

Comparison of Native and Non-native Perception of L2 Japanese Speech Varying in Prosodic Characteristics

Chiharu Tsurutani¹, Kimiko Tsukada², Shunichi Ishihara³

¹School of Languages and Linguistics, Griffith University, Australia

²Department of International Studies, Macquarie University, Australia

³School of Culture, History and Language, Australian National University, Australia

c.tsurutani@griffith.edu.au, kimiko.tsukada@mq.edu.au, shunichi.ishihara@anu.edu.au

Abstract

This study investigates non-native listeners' perception of prosodic variation of Japanese utterances. A previous study [12] reports that there is no significant difference in the assessment of the accented speech by two groups of native Japanese-speaking judges, i.e., between those who have been exposed to non-native speech and those who have not, in that both of them relied on the correctness of timing for their judgement. The same second language (L2) Japanese utterances as those used in [12] were presented to two groups of learners (beginners and advanced) whose L1 is English. The listeners were asked to assess the pronunciation of those utterances using a 7-point scale. It was found that beginners and advanced learners use different criteria for their judgement from native speakers depending on the error types in the speech samples.

Index Terms: perception, L2 speech, non-native listeners, Japanese

1. Introduction

The perception criteria of native listeners are fairly stable and consistent since the listener and speaker share the phonology in their internal perceptual system. Thus, they can tell whether the speaker is native or non-native upon hearing just one word or phrase [3]. Native listeners must rely on some salient cues in speech for their consistent judgment. It has been shown that native judges weigh prosody (e.g., pitch, length and intensity) more heavily than segmental features in their evaluation of non-native speech [8]. Some researchers suggest that the perception of native listeners is sensitive to temporal distortion [10, 12] which could differ depending on the language. In addition, native listeners' assessment of accented speech is not influenced substantially by whether they are exposed to L2 speech or not [13]. Tsurutani [12] investigated the perception of 80 Japanese university students who reside in Japan and have not had regular contact with foreigners, using L2 Japanese utterances as the stimuli of the research. Students' perception did not differ from that of experienced Japanese language teachers who live overseas. Japanese native listeners' perceptual criteria are expected to be deeply based on mora timing, which requires equal length for each mora (= sub-syllabic units mainly formed by one consonant and one vowel).

Japanese is also a pitch-accent language and some words have a different meaning depending on their pitch pattern. In addition, regional and generational differences [6] exist in pitch-accent patterns (e.g. kurabu – LHH = a night club, HLL = other group activities; a newly created distinction by young

people). Thus, the significance of pitch in assessment has been questioned [12].

As for non-native listeners' perception, sociolinguistic research [2, 7] reported that non-native listeners' judgments of accented speech were harsher than those of native listeners. However, the researchers do not provide the exact phonetic factors for the judgement other than participants' attitude or view. Through exposure to and practice of the target language, L2 learners' phonology will approximate to (but may not exactly match) the sound system of the target language, and the ability to detect the goodness of pronunciation will develop as learners establish prototypes for phones and prosody in L2. To date, studies of non-native listeners' perception of L2 speech are scarce, compared with those of native listeners. In fact, we do not know exactly on what basis L2 learners judge the goodness/naturalness of pronunciation, and how their judgment changes along their learning path. This leads us to the following questions with regard to L2 perception:

- Can L2 learners acquire criteria similar to those of native listeners or is their judgment more lenient or harsher towards accented speech?
- Is advanced learners' judgment closer to native listeners' than to beginners'?
- Do L2 listeners also use non-native speakers' timing errors for their judgment?

In this study, the perception of English-speaking learners of Japanese from two different proficiency levels is examined and compared with that of native listeners.

2. Experiment

2.1. Materials

The stimuli were extracted from a recording of Australian English speakers who had been studying Japanese for 160 hours at university. The recording was taken from a computer exercise that was developed as a self-assessment of pronunciation. Four types of sentences were extracted from a large sample of speech data as follows:

1. correct pitch, correct timing -- O
2. incorrect pitch, correct timing -- Px
3. correct pitch, incorrect timing -- Tx
4. incorrect pitch, incorrect timing -- X

These four types of utterances did not have any obvious segmental errors. The judgment of errors was made by the first author and two other native speakers who had been teaching Japanese for many years. When two out of three listeners agreed, the judgment was accepted. In 80 native listeners' judgments, the four patterns were clearly distinguished in the order of 1 > 2 > 3 > 4 from the best to the worst [12].

Six different sentences were selected for each of the four utterance types. Thus, the list of stimuli used for this study contained 24 utterances (= 6 sentences x 4 utterance types). They were presented in 6 blocks (one per sentence) within which the order of error patterns (1-4 listed above) was randomized (cf. Table 1). Since the stimuli were natural utterances, it was not possible to control the degree of incorrectness as accurately as in synthesized speech. Efforts were made to select utterances with similar speech rate and number of errors. Errors at both word and phrase levels were counted inclusively, but a slight phrase final pitch raising was not included in error counting for correct samples, as it can be regarded as natural speech behavior [7]. None of the utterances had pauses longer than 300 ms and all sounded reasonably fluent. The following is a list of the sentences and the description of speakers' errors is presented in Table 1.

- 1) *Shachoo-no kekkonshiki-ni okyakusan-ga sennin kita.*
(1000 people attended the president's wedding reception.)
- 2) *Tsugi-no jugyoo-no suugaku-wa chotto muzukashii desu.*
(Mathematics in the next class is a bit hard.)
- 3) *Watashi-no kookoo-de isshoni shashin-o torimashoo.*
(Let's take a photo together at my high school.)
- 4) *Otooto-no okusan-wa ryokoo-ni ikuno-ga suki-desu-yo.*
(My younger brother's wife likes traveling.)
- 5) *Tanjoobi-ni tomodachi-kara kireena hana-o moratta.*
(I received beautiful flowers from my friend on my birthday.)
- 6) *Shuumatsu-kara futarino hito-to shigoto-o suru yotee desu.*
(From the weekend I'm planning to work with 2 people.)

In order to provide some basic durational characteristics of the stimuli presented to the listeners, the underlined vowels (i.e., /u-u:/, /o-o:/) in the six sentences were measured for their acoustic duration. The beginning and end of each vowel token was identified by inspection of wide-band spectrograms and time-domain waveforms following the criteria used in previous research [5]. The results are plotted in Figure 1 below. The same vowels produced by a native speaker (the first author of this paper) are also shown in the figure.

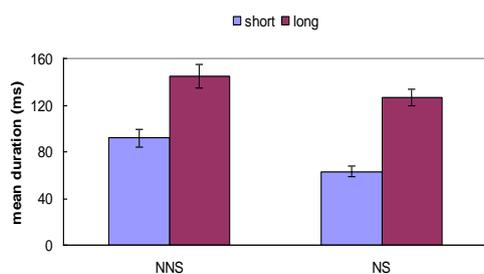


Figure 1: Mean vowel duration (in ms) by non-native (n=24) and native Japanese (n=1) speakers. The brackets enclose ± 1 standard error.

Long Japanese vowels tend to be more than twice as long as their short counterparts when spoken at a fixed speaking rate. Within the same speaking rate (i.e., fast, normal and slow), overlap between the two length categories is negligible [5]. As expected, a long-to-short ratio was larger for the native speaker at 2.0 (127 ms vs. 63 ms) than for non-native speakers at 1.57 (144 ms vs. 92 ms). Further, as is seen in Figure 1, the non-native speakers differed from the native speaker (i.e., first author of this paper) to a greater extent in the production of short than long vowels. This suggests that the timing control for phonemic contrasts in the L2 learners' speech may lack

precision and could be perceptually confusing to the native and possibly non-native Japanese listeners.

Table 1: Error types and pitch range of 24 sentences.

S#	ID & Gender	Error types	Number of errors			f0 range (Hz)
			Word Timing	Word pitch	Phrase final	
1	1F	Px		2	1	114.4
	2F	X	3		1	150.0
	3M	O			1	58.4
	4F	Tx	1		1 L	173.2
2	5F	X	4	3	1 L	199.8
	6M	O				97.1
	7M	Tx			1 L	86.4
	8F	Px		2		149.9
3	9F	Px		2	1 SFL	249.1
	10M	O			1	77.1
	11F	X	2	1		172.2
	12F	Tx	2	1		134.5
4	13F	X	4	3		144.9
	14M	Px	1		4	85.5
	15M	O				90.0
	16M	Tx	2			83.6
5	17M	O				52.7
	18F	Px		3		165.3
	19M	Tx			3 L	42.5
	20F	X	1	3		128.1
6	21F	Tx	3			74.6
	22M	O			1	76.0
	23F	X	3	2	2	119.4
	24F	Px		4	1	155.2

S# = sentence number, L= Phrase final lengthening, SFL= sentence final lengthening

Get_f0 function of Snack Sound Toolkit [9] was used to extract the fundamental frequency (f0) from each sample and measure its pitch range. The get_f0 function is an implementation of 'the Robust Algorithm for Pitch Tracking (RAPT)' proposed by Talkin [11], which is widely used for estimating f0. In this study, f0 was sampled only from the phonetically voiced segments of speech. The f0 values of each speech token were visually inspected, and those tokens which contained obvious f0 tracking mistakes were re-analysed by setting different tracking thresholds. The measurement shows that the pitch range varied from 74 to 249 Hz for female and from 43 to 90 Hz for male speakers.

2.2. Participants

Two groups of English-speaking learners of Japanese (15 beginners, 11 advanced) participated in the task for a small payment or on a voluntary basis depending on their university. The beginners had been studying Japanese for about 170 hours at Griffith University and the advanced learners at ANU for 4-18 years, including the residency in Japan at the time the task was conducted. All of the learners in the advanced group had passed the level 1 of the Japanese Language Proficiency Test

(JLPT) or were judged by the third author to possess the equivalent level of proficiency.

2.3. Procedure

Twenty-four utterances produced by the L2 Japanese speakers (see 2.1. Materials) were played to the two groups of L2 listeners for their naturalness judgements. These stimuli were played in 6 blocks as shown in Table 1 for ease of comparison. Each stimulus was played twice with a 4-second interval, and an inter-trial interval of 8 seconds. Before the task, three practice sentences were played for listeners to become accustomed to the task and proficiency level of L2 speakers. The participants were asked to rate the naturalness of utterances on a Likert scale ranging from 1 (native-like) to 7 (not at all native-like). The listening task took about 15 minutes, including the time for instructions.

3. Results

We compared the average scores of naturalness judgments of the four error types by the two groups of L2 listeners with those of native listeners which were collected in the previous study [12]. Figure 2 contains the average scores given to the four error types of O, Px, Tx and X separately by the three groups of listeners (native, advanced learners, beginners). The number on y-axis indicates the degree of naturalness, with higher scores representing greater *unnaturalness*.

As can be seen in Figure 2, the three participant groups exhibit the same ranking order for Px, Tx, and X, i.e., the judgments of naturalness descend in the order Px < Tx < X. In particular, considering the fact that the assessment of Tx, (incorrect timing) is worse than Px (incorrect pitch) for all three groups, it appears that timing errors contributed to the listeners' judgments of naturalness to a greater extent than did pitch errors. When the samples are free from obvious speech errors (i.e., the O type), L2 listeners may get distracted by paralinguistic information, such as voice quality, speech rate, fluency etc., which is irrelevant to naturalness judgments for native listeners. The judgments of the three groups were very similar with regard to the error type X. However, advanced learners were most critical (more 'not at all native-like' responses compared to other groups) in all four error types.

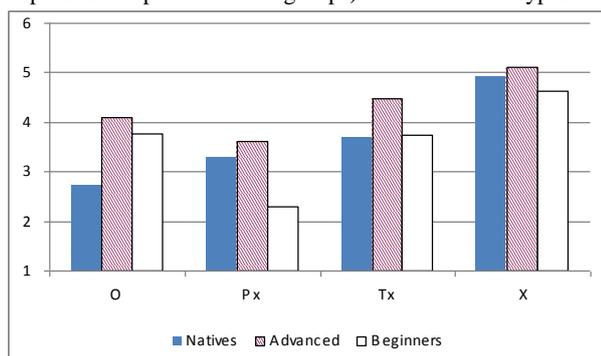


Figure 2: Average scores for four error types by three groups of listeners.

The four types of the speech samples, O, Px, Tx and X, were further investigated for their paralinguistic factors. In Figures 3-6, the y-axis refers to average scores given by the three groups of listeners and the numbers on the x-axis are sentence ID.

What is most noticeable in Figure 2 is that the non-native judgment of Error type O is harsher than that of native

listeners (Figure 3). The six sentences in Error type O were all read by male learners and thus their f0 ranges (their average = 75.2 Hz) were narrower than the average f0 range of the 24 utterances (=120.0 Hz) (See also Table 1 for each speaker's f0 range). This seems to be the reason that L2 learners' judgment scores were higher than native listeners' for all six sentences.

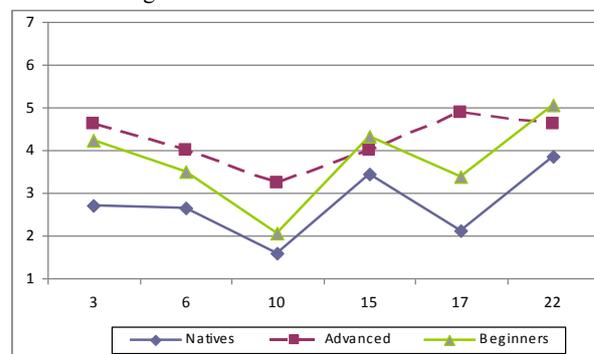


Figure 3: Average scores for Error type O: Correct pitch, correct timing.

Figure 4 shows scores for L2 utterances which had only pitch errors. Beginners gave better scores for the utterances only with pitch errors, unlike the native listeners and advanced learners who gave similar scores to these six utterances. This indicates that beginners do not take pitch errors much into consideration when they judge L2 utterances, which was also obvious from Figure 2. For example, beginners gave a good score to Sentence 24F that had pitch range of 155 Hz and sounded very 'perky'. This influence of paralinguistic factor was also observed in the L2 utterances of the error pattern Tx (Figure 5). Sentence 19 was very monotonous (pitch range 42.5Hz) by a male speaker with three instances of phrase final lengthening, which probably gave an impression of sloppiness.

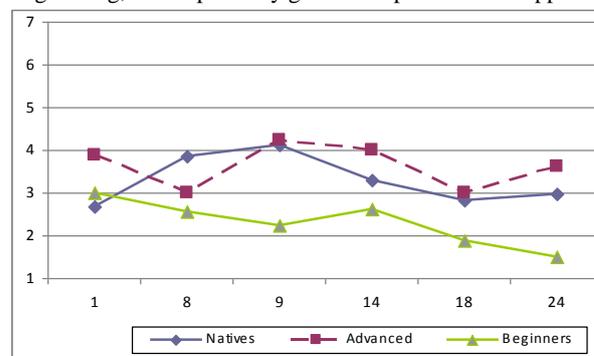


Figure 4: Average scores for Error type Px: Incorrect pitch, correct timing.

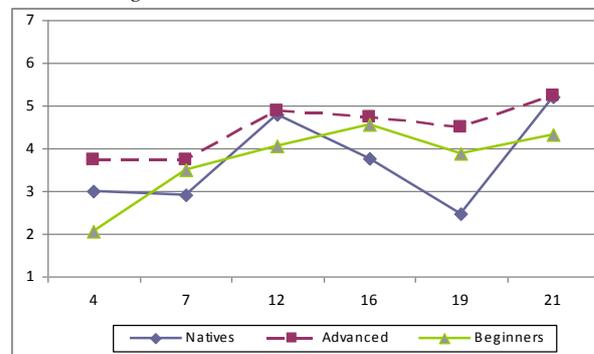


Figure 5: Average scores for Error type Tx: Correct pitch, incorrect timing.

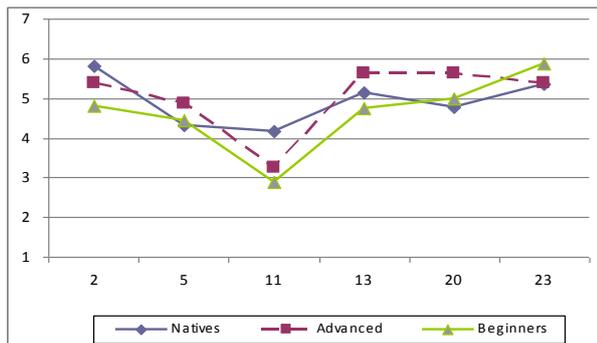


Figure 6: Average scores for Error type X: Incorrect pitch, incorrect timing.

Figure 6 shows listeners' naturalness judgments for L2 utterances that contained both pitch and timing errors. In this error type, the judgment by three groups matched very closely. This means that L2 speech that contains many apparent errors is perceived incorrect by both native and non-native listeners.

4. Discussion

In the previous study [12], it was found that native speakers relied on temporal information more heavily than on pitch to judge the naturalness of L2 speech. In this study, we observed that L2 learners also used temporal information for their judgment and showed the same ranking for the incorrect speech samples. The native vs. non-native difference in naturalness judgments was evident in pitch errors only (Px) and correct speech samples (O). Both groups of L2 learners were often distracted by the voice quality and fluency, and a slightly slow and quiet reading was judged as unnatural. On the contrary, perky and confident female speech which comprised most of the Px speech samples received a good score, particularly from beginners.

Notably, advanced learners' judgment was harsher than that of beginners, and even more critical than that of native listeners for all four types, which corresponds with [2]. When the proficiency level increases, learners seem to have critical ears towards incorrect speech, which presumably helped them improve their L2 speech. However, the overall pattern of judgment was almost the same for beginners and advanced learners, i.e., $Px > O > Tx > X$, except that O (3.8) and Tx (3.7) were almost identical in beginners' judgment. This suggests that even at the advanced level, learners' perception criteria are vulnerable to paralinguistic information, and misjudgment of correct timing can happen. This point needs to be taken into consideration in L2 classrooms and teachers have to draw learners' attention to the truly critical factors for good pronunciation, which are timing and pitch control.

5. Conclusions

We confirmed that both native and non-native Japanese listeners were sensitive to timing information. However, non-native perception had the following characteristics:

- Advanced learners were very critical towards L2 performance in general.
- Beginners do not pay much attention to pitch errors.
- Both levels of non-native listeners are distracted by paralinguistic information, such as pitch range and individual voice characteristics.

Although paralinguistic factors may not be the sole cause of misjudgment by L2 listeners, it is indeed the case that the model pronunciation L2 learners listen to in their classroom

quite often sounds confident, perky and cheerful. It would be beneficial for them to listen to sample speech by many different types of speakers and teachers in order to help them train their perception and familiarize themselves to a wide variety of speech samples. The general belief, "Beginners won't know which pronunciation is good or bad, they are guessing" is not necessarily accurate. They are actually using similar perceptual cues to native listeners. However, their perception is susceptible to paralinguistic factors. Findings of the present study provide useful information for L2 teaching and speech science, particularly for characterizing and synthesizing L2 speech. Since the stimuli were not controlled in all linguistic and non-linguistic aspects, there is a possibility that some non-linguistic aspects of the stimuli, such as voice quality, speech rate and fluency, may have influenced the judgment of the listeners. Speech synthesis will enable more rigorous selection of the stimuli.

In this study, we examined the perception of Japanese learners whose L1 is Australian English. We observed that, despite their incomplete mastery of L2 phonology, even beginners utilized timing information to make perceptual judgments. This may mean that timing information is generally more salient, accessible and, hence, easier to acquire than pitch information in Japanese. Alternatively, this pattern of results may be specific to speakers of Australian English, in which duration is known to play a more important role than in other varieties of English [1, 4]. In order to gain a better understanding of learners' access to timing vs. pitch information in Japanese, it would be necessary to examine the perception of Japanese learners from different L1 backgrounds.

6. References

- [1] Cox, F., "The acoustic characteristics of /hVd/ vowels in the speech of some Australian teenagers", *Australian J. Linguistics*, 26, 147-179, 2006.
- [2] Fayer, J. M. and Krasinski, E. K., "Native and nonnative judgment of intelligibility and irritation", *Lang.*, 37, 311-326, 1987.
- [3] Flege, J. E., "Factors affecting degree of perceived foreign accent in English sentences", *J. Acoust. Soc. Am.*, 84, 70-79, 1988.
- [4] Harrington, J. and Cassidy, S., "Dynamic and target theories of vowel classification: evidence from monophthongs and diphthongs in Australian English", *Lang. and Speech*, 37, 357-373, 1994.
- [5] Hirata, Y., "Effects of speaking rate on the vowel length distinction in Japanese", *J. Phonetics*, 32, 565-589, 2004.
- [6] Inoue, F., "The significance of new dialects", *Dialectologia et Geolinguistica (SIDG)*, 1, 3-27, 1993.
- [7] Riches, P. and Foddy, M., "Ethnic accent as a status cue", *Social Psychology Quarterly*, 52, 197-206, 1989.
- [8] Sato, T., "A comparison of phonemes and prosody in the evaluation of spoken Japanese", *Japanese Lang. Education around the Globe*, 5, 139-154, 1995.
- [9] Sjölander, K., *The Snack Sound Toolkit*. <http://www.speech.kth.se/snack/>, 2006.
- [10] Tajima, K., Port, R. and Dalby, J., "Effects of temporal correction on intelligibility of foreign-accented English", *J. Phonetics*, 25, 1-24, 1997.
- [11] Talkin, D., "A robust algorithm for pitch tracking (RAPT)", in W. B. Klejin & K. K. Paliwal [Eds.], *Speech Coding and Synthesis*, 495-518, Elsevier, 1995.
- [12] Tsurutani, C., "Foreign accent matters only when timing is wrong" (to appear).
- [13] Warren, P., Elgort, I. and Crabbe, D., "Comprehensibility and prosody ratings for pronunciation software development", *Lang. Learning & Technology*, 13, 87-102, 2009.