2012. Proceedings of the Annual Meeting of the Berkeley Linguistics Society 38. DOI: http://dx.doi.org/10.3765/bls.v38i0.3348
Published for BLS by the Linguistic Society of America

Perceiving pitch accent in the absence of F0

YUKIKO SUGIYAMA
Keio University, Japan

1 Introduction

In Tokyo Japanese[1], the suprasegmental property of fundamental frequency (F0) is used to distinguish words in addition to segmental information. For example, the phoneme sequence of /ame/ means 'rain' when its first syllable is on a high pitch and its second syllable is on a low pitch.[2] By contrast, it means 'candy' when its first syllable is on a low pitch and its second syllable is on a high pitch. When there is a pitch fall as observed in from the end of the first syllable into the second syllable of 'ame,' the syllable immediately preceding the fall is said to have pitch accent. While studies to date have shown that the F0 is the most dominant cue for pitch accent, it is not certain if secondary cues exist. Past production studies measured duration, and properties related to amplitude and devoicing (e.g. Beckman 1986, Kaiki, Takeda, and Sagisaka 1992, Lovins 1976, Weitzman 1970, Yoshida 2002), but their results as a whole do not present a consistent picture as to whether secondary cues exist. Perception studies that used naturally produced whispered speech suggest that listeners can perceive accent information even when words are produced without vocal fold vibration. Sugito, Higashiyama, Sakakura, and Takahashi (1991) found that listeners were able to identify the words produced in whisper with roughly 90 percent accuracy. In a similar vein, Liu and Samuel (2004) found that monosyllabic Mandarin words produced in whisper were identified fairly accurately. However, this study also found that when Mandarin speakers spoke the words in whisper, they had a tendency to enhance secondary cues compared to when they produced the words normally. Liu and Samuel's findings are informative when one tries to examine secondary cues to pitch accent. While most studies that examined secondary prosodic cues dealt with Indo-European languages that have stress-accent, Liu and Samuel's study showed that secondary cues can exist in a tone language as well. In addition, their findings show that the properties that are present in whispered speech are not necessarily present in speech produced normally. In other words, one cannot examine whispered speech to determine the existence of secondary cues in normal speech. For this reason, the present

[1]Tokyo Japanese is a variety of Japanese which is often associated with standard Japanese. Since this study deals with only Tokyo Japanese, it will be simply referred to as Japanese hereafter.

[2]Although F0 and pitch are not the same, the terms F0 and pitch will be used interchangeably in this paper.

study investigates secondary cues to pitch accent by using speech stimuli whose F0 had been artificially removed from words produced normally and replaced by white noise.

2 Method

2.1 Stimuli

The target words were minimal pairs of final-accented words and unaccented words that differ only in accent, such as /haná/, which means 'flower' and /hana/, which means 'nose.'[3] Two words from a minimal pair have the same phoneme sequence and the pitch pattern of low-high. The only difference between them is that, at least at the phonological level, while final-accented words have accent on its final syllable, unaccented words have no accent. Using an electronic dictionary (Amano & Kondo, 1999), minimal pairs analogous to <hana>[4] were thoroughly searched, resulting in 14 minimal pairs.[5] See Appendix for a complete list of words.

The original speech stimuli were produced by a female speaker who grew up in the Tokyo area and whose parents were also from the area. The target words were spoken in the following carrier sentence:

(1) *Kare wa* ___ *ga* *ii.*
he TOIPC NOMINATIVE good
"he wants ___ ." or "he has a sensitive ___ ."

The sentence can have either of the two meanings indicated in (1) above depending on the meaning of the target word embedded in the sentence.

Twenty-eight words were naturally produced twice in the carrier sentence, which were used as the natural speech stimuli in the perception experiment. They were recorded to disk on a computer at the sampling rate of 44.1 kHz with 16-bit resolution and then normalized for peak amplitude. Based on these natural speech stimuli, "whispered"[6] speech stimuli were generated by running a script on the Praat speech analysis software (Boersma & Weenink, 2011). The default parameter settings for LPC analysis-resynthesis were used with the window length of 25 ms, and the time step of 5 ms. The periodicity of the F0 in the

[3]The symbol " ´ " indicates that the syllable is accented.

[4]Angled brackets "< >" are used here to refer to both of the words from a minimal pair which has the same phoneme sequence indicated between them.

[5]Originally, 20 minimal pairs were found by searching disyllabic minimal pairs that had a relatively high familiarity rating in (Amano & Kondo, 1999). However, six pairs were removed from the list because they are usually used as a part of compound words and do not occur by themselves.

[6]The term "whispered" is in double quotation marks here because the stimuli used in this study are not real whispered speech produced by a human. Rather, as already explained, they were whisper-like stimuli that were created artificially by replacing the F0 in natural speech by random noise. This nature of stimuli should be emphasized for reasons discussed in Introduction. However, for the sake of simplicity, the term will be used without quotation marks hereafter.

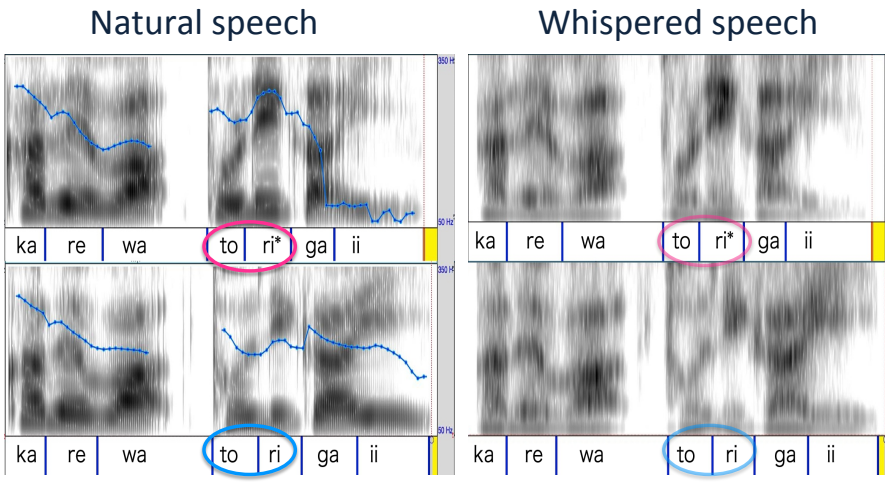

Figure 1: Spectrograms of the final-accented word /torí/(an upper panel) and the unaccented word /tori/(a lower panel) produced in the carrier sentence. In the annotation below the spectrograms, the target words are circled and accented syllables are marked with asterisks. The circles in pink indicate accented words and those in blue indicate unaccented words.

original natural speech was removed and replaced with white noise. Figure 1 shows snapshots of spectrograms on Praat. The two snapshots on the left are original normal utterances of the final-accented word /torí/ 'the last person to perform on the stage'(an upper panel) and the unaccented word /tori/ 'bird'(a lower panel) spoken in the carrier sentence. The two snapshots on the right are their whispered versions. As seen in the figure, while blue lines with dots that track F0 can be confirmed in the original speech, they are not present in their whispered counterparts.

2.2 Listeners

The listeners were twenty-two native speakers of Tokyo Japanese who were between 18 and 21 years old. They were recruited at Keio University in Yokohama. The participants grew up in the Tokyo area where Tokyo Japanese is spoken. In addition, both of their parents were also from the Tokyo area. None of them reported any history of a hearing or speaking disorder. The experiment lasted about for an hour for each listener.

2.3 Procedure

Each listener was run individually in a quiet room. The stimuli were presented using the SuperLab stimulus presentation software on a MacBook Pro, to which a Cedrus response pad RB-730 and SONY MDR-CD900ST headphones were connected. The stimuli were played at a comfortable listening level.

Before the actual trials started, the participants were shown flash cards on which the target words that would be presented to them were written. The words were written in Chinese characters, in *hiragana*, a Japanese syllabary[7], or in a combination of the two, depending on how they were commonly written in Japanese. Even though all the words were familiar to Japanese speakers, the subjects went through them, as many Chinese characters can be read in more than one way. In order for the target words to make minimal pairs, they had to be read in a certain way. In the actual experimental trials, the listeners' task was to identify the words they heard from the two alternatives provided (forced choice). At each trial, two alternatives, a final-accented word and its unaccented counterpart, appeared on the computer screen. One alternative appeared on the right side of the screen and the other appeared on the left side of the screen. After the presentation of an audio stimulus, the listeners pressed a button that corresponded to the word they think they heard. When a subject failed to respond within four seconds, the trial was treated as a missed trial and the next stimulus was presented.

Each listener received eight blocks of natural speech stimuli and eight blocks of whispered speech stimuli. Half the listeners heard eight blocks of natural speech first and then heard eight blocks of whispered speech. The remaining half heard eight blocks of whispered speech first and then heard eight blocks of natural speech. One block consisted of 28 words (14 pairs) of either natural speech or whispered speech presented in a random order. In presenting two alternatives on the computer screen, two versions were created. For one version, a final-accented word appeared on the right side of the screen and its unaccented counterpart appeared on the left. The sides on which the two alternatives appeared on the screen were switched for the other version. In addition, in order to avoid any idiosyncratic properties of a given stimulus token to affect the listeners' judgment, two repetitions of the same word were used. Since two tokens of a word were presented twice with two versions of presenting the alternatives ($2\times2\times2$), the listeners heard a total of eight tokens for each word.

3 Results and discussion

The data for one listener were omitted from the analysis. It turned out after the data were collected that neither of his parents was from the Tokyo area and the listener himself spent a few years of his childhood in an area where a variety other than Tokyo Japanese was spoken.

[7]Strictly speaking, not all *hiragana* characters correspond to one syllable. However, it will be sufficient to say so for the purpose of the present study.

Each listener heard a total of 448 trials (28 words×2 tokens×4 repetitions=224 tokens each of both natural and whispered speech). Out of 9408 trials (448 trials×21 listeners), 25 trials had no responses, of which eight were from natural speech stimuli and 17 were from whispered speech. In terms of percentage, the number of missed trials accounted for only 0.3 percent of all trials presented. In addition, these missed trials did not concentrate on certain words or listeners. Based on the responses of 9383 trials, the mean accuracy was computed for each pair of words separately for natural speech and whispered speech. The listeners' accuracy of word identification for natural speech provides an informative baseline in interpreting their performance on the whispered speech stimuli. As Table 1 shows, the listeners' accuracy exceeded 90 percent for all pairs, except <moti>, with which the accuracy was below 60 percent. Since it is difficult to interpret the whispered stimuli data when the accuracy is so low for the natural speech stimuli, <moti> was be left out of further analysis. For the rest of the words, final-accented words and unaccented words were identified fairly well for natural speech with the mean accuracy of 94.4 percent. In fact, many of the pairs were over 95 percent correct, indicating that the final-accented words and unaccented words were quite intelligible, even though they differ only in accent. Not surprisingly, whispered speech had much lower accuracy with the mean of 64.8 percent. However, the fact that the accuracy was over 50 percent for all the pairs suggests that the listeners' performance was not at random. In other words, there was some acoustic information in the whispered stimuli that the listeners utilized as cues to pitch accent.

Once the accuracy was computed for each word in the whispered speech, a planned one-sample t-test was conducted to determine if final-accented and unaccented words were identified reliably better than chance. In order to conduct the analysis, first, the mean accuracy was calculated for the final-accented words and unaccented words for each listener as shown in Table 2. Then, the mean accuracy for each pair of words was compared against the chance level of 50 percent. The t-test found a significant result ($t(20) = 7.58$, $p < 0.001$), indicating that the listeners' performance on whispered speech was reliably above chance.[8] As explained earlier, the whispered stimuli used in the present study were created artificially by removing only the periodicity in the original natural speech. Since the remaining acoustic properties were preserved, the result strongly suggests that some acoustic properties other than the F0 were present in the stimuli, which enabled the listeners to distinguish final-accented and unaccented words.

The data collected were further assessed with a repeated measures ANOVA. The factors examined were accent (accented words *vs.* unaccented words), speech style (natural speech *vs.* whispered speech), and order (whether the listeners heard natural speech first or whispered speech first). The first two factors were within-subjects factors and the third was a between-subjects factor. As expected, the analysis revealed a significant main effect of speech style, $F(1,19)$

[8]When a t-test was conducted including the pair <moti>, the result was still significant ($t(20) = 7.47$, $p < 0.001$).

Table 1: Correct responses (%) of each pair heard in natural and whispered speech (standard errors in parentheses)

Words	Natural speech	Whispered speech
<haji>	95.7 (1.85)	64.1 (2.57)
<hana>	96.4 (1.59)	66.4 (2.84)
<hane>	99.1 (0.65)	65.0 (3.59)
<hasi>	94.0 (1.90)	62.5 (3.05)
<hati>	98.2 (0.76)	54.1 (2.65)
<mame>	97.4 (1.02)	62.9 (2.74)
<moti>	58.4 (7.80)	50.4 (4.04)
<nami>	89.6 (3.88)	66.3 (4.41)
<nori>	95.5 (1.37)	69.3 (3.11)
<osu>	97.0 (1.59)	65.5 (3.41)
<sita>	94.6 (1.69)	66.8 (2.77)
<tama>	92.2 (2.55)	64.1 (4.75)
<tori>	96.3 (1.26)	77.3 (3.02)
<tume>	95.1 (1.92)	59.3 (2.40)

Table 2: Accuracy (%) of final-accented and unaccented words in whispered speech by each listener (standard error in parentheses)

	Final-accented	Unaccented		Final-accented	Unaccented
1	40.4 (4.93)	76.5 (5.68)	11	55.8 (7.83)	43.5 (7.15)
2	75.0 (4.93)	47.6 (4.67)	12	83.7 (3.58)	31.7 (6.88)
3	57.3 (7.65)	75.7 (4.79)	13	59.6 (8.39)	66.0 (8.54)
4	78.8 (7.67)	78.8 (7.41)	14	53.5 (5.18)	62.5 (6.00)
5	72.1 (4.51)	56.7 (6.58)	15	70.2 (5.40)	71.6 (7.11)
6	62.5 (6.17)	80.8 (5.59)	16	54.8 (5.21)	65.4 (4.93)
7	87.5 (4.90)	64.4 (8.11)	17	56.7 (7.02)	60.6 (4.87)
8	57.7 (6.87)	84.6 (6.03)	18	51.9 (5.81)	63.5 (5.91)
9	42.3 (7.69)	68.0 (7.44)	19	60.6 (4.66)	70.2 (5.40)
10	51.9 (7.05)	97.1 (1.52)	20	84.6 (4.93)	82.7 (5.01)
			21	51.5 (9.33)	59.8 (7.01)

= 462.8, $p < 0.001$, indicating that subjects were more accurate with natural speech than whispered speech. The interaction between order and speech style was also reliable, $F(1,19) = 8.5$, $p < 0.01$. The other factors produced no main effects or interactions. There was no main effect of either accent or order, $F(1,19) = 2.4$, $p > 0.10$; $F(1,19) = 0.4$, $p > 0.10$ respectively, and neither were there significant interactions between accent and order or accent and speech style, $F(1,19) = 1.3$, $p > 0.10$; $F(1,19) < 0.1$, $p > 0.10$ respectively. There was no three-way interaction of accent, order and speech style either, $F(1,19) = 1.2$, $p > 0.10$.[9] The effect of speech style can be seen in Figure 2, where the accuracy was close to 100 percent for natural speech (the two boxes on the left side) while the accuracy was clearly lower for whispered speech (the two boxed on the right side).

Because the interaction between order and speech style was reliable, two-way repeated measures ANOVAs were conducted separately for the listeners who heard natural speech first and those who heard whispered speech first. The factors analyzed were speech style (natural speech *vs.* whispered speech) and accent (final-accented *vs.* unaccented words). For the group of listeners who heard natural speech first (the natural speech group), there was a significant main effect of speech style, $F\ (1,\ 9) = 300.4$, $p < 0.001$, while the effect of accent was not significant, $F\ (1,\ 9) = 2.5$, $p > 0.10$. There was no interaction between speech style and accent, $F\ (1,\ 9) = 0.77$, $p > 0.10$. The results were similar for the group of listeners who heard whispered speech first (the whispered speech group). While the main effect of speech style was significant, $F\ (1,\ 10) = 218.6$, $p < 0.001$, the effect of accent was not significant, $F\ (1,\ 10) = 0.20$, $p > 0.1$. The interaction of speech style and accent was not significant, $F\ (1,\ 10) = 0.36$, $p > 0.1$. The analyses of the natural speech group and the whispered speech group indicate that, within each group, the type of stimuli (natural speech or whispered speech) was the only factor that had a consistent effect on the listeners' performance. As already mentioned, word identification was much better for natural speech than whispered speech. In addition, the listeners' performance appears to have varied to a greater extent for whispered speech than natural speech. In Figure 2, the data values are more widely distributed for whispered speech than for natural speech. It suggests that, in the absence of the primary cue to pitch accent, some listeners were better at picking up secondary cues to pitch accent than others.

Since accent had no main effect or interaction, the effects of order and speech style were further assessed with a repeated measures ANOVA with the factor of accent collapsed. The analysis found a significant main effect of speech style

[9]Because the set of words used in this study are a unique and exhaustive set of words that met the criteria discussed in the stimuli section, items analysis is not necessary. However, a three-way ANOVA was performed with words as repeated measures only for this analysis in order to ensure that the results of two types of analyses do not diverge substantially. The results were similar, except, in addition to significant main effect of speech style and interaction of speech style and order ($F(1,12) = 316.7$, $p < 0.001$; $F(1,19) = 24.9$, $p < 0.001$ respectively), which were also observed in the subjects analysis, the main effect of order and the three-way interaction of accent, order and speech style also became significant ($F(1,12) = 5.9$, $p < 0.05$; $F(1,12) = 8.9$, $p < 0.05$ respectively).

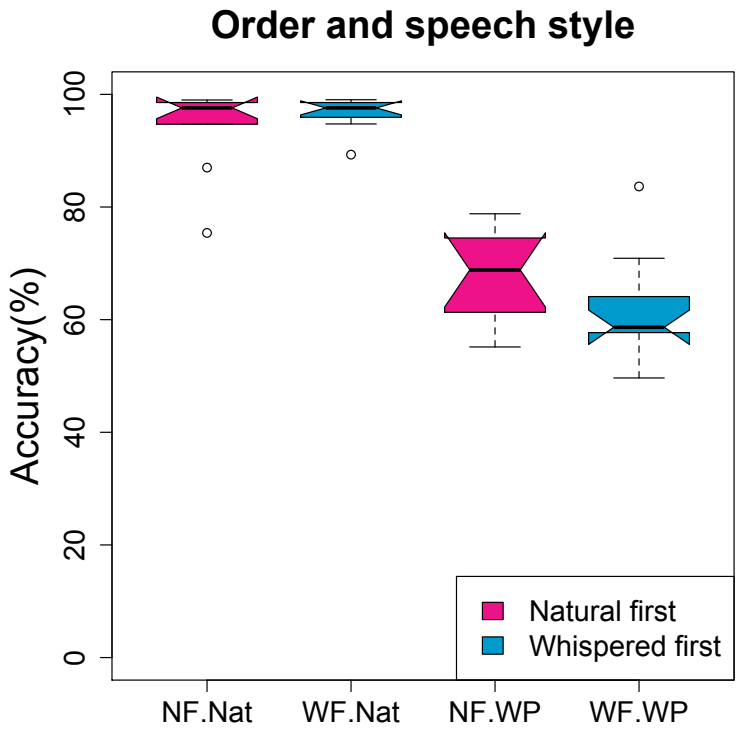

Figure 2: Accuracy for listeners who heard natural speech first and for listeners who heard whispered speech first. In the horizontal axis, "Nat" stands for natural speech and "WP" whispered speech.

and a significant interaction of order and speech style, $F(1,19) = 462.8$, $p < 0.001$; $F(1,19) = 8.5$, $p < 0.01$, respectively. The main effect of order did not reach significance, $F(1,19) = 0.40$, $p > 0.10$. The results indicate that listeners' performance was affected not only by the type of speech stimuli but also by whether they heard natural speech first or whispered speech first. Regardless of whether the listeners heard natural speech or whispered speech first, both speech groups seem to have taken advantage of the experience of being exposed to the first type of stimuli, whichever that may have been, when they heard the second type, although the extent to which they exploited the experience of perceiving the first type of stimuli seems to vary between the groups. Admittedly, the difference is very small, but the whispered speech group did slightly better on natural speech than the natural speech group. Similarly, the natural speech group did better on whispered speech than the whispered speech group. The whispered speech group did not show much improvement on natural speech probably because of a ceiling effect. On the other hand, an exposure to the natural speech stimuli seemed to have helped the natural speech group perform their task with the whispered speech stimuli.

4 Conclusions

The present study aimed at examining whether or not acoustic properties other than the F0 exist in normal speech as secondary cues to Japanese pitch accent. The method adopted in the study ensured that the only difference between natural and whispered speech would be the presence or absence of periodicity in F0. The results of whispered speech found that the listeners were able to distinguish final-accented and unaccented words reliably better than chance, which supports the evidence of secondary cues to Japanese pitch accent. It also suggests that previous studies were not able to identify secondary cues because they manifest themselves in forms other than duration, devoicing, or intensity. In addition, further analysis found that the listeners' performance with whispered speech was better when they were first exposed to natural speech than when the first stimuli they received was whispered speech. Further research is needed to understand exactly what aspects of hearing natural speech facilitated the listeners to perceive pitch accent in whispered speech. In addition, acoustic analysis needs to be done in order to determine what acoustic property in the whispered stimuli served as cues for the listeners to distinguish final-accented and unaccented words.

References

Amano, S., & Kondo, T. (1999). *Nihongo-no goitokusei* [Lexical Properties of Japanese]. Tokyo: Sanseido.

Beckman, M. E. (1986). *Stress and Non-stress Accent.* Dordrecht, Holland: Foris.

Boersma, P., & Weenink, D. (2011). *Praat: doing phonetics by computer*. [Computer program] retrieved from http://www.praat.org/.

Kaiki, N., Takeda, K., & Sagisaka, Y. (1992). Vowel duration control using linguistic information. *The Journal of IEICE*, *J75-A*, 467-473.

Lovins, J. B. (1976). Pitch accent and vowel devoicing in Japanese. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, *10*, 113-125.

Sugito, M., Higashikawa, M., Sakakura, A., & Takahashi, H. (1991). Perceptual, acoustical, and physiological study of Japanese word accent in whispered speech. *IEICE technical report*, *91*, 1-8. Retrieved from `http://db.ieice.org/gakkai/show.php?id=23318`

Weitzman, R. S. (1970). *Word accent in Japanese*. Detroit, MI: Management Information Services.

Yoshida, N. (2002). The effects of phonetic environment and vowel devoicing in Japanese (in Japanese, abstract in English). *Kokugogaku*, 34-47, 109. Retrieved from `http://ci.nii.ac.jp/naid/110002533186/`

Appendix

		Final-accented	**Unaccented**
1.	/haji/	恥 ‘shame’	端 ‘edge’
2.	/hana/	花 ‘flower’	鼻 ‘nose’
3.	/hane/	跳ね ‘jump’	羽 ‘feather’
4.	/hasi/	橋 ‘bridge’	端 ‘edge’
5.	/hati/	八 ‘eight’	蜂 ‘bee’
6.	/mame/	豆 ‘bean’	まめ ‘hardworking’
7.	/moti/	持ち ‘durability’	餅 ‘rice cake’
8.	/nami/	波 ‘wave’	並 ‘mediocre’
9.	/nori/	海苔 ‘seaweed’	乗り ‘ride’
10.	/osu/	雄 ‘male’	お酢 ‘vinegar’
11.	/sita/	舌 ‘tongue’	下 ‘below’
12.	/tama/	玉 ‘ball’	たま ‘infrequent’
13.	/tori/	取り ‘share’	鳥 ‘bird’
14.	/tume/	詰め ‘stuffing’	爪 ‘nail’

Precisely speaking, the symbol “r” is a tap “ɾ”. However, “r” is used instead in this paper.

Yukiko Sugiyama
Keio University, Faculty of Science and Technology
4-1-1 Hiyoshi, Kohoku-ku, Yokohama, Japan 223-8521

sugiyama@hc.st.keio.ac.jp