BLS 34, No 1 2008. DOI: http://dx.doi.org/10.3765/bls.v34i1.3579
(published by the Berkeley Linguistics Society and the Linguistic Society of America)

Production and Perception of Pitch Accent in Japanese[1]

YUKIKO SUGIYAMA
University at Buffalo, The State University of New York

0. Introduction

Word prosody of Tokyo Japanese (simply Japanese hereafter) is often labeled as pitch accent, characterized by a steep F_0 fall from the accented mora to the following one (e.g. McCawley 1968). For example, /hana/ with a low-high (LH) tone sequence means 'flower' when the final mora is accented but 'nose' when there is no accent. The most notable difference between the two accent types is manifested in the following particle, when there is one. It has a low pitch after a final-accented word (thus, /hana^ ga/ LH L 'flower NOMINATIVE', '^' indicates accent on the preceding mora) whereas it has a high pitch after an unaccented word (/hana ga/ LH H 'nose NOMINATIVE'). According to the traditional account of Japanese pitch accent, the tone difference on the following particle is the only difference between the two accent types (Kindaichi 1947). However, results from more recent experimental studies suggest that this may not be the only difference. At the same time, since they focused on a very few minimal pears of final-accented and unaccented words, their results are not totally inconsistent. It is still unresolved what exactly the difference is between final-accented and unaccented words, whether the contrast between the two accent types appears even when words are produced in isolation, and what perceptual cues distinguish the two accent types. The present study was designed to address these issues. First, a database of Japanese words was used to thoroughly search for minimal pairs (such as /hana^/ 'flower' and /hana/ 'nose') that exist in Japanese. Then, using these minimal pairs, production and perception experiments were conducted to establish general properties of Japanese pitch accent. Since word familiarity is known to influence word production and recognition (Amano, Kondo, and Kato 1999, Wright 1997), only familiar words were used.

[1] I would like to thank Karin Michelson for valuable discussions and comments on this paper. I am also thankful to Jim Sawusch for fruitful discussions. I am obliged to Haruo Kubozono for his assistance in designing the production experiment, Doug Roland for his assistance with Perl scripts, and Mitsu Shimojo for helping me recruit subjects.

1. Past Work

For words produced in isolation, results from previous studies are not consistent (Han 1962, Poser 1984, Sugito 1979, Vance 1995). Sugito (1979) analyzed the minimal pair /hana^/ 'flower' and /hana/ 'nose' produced by 14 males in isolation and found that two talkers produced the two words significantly differently. This was confirmed by a higher F_0 peak value on the second mora and a greater F_0 rise value from the first to the second mora. Poser (1984) measured the F_0 peaks of the same words produced by a single male talker, but no significant difference was found. For words produced in a sentence, contrary to the traditional account, experimental studies have consistently found that final-accented and unaccented words are different even within words in addition to the F_0 fall difference into the following particle (Poser 1984, Sugito 1979, Vance 1995). Specifically, final-accented words have a higher F_0 peak on the second mora than unaccented words. In addition, Sugito and Vance (1995) found that F_0 rise was greater for final-accented words than unaccented words. However, while Sugito seems to hold that the F_0 rise is important for distinguishing the two accent types, in Vance's data the two accent types were distinguished more clearly in terms of F_0 peak than F_0 rise. Since only one minimal pair was used in earlier studies, it is possible that inconsistent results were due to not having enough minimal pairs.

Studies that examined the perception of final-accented and unaccented words produced in isolation found that listeners' word identification was not very accurate overall (e.g. Neustupný 1978, Sugito 1979, Vance 1995). While Neustupný used two minimal pairs, the others used only one. They found that some tokens recorded by certain individuals had accuracy above chance, but others did not. The results suggest that listeners' performance was dependent not only on the listener's ability to identify words but also on whether or not the individual who recorded the stimuli maintained a clear distinction between the two accent types. Interestingly, most studies report that listeners had a tendency to respond that they heard final-accented words rather than unaccented words. Studies that examined the perception of lexical accent in a sentential context found that listeners' judgment was dependent on the size of F_0 fall (e.g. Hasegawa and Hata 1992, Kitahara 2001). Listeners perceived accent when a mora was followed by a relatively steep F_0 fall. However, since the presence or absence of accent was not the only property varied in most of the test items, they do not show if listeners can identify words based solely on the accent information. It also has to be noted that (re)synthesized speech was used in these studies. Thus, the stimuli that listeners heard may not necessarily correspond to what typically occurs in natural speech.

In short, previous studies on Japanese pitch accent used very limited numbers of minimal pairs. The research question in the present study was whether or not findings from previous studies on Japanese pitch accent could be extended to a larger set of words. This, in terms of production, was to examine if talkers consistently produce differences between the two accent types. In terms of perception, the question was whether or not listeners can use the acoustic information to

distinguish the two accent types, if any. While past research on Japanese pitch accent tended to study production and perception separately, given the communicative function of speech, it is important to understand the relation between them.

2. Production Experiment

2.1. Method

Speakers: The data were collected from ten native speakers of Tokyo Japanese (five male, five female, ages 28-33 years old) at the University at Buffalo.

Materials: Twenty pairs of bimoraic words that differed only in accent were selected using the database developed by Amano and Kondo (1999). Bimoraic words were necessary because they allow measuring the F_0 difference between the first and second moras. Trimoraic and longer words could not be used because minimal pairs comprising final-accented and unaccented words are extremely limited (Kitahara 2001). In selecting bimoraic minimal pairs, first, only words that were unambiguously final-accented or unaccented were selected. In the database, some words were indicated as having more than one possible accent type. Since it was important that the tone sequence was LH for all of the words, words with more than one possible accent type were left out, resulting in 55 minimal pairs. Then, since word familiarity has been found to have an effect on word recognition and recall in Japanese (e.g. Amano, Kondo, and Kato 1999), familiarity of the 55 pairs was checked using the database. In the database, the familiarity of each word was listed on a 7-point scale, with 7 indicating the highest familiarity. Out of the 55 pairs, 19 pairs that had a familiarity rating of 5.0 or higher were selected. To these pairs, one pair was added for comparison with previous studies. Thus, a total of 20 pairs were selected.

Procedure: The talkers were asked to produce the 40 words in three environments: 1) in isolation, 2) in the focus frame, and 3) in the non-focus frame. In the focus frame, target words were produced as focus of the sentence. In the non-focus frame, some other word was the focus of the sentences. Two types of carrier sentences were prepared because words under focus are known to have greater F_0 movement than words that are not under focus (Pierrehumbert and Beckman 1998). All the words were recorded twice in each environment. The total number of tokens measured was:

40 (words)×3 (environments)×2 (repetitions)×10 (speakers)–1 (missing)=2399.

Measurements: The values for three parameters were obtained for the words produced in a sentence: 1) F_0 maximum on the second mora, 2) F_0 rise from the first to second mora, which was obtained by subtracting the F_0 minimum on the first mora from the F_0 minimum on the second mora, and 3) F_0 fall from the second mora into the following particle, which was obtained by subtracting the F_0 minimum on the particle from the F_0 maximum on the second mora. For the words produced in isolation, only the values for the first two parameters were obtained. All acoustic analyses were done with Praat (Boersma and Weenink 2005). Once each of the talkers' mean F_0 peak, F_0 rise, and F_0 fall were obtained

for each word produced in each environment, the data for all the talkers were combined and submitted to statistical analyses.

2.2. Results & Discussion

Three-way analyses of variance (ANOVAs) were performed on F_0 peak, F_0 rise, and F_0 fall, with either talkers (*F1*) or final accented and unaccented words (*F2*) as the repeated measures. The between-subjects factor was sex (male vs. female). The within-subjects factors were environment (isolation, focus frame, non-focus frame) and accent (final-accent vs. no accent). In the following, the complete statistical results will be presented to show the exact nature of the data. However, due to limitations of space, the discussions will be made with reference to mainly the effect of accent (whether or not final-accented and unaccented words showed any difference) and in which environment (isolation and sentence) the difference between the two accent types appeared. The effects of sentence type (focus and non-focus frames) and sex will not be specifically discussed.[2] Readers are referred to Sugiyama (2008) for the thorough presentation and discussion of the data.

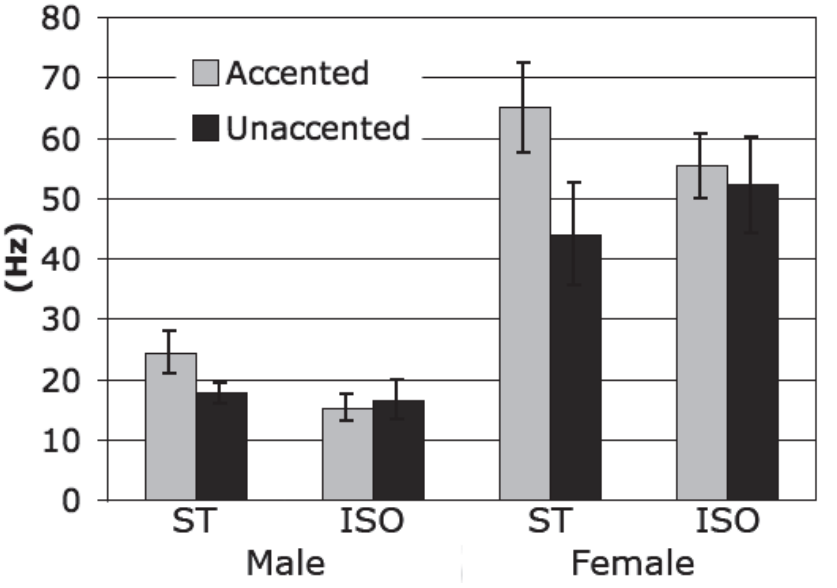
2.2.1. F_0 Rise

A three-way ANOVA on F_0 rise revealed significant effects of accent, $F1(1,8) = 36.581$, $p = .0003$; $F2(1,31) = 46.841$, $p < .0001$, and sex, $F1(1,8) = 33.192$, $p = .0004$; $F2(1,31) = 41.092$, $p < .0001$. The main effect of environment was reliable by words, $F2(2,62) = 19.954$, $p < .0001$, but only marginal by talkers, $F1(2,16) = 3.093$, $p = .0732$. The two-way interactions between accent and environment, and accent and sex were significant both by talkers and words: $F1(2,16) = 12.320$, $p = .0006$; $F2(2,62) = 8.787$, $p = .0004$, and $F1(1,8) = 12.324$, $p = .0080$; $F2(1,31) = 8.102$, $p = .0078$, respectively. The two-way interaction between environment and sex approached significance by words, $F2(2,62) = 3.065$, $p = .0538$, but it was not significant by talkers, $F1(2,16) = 1.859$, $p = .1879$. The three-way interaction of accent, environment, and sex was not significant either, $F1(2,16) = 2.612$, $p = .1043$; $F2(2,62) = 1.086$, $p = .3439$. The figure in (1) shows the F_0 rise values for final-accented and unaccented words. In order to illustrate the effects of accent in sentence and isolation, the data for the focus frame and non-focus frame were collapsed and shown as words produced in sentence. As seen in (1), the F_0 rise for final-accented and unaccented words did not differ reliably when they were produced in isolation. The error bars that indicate standard error of the mean overlap considerably for male and female talkers. By contrast, when words were produced in sentence, the F_0 rise was greater for final-accented words than unaccented words for both male and female talkers. The results show that, in spite of the traditional account of Japanese pitch accent that the two accent types are

[2] No clear, consistent effects of sentence type were observed in any of the analyses of F_0 rise, F_0 peak, or F_0 fall. Since males and females typically differ in their normal F_0 and therefore in F_0 range, significant main effects of sex were found in all analyses.

identical within words and differ only on the following particle (e.g. Kindaichi 1947, McCawley 1968), the two accent types were not the same within words when they were produced in sentence. The results were consistent with experimental findings (e.g. Poser 1984, Sugito 1979, Vance 1995). The figure in (2) shows the F_0 contours of the final accented word /mame^/ 'bean' and the unaccented word /mame/ 'diligence' and the following particle /to/, a citation marker, produced by a female talker, as created by Praat. The solid line shows the F_0 contour of the final-accented word and the dashed line shows the F_0 contour of the unaccented word. These F_0 tracks illustrate that F_0 rose steeply throughout the second mora for the final accented word and then dropped steeply into the particle. By contrast, the F_0 was relatively flat for the unaccented word. In fact, the F_0 was higher on the particle than on the second mora of the target word, which was not uncommon for unaccented words. The figure clearly shows that the two accent types differed within words as well as on the following particle.

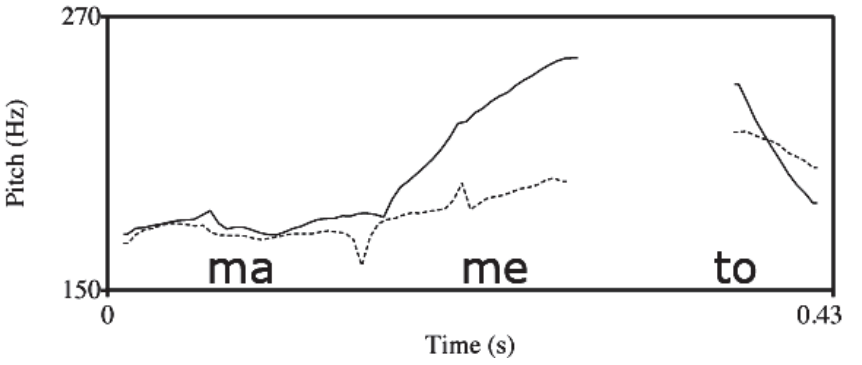
(1) Mean F_0 rise values for words produced in sentence (ST) and in isolation (ISO) for male and female talkers. In the legend, Accented stands for final-accented words. Error bars indicate standard error of the mean.



The results from the current study were consistent with the traditional claims (e.g. Kindaichi 1947, McCawley) that final-accented and unaccented words were virtually the same when they were produced in isolation. On the other hand, for words produced in sentence, the results were consistent with findings by Sugito

(1979) and Vance (1995), with final-accented words showing a greater rise than unaccented words.

(2) The pitch contours of a final-accented word (solid line) and an unaccented word (dashed line) produced by a female talker.



2.2.2. F_0 Peak

The statistical results for F_0 peak showed a very similar pattern to those for F_0 rise. In short, while final-accented words and unaccented words did not differ significantly when they were produced in isolation, the F_0 peak was significantly higher for final-accented words than unaccented words when they were produced in sentence. As mentioned earlier, Sugito (1979) and Vance (1995) had different views as to whether F_0 rise or F_0 peak was a better correlate of accent. Since reliable effects of accent were found on both measures in the present data, they were equally good acoustic correlates of accent.

2.2.3. F_0 Fall

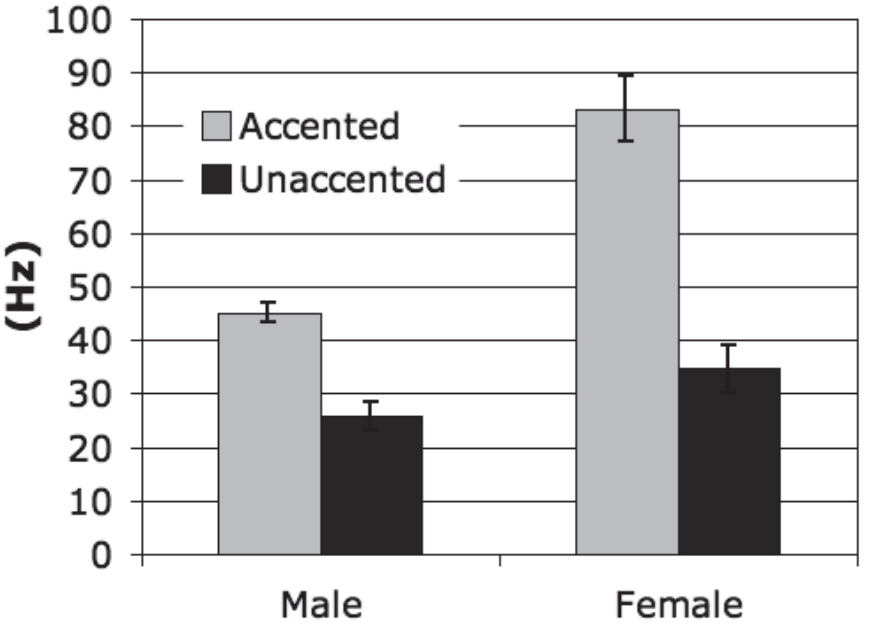
Traditionally, the contrast between final-accented and unaccented words has been thought to be found only in the F_0 fall. It remains high for unaccented words whereas it falls on the following particle. The F_0 fall was primarily measured to replicate this finding and thus to verify that the data collected in this study were not off the track.

As expected, a three-way ANOVA on F_0 fall found significant effects of accent, $F1(1,8) = 47.336$, $p = .0001$; $F2(1,35) = 180.844$, $p < .0001$, and environment, $F1(1,8) = 7.650$, $p = .0245$; $F2(1,35) = 11.653$, $p = .0016$. The effect of sex was significant by words, $F2(1,35) = 29.413$, $p < .0001$, although it was only marginal by talkers, $F1(1,8) = 4.548$, $p = 0.0655$. Accent reliably interacted with environment, $F1(1,8) = 18.303$, $p = .0027$; $F2(1,35) = 6.781$, $p = .0134$. The two-way interaction between accent and sex was significant by words, $F2(1,35) = 21.329$, $p < .0001$, but it was only marginal by talkers, $F1(1,8) = 4.552$, $p = .0654$. The two-way interaction between environment and sex was not reliable by either

talkers or words, $F1(1,8) < 1$; $F2(1,35) < 1$. The three-way interaction of accent, environment, and sex was significant by both talkers and words, $F1(1,8) = 16.853$, $p = .0034$; $F2(1,35) = 6.847$, $p = .0130$.[3] The effect of accent on F_0 fall is illustrated by the figure in (3).

As claimed in the traditional account and found in previous experimental work, the effect of accent was quite substantial, with final-accented words having a much greater F_0 fall than unaccented words (Kindaichi 1947, McCawley 1978, Poser 1984, Sugito 1979, Vance 1995). This can also be confirmed in the pitch track in (2). At the onset of the particle, the F_0 is higher for the final-accented word than for the unaccented word. However, the F_0 falls rapidly for the final-accented word to end up being lower than that for the unaccented word.

(3) Mean F_0 fall values for words produced sentence for male and female talkers. Error bars indicate standard error of the mean.



[3] Since two- and three-way interactions involving environment (only two levels, focus frame and non-focus frames, for F_0 fall) were reliable, a brief explanation is warranted here. These interactions seemed due to male and female talkers showing somewhat different patterns. Final-accented words had a greater F_0 fall than unaccented words for both speakers regardless of the sentence type. However, for females, the size of F_0 fall for unaccented words was not as much affected by the sentence type as that for males.

3. Perception Experiment

The production study found that final-accented and unaccented words did not differ in F_0 rise or F_0 fall when they were produced in isolation. By contrast, the two accent types differed significantly in F_0 rise and F_0 fall when they were produced sentence-medially followed by a particle. The perception experiment was conducted to examine perceptual cues for distinguishing the two accent types.

3.1. Method

Listeners. Thirty-two native speakers of Tokyo Japanese between the ages 19 and 35 years old were recruited at the University at Buffalo.

Stimuli. Recordings of one male and one female talkers collected in the production experiment were used to create three types of stimuli: 1) final-accented and unaccented words produced in isolation (isolation speech), 2) final-accented and unaccented words alone excised from a sentence (no particle speech), 3) words and the following particle excised from a sentence (particle speech). The stimuli for the particle speech were created by editing out the target word and the following particle from a sentence. The stimuli for the no particle speech were created by removing the particle from the particle speech. If listeners identify words in the isolation speech better than chance, it indicates that some acoustic cue(s) other than F_0 rise or F_0 peak are present, signaling the accent type. If listeners' performance is better than chance for the no particle speech, it shows that listeners can use the differences in F_0 rise and F_0 peak to distinguish the two accent types. Judging from the literature on Japanese pitch accent, it is likely that listeners can identify words with high accuracy for the particle speech.

Procedure. The listeners were told that they would hear Japanese words. At each trial, two alternatives appeared on a computer screen and the listeners' task was to choose the word they heard by pressing a key on a computer board (two-alternative forced choice task). Each listener heard the forty words once in each stimulus type in male and female voices. The total number of stimuli presented to each listener was: 40 (words)×3 (stimuli types)×2 (voices)=240.

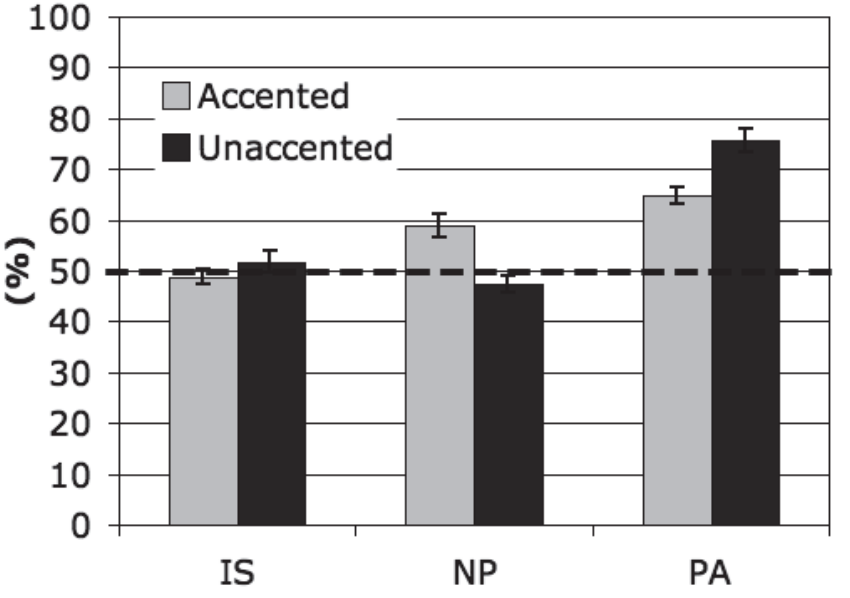
3.2. Results & Discussion

A three-way ANOVA was performed for correct responses, with either listeners (*F1*) or final-accented and unaccented words (*F2*) as the repeated measures. All of the three factors examined were within-listeners factors: accent (final-accent or no accent), stimulus type (isolation speech, no particle speech, particle speech), and voice (male or female). There were no between-listeners factors. After all the data were collected, the percentage of correct responses was calculated by dividing the number of correct responses by the number of responses provided by the listeners. The trials that listeners failed to respond to within the four seconds of time limit were treated as missing data and not included in the statistical analyses. Out of 5520 trials presented to the participants, only 18 trials resulted in missing data.

The three-way ANOVA revealed a significant effect of stimulus type, $F1(1,22) = 64.504$, $p < .0001$; $F2(2,38) = 29.518$, $p < .0001$. The effect of voice

was marginal by listeners, $F1(1,22) = 2.954$, $p = .0997$, but it was not significant or marginal by words, $F2(1,19) = 1.215$, $p > .10$. The main effect of accent was not significant, $F1(1,22) = .175$, $p > .10$; $F2(1,19) = .008$, $p > .10$. A significant two-way interaction was found between accent and stimuli, $F1(2,44) = 21.541$, $p < .0001$; $F2(2,38) = 5.852$, $p = .0061$. Neither the interaction between stimuli and voice nor accent and voice was reliable, $F1(2,44) = 2.346$, $p < .1076$; $F2 < 1$, and $F1 < 1$; $F2 < 1$, respectively. There was also a three-way interaction of accent, stimuli, and voice, $F1(2,44) = 5.592$, $p = .0069$; $F2(2,38) = 3.609$, $p = .0367$. The results are illustrated by the figure in (4). In the figure, the results for male and female voices are collapsed to show the effect of stimulus type on the number of correct responses. Keep in mind that what is at issue here is whether or not word identification was better than 50 percent for both accent types. Recall that the listeners' task was two-alternative forced choice. Thus, if one accent type has high accuracy well over 50 percent but the other well below 50 percent, it only means that listeners' responses are biased. It is only when both final-accented and unaccented words have reasonably high accuracy that the words are considered to be intelligible.

(4) IS, NP, and PA on the x-axis stand for isolation speech, no particle speech, and particle speech, respectively. Error bars indicate standard error of the mean.



For the isolation speech, correct identification was chance for both accent types. This suggests that no acoustic properties other than F_0 rise or F_0 peak were present in the signal to convey accent information, making the two accent types unintelligible in terms of accent. Unlike earlier studies, no response bias for final-accented words was observed. The accuracy was slightly higher for the no particle speech than the isolation speech, but on average, the accuracy was only slightly higher than 50 percent. This result indicates that the F_0 peak and F_0 rise differences observed for words produced in sentence were not useful to listeners for identifying words. As expected, the percentage of correct responses exceeded 50 percent for both accent types for the particle speech. The accuracy was about 65 percent for final-accented words and 76 percent for unaccented words. However, it should be added that the listeners' performance was not as good as suggested in the literature, even when the F_0 fall information was present.

4. Conclusion

The primary goal of this study was to establish the acoustic correlates and perceptual cues to distinguish final-accented and unaccented words. Twenty minimal pairs of final-accented and unaccented words that have a high familiarity searched from a database of Japanese words were used as test items.

The results of this study show that, in order for final-accented and unaccented words to be realized differently, they have to be produced with a following particle. Put differently, realizing the contrast is, as it were, all or nothing. When it is realized at all, multiple correlates are observed: F_0 rise, F_0 peak, and F_0 fall. Otherwise the contrast is neutralized, at least in terms of these three measures. The results are summarized in (5). Among the three correlates, the largest difference appears in the F_0 fall from the second mora into the following particle. Contrary to the traditional account, final-accented and unaccented words themselves are produced differently when they occur sentence-medially before a particle, with final-accented words having a higher F_0 peak on the second mora and a greater F_0 rise from the first to second mora. By contrast, when the two types of words are produced in isolation, not only is there no particle that indicates accent information on the following words, but also the difference on the F_0 peak also disappears, making the two accent types identical.

(5) Results Summary

	Isolation		Sentence	
	Production	Perception	Production	Perception
F_0 peak	×	×	√	×
F_0 rise	×	×	√	×
F_0 fall	—	—	√	△

Note: The symbol × in the isolation columns indicates that no significant difference was found between final-accented and unaccented words. The symbols √/× in the sentence columns indicate good/poor accuracy in listeners' word identification. The symbol △ indicates that word identification was better than chance but was not as good as expected.

The perception study was conducted to examine if listeners actually use the acoustic differences between final-accented and unaccented words to identify words. The results of the no particle speech show that, in spite of the F_0 rise and F_0 peak differences between the two accent types, these differences alone are not enough for listeners to distinguish them. The listeners' performance improves with additional F_0 information from the following particle, but the accuracy reaches only about 70 percent. Considering that the F_0 on the following particle has been argued as if it is the "dead giveaway" for the accent type, its contrastive function is not as clear as has been suggested in the literature. This result suggests that, in normal conversation, listeners may rely on context to distinguish final-accented and unaccented words.

References

Amano, Shigeaki, and Tadanori Kondo. 1999. *Nihongo-no Goitokusei*. Tokyo: Sanseido.

Amano, Shigeaki, Tadahisa Kondo, and Kazumi Kato. 1999. Familiarity Effect on Spoken Word Recognition in Japanese. *Proceedings of the 14th International Congress of Phonetic Sciences* 2:873-876.

Boersma, Paul, and David Weenink. 2005. Praat: Doing Phonetics by Computer [Computer Program]. http://www.praat.org/.

Han, Mieko. 1962. *Japanese Phonology: An Analysis Based upon Sound Spectrograms*. Tokyo: Kenkyusha.

Hasegawa, Yoko, and Kaue Hata. 1992. Fundamental Frequency as an Acoustic Cue to Accent Perception. *Language and Speech* 35:87-98.

Kindaichi, Haruhiko. 1947. Tookyoogo ni okeru "hana" to "hana" no kubetsu – tookyoo akusento shinnidankan kyoochooron. [The distinction between "hana 'flower'" and "hana 'nose'" in the Tokyo dialect – the emphasis of the new two-level theory of the Tokyo accent]. *Kokugo* [The Japanese language], Summer.

Kitahara, Mafuyu. 2001. Category Structure and Function of Pitch Accent in Tokyo Japanese. Ph.D. diss., Indiana University.

Kubozono, Haruo. 1993. *The Organization of Japanese Prosody*. Tokyo: Kuroshio.

McCawley, James. 1968. *The Phonological Component of a Grammar of Japanese.* The Hague: Mouton.

Neustupný, Jirí. 1978. *Post-Structural Approaches to Language: Language Theory in a Japanese Context*. Tokyo: University of Tokyo Press.

Pierrehumbert, Janet, and Mary Beckman. 1988. *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Poser, William. 1984. The Phonetics and Phonology of Tone and Intonation in Japanese. Ph.D. diss., Massachusetts Institute of Technology.

Sugito, Miyoko. 1979. Tookyoo akusento ni okeru "hana" to "hana" no hatsuwa to chikaku ni tsuite. *The 79th Meeting of the Linguistic Society of Japan* (in Sugito 1998).

Sugito, Miyoko. 1998. *Nihongo onsei no kenkyu 5: "hana" to "hana."* Osaka: Izumi Shoin.

Sugiyama, Yukiko. 2008. The Nature of Japanese Pitch Accent: An Experimental Study. Ph.D. diss., University at Buffalo, The State University of New York.

Vance, Timothy. 1995. Final Accent vs. No Accent: Utterance-Final Neutralization in Tokyo Japanese. *Journal of Phonetics* 23:487-499.

Wright, Richard. 1997. Lexical Competition and Reduction in Speech: A Preliminary Report. *Progress report* 21, Speech Research Laboratory, Bloomington, IN.

Yukiko Sugiyama
University at Buffalo, The State University of New York
Department of Linguistics
609 Baldy Hall
Buffalo, NY 14260-1030

yukiko_sugiyama@mac.com