Naturalness on Japanese Pronunciation before and after Shadowing Training and Prosody Modified Stimuli

Rongna A 1, Ryoko Hayashi 2, Tatsuya Kitamura 3

1 Department of Rehabilitation for Sensory Functions, Research Institute of National Rehabilitation Center for Persons with Disabilities, Tokorozawa, Japan
2 Graduate School of Intercultural Studies, Kobe University, Kobe, Japan
3 Faculty of Intelligence and Informatics, Konan University, Kobe, Japan

hohchahar-aruna@rehab.go.jp, rhayashi@kobe-u.ac.jp, t-kitam@konan-u.ac.jp

Abstract

This study attempts to investigate the change of naturalness impression for the Japanese utterance by Japanese as foreign language learners (JFL) before and after pronunciation training (shadowing / repeating), and to discuss the crucial prosodic cue for the naturalness judgment.

The speech of 8 JFL learners before and after pronunciation training was used, and their durational pattern, pitch pattern, or both of them were replaced with those of the model speech. 52 Japanese native speakers (JNS) assessed the naturalness of these stimuli. The results showed JNS judge the duration AND F0 modified stimuli most natural. In addition, the shadowing trained group tended to have been valued highly than the repeating trained group after the training. Furthermore, the acoustical analysis of speech material showed a difference of moraic structure and pitch accent between the shadowing and repeating group.

Index Terms: Japanese as a foreign language learners, naturalness, shadowing, repeating, prosodic modified stimuli

1. Introduction

Japanese is known as a mora-timed language and has lexical pitch-accent. These characteristics cause the learners of Japanese (JFL) difficulty to control speech timing and realize correct word accents [1, 2]. Wrong timing control and pitch accent, however, causes misunderstanding of word meaning and unnaturalness of speech. The acquisition of mora and pitch accent is considered to be very important for JFL learners despite the existence of regional varieties of pitch accent [3, 4]. In recent years, the importance of pronunciation training has received remarkable attention in teaching Japanese. In our previous studies, we used shadowing training to practice JFL’s pronunciation [5, 6, 7]. Shadowing requires learners to listen to the model sentences while repeating almost simultaneously: The learners should repeat the sentences as exactly as possible while listening to the in-coming information [8]. Shadowing is said to improve prosodic features of learners’ pronunciation [9]. Our previous studies showed that the speech rate and accuracy of pitch-accent in the learners’ pronunciation rose radically during the shadowing training and this effect lasted also after the training [5]. However, it remained unknown how the naturalness of the learners’ pronunciation was changed by the shadowing training.

Furthermore, the crucial prosodic cue for Japanese native speakers’ (JNS) naturalness judgment is still in discussion. Sato [10] conducted a naturalness judgment test of Japanese utterances of a Korean native speaker and a Chinese native speaker and modified stimuli, i.e., one or all of the prosodic features were replaced by that or those of the native speaker’s utterance: pitch pattern, durational pattern or intensity of the sentence. As a result, he concluded that pitch pattern is the most important prosodic feature for JNS. On the other hand, Tsurutani [11] showed that Japanese native speakers put more weight on accuracy in timing (durational pattern) than in pitch when judging the naturalness of JFL learners’ speech. However, in [11] a Japanese-English bilingual speaker, with near native degree of fluency in both languages, was asked to utter the speech materials with perfect model pronunciation, and also with absolute beginner’s pronunciation containing all the classical errors. The two previous studies used different types of stimuli, and they have different results.

The purpose of this paper is, thus, to compare the naturalness judgments by JNS before and after shadowing training. As mentioned above, after the shadowing training, the pitch accent pattern and the speech rate (durational structure) are easily improved [5]. Our question is to what extent the corrected pitch accent and durational structure are important for JNS to perceive the speech as natural. To resolve this question, the naturalness judgment of the learners’ utterance was compared with the synthesized speech and the crucial prosodic cue for the naturalness judgment was also explored at the same time.

2. Methods

2.1. Speech materials

The natural stimuli were taken from our previous studies [6, 7], in which there were thirty three Chinese and Mongolian JFL learners who participated in two kinds of short term pronunciation training. All participants majored in Japanese language department at Inner Mongolian University. They learned Japanese for three years, and their level of Japanese was intermediate then. 19 of them trained to pronounce sentences with shadowing (shadowing group) and the other 14 with repeating (repeating group). In shadowing training, the JFL were instructed to imitate the model speech almost simultaneously as soon as possible. In the repeating training, the learners are required to repeat the model speech, presented as a short phrase, after listening to the end. In both training methods, the speech produced by a native male speaker of Japanese was used as the model speech. The text the participants read was always the same at pre-, post- oral reading and during training, and was chosen from [12], consisting of 656 morae, with the total...
duration of the model speech for 117.6 seconds. The data were recorded both before and after the training at a sampling rate of 48,000 Hz with 16-bit resolution using an IC recorder (Roland EDIROL, R-09) individually.

At our preliminary study [13], the naturalness of their utterances before and after training was evaluated with a Likert scale (ranging from 1: extremely unnatural to 5: extremely natural = native-like) by ten JNS. The sentences taken from the utterances by JFL were two sentences and consisted of 47 moras. Minasan wa okashi o yoku tabemasu ka? Moo otona dakara okashi wa amari tabenai toto hito mo ooi deshoo (Do you all eat snacks often? Perhaps someone says that he doesn’t eat snacks so much because he is already grown-up).

Ten JNS listened to the speech of JFL, and judged the naturalness from three viewpoints: overall impression, the correctness of speech and, the correctness of rhythm. The present study used the evaluation of overall impression, and chose four JFL learners’ utterance (two Chinese and two Mongolian) from each group of which naturalness was evaluated as almost the same (2.8 and 2.9 points) at pre-reading.

2.2. Stimuli

The following five types of stimuli were used:

pre: speech data recorded before the training,
post: speech data recorded after the training,
dur: speech data with the duration converted into that of the model speech,
F0: speech data with the F0 converted into that of the model speech,
durF0: speech data with the duration and F0 converted into those of the model speech.

Stimuli dur, F0, and durF0 were synthesized using Praat [14]. The source signal of dur, F0, and durF0 was pre. First, the speech data were labeled at the phoneme level. Stimulus dur was then synthesized so that the duration of each phoneme of the speech data was corresponded to that of the model speech. Time-Domain Pitch-Synchronous Overlap-and-Add (TD-PSOLA) method [15] was employed to the phoneme-level time warping. Stimulus F0 was synthesized by replacing the F0 of the speech data with that of the target speaker. The F0 of the target speaker was time-warped at the phoneme level to fit the time structure of the learner’s speech. Stimulus durF0 was synthesized by replacing the F0 of Stimulus dur with that of the model speech. The last stimulus thus has the duration and the F0 of the model speech and the spectral envelopes of the original speaker.

2.3. Naturalness judgment

Two types of natural stimuli (pre and post) and three types of modified stimuli (dur, F0 and durF0) were used in the present experiment. Each of the 5 type of stimuli includes 8 speakers’ stimuli: 4 from shadowing group and 4 from repeating group. Together with the model speech and five dummies, 46 stimuli in total were randomized and evaluated by 52 native Japanese speakers. Among them, 26 were Kansai dialect speakers (12 females, 14 males, average age 18.6) and 26 were Tokyo dialect speakers (15 females, 11 males, average age 21.3).

The JNS were required to assess the naturalness of stimuli using a Likert scale with potential responses ranging from 1 (extremely unnatural) to 7 (extremely natural = native-like). Before the real assessment experiment, we presented 5 sample stimuli for practice.

3. Results

There are no significant differences in pair-wise comparisons between Kansai dialect speakers and Tokyo dialect speakers (shadowing group: F (1, 6) = 0.06, n.s.; repeating group: F (1, 6) = 0.01 n.s.) at 5 type of stimuli. Therefore, we use the scores of naturalness from all of the 52 native Japanese speakers.

Figure 1 shows the mean evaluation score for each stimuli type. The naturalness judgment for the model speech was evaluated as 7 (SD=0). The shaded bars indicate the shadowing groups, and the white bars indicate the repeating groups. Two-way analysis of variance with group (shadowing and repeating group) and type of stimuli (pre, post, dur, F0, durF0) showed a significant main effect in type of stimuli (F (4, 24) = 29.51, p < .001) and marginal significance in interaction (F (4, 24) = 2.62, p = .06). The results of multiple comparisons of Bonferroni-test show the scores of durF0 were significantly higher than others in both training groups (p < .05). For stimuli post, the score of the shadowing group tends to be higher than that of the repeating group (F (1, 6) = 5.73, p = .05). The mean score of dur was higher than that of post in the repeating group (p < .05).

In other words, the JNS judged the stimuli durF0 most natural. The score became significantly higher only if both prosodic cues - duration AND pitch pattern - of the utterance were corrected. The results showed also that in the stimuli post, that is, the utterance after training, the shadowing group tended to get higher one point score than did the repeating training group.

4. Discussion

In the present study, it is revealed that in the natural speech after training (stimuli post) the shadowing group tended to get higher scores than the repeating group. In order to clarify the reason, the durational pattern and accent pattern of the natural stimuli pre and post was analyzed.

4.1. The duration of moras

As is well known, Japanese is a mora-timed rhythm language it has the characteristics of isochronal mora timing [16, 17]. At this present study, we measured the duration of each mora, but the...
moras prior to the pause (the end of bunsetsu) were excluded because of the prepausal lengthening [18]. The results show the mean duration of a mora in model speech was 98.2ms (SD=28.6, naturalness judgment in the former section was 7), while that in JFL speech before training was 133.7ms (SD=50.0, naturalness 3.2) in the shadowing group and 134.4ms (SD=45.5, naturalness 3.1) in the repeating group. After the training, the difference between the model speech and JFL’s speech became smaller in the shadowing group: the mean duration of a mora in the shadowing group was 117.1ms (SD=38.8, naturalness 3.8) and in the repeating group was 132.5ms (SD=45.2, naturalness 2.8). The mean duration and standard deviation of mora in the JFL’s utterance were longer than those in the model speech. However, after the shadowing training, the mean duration and standard deviation of mora had been got closer to model speech.

To see the deviation of durational patterns in the utterance by JFL from model speech, the duration of each mora was measured and the deviation of the moraic structure (DM) was calculated as follows:

\[ m: \text{number of moras in utterance} \]
\[ d: \text{duration of mora in model speech} \]
\[ d': \text{duration of mora in JFL speech} \]

Figure 2 shows the mean value and standard deviation of DM. Two-way repeated measures ANOVA was performed between groups (shadowing, repeating) and stimuli (pre, post). The result showed a significant main effect (F(1, 6) = 33.41, \( p < .01 \)) and interaction (F(1, 6) = 10.38, \( p < .05 \)). Post hoc analysis showed that the deviation of the shadowing group is smaller than the repeating group in the stimuli post (\( p < .05 \)). That is to say, after the shadowing training the durational deviation from model speech was reduced at the mora level.

![Figure 2: The deviation of moraic structure](image)

**4.2. Accuracy of word accent**

To investigate the importance of correct pitch accent, a native Japanese speaker who had received phonetic training (the second author) listened and judged if the pitch pattern of JFL was same with the model speech. The words to be judged were 13, e.g., all nouns, verbs, adjectives and adverbs underlined: Minasam wa okashi o yoku tabemasu ka? Moo otona dakara okashi wa amari tabenai toiu hito mo oai deshoo.

Mean accuracy of pre and post in both groups are shown in Figure 3. Two-way repeated measures ANOVA was performed between groups (shadowing, repeating) and stimuli (pre, post). The accuracy of word accent among the tasks showed no significant difference between shadowing group and repeating group (F(1, 6) = 1.05, n.s.), and between pre and post (F(1, 6) = 0.05, n.s.). However, the shadowing group tended to show higher accuracy than the repeating group after training (post) about 11.6 points.

![Figure 3: The accuracy of pitch accent](image)

Our previous study observed the effectiveness of shadowing training for JFL, they showed that speech rate and accuracy of pitch accent were substantially changed after training [5]. The present study attempted to see the improvement of JFL speech by shadowing training from the view point of the native speakers’ impression and to find the crucial prosodic cue for the naturalness judgment. The shadowing group got higher score of naturalness judgment after training than the repeating group. This result could refer to the difference of moraic duration and accuracy of pitch accent.

The present study showed a small tendency that dur stimuli were judged more natural than F0 stimuli, but there was no significant difference like [11]. The reason why Sato [10] concluded that pitch pattern is more important than duration pattern for JNS could probably be because of the difference of the learners level, which was intermediate. Therefore, the timing control in their utterances could have been better than that in the stimuli we used in the present study. In fact, in [10], the judgment for natural stimuli was 5.21 point (7 point scale) and higher than that of pre in the present study.

The other reason for the slight increase of the naturalness in dur could be thought to be the effect of pause. Analyzing the original speech of JFL learners, it was often found that the duration of pause was shorter than the model speech. This fact is in agreement with [19]. In addition, the number of pauses was quite different in each utterance. In the present study, however, we used two sentences and the pause between the sentences in dur stimuli was modified the same as that of the model speech.

The shadowing training requires the learners to imitate as soon as possible after the model speech was heard. This task can affect easily the durational structure (speech rate) including pauses.
Moreover, it is revealed that neither durational pattern nor F0 pattern alone caused significant increase of naturalness judgment scores by native speakers. The crucial prosodic cue is neither the durational and nor the F0 pattern, but both. For the JFL education it is important to pay attention to both prosodic cues and they cannot be treated separately.

5. Conclusions

The present paper reports about native speakers’ impressions of JFL utterances before and after shadowing/repeating training. Evaluation was also done for the synthesized stimuli in order to examine what the crucial prosodic cue might be. The results suggest that both pitch patterns and durational patterns are important in order for the utterances to be heard as natural Japanese and it is difficult to discuss the cues separately.

This results of the present study could provide useful suggestion for developing pronunciation training for JFL learners.

6. Acknowledgements

The authors would like to thank for staffs and students at Inner Mongolian University, Kokugakkin University and Kobe University. This research was supported by Grants-in-Aid for Scientific Research, project No. 21242013, No. 24652101 and No. 23242023, from the Japan Society for the Promotion of Science.

7. References