Comparison of Native and Non-native Evaluations of the Naturalness of Japanese Words with Prosody Modified through Voice Morphing

Shuhei Kato[1], *Greg Short*[1], *Nobuaki Minematsu*[1], *Chiharu Tsurutani*[2], *Keikichi Hirose*[1]

[1]Graduate School of Information Science and Technology, The University of Tokyo, Japan
[2]School of Languages and Linguistics, Griffith University, Australia
{kato,short,mine,hirose}@gavo.t.u-tokyo.ac.jp, c.tsurutani@griffith.edu.au

Abstract

Information on what kinds of mispronunciation cause the greatest loss in naturalness can be used in guidance to help learners achieve their goals of speaking with good pronunciation. We recorded Japanese words spoken by bilinguals in Tokyo Japanese and in foreign accents (English and Korean) and morphed them together with STRAIGHT emphasizing different acoustic parameters. By using STRAIGHT, we can simulate a wide variety of English-accented and Korean-accented pronunciations of Japanese. Then, we had native Japanese speakers and non-native speakers (Australian) evaluate the naturalness of these morphed utterances. The experimental results showed that Australian learners were very insensitive to the difference between native and non-native pitch although Japanese speakers were critically sensitive to these errors. Following these results, we give some pedagogical suggestions on what kind of pronunciation guidance is necessary for learners to become aware of what is unnatural to native speakers.

Index Terms: Foreign language learning, foreign accent, Japanese, naturalness, acoustic morphing, listening test

1. Introduction

For a language learner to achieve good and natural pronunciation, it is important to become aware of differences in vocalization and acoustics between the learner's native language (L1) and the target language (L2). Due to language transfer, though, this can be a large hurdle. One aspect that language transfer has a large effect on is prosody. This aspect can be especially difficult for learners to correct [1], so it is necessary to build awareness of the prosodic differences between L1 and L2. In college education for the Japanese language, however, due to time limitations, there are not enough chances for students to learn pronunciation in class. Prosody instruction is especially rare. One example of this is that it is not uncommon for college learners of Japanese to have little knowledge of the pitch accent. They are also unaware that the pitch accent will change when words are compounded together. Under such conditions, it is highly difficult for learners to notice the unnaturalness of their own prosody. In recent years, the importance of prosody in Japanese language education has gained a lot of attention [2].

As language transfer results in unnatural-sounding speech, it is necessary to examine what kinds of accents have the greatest impact in the loss of naturalness. To find this out, it is necessary to have native speakers subjectively assess the naturalness of speech samples with various kinds of foreign accents and degrees of accentedness. In [3–5], intelligibility or naturalness evaluations were conducted by having native speakers assess the utterances given by non-native learners. In these studies, learners' utterances themselves were used as stimuli. With this method of experimentation, however, it is difficult to cover a wide variety of accents and focus on different pronunciation features independently. Even if researchers focus only on one kind of accent, it will be difficult to collect speech samples accented at different degrees. What we need is a method of preparing speech samples in different kinds of accents at different degrees of accentedness.

One possible solution is using a speech morpher. A well-known speech morpher, STRAIGHT [6, 7], has been used in many applications such as transformation of age, gender, emotion, accent, etc. In [8], between two utterances of a minimal pair of words such as right and light, which were spoken by a single speaker, STRAIGHT could generate intermediate utterances between the two. Quantitative interpolation at different degrees is possible from the utterance of right to that of light.

Following [8], we generated interpolated speech stimuli between Japanese word utterances and their accented version, both of which were given by the same bilingual speakers, and had native Japanese speakers assess the naturalness of these stimuli [9]. Using STRAIGHT, we can morph speech quantitatively for specific acoustic parameters and deal with foreign accents better in terms of their acoustic properties. In [9], the experimental results showed that the degree to which language transfer affects naturalness perceived by native Japanese listeners varies from acoustic parameter to acoustic parameter and from L1 to L1, and found out what kinds of accents have the greatest impact in loss of naturalness.

However, to help understand how to remedy the problem of language transfer, it is important to know what acoustic characteristics learners are not aware of. If they are unaware of what is unnatural, it will be difficult to speak with natural Japanese. In this paper, then, we had Australian learners of Japanese listen to a part of the stimuli used in [9] and assess the naturalness of them in the same way as [9]. We compared the scores given by native Japanese and those of the learners.

2. Experiment

2.1. Recording of bilinguals' utterances

We recorded speech samples of Japanese words spoken by two speakers who are at native proficiency for both Japanese and another language. Their language backgrounds are shown in Table 1. They are recognized socially as having native proficiency of the two languages.

The words were selected in order to form a balanced word set emphasizing accent type, the number of morae, which is the rhythm-timing unit in Japanese, existence of heavy syllables, and their location(s) in the word. The total number of words was 162 (112 nouns, 30 verbs, 20 adjectives).

Each word was pronounced by each speaker both in Tokyo

Table 1: *The two speakers' language backgrounds*

sex	age	the other language	residential history except Japan	word list chosen by the speaker
F	early 30's	American English	6–14 years of age in California, the USA She attended local schools.	alphabet with language transfer
F	early 30's	Korean	0–23 years of age in Seoul, South Korea She attended local schools.	Hiragana

Table 2: *Acoustic parameters used for morphing*

parameter	abbreviation
fundamental frequency	F0
phonetic duration	dur
spectral envelope and aperiodicity	sp_ap
fundamental frequency and phonetic duration	F0_dur
all	all

Japanese and in a mimicking of foreign-accented Japanese in the other language she is proficient in. Like [8], we prepared word pairs spoken by the same speaker but unlike [8], one is native-sounding and the other is non-native sounding. This kind of preparation, though somewhat unnatural, of utterance pairs is because within-speaker morphing gives us speech samples of higher quality than cross-speaker morphing. For the native-like reading of Japanese, the speakers were given a reading sheet listing all the words in Hiragana as well as Kanji with word accents according to the NHK Accent Dictionary [10]. The nouns were pronounced in carrier sentences (*korewa [noun] desu.*). For the non-native sounding pronunciation, they were given the choice whether to use Hiragana, the Latin Alphabet (Hepburn style), the Latin Alphabet with language transfer considered in the notation, or in Hangul (as the standard notation [11]). We had the speakers make an effort to produce the words as phonemically accurate as possible in order for us to focus on the effect of prosodic transfer on naturalness.

2.2. Morphing between unaccented and accented samples

The stimuli were generated by morphing the Tokyo Dialect version of the word and the strongly accented version together from each speaker. With STRAIGHT, it is possible to select which acoustic features to morph. Namely, the values of four acoustic parameters (fundamental frequency, phonetic duration, spectral envelope, and aperiodicity) can be morphed independently. By morphing the spectral envelope, the spectrum shape as well as the energy of the spectrum will change. The aperiodicity parameter morphs the voicing degree of an input sound. In this paper, the five parameters shown in Table 2[1], and five morphing rates (0, 0.25, 0.5, 0.75, 1) were used. 0 indicates that the sample is spoken in Tokyo Japanese. 1 means it is morphed completely into the accented version.

Through the use of morphing in this way, it is possible to select which parameters to morph and the degree at which to morph them. From this, it is possible to determine at what degree of morphing there is a noticeable degradation in naturalness. Examples of speech morphing are shown in Fig. 1.

Lastly, the number of stimuli for each word was 42 (=(1+4 rates×5 parameters)×2 languages). Since the size of the balanced word set was 162, the total number of samples for the listening experiment was 6,804.

[1] sp and ap were morphed always synchronously, not independently because independent morphing often resulted in producing unnatural (non-humanlike) sounds. In the discussion section, we consider sp_ap as one acoustic parameter.

2.3. Subjects

2.3.1. Native Japanese speakers

42 native Japanese speakers ranging from 19–28 years of age, many of whom were university students, took part in the listening experiment. The stimuli were divided into four subsets, and each subset was assessed by 10–12 subjects. For each subset, the listening experiment was carried out over several sessions and the subjects were allowed to rest between sessions. For this experiment, we developed a web-based assessment system. The subjects listened to the stimuli through headphones and indicated the naturalness by clicking the corresponding checkbox in a window of the system.

Before the listening experiment, we presented 15 sample stimuli in order for the subjects to get a feel for the correspondence between seven-degree naturalness and stimuli.

2.3.2. Australian learners of Japanese

15 Australian learners of Japanese ranging from 18–30 years of age, who were studying Japanese at an university, took part in the listening experiment. They had studied Japanese for 3–10 years including classes in their university. Many words used in the experiment for native Japanese are considered to be beyond the proficiency level of the Australian learners. Then, the vocabulary chosen was limited to a part of the words which may appear on the Japanese-Language Proficiency Test Level 2 or lower [12]. The number of the selected words was 20, giving us 840 stimuli altogether. The stimuli were divided into two subsets and each subset was assessed by seven or eight subjects. Other conditions are the same as the experiment for native Japanese speakers.

2.4. Task

Two kinds of tasks were assigned to both Japanese and Australian subjects. The first task was to assess the naturalness of stimuli as Japanese (Tokyo Japanese) using a Likert scale with potential responses ranging from 1 (extremely unnatural) to 7 (extremely natural = native-like). The stimuli were presented in random order. Then, for Australian subjects to assess the naturalness in a condition similar to Japanese subjects, the pronunciation of the stimuli they heard was displayed in Hiragana without the word accent, and a sheet listing the words written in Hiragana with their word accent was prepared. The second task was to check whether a given stimulus was highly synthetic (extremely non-humanlike) as a result of the morphing with STRAIGHT, which outputs non-humanlike sounds on occasion. Although all the subjects were informed that the stimuli were artificially synthesized, if a stimulus was judged as highly synthetic, the judgement of that stimulus on nativeness will not be reliable. In the discussion of the following section, we disregarded those stimuli that were labeled as highly synthetic (non-humanlike) by more than half subjects.

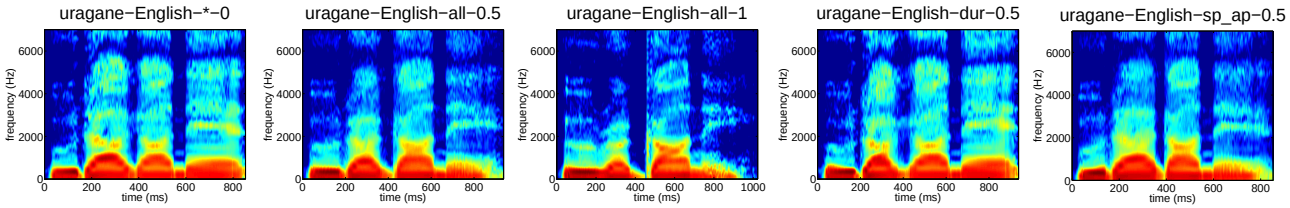

Figure 1: *The spectrograms of morphed utterances for* uragane *(dirty money)*

3. Results and Discussion

3.1. Results of the listening experiments

The experimental results are shown in Fig. 2 with the results at the top being for English-accented assessed by Japanese, by Australian, Korean-accented assessed by Japanese and the bottom for Korean-accented assessed by Australian. The horizontal axis gives the rates at which Tokyo Japanese words were morphed, and the vertical axis to the left of each graph shows the subjective naturalness as assessed by the subjects, and the right-hand vertical axis shows the p-values calculated with one-way ANOVA. The p-value at a morphing rate gives the minimum significance level at which the subjective naturalness differs significantly from that at a morphing rate of 0 (unaccented). In other words, the smaller the p-value at a particular morphing rate is, the larger the difference in the naturalness at that morphing rate is to that of Tokyo Japanese.

When a morphing rate, a parameter, and an accent are given, the subjective naturalness is calculated as follows.

1. Calculate the mean of the naturalness scores of each word over the subjects.
2. Calculate the average and the standard deviation of the mean scores over the words.

The average and the standard deviations ($\pm\sigma$) are shown here.

3.2. Discussion

We inspected which acoustic parameter has a larger effect on the naturalness for each accent. For this, the minimum morphing rate at which the subjective naturalness drops significantly is calculated for each graph. The smaller the minimum rate is, the larger the effect of that parameter on naturalness is.

First, we inspected the results by native Japanese speakers. For the English-accented stimuli, the minimum morphing rates at which the p-value is 0.01 are 0.5 for F0, dur, and sp_ap. For the Korean-accented stimuli, the minimum rates are 0.5 for F0 and dur, and 0.75 for sp_ap. Next, we looked into the results by Australian learners. For the English-accented stimuli, the minimum morphing rates at which the p-value is 0.01 are 1 for F0, 0.75 for dur, and 0.5 for sp_ap. For the Korean-accented stimuli, the minimum rates are none for F0 and dur, and 1 for sp_ap. From these results, F0 produced the largest discrepancy between the degrees of effect on the subjective naturalness assessed by native Japanese speakers and Australian learners, and dur followed. On the other hand, sp_ap showed similar naturalness patterns between native Japanese and Australian learners.

Looking at these results and observing the declining patterns of naturalness in the graphs of F0, dur, and sp_ap, we can say that, depending on the non-native accent, different acoustic parameters have different degrees of impact on subjective naturalness, and the degrees are different in the cases of native Japanese speakers and Australian learners. In the cases of English accents assessed by native Japanese speakers, F0, dur, and sp_ap, there is a similar decrease in subjective naturalness. In the cases of assessed by Australian learners, however, sp_ap decreases the naturalness the most drastically, and dur and F0 follows. In the cases of Korean, the results show a similar tendency.

For native Japanese speakers, F0 pattern changes have a great effect on the naturalness. For the Australian learners, however, they had a very small effect. This may be caused by the fact that learners had not much time devoted to study about Japanese pitch accents systematically, even though Japanese is a pitch accent language. It is known that many Japanese language textbooks do not have pitch accent markers. It will be necessary for learners to have more time in which to study Japanese pitch accents systematically and be taught to be aware of the unnaturalness of F0 patterns which is perceived by Japanese speakers.

For F0_dur and all, i.e. combination of multiple parameters, the results can be interpreted as the mixtures of the effects of each single parameter.

3.3. Pedagogical suggestions

The study revealed that learners were able to detect durational errors fairly accurately, though not as accurately as native Japanese, but not pitch errors at all. In general, the deviation of pitch is harder for learners to detect than that of timing, and untrained listeners often cannot tell whether pitch was raised or not [13]. In addition, many Japanese language textbooks do not have pitch accent markers on the new vocabulary and leave the learner to his/her own devices for the acquisition of the Japanese pitch accent despite the fact that Japanese is a pitch accent language. As a result, learners learn the correct pitch pattern by chance, through audio materials or native speaker friends. In order to encourage learners to communicate in the target language, language instructors tend to tolerate pitch errors as a minor issue when their focus is on grammar or contents. When learners have a face to face conversation with an instructor or native speaker, they should not be interrupted for each pronunciation error. However, the pitch pattern should be taught more clearly when a new word is introduced. One of the ideal tools for Japanese language classrooms would be a computer program which can point out learners' pitch error in a manner of a challenging quiz or entertaining game by measuring the pitch contour of their utterance. A visual representation of pitch will be particularly helpful when the learner does not know whether the pitch was raised or lowered in his/her utterance. Drawing learners' attention to a serious pitch error is the first step. It is reported that learners enjoyed using computer programs that can detect the segmental errors of their pronunciation [14]. The development of the program for pitch error detection is awaited in language classrooms.

4. Conclusions

In this paper, we investigated what kinds and degrees of foreign accents in Japanese utterances do and do not affect their naturalness perceived by native Japanese speakers and by Australian learners of Japanese, and their difference between the two listener groups. By using STRAIGHT morphing, we prepared speech stimuli with various kinds and degrees of Amer-

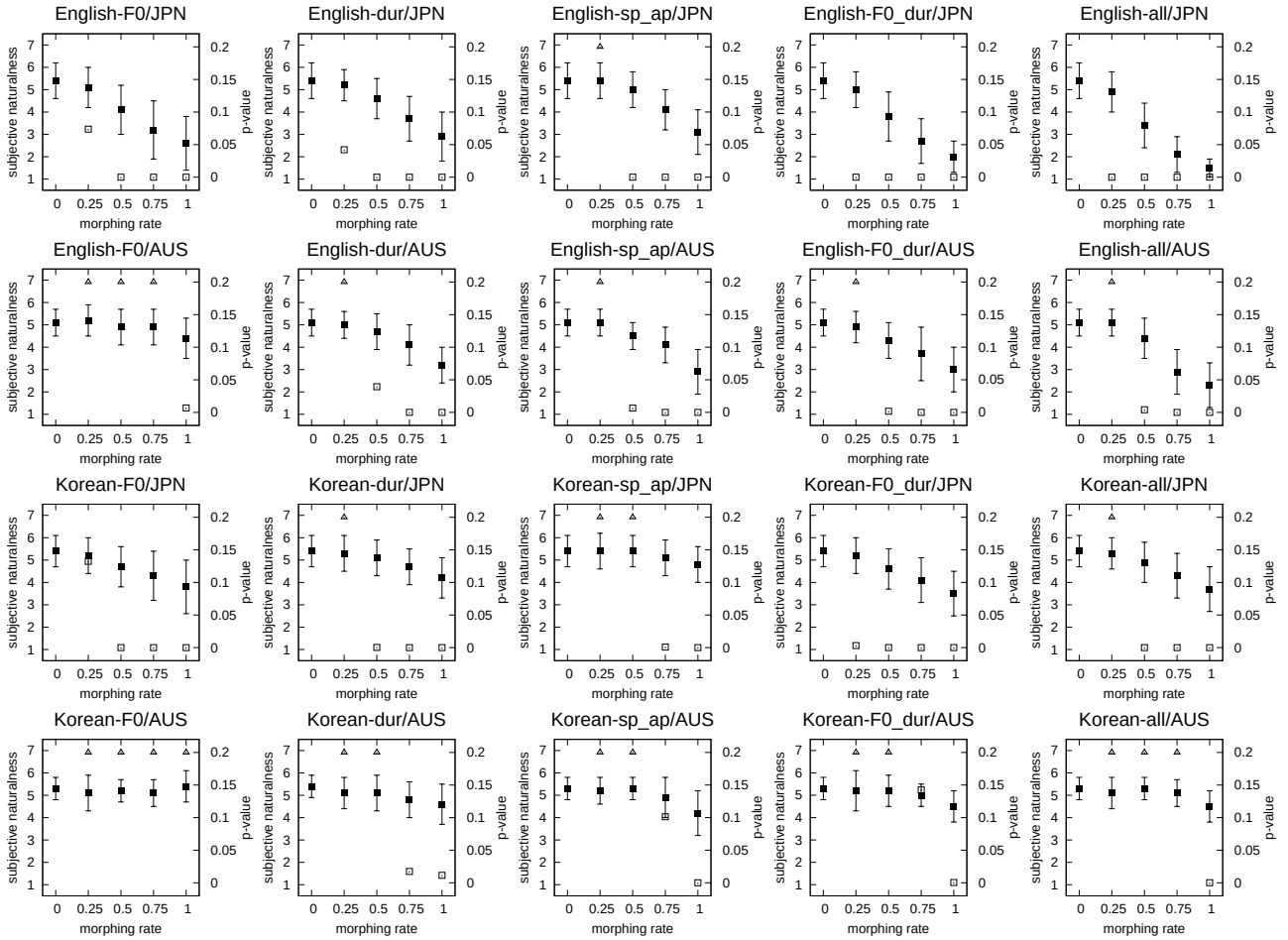

Figure 2: *Subjective naturalness for each accented Japanese and each acoustic parameter as a function of morphing rate.* ■ *: subjective naturalness;* □ *: p-value (*△ *means that the p-value* > 0.2*).*

ican Japanese and Korean Japanese. These stimuli were presented to both native Japanese listeners and Australian listeners, who were asked to judge the naturalness of the utterances as Tokyo Japanese. The experimental results showed that the degree to which language transfer affects naturalness varies from acoustic parameter to acoustic parameter and from language to language, and the degree was different between native Japanese and Australian learners. Specifically, Australian learners were insensitive to F0 pattern changes. Reasons why the Australian learners were unable to judge the naturalness of certain acoustic features with the same precision of natives was then discussed. Lastly, some ideas for how this research can be applied to the classroom were touched on.

In future work, we're planning to use learners of Japanese except Australian, such as Chinese or Korean. We're also interested in using more bilinguals because the results obtained in this paper may be dependent on the two bilingual speakers.

5. References

[1] T. Shibata *et al.*, "Prosody Acquisition by Japanese learners," In Zhaohong Han (Ed.), *Understanding Second Language Process*, Multilingual Matters, 2007.

[2] C. Nakamura *et al.*, *Japanese pronunciation training for advanced presentation*, Hitsuji Shobo, 2009.

[3] J. Bernstein, "Objective measurement of intelligibility," *Proc. ICPhS*, 1581–1584, 2003.

[4] N. Minematsu *et al.*, "Measurement of objective intelligibility of Japanese accented English using ERJ database," *Proc. INTERSPEECH*, 2011 (submitted).

[5] C. Tsurutani, "Foreign accent matters most when timing is wrong," *Proc. INTERSPEECH*, 1854–1857, 2010.

[6] H. Kawahara *et al.*, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Communication*, 27(3–4), 187–207, 1999.

[7] H. Kawahara *et al.*, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," *Proc. ICASSP*, 256–259, 2003.

[8] R. Kubo *et al.*, "/r/-/l/ perception training using synthetic speech generated by STRAIGHT algorithm," *Proc. Spring Meeting of Acoust. Soc. Japan*, 1-8-22, 383–384, 1998 (in Japanese).

[9] S. Kato *et al.*, "Perceptual study on the effects of language transfer on the naturalness of Japanese prosody for isolated words," *IEICE Technical Report*, SP2010-118, pp.19–24, 2011 (in Japanese)

[10] *NHK new Japanese accent dictionary*, NHK Publishing, 1998.

[11] The national institute of Korean language, *Foreign Language Notation*, http://www.korean.go.kr/09_new/dic/rule/rule_foreign_0104.jsp (in Korean)

[12] The Japan foundation & Japan educational exchange and services, *Japanese-Language Proficiency Test*, http://www.jlpt.jp/e/index.html

[13] C. Tsurutani, *Pronunciation and rhythm of Japanese as a second language*, Keisuisha, 2008 (in Japanese).

[14] C. Tsurutani, "A computer program for pronunciation training and assessment in Japanese language classrooms — experimental use of "*Pronunciation check*," *Journal of Japan Studies Association of Australia*, 305–315, 2008.