [3]," recently constructed under a national project on "Advanced utilization of multimedia to promote Higher Educational Reform," two non-Japanese female speakers' (F1 and F2, see section 3.1 for the detail) utterances were selected for analysis. For comparison, one Japanese female speaker's (N) utterances were selected from the ATR "Continuous Speech Database for Research [4]." Speech samples used for the analysis have the same (linguistic) content for the 3 speakers. First the speech samples were segmented in

phonemes by the forced alignment using speaker independent mono-phone models (HMMs) included in the Japanese speech recognizer Julius [5]. The models were trained without separating male and female utterances, and their mixture number was 16. We used the segmentation results without any manual corrections. Then, the vowel durations were analyzed separately for normal vowels and long vowels. Here, the long vowel is the vowel uttered in two morae, such as "oo" and "ii" in "ookii (big)." This is because the mora boundary in the long vowel cannot be detected by the forced alignment and the long vowels may have different durational feature from normal vowels.

Table 1 shows mean μ, standard deviation σ and normalized standard deviation σ/μ of the vowel duration for each speaker. n is sample numbers. It is clear from the table that the (normalized) standard deviation is larger for non-Japanese speakers. This result inspired us to use the standard deviation of vowel length as an index of unnaturalness of non-Japanese speech.

Table 1: *Comparison of vowel duration between Japanese and non-Japanese speakers.*

speaker	vowel	n	μ[ms]	σ[ms]	$\frac{\sigma}{\mu}$
N	normal	1307	61.9	29.9	0.483
F0			97.2	58.3	0.600
F4			115	64.5	0.561
N	long	111	118	36.3	0.308
F0			170	56.5	0.332
F4			155	60.6	0.391

3. Correlation between variation of vowel duration and pronunciation rating

3.1. Pronunciation ratings by Japanese speakers

Utterances of 5 sentences from 49 sentence-set (Set C [7]) were first selected for 10 non-Japanese female speakers from "Japanese Learner's Database [3]." Then, 10 Japanese speakers were asked to rate each speaker's pronunciation skill by listening these 5 utterances. The ratings could be only on the mora rhythm, but we asked the raters to evaluate the pronunciation skill as a whole. This is because of two reasons; one is the ratings focusing on one feature and neglecting others are usually difficult. Sometimes this kind of request to raters causes large distortions in the results. The other is that our goal is to find a ruler for rating non-Japanese pronunciation skill and to change the training content by referring to the ratings.

The raters were asked to choose one from 11 ranks, from very poor to excellent. Since rather large variations were observed in the ratings among raters, we first normalized them so that ratings of each rater had the mean value of 0 and the standard deviation of 1. Then we took the average of the ratings of the raters for each non-Japanese speaker. The results were summarized in Table 2.

3.2. Vowel duration and ratings

Vowel variation can be viewed in various ways: only vowel part or including consonants, to calculate all the vowels together or each vowel, to calculate separately for preceding consonants or not, etc. Although a lot of analysis is necessary before deciding the best vowel duration definition for the purpose, here we simply decided to view the duration for each vowel separately. In the current paper, we only compared two choices: to view CV (consonant+vowel) syllable duration, or to view duration of V (vowel) part. The result is shown in Table 3 as the correlation of durational variations and pronunciation ratings. Figure 1 shows how theses two features relate with CV duration of vowel /e/. Since duration of a syllable differs a lot with and without preceding consonant, syllables without consonant were excluded from the calculation of variation of CV duration. From a similar reason, long vowels are not counted. Also, to avoid the phrase final lengthening affecting the result, vowels immediately before short pauses or the utterance end were excluded. In the table, correlation was also calculated for variations normalized by the segment duration. The better (negative) correlations were obtained for V for vowels /i/, /u/, and for CV for other vowels. It is interesting that the normalization degrades the correlation. This may imply that the absolute value of duration has some relation with the ratings, though further research is required to reach to the conclusion.

Table 2: *Ratings for non-Japanese speakers.*

speaker	mother tongue	rate
F0(IWA_F01)	Chinese	0.208
F1(KYO_F03)	Russian	-0.433
F2(KYO_F09)	Chinese	-0.207
F3(OSA_F03)	Korean	1.99
F4(TKT_F05)	Chinese	-1.18
F5(TOH_F01)	Thai	1.01
F6(TOH_F07)	Korean	0.143
F7(TOH_F11)	Korean	-0.266
F8(TSU_F01)	Chinese	-1.34
F9(TUT_F05)	Tagalog(not confirmed)	0.0818

Table 3: *Correlational coefficients between vowel durational variations and ratings by Japanese speakers.*

label	V(σ)	CV(σ)	V$\left(\frac{\sigma}{\mu}\right)$	CV$\left(\frac{\sigma}{\mu}\right)$
/a/	-0.631	**-0.730**	-0.0970	-0.537
/i/	**-0.687**	-0.443	-0.404	-0.176
/u/	**-0.778**	-0.662	-0.749	-0.565
/e/	-0.646	**-0.814**	-0.259	-0.612
/o/	-0.545	**-0.596**	-0.462	-0.466

3.3. Pronunciation ratings from variations in vowel duration

From the result in the previous section, we constructed a method of automatic rating of non-Japanese pronunciation. The method first normalizes standard deviation of CV or V duration of each vowel using the distribution of all the non-native speakers' deviations. Here, selection of CV or V is done following the result in section 3.2: CV for /a/, /e/, /o/, V for /i/, /u/. The normalized standard deviation for /a/ is given as follows:

$$s_{/a/\mathrm{CV}}(m) = \frac{\sigma_{/a/\mathrm{CV}}(m) - \mu(\sigma_{/a/\mathrm{CV}})}{\sigma(\sigma_{/a/\mathrm{CV}})} \quad (1)$$

where, $\mu(\sigma_{/a/\mathrm{CV}})$ and $\sigma(\sigma_{/a/\mathrm{CV}})$ respectively mean the average and standard deviation of the standard deviations $\sigma_{/a/\mathrm{CV}}$ of the vowel /a/ duration (CV duration, for /a/) among the 10 non-Japanese speakers. m is the speaker number. Then the weighted average of the normalized standard deviations for 5 vowels is calculated

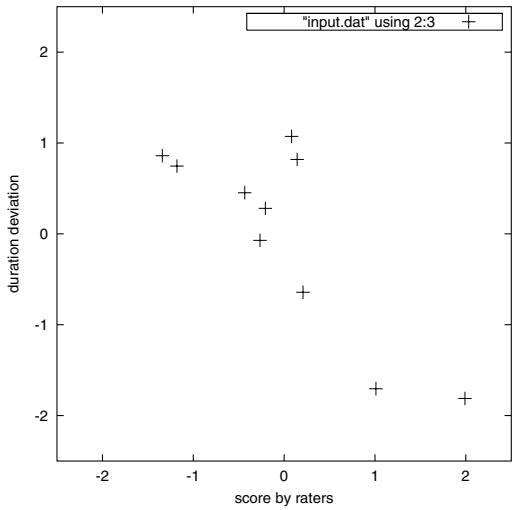

Figure 1: *An example of correlation between raters' ratings and vowel durations.*

and denoted as the estimated rating:

$$\begin{aligned} r_{\text{estimated}}(m) \equiv & -\{s_{\text{/a/CV}}(m)n_{\text{/a/}} + s_{\text{/i/V only}}(m)n_{\text{/i/}} \\ & + s_{\text{/u/V only}}(m)n_{\text{/u/}} + s_{\text{/e/CV}}(m)n_{\text{/e/}} \\ & + s_{\text{/o/CV}}(m)n_{\text{/o/}}\} \\ & /(n_{\text{/a/}} + n_{\text{/i/}} + n_{\text{/u/}} + n_{\text{/e/}} + n_{\text{/o/}}) \end{aligned} \quad (2)$$

where $n_{\text{/a/}}$ means number of samples of vowel /a/ in the 5 utterances, and so on. The sign is changed so that the correlations between the estimated ratings and the raters' ratings take positive values.

Table 4 shows the correlations among the raters' ratings. It also shows the correlations between the estimated rating and the raters' ratings. In the table, the ratings by the Japanese raters are represented as r1, r2, and so on. Rather high correlations, similar to those among Japanese raters, are observable between the estimated ratings and the raters' ratings. This indicates that the automatic rating can be used as an alternative to Japanese speaker's rating.

4. Speech modification by vowel length manipulation

The results in the preceding sections indicated the importance of vowel duration control in Japanese, though the role of phoneme duration for conveying prosodic features is said to be minor as compared to other languages, like English. So we investigated how the non-Japanese utterances would be improved by copying vowel durations of native utterances.

4.1. Method of modification

The vowel durations of non-Japanese utterances were modified to those of Japanese (N) utterances by the TD–PSOLA algorithm [6]. The Hanning window was applied to the speech waveform centered to the pitch marks to obtain speech segments, which were used for the overlap-add process:

$$w(i) = 0.5 + 0.5\cos\left(\frac{2\pi i}{N}\right)\left(-\frac{N}{2} \le i < \frac{N}{2}\right) \quad (3)$$

where N is sample numbers in a pitch period. The entire modification process is shown in Figure 2.

Since deletion and duplication of the segments deform the power and F0 contours of original waveform, their timings were selected sparsely for the entire period of the vowel; for instance, if one third of the pitch periods should be deleted, every third segment was deleted. For the current modification, the process was mostly deletion. To avoid the excessive deletion or duplication degrading the speech quality, the duration modification for each vowel was limited between half and double of the original length.

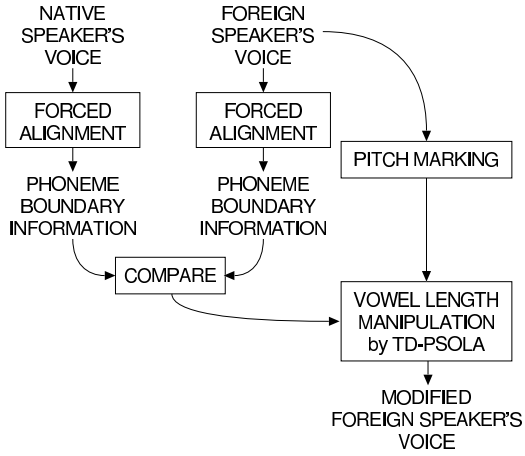

Figure 2: *Algorithm for vowel duration modification.*

4.2. Ratings after modification

Three sentences (different from those selected in the experiment in section 3) were selected from the 49 sentences of set C (see section 3.3), and their utterances by the 10 non-Japanese speakers were modified through the procedure in section 4.1. The 10 Japanese raters were again asked to evaluate the pronunciation of the utterances before and after the modification in the same 11 ranks as mentioned in section 3.3. So, each Japanese speaker rated 20 sets (10 non-Japanese before and after modification, each set with 3 sentence utterances) of utterances. To avoid the order of speech stimuli affecting the result, these sets were randomized before presenting to the raters. The resulting ratings were also normalized in the same way. Figure 3 shows the results: solid line for the original speech and dotted line for the modified speech. Each vertical bar indicates the one σ region. The average rating for the original speech was -0.065, while that for the modified speech was 0.065. The largest improvement was obtained for F6, whose utterances were modified most among 10 non-Japanese speakers. The standard deviation of (normal) vowel duration decreased from 57.8 ms to 34.2 ms by the modification in her case.

5. Discussion

Here, we should note that to simply decrease the standard deviations in vowel durations would not lead to the improvement of the pronunciation skill. As widely known, duration of a phoneme changes by various factors, such as preceding and succeeding phonemes, number of phonemes of the word to which the phoneme in question belongs, location of accent nuclei with respect to the phoneme, and so on. The good correspondence

Table 4: *Correlational coefficients between raters including automatic estimation.*

rater	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	estimated
r1	1	0.808	0.796	0.675	0.816	0.833	0.899	0.695	0.889	0.721	0.907
r2		1	0.922	0.831	0.905	0.959	0.939	0.76	0.893	0.889	0.842
r3			1	0.772	0.849	0.91	0.949	0.772	0.946	0.899	0.845
r4				1	0.677	0.814	0.755	0.461	0.651	0.668	0.568
r5					1	0.964	0.899	0.787	0.895	0.954	0.804
r6						1	0.936	0.737	0.904	0.932	0.818
r7							1	0.829	0.984	0.898	0.895
r8								1	0.844	0.868	0.678
r9									1	0.902	0.915
r10										1	0.722
estimated											1

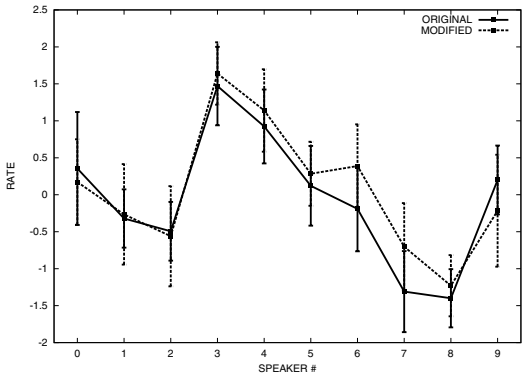

Figure 3: *Ratings by native Japanese before and after vowel duration modification.*

between standard deviations of vowel durations and pronunciation ratings by Japanese may be true only when the non-Japanese speakers are beginners of Japanese learning. When their skill increases, a more sophisticated rating scheme is necessary.

6. Conclusion

From the analysis of vowel durations in non-Japanese utterances, a method was developed to rate the pronunciation skill from their standard deviations. The obtained ratings were found to have a good correspondence with those by Japanese speakers. We also conducted a speech modification of vowel durations for non-Japanese utterances by copying those of Japanese utterances. The result of listening experiment showed the positive effect of the modification on the ratings.

For the future work, we are planning to combine the developed method with others developed by the authors, such as lexical accent pronunciation training method, to construct a unified CALL system teaching Japanese prosody to non-Japanese learners.

7. Acknowledgements

This work was partly supported by Grant in Aid for Scientific Research of Priority Areas (#120).

8. References

[1] Goh Kawai and Keikichi Hirose, "Teaching the Pronunciation of Japanese Double-mora Phonemes using Speech Recognition Technology, " *Speech Communication*, 30, pp. 131–143, 2000.

[2] Carlos Toshinori Ishi, Goh Kawai, and Keikichi Hirose, "A CALL System for Japanese Word Pitch Accent Types," *Proc. Acoustic Society of Japan*, I, pp. 245–246, Mar. 1999.

[3] Kikuko Nishina, Yumiko Yoshimura, Izumi Saita, Yoko Takai, Kikuo Maekawa, Nobuaki Minematsu, Masatake Dantsuji, Seiichi Nakagawa, and Shozo Makino, "Speech Database for Japanese Pronunciation Education," The Institute of Electronics, Information and Communication Engineering System Society Conference, SD-2-1, pp. 297–298, 2001.

[4] Shuichi Itabashi *et. al*, "Continuous Speech Database for research," Acoustic Society of Japan, 1991.

[5] Akinobu Lee and Kiyohiro Shikano, "Recognition Engine Julius/Julian-3.3 swaps multi-grammar simultaneously and dynamically," *Proc. Acoustic Society of Japan*, I, pp. 153–154, Sep. 2002.

[6] Francis Charpentier and Eric Moulines, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones," *Speech Communication*, 9 Nos.5–6: pp. 453–467, 1990.

[7] Masanobu Abe, Yoshinori Sagisaka, Tetsuo Umeda, and Hisao Kuwabara, "Japanese Speech Database for Research", ATR Interpreting Telephony Research Laboratories, Sep. 1990.