Reaction Time to Unnatural and Natural Pronunciation by Native and Non-Native Speakers of Japanese

Ian Wilson, Jeremy Perkins, Julián Villegas, and Ayaka Orihara (University of Aizu)

1 Introduction

When native listeners listen to *native* speakers speaking with *non-standard* phonological conventions, their reaction time (RT) is slower than when they listen to them speaking with *standard* phonological conventions. A phonological process that can be used to test this is high-vowel devoicing. Japanese high-vowel devoicing is a phonological process where high vowels /i, u/ are devoiced when they occur between voiceless consonants or between a voiceless consonant and the end of a word [1]. There are many factors that influence the likelihood of high-vowel devoicing, such as phonological environment, stress and accent, speaking tempo and style, sociolinguistic factors, and lexical, syntactic, and semantic constraints [2]. In Japanese, it is also a process that varies by dialect and age, with younger speakers being more likely to devoice high vowels [3]. In standard Japanese, high-vowel devoicing is normal. Ogasawara and Warner [4] asked native listeners from Tokyo to identify words as quickly and accurately as possible. All the words had been produced by a native speaker, some of them with high vowels that were devoiced and others with fully voiced vowels (i. e., intentionally unnatural pronunciation). Their results showed that native listeners were faster at identifying words with natural pronunciation (i. e., with high-vowel devoicing), even though those words contained less acoustic information. The researchers attributed this to the fact that words with the expected vowel devoicing are much more frequent (because they are natural for those speakers).

Of course, pronunciation can be unnatural because of phonological reasons like the example given above, but it can also be unnatural due to the phonetic characteristics of a foreign accent (e. g., f0, VOT, vowel formants, etc.) A question that arises, then, is "What happens to RT when people listen to *non-native* speakers speaking with *non-standard* compared to *standard* phonological conventions?" When considering non-native speech, one can no longer say that vowel devoicing is more frequent. Varden and Sato [5] found that Taiwanese learners who had been studying Japanese for 3 to 4 years still had very low rates of high-vowel devoicing. If a listener can hear that the speaker is non-native and expects phonetic and phonological errors, is RT unaffected, or do those multiple errors slow RT even more than when the speaker is native?

To measure what happens to RT when native listeners hear non-natives making phonological errors, we measured the RT of Japanese participants who did a forced-choice word recognition experiment using natural and unnatural stimuli recorded by native and non-native speakers of Japanese. Each stimulus was a simultaneous combination of visual (two side-by-side images) and audio (one spoken word) prompts.

2 Method

2.1 Participants

Thirty Japanese listeners participated in the RT experiment (21 male, 9 female). They ranged in age from 19 to 23, and none of them reported any hearing impairment. We did not inquire about the handedness of participants. Although most of the participants were from the Southern Tohoku or Kanto areas of Japan, where high-vowel devoicing is the norm, 3 participants (10%) were from dialect areas with a lower likelihood of high-vowel devoicing in the local dialect. Even so, those 3 participants would have been exposed to speech containing high-vowel devoicing on television and in the speech of most of their peers at university, and previous research [6] has shown that participants from areas where high-vowel devoicing is *not* the norm have RTs that are similar to other participants.

2.2 Procedure

2.2.1 Stimuli preparation

As mentioned above, each stimulus was a simultaneous combination of visual and audio prompts. Eight pairs of colour images were created, each pair corresponding to the vowel-devoicing and non-vowel-devoicing conditions in Japanese. For example, *KISHI (coast)* where the first /i/ is normally devoiced and *KIJI (pheasant)* where the first /i/ is normally fully voiced. All stimuli were common Japanese words.

Each word was exactly two mora of the form consonant-vowel-consonant-vowel ($C_1V_1C_2V_2$). In all words, C_1 was voiceless, and in each pair of words C_2 was voiceless in one word (always the image on the left) and voiced in the other (always the image on the right). Table 2.2.1 shows the eight pairs of stimuli used, with the high pitched vowel marked (optional pronunciation in pairs 1 and 4). For each image pair, three audio prompts were created: the natural (devoiced V_1) pronunciation of the image on

the left, the unnatural (voiced V_1) pronunciation of the image on the left, and the natural pronunciation of the image on the right. Thus, a total of 24 audio-visual stimuli prompts were created (8 image pairs × 3 audio prompts). The pitch accent of all words was checked in a Japanese accent dictionary [7]

Table 1 Japanese CVCV stimuli used.

Image pair	C_2 voiceless	C_2 voiced
1	FÚKU/FUKÚ (blow)	FÚGU (blowfish)
2	KISHÍ (coast)	KIJÍ (pheasant)
3	KUKÍ (stem)	KUGÍ (nail)
4	SHÍKA/SHIKÁ (dentist)	SHÍGA (prefecture)
5	SHÍSHI (lion)	SHÍJI (instruction)
6	SUKÍ (like)	SUGÍ (cedar)
7	SÚSU (soot)	SUZÚ (bell)
8	TSUKÍ (moon)	TSUGÍ (next)

Three speakers were recorded reading the 24 audio prompts: one 22-year-old Japanese native speaker from Ishinomaki, Miyagi prefecture (Tohoku region), and two non-native speakers, both Canadian-English speakers, a 47-year-old who had lived in Japan for over 15 years (high-proficiency) and a 35-year-old who had lived in Japan for less than 1 year (low proficiency). The recordings were made in a sound-proofed recording studio using a Korg MR-1000 1-Bit recorder with a DPA 4080-BM miniature cardioid lavalier microphone at a 44.1 kHz/16 bits quantization. After recording, Audacity 1.3.12 was used to manually extract each word for use as a stimulus in the E-Prime program.

The native speaker produced all words with dictionary pitch accent, except for *SUSU* , which he produced as LH for both the devoiced and fully-voiced versions. The non-native speakers purposely produced all words as HL, i. e., they produced only 7 out of the 16 words with a pitch accent matching the dictionary. This is taken into consideration in the statistical model used in the data analysis.

The audio and visual stimuli were used to create a forced-choice word identification experiment using E-Prime 2.0.10.242.

2.2.2 Data collection

The RT experiment was done in a quiet room, one participant at a time seated in front of the computer that displayed the images. The participants wore headphones and could adjust the volume to a comfortable level before starting. Participants had to push “1” or “2” on the keyboard when they could identify the sound that they heard. The image on the left was always “1” (and had the number “1” displayed under it) and vice versa for the image on the right.

Before starting the RT experiment, all 16 images were shown to the participants to ensure that they could correctly identify them all, but no practice was given doing the actual task with the audio prompts. Participants were asked their hometowns and to indicate any hearing problems.

The order of presentation of the stimuli was randomized by E-Prime. Although changing the speaker within the experiment slows down RT [8], this would affect all stimuli and speakers equally. The images were displayed at exactly the same time as the audio was played through the headphones, and RT was measured from the beginning of C_1. The images remained visible until a response was given, or until the response deadline of 5 s was reached.

2.2.3 Data analysis

First, we determined the error rate of participants (i. e., cases of misidentification of the spoken word). Out of all 2160 cases (72 stimuli × 30 participants), only 21 cases (0.97%) were errors and these were excluded from further analysis, leaving 2139 cases. This low error rate was probably due to the fact that the audio stimuli were clear recordings, not presented in noise.

Because RT distributions are usually skewed towards lower RTs, they are not normal distributions. Transforming the data can minimize the effects of skew and outliers [9] and so we transformed the remaining RTs by taking the logarithm of their reciprocal. We scaled the result by −1000 so that the transformed RTs would be on the same order of magnitude as the original RT data. Thus, the operation we performed to transform and scale the data was $-1000 \log \frac{1}{RT}$. Additionally, we excluded from the analysis data that were three times the interquartile range below or above the first and third quantile, respectively. This outlier exclusion was performed independently per subject and stimulus. The amount of data removed from the analysis was 0.9% of the original data.

R and ‘lme4’ were used to perform a linear mixed effects analysis of the relationship between RT and naturalness as well as the Japanese proficiency of the speakers (we call this independent factor “speaker”). As fixed effects, we considered naturalness, speaker, RT of the previous stimulus for a given participant (called “previous RT”), the order of presentation of a stimulus (called “OrderSubject”) from 1 to 72, word duration in seconds, mean intensity in dB, lexical pitch accent, the pitch-accent pattern actually produced, and whether the lexical pitch accent was correctly produced. In addition, we included a factor meant to capture the f0 range of a given speaker. Since many tokens had a devoiced V_1, this was done by calculating the f0 in semitones relative to the mean f0 of V_1 in HL tokens for each speaker. This factor, together with intensity, and word duration explained the effects of speaker difference on

RTs. As a result, these three continuous factors were used in the models instead of the categorical factor, "speaker."

In addition to a grand model for the entire data set, separate LMM analyses were investigated for subsets of the data from tokens produced by only non-native speakers and by only native speakers. This enabled an evaluation of whether the non-native pronunciations with English stress (HL) on Japanese words that have LH pitch-accent patterns would affect RTs. It also allowed an unbiased assessment of the effect on RTs of the two pitch-accent patterns (LH and HL) among tokens produced by a native Japanese speaker.

3 Results and Discussion

As predicted, transformed RTs were slower in unnatural stimuli relative to natural stimuli ($\beta = 16.9$, $t(1329) = 3.12$, $p < 0.01$) in the analysis of the entire data set (see Figure 1), corroborating the results of [4]. Analyzing native speaker tokens, we found an interaction between naturalness and the lexical pitch accent of a token (see Figure 2). For tokens listed as LH and listed as variably LH or HL, RTs were slower for unnatural stimuli (for LH: $\beta = 47.2$, $t(366) = 4.15$, $p < 0.001$; for LH/HL: $\beta = 46.3$, $t(366) = 2.91$, $p < 0.01$), as expected. However, in tokens with a HL pitch-accent pattern, unnatural tokens resulted in faster RTs ($\beta = -125.0$, $t(366.2) = -5.48$, $p < 0.001$), a result we are unable to explain.

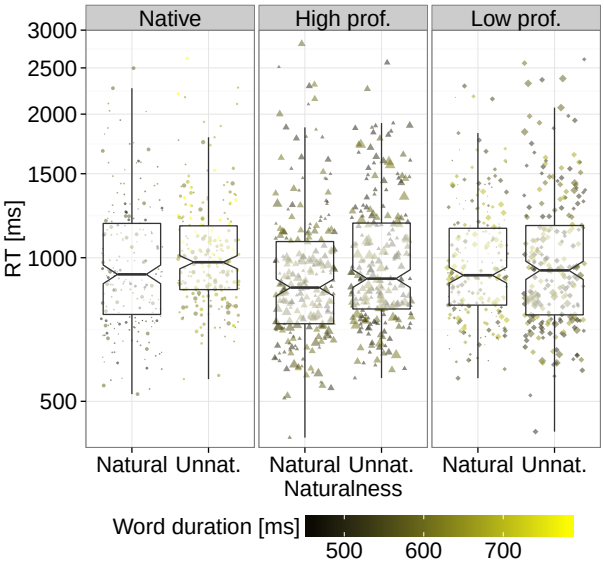

Fig. 1 RTs by naturalness and speaker (native, high-proficiency non-native, low-proficiency non-native). Word duration is shown by shading (shorter = darker) and intensity is shown by the size of each mark (bigger is more intense).

Regarding RT differences in tokens produced by natives versus non-natives, no significant effect was discovered. Instead, differences in RT effects between the three speakers could be explained via mean intensity, word duration and relative pitch of V_2. All three factors had significant effects. Longer word duration led to slower RTs across all tokens ($\beta = 218.6$, $t(568.8) = 5.01$, $p < 0.001$), not surprising given that RT was measured from the beginning of each word. Greater intensity led to faster RTs ($\beta = -3.67$, $t(936.7) = -4.21$, $p < 0.001$), again not surprising – if one can hear a stimulus more easily, RT will be faster. A higher pitch in V_2 relative to V_1 resulted in faster RTs ($\beta = -3.26$, $t(1264) = -2.76$, $p < 0.01$).

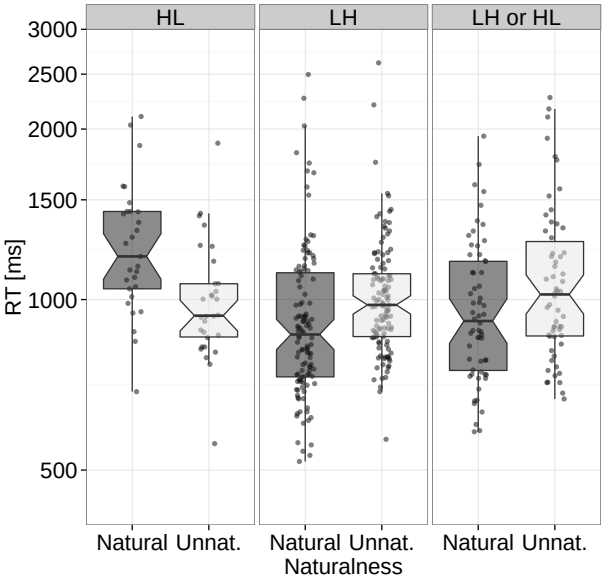

Fig. 2 RTs to native speaker's tokens by naturalness and lexical pitch accent.

Intensity and word duration differed for each speaker. Strangely, mean RTs for the native speaker's tokens appear slower than for the non-native speakers. However, the linear mixed model reveals that this is due to the fact that the native speaker spoke more slowly and softly.

In addition, naturalness was highly correlated with duration for the native speaker. Unnatural tokens had a mean duration of 663 ms, about 120 ms longer than natural tokens, which had a mean duration of 547 ms. This is indicated by shading in Figure 1, and it seems to indicate hesitation or extra care by the native speaker when pronouncing unnatural tokens. This correlation created a confound between effects of word duration and naturalness for the native speaker.

We also considered effects of the pitch-accent pattern of a token. We explored both the lexical pitch accent, and the produced pitch-accent pattern, which differed because some tokens had more than one possible pitch-accent pattern. We were particularly concerned with the systematic choice by the non-native speakers to produce only HL patterns,

consistent with default stress in their native language. We added a two-level factor, Pitch Accent Correctness corresponding to whether the produced pitch-accent pattern matched that of a Japanese dictionary. In the analysis of the non-native speaker tokens, Pitch Accent Correctness was not in the final model. Instead, the lexical pitch-accent pattern had a significant effect on RT. This effect was consistent with our expectation, in that tokens with LH words (which were produced incorrectly) had slower RTs than HL words ($\beta = 26.7$, $t(882.7) = 2.98$, $p < 0.01$). However, the two words with optional LH or HL patterns were produced with significantly slower RTs as well ($\beta = 67.9$, $t(882.9) = 6.30$, $p < 0.001$). This was unexpected, since we considered these tokens to be pronounced correctly by the non-native speakers. Perhaps the RTs are slower for these words because having more variability in the accepted pitch accent increases processing load in general.

In an analysis of the native speaker tokens, an interaction between the lexical pitch accent and naturalness was discovered. In general, RTs were faster for LH ($\beta = -156.1$, $t(11.7) = -6.05$, $p < 0.001$) and for optional LH/HL relative to HL tokens ($\beta = -132.0$, $t(11.6) = -4.68$, $p < 0.001$). However, for HL pitch accents with unnatural voicing on the vowel, these RTs were even faster than expected ($\beta = -125.0$, $t(366.2) = -5.48$, $p < 0.001$). The expected effect where unnatural tokens have slower RTs is reversed for HL tokens then. When considering LH tokens and LH/HL tokens with unnatural voicing in the vowel, a relative increase in RTs is seen, as expected (for LH: $\beta = 47.2$, $t(366) = 4.15$, $p < 0.001$; for LH/HL: $\beta = 46.3$, $t(366) = 2.91$, $p < 0.01$).

In all three analyses, significant additional effects were discovered for previous RT (for entire data set: $\beta = 243.2$, $t(1372.1) = 10.06$, $p < 0.001$) and Order-Subject ($\beta = -1.29$, $t(1355.5) = 10.06$, $p < 0.001$). These effects can be seen in Figure 3. RTs were positively correlated with the RT of the immediately preceding stimulus item (correlation, r = 0.53), a priming effect apparent in the shading of Figure 3, where lower tokens in the graph are generally darker and higher tokens are lighter. RTs also decreased (correlation, r = -0.31) over the course of the 72 trials of an experimental session, a learning effect, and the sloping lines show this. The curvature of these lines shows that for tokens produced by either the native speaker or the high-proficiency non-native speaker, the learning effect disappears after about the 40th token. For tokens produced by the low-proficiency non-native speaker, though, RTs keep getting faster throughout the whole experiment. In other words, it seems to take the listeners longer to get used to the voice of the low-proficiency non-native speaker. One reason for this could be the fact that his non-native pronunciation has more phonetic distractions for the listener – indeed, the range of f0 he used was much greater than that of the other two speakers. Speaker familiarity could also account for this, though – many of the participants knew the native speaker and had probably taken a class from or met the high-proficiency non-native speaker.

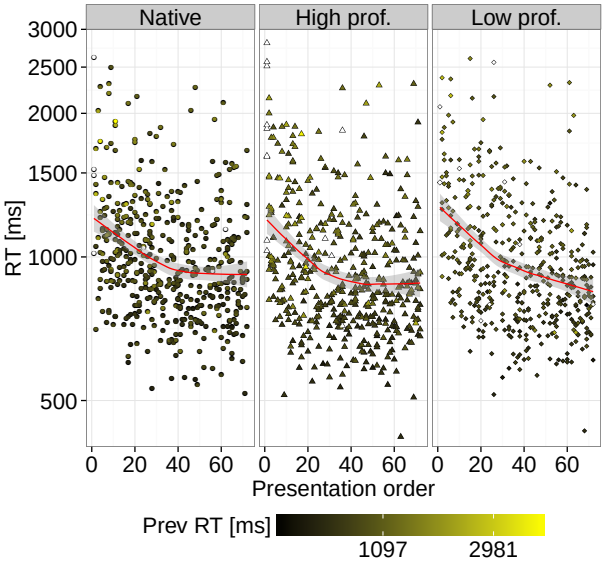

Fig. 3 RTs by stimulus presentation order and speaker. Shading indicates RT of previous stimulus (darker = faster).

参考文献

[1] T. J. Vance, *An introduction to Japanese phonology*. Albany, NY: SUNY Press, 1987.

[2] M. Kondo, "Mechanisms of vowel devoicing in Japanese," Ph.D. dissertation, University of Edinburgh, 1997.

[3] H.-G. Byun, "Nihongo semaboin no museika: Kyoutsuugo fukyuu no shihyou to shite [Vowel devoicing in Japanese: As an indicator of standardization of dialect]," Ph.D. dissertation, University of Tokyo, 2012.

[4] N. Ogasawara and N. Warner, "Processing missing vowels: Allophonic processing in Japanese," *Language and Cognitive Processes*, vol. 24, no. 3, pp. 376–411, 2009.

[5] K. J. Varden and T. Sato, "Devoicing of Japanese vowels by Taiwanese learners of Japanese," in *Proc. of International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, PA, Oct. 1996, pp. 618–621.

[6] N. Ogasawara, "Processing of speech variability: Vowel reduction in Japanese," Ph.D. dissertation, University of Arizona, 2007.

[7] T. Hirayama, *Zenkoku Akusento Jiten [Japanese Accent Dictionary]*. Tokyo: Tokyodo Publisher, 1960.

[8] J. W. Mullennix, D. B. Pisoni, and C. S. Martin, "Some effects of talker variability on spoken word recognition," *Journal of the Acoustical Society of America*, vol. 85, no. 1, pp. 365–378, 1989.

[9] R. Whelan, "Effective analysis of reaction time data," *The Psychological Record*, vol. 58, no. 3, pp. 475–482, 2008.