Effects of Learners' Language Transfer on Native Listeners' Evaluation of the Prosodic Naturalness of Japanese Words

Shuhei Kato, Greg Short, Nobuaki Minematsu, Keikichi Hirose

Graduate School of Information Science and Technology, The University of Tokyo, Japan
{kato,short,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract

For language learners it can be very difficult to speak without a non-native accent. This is due to the phenomenon called language transfer. At times, this can decrease the intelligibility of the speech and make it difficult to convey the correct message to the listener. Thus, many learners have a desire to speak more naturally in order not to be misjudged due to their pronunciation. For this, learners should be provided with some knowledge on what kinds of accents cause greater loss in naturalness when native speakers hear learners. In this paper, word utterances with various kinds and degrees of foreign accentedness were synthesized using a technique to morph the prosodic aspect of an utterance. These variously accented utterances were presented to native speakers, who were asked to judge the naturalness. This paper describes what degrees of prosodic morphing from native Japanese to American Japanese, Chinese Japanese, and Korean Japanese do and do not affect the naturalness perceived by Japanese and what kinds of words produced the least loss in naturalness for the non-native speakers.

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

1. Introduction

For a language learner to achieve good, natural pronunciation, it is important to become aware of differences in vocalization and acoustics between the learner's native language and the target language. One aspect that language transfer can greatly affect is prosody. This aspect can be especially difficult for learners to correct [1, 2], so it is necessary to build awareness of the prosodic differences between the two languages. In college education for the Japanese language, however, due to time limitation, 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 language learners of Japanese to have little knowledge of the pitch accent. They are also unaware that the pitch accent will change when words are concatenated 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 [3].

As language transfer results in unnatural-sounding speech, it is necessary to examine what kinds of accents have the greatest impact in loss of naturalness. To understand this, it is necessary to have native speakers subjectively assess the naturalness of speech samples with various kinds and degrees of foreign accents. For example, in [4, 5, 6], intelligibility evaluations were conducted by having native speakers of American English listen to English utterances read by non-native speakers [4] and by Japanese speakers [5, 6]. In [7, 8], listening experiments were carried out by having native speakers of Japanese assess the naturalness of Japanese utterances spoken by Chinese university students [7] and by Australian university students [8].

In these researches, utterances read out or spoken by learners were used themselves as stimuli for the listening experiments. In this method of experimentation, however, it is difficult to cover all possible kinds of foreign accents that might result due to language transfer. Even if researchers focus only on one kind of accent, it will be difficult to collect speech samples affected by that accent in different degrees. What we need is a method of preparing speech samples affected by different kinds of accents even in different degrees.

One possible solution is using a speech morpher. In recent years, analysis-resynthesis technologies have made rapid progress and they can morph a speech sample in many ways. A very well-known speech morpher, STRAIGHT [9, 10], has been used in many applications such as transformation of age, gender, emotion, accent, etc. In [11], between two utterances of a minimal pair of words such as right and light, spoken by a single speaker, STRAIGHT could generate intermediate utterances between the two. Quantitative interpolation at different degrees is such as a morphing of the word "right" to that of "light." This morphing will produce phones that fall in between [r] and [l].

In this paper, speech stimuli with various kinds and degrees of accents were generated from a small number of utterances through the use of morphing with STRAIGHT. These stimuli were then used for listening experiments, in which the naturalness was assessed by native listeners. Using STRAIGHT, we can morph speech quantitatively by selected acoustic parameters and deal with foreign accents better in an acoustic sense.

2. Experiment of assessing the naturalness

2.1. Recording of bilinguals' utterances

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

The words were selected 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 Japanese and in a mimicking of foreign accented Japanese in the other language he/she is proficient in. Like [11], we prepared word pairs spoken by the same speaker but unlike [11], one is native sounding and the other is non-native sounding. This somewhat unnatural preparation of utterance pairs is because

Table 1: *The three 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
M	early 20's	Chinese	0–10 years of age in Shanghai, China He attended local schools.	Hiragana
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

within-speaker morphing gives us speech samples of higher quality than cross-speaker morphing. For the native reading of the word 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 [12]. 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 Romaji), the Latin Alphabet with language transfer considered in the notation, or in Hangul (as the standard notation [13]). 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 used in the listening experiment were generated by morphing the Tokyo Dialect version of the word and the accented version together from each speaker. This operation was done by using STRAIGHT. As mentioned in Section 1, by using STRAIGHT, it is possible to select which parameters 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 the five morphing rates (0, 0.25, 0.5, 0.75, 1) were used. For example, "F0 at morphing rate 0.25" means a stimulus is morphed in F0 at morphing rate 0.25 and the other parameters at morphing rate 0. Morphing rate 0 indicates that the sample is spoken in Tokyo Japanese. A morphing rate of 1 means it is morphed completely into the accented version of the word.

Lastly, the number of stimuli for each word was 63 (=(1+4 rates × 5 parameters) × 3 languages). Since the size of the balanced word set is 162, the total number of word utterances for the listening experiment became 10,206.

2.3. Subsets of stimuli

Since it was practically impossible for a subject to assess the naturalness of all the 10,206 stimuli, they were divided adequately into four subsets with respect to heavy syllables. Subset 1 was composed of only words without heavy syllables and subsets 3 and 4 are of only words with heavy syllables. Subset 2 is a mixture of two kinds of words. As for accent type and word length, these subsets were made so that they had unbiased distribution for these two parameters. The numbers of words were 34, 34, 44, and 50 for subsets 1, 2, 3, and 4, respectively. By morphing these words, we obtained the stimuli. The numbers were 1,734, 2,550, 2,772, and 3,150 for the four subsets.

[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.

2.4. Task

Two kinds of tasks were assigned to subjects. The first was to assess the naturalness of the stimuli as Japanese (Tokyo Japanese) using the Likert scale with potential responses ranging from 1 (extremely unnatural) to 7 (extremely natural = native-like). The stimuli were presented in random order. The second task was to check whether a given stimulus was highly synthetic (extremely non-humanlike) due to the synthesis process in STRAIGHT, which outputs non-humanlike sounds very occasionally. Although all the subjects were informed that the stimuli were artificially synthesized, if a stimulus was judged as highly synthetic, judgement of that stimulus on nativeness would not be reliable. In the discussion of the following section, we ignored those stimuli that were labeled as highly synthetic (non-humanlike) by more than half subjects.

2.5. Subjects

42 native Japanese speakers ranging from 19–28 years of age, many of whom were university students, took part in the listening experiment. Subset 1 was assessed by 12 subjects and subsets 2 to 4 were assessed by 10 subjects each. For any subset, the listening experiment was carried out through several sessions and the subjects could have a break between sessions if they wanted. To complete all the sessions, it took between 3–5 hours. 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 had presented 15 sample stimuli in order for the subjects to get to understand the correspondence between seven-degree naturalness and stimuli.

3. Results and Discussion

3.1. Discussion for each accent and each parameter

3.1.1. Results of the listening experiments

The experimental results are shown in Fig. 1 with the results at the top being for English-accented, the middle for Chinese-accented, and the bottom for Korean-accented. 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 particular morphing rate gives the minimum significance level at which the subjective naturalness is different significantly from that at a morphing rate of 0 (unaccented). In other words, the smaller the p-value at a morphing

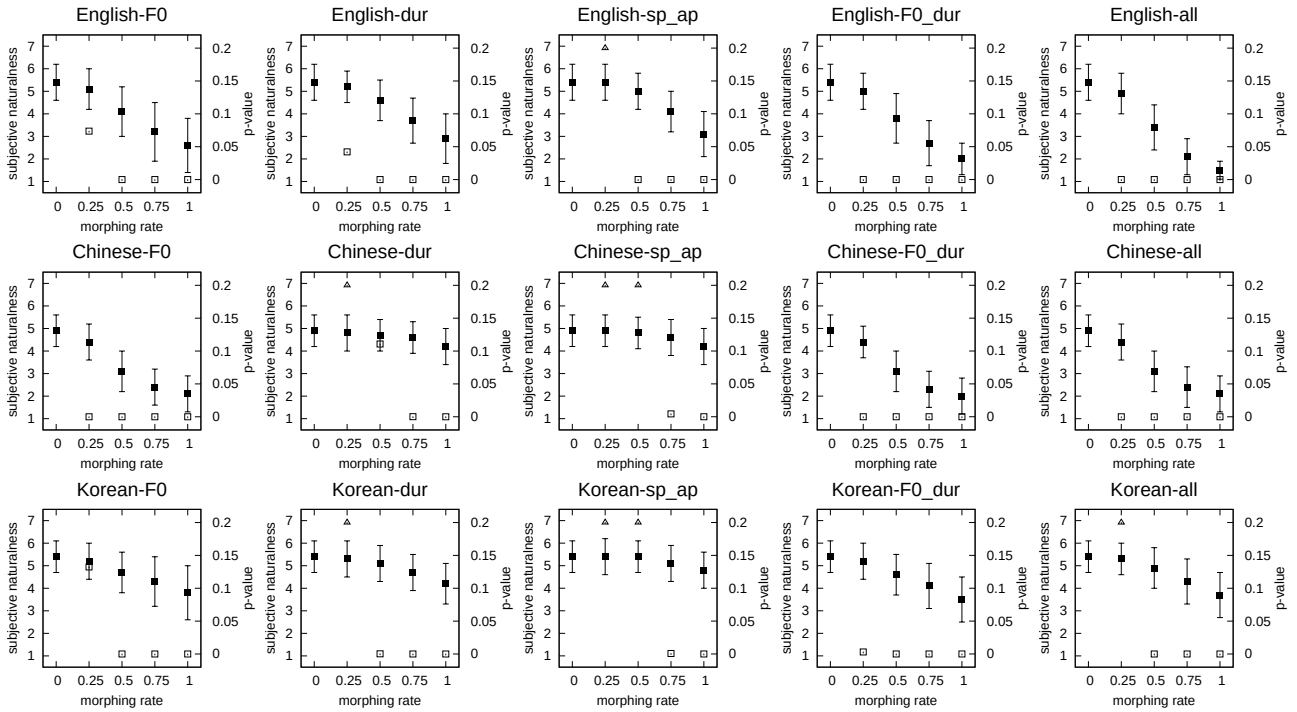

Figure 1: *Subjective naturalness for each accented Japanese and each acoustic parameter as a function of morphing rate.* ■ *: subjective naturalness;* □ *: p-value (*△ *means that the p-value* > 0.2*).*

rate is, the larger the difference in the naturalness at that morphing rate 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 ($\pm$) are shown here.

3.1.2. Discussion for each non-native accent

We inspect 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. 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 Chinese-accented stimuli, the minimum rates are 0.25 for F0 and 0.75 for 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.

Looking at these results and observing the naturalness declination patterns in the graphs of F0, dur, and sp_ap, we can say that, depending on non-native accent, different acoustic parameters have different levels of effects on the subjective naturalness. Although in the case of English-accented speech, F0, dur, and sp_ap decrease the naturalness similarly, in the case of Chinese-accented speech, only F0 decreases the naturalness drastically and the declination patterns of dur and sp_ap are very flat compared to the F0 pattern. It seems that American learners have to pay attention to every prosodic aspect of their pronunciation but Chinese learners have only to focus on their F0 patterns. In the Korean graphs, the declination pattern in sp_ap is more flat compared to the others.

For F0_dur and all, i.e. combination of multiple parameters, it is much to be expected that the minimum morphing rates are 0.25 almost for every graph and the declination pattern is clear.

3.1.3. Discussion for each acoustic parameter

To understand the naturalness of which accented Japanese is more easily affected by each acoustic parameter, we inspect the naturalness declination patterns and the p-values over the three accents for each parameter. According to the p-values in Fig. 1, F0 has the largest effect for the Chinese-accented stimuli, dur the largest for the English-accented and Korean-accented stimuli, and sp_ap for the English-accented stimuli. We consider that these findings are reasonable. Since Chinese is a tonal language, it is well-known that Chinese learners tend to have multiple phrasal and lexical tones in a word although, in Tokyo Japanese, only one pitch nucleus can be found at most in a word. Then, Chinese-sounding pitch patterns are considered very easy for Japanese to detect. Being different from Chinese, Korean, and Japanese, word accents of American English are acoustically realized by multiple factors of pitch, intensity, duration, and timbre. Thus it is to be expected that larger effects are found in the American graphs of dur and sp_ap.

For F0, F0_dur and all, the difference between the subjective naturalness scores at morphing rates of 0 and 1 for the English-accented and Chinese-accented stimuli is much larger than that for the Korean-accented stimuli. For dur and sp_ap, that of the English-accented stimuli is relatively large compared to that of the Chinese-accented and Korean-accented stimuli.

Next, we inspect phenomena specific to particular stimuli groups: words of type-1 accent and those of 2-mora length.

3.2. Discussion for concerning Japanese accent types

Japanese word accents are classified by the location of the accent nucleus in a word. In this section, we inspect differences in

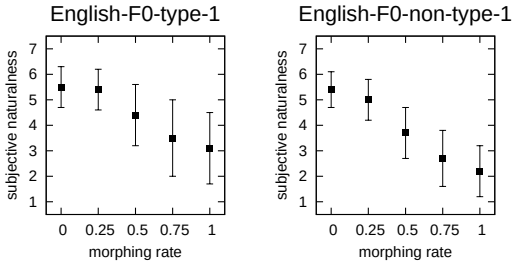

Figure 2: *Naturalness differences between type-1 words and non-type-1 words*

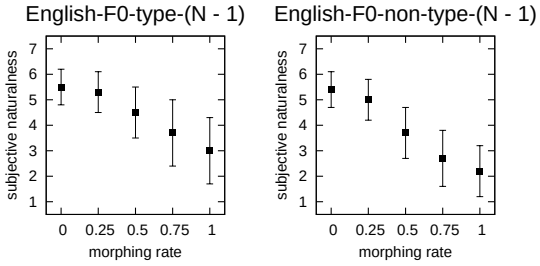

Figure 3: *Naturalness differences between type-(N 1) words and non-type-(N 1) words*

the naturalness declination patterns between words of a specific accent type and those of the other types. In the English-accented stimuli, significant differences in the naturalness are found only between words of type-1 and those of the others and between words of type-(N 1) and those of the others, where N is the number of morae of the word. Fig. 2 shows the former and Fig. 3 shows the latter. In both figures, at a morphing rate of 1, words of type-1 accent and those of type-(N 1) accent are found to be significantly more natural than those of the other accent types. We consider that this finding is due to the accent rules of American English. According to the rules, words tend to have their stressed syllable at the first syllable or the second-to-last syllable. If these rules are applied directly to Japanese words, their accent (pitch) nucleus will tend to be found at the first mora (type-1 accent) or at the second-to-last mora (type-N 1 accent) irrespectively of the original position of the pitch nucleus in the words. This is considered to be why a drop of the naturalness is not so large even for completely accented stimuli in the case of type-1 and type-(N 1) accents.

3.3. Discussion for each word length

In the Chinese-accented stimuli, 2-mora words are found to bring a very unique pattern of the naturalness declination, shown in Fig. 4. In 2-mora words, the naturalness does not drop much. As mentioned in Section 3.1.3, Chinese speakers tend to produce multiple accent and phrase tones even within a single word. This habit decreases the naturalness of their pronunciation. In the case of 2-mora words, however, the number of pitch nuclei has to be one at most even when the words are pronounced by Chinese learners. We consider that this fact is a reason for a small declination of the naturalness in the case of 2-mora words.

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 Japanese. By using STRAIGHT morphing, we prepared speech stimuli with various kinds and degrees of American Japanese, Chinese Japanese, and Korean Japanese. These stimuli were presented to Japanese listeners, who were asked to judge the naturalness 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.

For future work, we're planning to use more bilinguals because the results obtained in this paper may be dependent on the three bilingual speakers. We're also interested in investigating differences between the naturalness perceived by native speakers and that perceived by language learners and some experimental results are shown in [14].

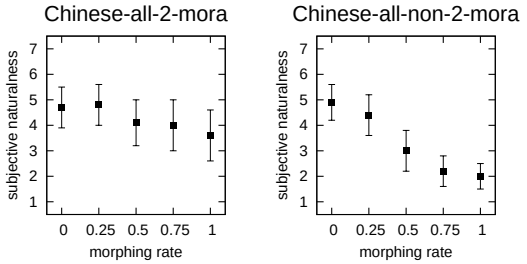

Figure 4: *Naturalness differences between 2-mora words and the others*

5. References

[1] G. Ioup *et al.*, "The acquisition of tone: a maturational perspective," In Georgette Ioup & Steven H. Weinberger (Eds.), *Interlanguage Phonology: the Acquisition of a Second Language Sound System*, Cambridge, MA: Newbury House Publishers, 1987.

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

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

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

[5] N. Minematsu *et al.*, "CART-based factor analysis of intelligibility reduction in Japanese English," *Proc. EUROSPEECH*, 2069–2072, 2003.

[6] N. Minematsu *et al.*, "Measurement of objective intelligibility of Japanese accented English using ERJ database," *Proc. INTERSPEECH*, 1481–1484, 2011.

[7] H. Hirano *et al.*, "F0 models show Chinese speakers of Japanese insert intonational boundaries and drop pitch," *Proc. INTERSPEECH*, 1885–1888, 2007.

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

[9] 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.

[10] 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.

[11] 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).

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

[13] The National Institute of Korean Language,"Foreign Language Notation" http://www.korean.go.kr/09_new/dic/rule/rule_foreign_0104.jsp

[14] S. Kato *et al.*, "Comparison of native and non-native evaluations of the naturalness of Japanese words with prosody modified through voice morphing," *Proc. SLaTE*, 2011.