Effects of signal degradation and speaker sex on recognizing Japanese disyllabic words

Yukiko Sugiyama*

*Department of Foreign Languages and Liberal Arts, Faculty of Science and Technology, Keio University,
4-1-1 Hiyoshi, Kohoku-ku, Yokohama, 223-8521 Japan*

(Received 30 September 2018, Accepted for publication 28 May 2019)

Keywords: Japanese pitch accent, Acoustic correlates, Speech production, Speech perception

PACS number: 43.70.Fq, 43.70.Mn [doi:10.1250/ast.41.416]

1. Introduction

Phonological contrasts are typically realized by multiple acoustic properties. For example, in many languages, voice onset time (VOT) is the primary cue to a voicing contrast for stop consonants in the word-initial position [1]. At the same time, the fundamental frequency (f_0) and the onset frequency of the first formant (F_1) co-vary with VOT. Because these acoustic properties regularly co-occur, listeners can perceive the intended phoneme, even when one cue is missing, using the remaining cues. Such phonetic redundancy is also known to exist for prosodic contrasts such as stress in English [2] and tone in Mandarin Chinese [3]. Fry [2] manipulated duration, intensity, and f_0 to examine whether they affected the perception of stress in English and found that all three were relevant in perceiving stress in English. Similarly, Fu and Zeng [3] manipulated duration, amplitude contour, and f_0 to examine how they affected the perception of four tones in Mandarin Chinese. They found that the amplitude contour of each tone was considerably similar to the f_0 contour of the corresponding tone and that listeners' tone identification was in fact slightly more accurate with the amplitude cue alone than with the f_0 cue alone. In Tokyo Japanese (simply Japanese hereafter), the primary cue to pitch accent is f_0 [4], but less is understood about its secondary cues. In the present study, noise-excited vocoder speech was used (1) to examine the existence of secondary cues by conducting a perception experiment and (2) to determine their acoustic identity by acoustic analysis.

2. Perception experiment

2.1. Method

Five minimal pairs of disyllabic final-accented words and unaccented words (e.g., /haná/ “flower” vs. /hana/ “nose”) that had a relatively high familiarity rating (i.e., 5.0 or higher on a 7-point scale) in Amano and Kondo [5] were used as test words. These words were produced in the middle of a carrier sentence /watasi wa ____ ga suki/ (I TOPIC ____ NOMINATIVE like) by a male and a female native Japanese speaker, which were then noise-vocoded using Praat software [6] at a sampling rate of 44.1 kHz with 16-bit resolution. The minimal pairs between the final accented words and unaccented words were chosen because they have the same pitch pattern of low-high with the first syllable having a low pitch

and the second syllable having a high pitch. The most prominent difference between the two accent types appears in the following syllable, when there is one [7,8]. The pitch on the following syllable remains high when the syllable follows an unaccented word whereas it falls rapidly after an accented word. Disyllabic words were chosen because the number of minimal pairs between final-accented words and unaccented words is very limited for trisyllabic and longer words. Two types of noise-vocoded stimuli spanning 50–6,805 Hz were created for both male and female voices. The original recordings were divided into ten 3-ERB_N-wide bands in one type of stimuli and into fifteen 2-ERB_N-wide bands in the other type of stimuli following the equation for deriving the ERBs proposed by Glasberg and Moore [9].

Forty Tokyo Japanese speakers who were raised in Tokyo or its neighboring areas where Tokyo Japanese is spoken participated in the perception experiment. In order to minimize the influence of dialects that have different accent patterns from Tokyo Japanese, it was ensured that the parents of the participants were also from an area where Tokyo Japanese is spoken.

The forty participants heard the minimal pairs embedded in the carrier sentence in either the 10-band noise-vocoded speech or 15-band noise-vocoded speech through Audio-Technica ATH-M50x headphones at a comfortable listening level and were asked to identify the words they thought they heard. In other words, the number of bands was a between-listener factor. Half the listeners who heard 10-band noise-vocoded speech (i.e., 10 listeners) heard the male voice stimuli first and the remaining half heard the female voice stimuli first. Similarly, half the listeners who heard 15-band noise-vocoded speech heard the male voice stimuli first and the remaining half heard the female voice stimuli first. Each listener went through 200 trials (10 words $\times$ 2 repetitions $\times$ 5 blocks $\times$ 2 voices).

In order to assess how well listeners discriminated final-accented words and unaccented words, d' for each listener was calculated for each word for each of the voices. Then, the d' values obtained were analyzed using one-sample t -tests against the null hypothesis $d' = 0$, corresponding to the case when final-accented words and their unaccented counterparts were not reliably discriminable. It was found that d' was significantly above zero for listeners ($t(39) = 13.1$, $p < 0.0001$), with a mean d' of 1.49 (SE = 0.11), indicating that listeners were able to distinguish the two types of words

*e-mail: yukiko@keio.jp

Fig. 1 Intensity difference between the word-final syllable and the following particle.

significantly better than chance. This result suggests that they used some acoustic properties other than f_0 or its harmonics as a cue to pitch accent.

3. Acoustic analyses

In order to examine what acoustic cues were available to listeners to distinguish final-accented and unaccented words, acoustic analysis was conducted on the noise-vocoded stimuli that they heard. While studies of the examination of acoustic manifestations of pitch accent exist [7,8], they mainly focused on f_0 contours and it was unclear whether listeners can use other cues to distinguish the two types of words. Since f_0 was absent in the present study, other acoustic properties were measured to investigate possible cues the listeners used to distinguish final-accented and unaccented words. In comparing the two types of words, five types of measurements were made: the word-final syllable duration and word-final vowel duration of the final-accented and unaccented words, the intensity difference between the word-final syllable and the following particle /ga/, and the first and second formant frequencies of the following particle. The formant frequencies were measured for the following particle because, in normal speech, f_0 contours differ the most clearly in that location between the two types of words. If any acoustic properties covaried with f_0 , they were also expected to differ the most clearly in the same location. Because the number of tokens was limited, no statistical analysis was conducted.

Of the five types of measurements made, the syllable duration and second formant frequencies showed no clear effect of accent for either the male or female voice. While vowel duration showed no consistent pattern for the female voice, it was consistently longer for unaccented words than for final-accented words for the male voice, although the difference was variable. The intensity difference was greater for accented words than for unaccented words for both voices (Fig. 1). In addition, the first formant frequencies appeared to be higher when the particle followed final-accented words than when it followed unaccented words (Fig. 2), suggesting a lowered tongue position for final-accented words.

Fig. 2 First formant frequencies for the male (left) and female (right) voices taken from the middle 100 ms portion of the vowel in the particle /ga/.

4. Summary

The present study explored secondary cues that listeners could use when the f_0 is not available. Comparing the results obtained in the perception experiment and the subsequent acoustic analysis suggest that listeners might make use of the intensity, the first formant frequencies, and possibly the duration information to discriminate the two types of words that differed only in accent. Further investigation with more controlled stimuli would help to exactly determine which acoustic properties listeners use to perceive pitch accent in Japanese.

References

- [1] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, **20**, 384-422 (1964).
- [2] D. B. Fry, "Experiments in the perception of stress," *Lang. Speech*, **1**, 126-152 (1958).
- [3] Q. Fu and F. Zeng, "Identification of temporal envelope cues in Chinese tone recognition," *Asia Pac. J. Speech Lang. Hear.*, **5**, 45-57 (2000).
- [4] M. E. Beckman, *Stress and Non-stress Accent* (Foris, Dordrecht, 1986).
- [5] S. Amano and T. Kondo, *Nihongo-no Goitokusei (Lexical Properties of Japanese)* (Sanseido, Tokyo, 1999).
- [6] P. Boersma and D. Weenink, *Praat: Doing Phonetics by Computer* [Computer program]. Version 5.2.24 (2014). <http://www.praat.org/>
- [7] M. Sugito, *Nihongo Onsei no Kenkyu 5: "Hana" to "Hana" (Study on Japanese Sounds 5: "Hana (Flower)" and "Hana (Nose)")* (Izumi Shoin, Osaka, 1998).
- [8] S. Kawakami, *Nihongo Akusento Ronshu (Collection of Papers on Japanese Accent)* (Kyuko Shoin, Tokyo, 2005).
- [9] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, **47**, 103-138 (1990).

Yukiko Sugiyama received her bachelor's degree in English from Keio University, Japan, and her PhD degree in linguistics from University at Buffalo (SUNY). She is currently an assistant professor in the Department of Foreign Languages and Liberal Arts in the Faculty of Science and Technology at Keio University.