Naturalness Judgement of Prosodic Variation of Japanese Utterances with Prosody Modified Stimuli

Chiharu Tsurutani[1], *Shunichi Ishihara* [2]

[1]School of Languages and Linguistics, Griffith University, Brisbane, Australia
[2]School of Culture, History and Language, Australian National University, Canberra, Australia
[1]c.tsurutani@griffith.edu.au, [2]shunichi.ishihara@anu.edu.au

Abstract

This study aims to identify the crucial prosodic factor for native speakers' naturalness judgement of L2 pronunciation. Prosodic features are known to have more impact on the naturalness of L2 learners' pronunciation than segmental features do. Among prosodic features, timing and pitch are looked at in this study as major prosodic factors which affect native speakers' naturalness judgement of L2 learners' pronunciation. To examine the relative importance of timing and pitch, we synthesized stimuli using STRAIGHT to produce well-controlled timing and pitch errors. Native Japanese listeners assessed the naturalness of these stimuli and the result was compared with the one obtained using natural speech stimuli. The current study obtained the same result as the previous study: that timing is more important than pitch in improving the naturalness of L2 speech.

Index Terms: **naturalness, morphing, prosodic features, Japanese**

1. Introduction

Language researchers have often stated the belief that prosodic errors impair the intelligibility of L2 learners' speech more than do segmental (phonemic) errors [1][6][7][8]. In general, three major prosodic features – timing, pitch and intensity – are coordinated to constitute the rhythm of languages by their phonological rules. However, the relative importance of each prosodic feature in L2 pronunciation has not been fully explored across languages. This study investigates the influence of prosodic features on native listeners' judgment of synthesized L2 Japanese utterances by English-speaking learners. These stimuli are controlled in terms of pitch and timing errors with other acoustic features being constant. Thus, using these stimuli, we can investigate how the deviation of pitch and timing affect the judgment of the naturalness of L2 speech. This is the aim of our study.

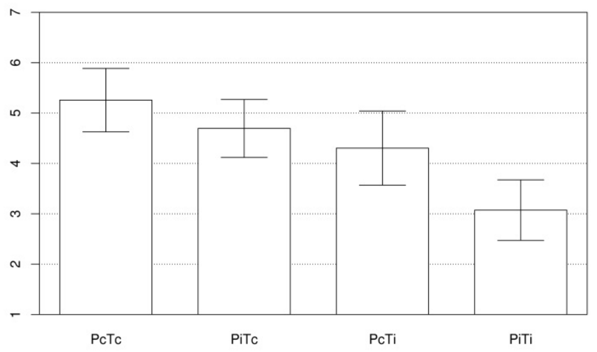
In the previous studies, natural speech was used to test the respective roles of each prosodic feature in the judgment of naturalness [12] [13] [14]. While natural speech is ideal for the studies of L2 speech as it contains genuine errors, researchers could not accurately control the amount and number of errors in each stimuli. Even in the studies which used synthesized speech, it was not easy to include adequate amounts of incorrectness in the original speech [4][8] depending on the proficiency level of the learner informant or the experience of a bilingual speaker who can mimic possible errors and record the original speech. In this study the researcher, who is a phonologist and experienced language teacher, thoroughly studied possible errors in the prepared stimuli, and recorded original speech for speech synthesis through morphing.

As our focus was on identifying the importance of timing and pitch, timing and pitch errors in the original speech were separately manipulated. To create stimuli which met our needs, we used a speech morpher, STRAIGHT, and controlled the proportion of pitch and timing errors in the stimuli. The following section discusses the background of the study, the experiment, the result and comparison with the outcome of the experiment which used natural speech.

2. Background

In Japanese, pitch-accent is a lexical property of a word, and is not affected by the prosodic organization at the higher prosodic level. Japanese intonation is formed by incorporating the accent patterns of words into phrase pitch pattern [11]. Because Japanese is a pitch-accent language, intonation, which is a feature of phrases or sentences expressed by pitch, is expected to have a major influence on the acceptability of Japanese utterance. At word level, it has been reported that pitch is the most dominant cue to accent patterns in the prosodic features, such as pitch, duration, intensity and spectral coefficient patterns [2]. In other words, pitch is the most significant feature to determine the word accent for Japanese (and also for English). However, the influence of prosodic features on overall performance needs to be tested above word level. Tajima et al. [9] found the importance of timing in the improvement of English phrases spoken by Chinese-speaking learners of English. Cheng [3], who investigated read-aloud English passages spoken by 126 L2 speakers with different L1 backgrounds, also reported that duration information was the best predictor of human prosody ratings, while pitch and energy contours were also strong predictors. Maier et al. [5] used time alignment information from the structure of speech data of German and Japanese to evaluate the prosody, and they obtained a high correlation between the score by native speakers of both languages and the computer program they created. It can be predicted from the results of the previous studies that timing is the most important factor for assessment of L2 speech in many languages in the world.

There is also a study of L2 Japanese speech to test the importance of prosodic features. Sato [8] reported the predominance of prosodic factors over segmental factors, and that pitch was most influential among prosodic factors, such as pitch, timing and intensity for native listeners in assessing L2 Japanese speech. In his experiment, however, the original speech samples for synthesis were recorded by Korean- and Chinese-speaking learners whose level of Japanese was intermediate. Due to their proficiency level, their speech did not contain enough timing errors compared with pitch errors. It is obvious that the L1 background and L2 proficiency of learners affect the appropriateness of L2 speech as an original speech for morphing. Tsurutani [12] examined Japanese native listeners' perceptions of L2 speech, using English-speaking learners' natural speech and found that correctness in timing was more important than that in pitch for the L2 speech to be perceived as more natural by Japanese native listeners. She used four types of natural speech samples which were, respectively, correct in both pitch and timing (PcTc), correct in timing only (PiTc), correct in pitch only (PcTi), and incorrect in both pitch and timing (PiTi). The scores received for samples that were incorrect in timing (PcTi) were worse than those for samples that were incorrect in pitch (PiTc), indicating that correctness in timing is more important than that in pitch for the naturalness. Her result is shown in Figure 1 below.



SD 0.75 0.7 0.65 0.67

Figure 1: *The average score (+/- standard deviation) for four types of errors (7=Native like, 1 = Non-native like).* [12]

Tsurutani's finding which is based on natural speech needs to be confirmed, using thoroughly controlled synthesized speech. Thus, in the current study focusing on pitch and timing, experimental stimuli were synthesized by morphing two sets of Japanese utterances: one is the utterances spoken by a native Japanese speaker and the other is the English-accented utterances spoken by the same speaker, mimicking the typical pronunciation of English speakers when they speak Japanese. This speaker is a Japanese-English bilingual speaker, being able to manipulate each language equally. With these stimuli, the influence of pitch and timing errors on native listeners' judgement will be examined in order to determine their relative importance in Japanese.

3. Experiment

The STRAIGHT (http://www.wakayama-u.ac.jp/~kawahara/PSSws/), a speech manipulation and morphing system, was employed to control different prosodic variables and produce speech samples of high quality. Five stimuli sentences were created to cover difficult phonemes for L2 English learners of Japanese. The length of sentences was kept between 13-14 mora as the naturalness of the synthesized stimuli by the STRAIGHT will decrease as an original sentence gets longer. Participants were asked to judge the naturalness of utterances using a Likert scale (1 to 7).

3.1. Materials

Japanese sentences which contain problems in pitch and timing for English speaking learners of Japanese were created. Segmental errors will not be examined in this study, however, phonemes which could induce prosodic errors (e.g. diphthongization, consonant cluster reduction or rhoticization affects duration) were included in the stimuli sentence. In order to synthesize perceptually acceptable stimuli by morphing, a near bilingual speaker is required who can utter the speech materials with perfect model pronunciation, and also with absolute beginner's pronunciation containing all required errors. This requirement is due to the nature of the STRAIGHT. Synthesizing two different voices will increase the unnaturalness of the resulting synthesized speech. The first author performed this task.

Out of 10 sentences the researcher produced for the L2 speakers' model, five sentences which Japanese native listeners could not identify clearly were chosen as stimuli. The stimuli sentences, its length in terms of mora and the pitch pattern are given in Table 1. Table 2 presents the errors that were included in the sentences.

Table 1: *Stimulus sentences*

Sentences	No. of mora	Pitch pattern
Sen-en kuretara kaemasuyo. If you give me 1000en, I can buy it.	12	HHHH LHHL LHHLH
Tsumaranai hon o kattekita. I bought a boring book.	13	LHHLL HLL LHHHL
Watashino iega miemasuka. Can you see my house?	12	LHHH LHL LHHLH?
Gaikoku ni ryokoo shimashoo. Let's travel overseas.	12	LHHHH LHH LHHL
Daigakuno sotsugyoo shashin desu. This is my graduation photo of university.	14	LHHHH LHHH HLL LL

Table 2: *Errors included in the stimuli*

*R= a flap was replaced with a central approximant,
EI, AI = English diphthongs, gIO = "gyo" was simplified

Timing errors (in shadow*)- Number of timing errors	Pitch pattern errors- Number of pitch errors in the utterance
Senen kuRetaRa kaemasyo. -3	HHH HLLL HHLLH -2
TsumaranAI hono katekita.-2	LLLHL HL HLLL -2
Wataashino IEga mIImasuka. -1	LHLL HLL HHLLH -3
GAIkokku ni Rookoo shiimashIO. -3	LLHLL HHLL HLLL -3
DEIgakkuno sotsugIo shaashin desu.-2	LLHLL HHLL HHLL LL -2

The number and amount of errors included in the non-native speaker model required careful consideration. First of all, it is difficult to decide how many and what type of errors should be contained in the original speech. If we use natural utterances by

L2 learners, sufficient errors of certain type might not be found in the original speech, depending on the level of learners. For this study, the number of errors in the original speech was at our discretion as the researcher produces the erroneous utterances. However, making errors deliberately is not an easy task. The researcher needed the production of an English native speaker as guidance, who does not know Japanese. In addition, making errors on every single word was impossible and unrealistic, and yet we needed a reasonable number of both pitch and timing errors. Efforts were made to change a flat pitch pattern to a HL pitch pattern which is a common error observed in L2 English learners' production [10] and to make durational errors between long and short vowels and consonants. Durational errors at segmental level and at higher levels of linguistic description could not be separated since both would be judged as temporal deviation by native listeners. Incorrect flap sounds, English-sounding vowels and diphthongization of two vowels were also added to make the stimuli sound authentic. The researchers' performance achieved a reasonable level of non-nativeness. The participants did not realize that all the speech samples were recorded by a Japanese native speaker until the researcher revealed it after the task.

3-2.Participants

A total of 45 Japanese native listeners who are university students were recruited and participated in the task for a small payment at the authors' home institution.

3-3.Procedure

3-3-1. Constructing Stimuli

First, the original speech samples were separately decomposed into three independent acoustic parameters of segment, pitch and duration. With the segmental parameters being constant, only the pitch and duration features of the native Japanese samples were morphed with the corresponding pitch and duration features of the non-native samples with different proportions (0%, 50% and 100%). This was done using STRAIGHT.

All possible combinations of the morphing features with respect to their percentages are listed in Table 3. For example, stimulus p000t000 means the pitch and timing of the original native Japanese samples were morphed with those of the non-native speaker with the rate of 0%, which means p000t000 is identical to the original native speaker Japanese sample. Stimuli p050t100 is a morphed sample with 50% pitch and 100% timing of the non-native sample.

Table 3: *The possibly permutations of the morphing parameters.*

Native			Non-native		Stimuli
segment	pitch	timing	pitch	timing	type
100%	100%	100%	0%	0%	p000t000
100%	100%	0%	0%	100%	p000t100
100%	0%	100%	100%	0%	p100t000
100%	0%	0%	100%	100%	p100t100
100%	50%	0%	50%	100%	p050t100
100%	0%	50%	100%	50%	p100t050
100%	50%	50%	50%	50%	p050t050

3-3-2. Listening task

A listening task was conducted online, using the following instruction.

"You will listen to short Japanese sentences recorded by one female speaker. Some of them are natural and some of them are foreign-accented. Evaluate the naturalness of utterance using a Likert scale with potential responses ranging from 1 (Non-native like) to 7 (Native like). There is no right and wrong answer. Don't have to listen to the same speech sample more than twice. Just follow your intuition as a native listener."

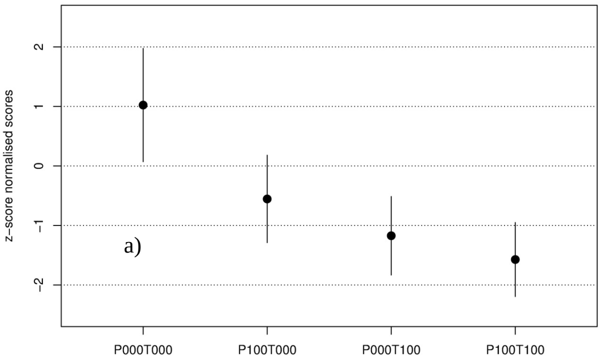
It took around 15 minutes for the participants to complete the task including the practice session with four sentences. The order of 35 sentences was automatically randomised in order to avoid order of presentation effect.

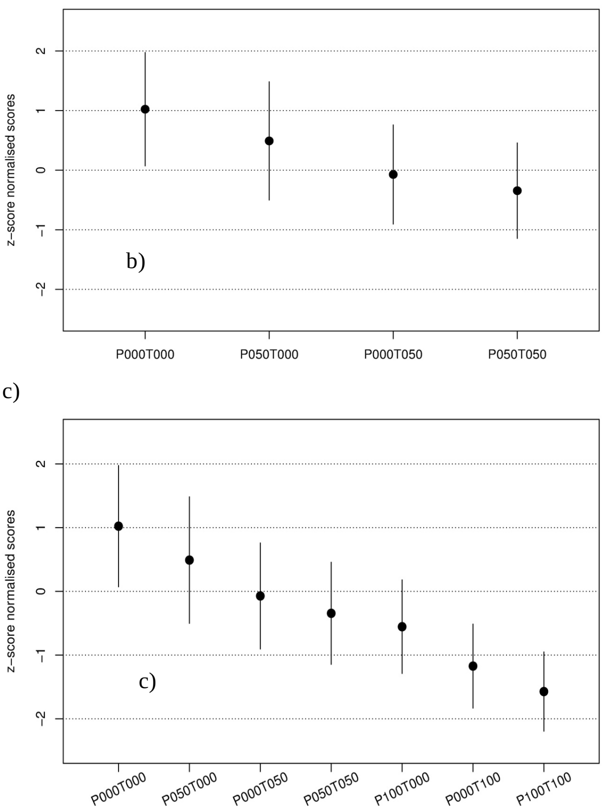
4. Results

All scores were z-score normalised in order to factor out the between-rater differences as different raters used different criteria to assess the stimuli. The normalised scores were then pooled together separately according to the stimuli types. The mean normalised score for each stimulus type is plotted in Figure 2 together with one standard deviation above and below the mean. The numerical values of Figure 2 are given in Table 4. For Figure 2a), the ranking of scores, p000t000 > p100t000 > p000t100 > p100t100 was statistically confirmed by the Tukey HSD test (p < 0.01). It is not surprising that the stimulus which has correct prosodic features (p000t000) received the highest score while the stimulus which has 100% incorrect prosodic features (p100t100) had the lowest score. The important point is that there is a statistically significant difference between p100t000 and p000t100 (p100t000 > p000t100). The result indicates that the native Japanese speakers put more weight on accuracy in timing than in pitch when judging the naturalness of speech, which confirms the findings in Tsurutani [12] which used natural speech as stimuli.

In Figure 2b), the naturalness ranking of p050t000 > p000t050 was statistically confirmed by the Turkey HSD test (p < 0.0001) as well. However, in Figure 2b, the difference between p000t050 and p050t050 was not statistically significant (p000t050 = p050t050). This indicates that the addition of pitch errors to the stimuli with timing errors does not significantly decrease the degree of naturalness. That is, the impact of pitch error is not as extensive as that of timing error on the judgement of the naturalness.

In Figure 2c), the results of all stimulus types are given. Statistically, all the stimulus types have the ranking order of p000t000 > p050t000 > p000t050 = p050t050 = p100t000 > p000t100 > p100t100. The equality of p000t050 = p050t050 = p100t000 shows that the addition of pitch error does not influence the raters' judgement extensively, which indicates the correctness in pitch does not play a crucial role as much as the correctness in timing.





c)

Figure 2: *Results of the perception experiments. Panel (a) is for the stimuli with 100% morphing in pitch and timing; (b) for the stimuli with 50% morphing; and (c) for all stimuli types together.*

Table 4: *Numerical information of Figure 2*

Type	Mean	sd
p000t000	2.02	1.90
p050t000	1.49	1.98
p000t050	0.92	1.66
p050t050	0.65	1.60
p100t000	0.44	1.46
p000t100	-0.17	1.31
p100t100	-0.57	1.24

5. Conclusion

Among prosodic features, it was confirmed that timing was the most crucial factor for naturalness judgement by Japanese native listeners. This point has to be taken into consideration when CAPTA (Computer Assisted Pronunciation Training and Assessment) programs are developed. This study was conducted on English-accented Japanese, however, the importance of timing over pitch could be found in languages other than Japanese. It is considered that native listeners are more sensitive to timing than pitch [11]. People can easily and accurately tell whether a sound in their native tongue was short or long, but cannot as easily determine pitch. The significance of timing in naturalness judgement needs to be tested in other languages as well.

Since each component of spoken language can be separately tested and evaluated, rather than scoring prosodic performance of L2 learners as a whole, it might be desirable to test one prosodic feature at a time. It is quite often the case that the improvement of one phonetic feature reflects the progress of the learner's overall language proficiency. The correctness of timing has a great potential as an assessment scale for L2 speakers' proficiency level.

6. Acknowledgements

This study was supported by Griffith University Research Grant 2011.

7. References

[1] Anderson-Hsieh, J., R. Johnson and K. Koehler (1992). "The relationship between native speakers' judgements of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure," *Language Learning* 42(4), 529-555.

[2] Beckman, M. and Pierrehumbert, J. (1986). "Intonational structure in Japanese and English", *Phonology Yearbook.* 155-300.

[3] Cheng, J. (2011). Automatic Assessment of prosody in high-stakes English tests *INTERSPEECH* 2011, 1589-1592

[4] Kato, S., Short, G., Minematsu, N., Tsurutani, C., and Hirose, K. (2011) "Comparison of native and non-native evaluations of the naturalness of Japanese words with prosody modified through voice morphing" *SLATE 2011*

[5] Maier, A., Honig,F.,Zeissler, V,, Batliner, A. ,Korner,E., Yamanaka, N.,Ackermaqnn, P. and North, E. (2009). A Language-independent feature set for the automatic evaluation of prosody. *Interspeech 2009*, 600-603.

[6] Mouri, T., K. Hirose, and N. Minematsu (2003). "Consideration on vowel durational modification for Japanese CALL system," *Proceeding. EUROSPEECH*, 3153-3156.

[7] Munro, S. and Derwing, T. (1997) Accent, Intelligibility and Comperehensibility, *Studies in Second Language Acquisition*, 19, 1-16.

[8] Sato, T. (1995). "Comparison of the influence of segmental information and prosody on the assessment of Japanese pronunciation," *Sekai no Nihongo kyoiku* 5, 139-154.

[9] Tajima, K. Port, R. and Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English *Journal of Phonetics*, 25, 1-24.

[10] Tsurutani, C. (2008). *Pronunciation and Rhythm of Japanese as a second language* Hiroshima, Keisuisha.

[11] Tsurutani, C. (2009) Intonation of Japanese sentences spoken by English speakers *INTERSPEECH 2009*, 692-695.

[12] Tsurutani, C. (2010). Foreign accent matters most when timing is wrong, *Interspeech 2010* 1854-57.

[13] Ishihara, S., Tsurutani, C. & Tsukada K. (2011). What Constitutes 'Good Pronunciation' from L2 Japanese Learners' and Native Speakers' Perspectives? A Perception Study. Electronic *Journal of Foreign Language Teaching*, 8 (S1). 277–290.

[14] Tsurutani, C., Tsukada, K. and Ishihara, S. (2010). Comparison of Native and Non-native Perception of L2 Japanese Speech Varying in Prosodic Characteristics. In *Proceedings of 2010 Australasian International Conference of Speech Science and Technology* 122-125.