Proceedings of the 21st Annual Meeting of the Berkeley Linguistics Society, 141-51, 1995.

The function of F0-peak delay in Japanese

Yoko Hasegawa
University of California, Berkeley

Kazue Hata
Speech Technology Laboratory, Panasonic Technologies, Inc.

I. INTRODUCTION

There is a perceptual difference between male and female speech in fundamental frequency (F0) and, less significantly, in formant frequencies. Many, naturally, assume that these characteristics are physiologically determined to a great extent, and thus speakers exert little control over these characteristics in their normal utterances. F0 is certainly influenced by anatomical differences between male and female speakers. Nonetheless, observed F0 differences are much greater than those that can be attributed to anatomy: F0 is also manipulated by speakers to display their demeanor, for example in conforming to social norms.

The present study reports on one such F0 manipulation, the paralinguistic function of F0-peak delay in Japanese. We claim that F0 in Japanese is not only modified to indicate accentual distinctions but can also be used to convey femininity.

In this paper, we first provide a brief review of previous studies on male-female speech differences, and then explain how Japanese pitch accent is realized. Section II reports on our experiment to test the hypothesis that F0-peak delay and femininity are associated. Section III discusses the similarities and differences in functions of F0-peak delay between pitch-accent and intonation languages.

1.1 Comparisons of Male and Female Speech

Male speech and female speech are sufficiently different that people usually have no difficulties in identification of a speaker's gender. Usually, female speakers have higher F0 (Peterson & Barney, 1952) and higher formant frequencies (Chiba & Kajiyama, 1941), with the F0 difference being more significant than the formant-frequency differences between male and female groups (Coleman, 1976). These acoustic differences are partly due to anatomical differences between males and females. For example, the length of adult male vocal folds (cords) is about 15 to 20 mm, whereas the length of adult female's is 9 to 13 mm (Sundberg, 1987). Furthermore, the average vocal tract length of an adult male is 17.5 cm (Pickett, 1980), and that of an adult female is about 15% shorter (Nordström, 1977).

Although there is no dispute as to whether these anatomical differences are a cause of different F0 ranges observed between male and female speakers, the actual F0 differences cannot be explained solely in anatomical terms. F0 is also

manipulated, whether consciously or unconsciously, by both gender groups (Sachs et al., 1973).[1] Ohala (1983) contends from an ethological point of view that not only humans but also other species manipulate F0 to convey and clarify their intentions and attitudes. He claims that a confident or dominant individual utters low-pitched and often harsh sounds, while a submissive or subordinate individual utters high-pitched and tone-like sounds — he calls this phenomenon the *frequency code*, which evolved originally from the association between body size and emitted frequency for non-linguistic communication. This frequency code is used in linguistic communication as well. Human males frequently use low F0, which is associated not only with a large body size but also with character traits such as aggressiveness, assertiveness, self-confidence, and so forth. By contrast, high F0 suggests that the speaker has a small body, is non-threatening, submissive, subordinate, and in need of others' cooperation and good will (ibid.). In addition, when humans ask a question, a high pitch is a natural choice to signal their dependency on others. But when they make a statement, a low pitch would be more appropriate to show their confidence (ibid.).

Other researchers have postulated that F0 differences are more of a social phenomenon, reflecting different social norms laid down for men and women (Trudgill, 1974; Brend, 1975; Lakoff, 1975; Brown & Levinson, 1978; Edelsky, 1979; Jugaku, 1979; Loveday, 1981). Society assigns different social roles to men and women and expects different behavioral patterns from each group; language simply reflects this social fact (Trudgill, 1974). American females, for example, make great use of rising intonation (Brend, 1975), and, when using formal and polite register, Japanese female speakers use higher F0 than female speakers of other languages (Jugaku, 1979; Loveday, 1981). The frequent use of higher F0 and rising intonation are non-physiological aspects of female speech.

1.2 Ososagari: Delayed F0 Fall

Some Japanese speakers, especially female speakers, tend to delay the F0 peak that signals a lexical accent. This phenomenon is known as *ososagari* (*delayed F0 fall*), which is briefly explained in this section.

The Tokyo dialect of Japanese is a prototypical pitch-accent language in which accent is realized solely by a change in pitch, not by a change in loudness or duration such as found in English. Phonologically, it is widely assumed that the accented syllable in Japanese has a high tone, and the post-accent syllable a low tone; phonetically, the accentual high tone is realized by a higher F0 value on the accented syllable than on surrounding syllables (Pierrehumbert & Beckman, 1988). This accentual F0 peak, however, occasionally occurs on the post-accent syllable, without listeners detecting any change in accent placement. For example, in the word /námida/ 'tears', although the lexical accent falls on the first syllable /na/, the actual F0 peak may occur on the second syllable /mi/, as shown in Figure 1.

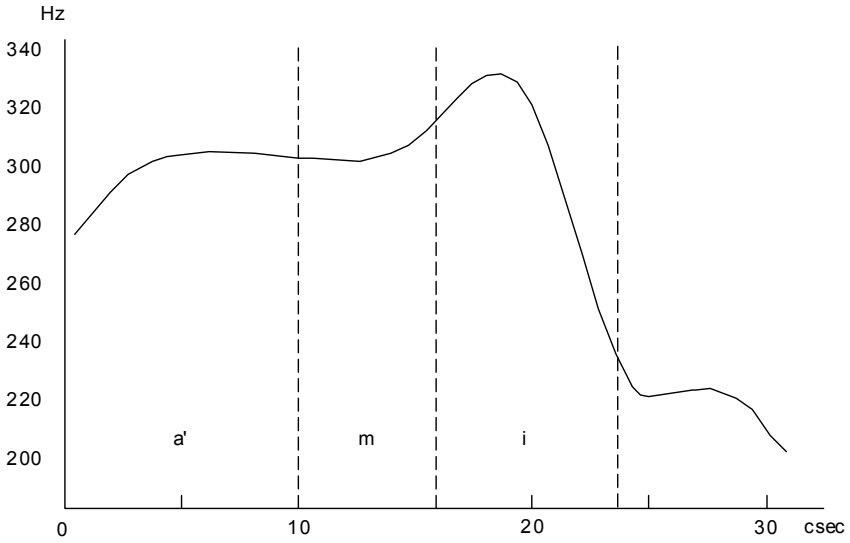

Figure 1: F0 Contour of /ámi/ in the word *námida* 'tears'

However, the first syllable /a/ is still perceived as accented in tokens with F0-peak delay, even though the actual F0 peak does not occur on it.[2] Listeners are not consciously aware of the delay and do not consider that the accent is shifted.

Investigating the ososagari phenomenon, Sugito (1968) discovered that the most significant acoustic correlate of the Japanese accent is a falling F0 contour of the post-accent syllable, rather than the F0-peak location itself; i.e., native speakers of Japanese perceive an accent on a syllable when it is followed by a falling F0 contour. In the /námida/ example above, if the post-accent syllable /mi/ contains an F0 fall, /na/ is perceived as accented, even when the F0 peak occurs on /mi/.

Although the frequency of ososagari varies between speakers as well as between words uttered by the same speaker, it is more commonly observed in words with an accent in certain positions or certain segmental environments. In Sugito's data, about 36% of 3-, 4-, 5-syllable words with the accent on the initial syllable had delayed F0 peak. In Hasegawa and Hata (1988), we reported that 37% of the words beginning with /(C)VmV/ were uttered with delayed F0 peak.[3] We also found in the same experiment that the phenomenon occurs much more frequently in female speakers' utterances than those of male speakers (38% vs. 5% of the time). A subsequent experiment confirmed the same tendency (Hata & Hasegawa, 1992).

Based on the results of our previous experiments, we hypothesized that listeners associate F0-peak delay with a feminine speech style in Japanese and conducted a perceptual experiment to test this hypothesis. We presented

synthesized sentences with and without peak delay to native speakers of Japanese, and the subjects judged which utterance in each pair sounded more female-like.

II. EXPERIMENT

2.1 Stimuli

The following sentences were prepared with a speech synthesizer. Each sentence contains a target word (shown in italics), which has the lexical accent on the first syllable.

A.	*námida* ga deru.	'Tears came into my eyes.'
B.	*túmari* kore desu ne.	'You mean this, don't you?'
C.	*káta* ga itai.	'I have sore shoulders.'
D.	*kánari* omosiroi.	'It's fairly interesting.'

The F0 contour of sentences A, C, and D was a rise-fall shape with the F0 starting at 230 Hz and ending at about 150 Hz. Sentence B contains a tag question, and thus had a slight rise at the end of the sentence. The global F0 peak always occurred within the target word.

Each sentence had two variations: in one the target word was made without an F0-peak delay (non-delayed token), and the other variation there was a delay (delayed token). Figures 2 and 3 represent these two types of F0 contour.

As shown in Figure 2, the non-delayed tokens had the peak (270 Hz) in the middle of the vowel of the lexically-accented first syllable, decreasing to 200 Hz at the onset of the third vowel. Once the fall reached 200 Hz, the F0 was sustained into the third syllable.

In the delayed tokens, as shown in Figure 3, the peak (300 Hz) occurred at 30 msec into the second vowel, followed by a 7-semitone F0 fall. As with the non-delayed tokens, once the fall reached 200 Hz, the F0 was leveled throughout the third syllable.[4]

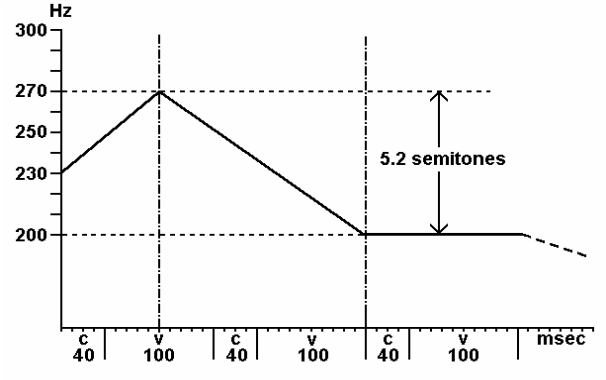

Figure 2: F0 contour of a target word
without a delayed F0 peak (non-delayed token)

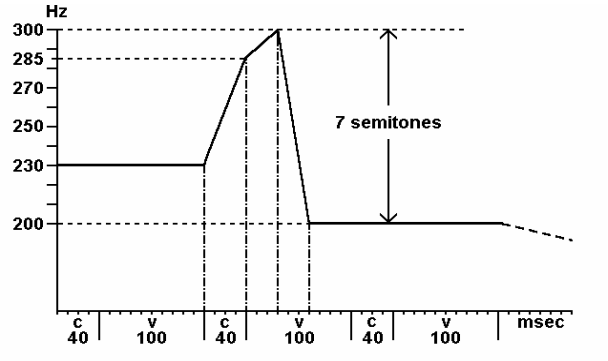

Figure 3: F0 contour of a target word
with a delayed F0 peak (delayed token)

Having obtained eight distinct sentence-tokens (4 sentences x 2 variations), we coupled the non-delayed token (ND token) and delayed token (D token) of each sentence in two orders: (1) the non-delayed token first and then the delayed token and (2) the delayed token first and then the non-delayed token, as shown below.

A1.	ND-D	A2.	D-ND	*námida* ga deru.
B1.	ND-D	B2.	D-ND	*túmari* kore desu ne.
C1.	ND-D	C2.	D-ND	*káta* ga itai.
D1.	ND-D	D2.	D-ND	*kánari* omosiroi.

These eight pairs were duplicated and randomized. The general instructions of the experiment were given to the subjects in written form as well as in spoken form using the synthetic voice. This precaution was taken in order to familiarize the subjects with the synthetic voice quality.

2.2 Procedure

Thirty-two subjects (19 males and 13 females, all native speakers of Japanese) participated in this experiment. They were told that one of the tokens in each pair was uttered by a male speaker and the other by a female speaker, and that the recorded voice was normalized with respect to pitch and length. The subjects then listened to the stimuli and decided which one in each pair sounded more female-like. Using answer sheets they circled 1 if they thought that the first token was likely to be uttered by a female speaker, and circled 2 if the second token sounded more female-like.

2.3 Results

Figure 4 summarizes the results. The abscissa shows how many delayed tokens each subject identified as uttered by a female speaker, and the ordinate shows how many subjects obtained the indicated score.

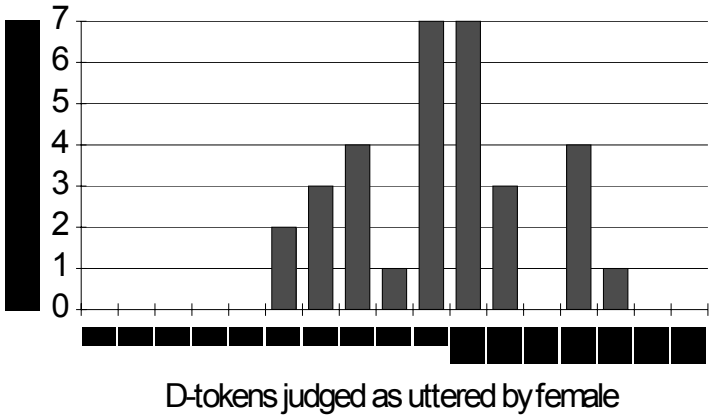

Figure 4: Results of the experiment

There were 16 token-pairs, half of which had the ND-D order and the other half the D-ND order. Therefore, if there is no perceptual difference between the non-delayed tokens and delayed tokens, we expect the average of the subjects' judgments of delayed tokens as more feminine to be close to 8 (50%). The result, however, shows that the responses are skewed: the average is 9.47 (59.2%) with a standard deviation of 2.21. The difference is highly statistically significant (t (31) = 3.7526, two-tailed $p < 0.001$).

This result was surprising to us because all subjects stated after the experiment that they could not hear any difference between the two tokens in each pair, and that they circled numbers randomly. The statistics, however, strongly suggest that the difference was perceived, although the subjects were not consciously aware of it.

It has been reported in the literature that not all listeners utilize the same strategy in detecting the prominent syllable. For example, among four cues for English accent (F0, duration, amplitude, spectral patterns), F0 is reported to be most significant (Fry, 1958), and yet, other cues being equal, some native English listeners do not respond to F0 changes alone in determining pitch peaks (Hata & Hasegawa, 1991). Therefore, it can be useful to separate the subjects according to their strategies in F0-perception experiments (Bartels & Kingston, 1994).

As seen in Figure 4, there is a great variability in the between-speaker results in our experiment. For example, one subject associated delayed tokens with a female speaker 14 times out of 16. It is unlikely that the subject obtained this score by mere guessing. We thus investigated how consistent the subjects' responses were. Because each order of tokens was duplicated, if subjects unconsciously perceived delayed tokens as more feminine, they would likely give the same responses for the identical pairs.

For this purpose, we counted only those responses that indicated (A) the delayed token as more feminine both times (consistent D-feminine responses) or (B) the non-delayed token as more feminine both times (consistent ND-feminine responses) for each identical pair of stimuli. We then divided (A) by the total consistent responses (A+B) and obtained the result shown in Figure 5.

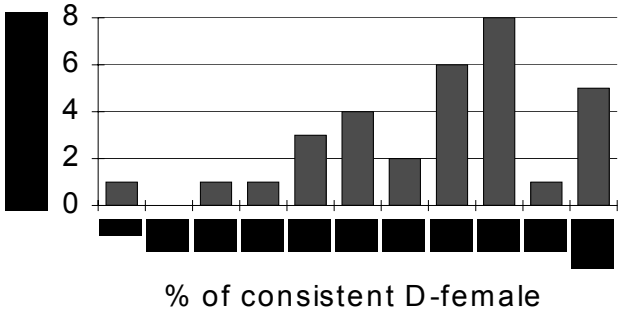

Figure 5: Consistent D-female Responses

The average of consistent D-feminine responses was 67%; the mode was 80%; and five of the 32 subjects gave 100% consistent D-feminine responses. On the other hand, one subject each delivered 0, 20, and 30% consistent D-feminine responses. This result supports the hypothesis that delayed F0 peak is a significant cue to femininity in Japanese, but not all native listeners are equally sensitive to it.

III. DISCUSSION

F0-peak delay with respect to the accented/stressed syllable has been discussed in the literature for decades to account for intonational meanings (Bolinger, 1958; O'Connor & Arnold, 1961; Ladd, 1983; Gussenhoven, 1984; Pierrehumbert & Steele, 1989; *inter alia*). For example, when an adult speaker says to a sobbing child 'Would you rather have your *Mommy* take you to the hospital?', the speaker may utter *Mommy* with a rise-fall contour (i.e. with a peak delay), rather than with an unmodified simple fall on *Mommy*, to indicate his/her concern for being understood by the inexperienced and possibly inattentive child (Gussenhoven, 1984). To give another example, when the speaker is certain that the person in question is a millionaire, *millionaire* is pronounced with a peak on the lexically stressed first syllable, but when the speaker is in doubt, the F0 peak shifts to the second syllable (Pierrehumbert & Steele 1989).

These previous works have demonstrated that delay in the alignment between the F0 peaks and accented syllables can have a significant function in human languages. However, what this significance actually designates may differ according to language types. In intonation languages, e.g. English, it is possible for the F0-peak delay to stretch over many syllables, and the listener can easily hear the delay and interpret it to be conveying some intonational meaning, e.g. the speaker's attentiveness, incredulity or uncertainty.

In pitch-accent languages, on the other hand, if the F0 peak is detected on a post-accent syllable, the perceived accent will inevitably shift to that syllable, resulting in an anomalous pronunciation. Therefore, the magnitude of delay in pitch-accent languages is much more limited than that in intonation languages (cf. Hata & Hasegawa, 1988). As a consequence, the conveyed meaning in pitch-accent languages is likely to be subtler than the meanings found in intonation languages. In the case of Japanese, as the present experiment has demonstrated, the paralinguistic function of the delay is to convey femininity.

The meanings of F0-peak delay in both types of languages are not arbitrarily chosen, however. It appears that what is underlying those meanings is indecisiveness (or hesitancy), which normally invites the participation by the hearer in the conversation. Female speakers are reported to prefer more collaborative conversations than male speakers do (Lakoff, 1990; Tannen, 1990). We conjecture that peak delay signals a degree of indecisiveness, which is frequently associated with females, and in turn, is further conventionalized to convey a sense of femininity in Japanese.

NOTES

* We would like to thank the following individuals for their comments and criticism: Mary Beckman, Gregory De Haan, Carlos Gussenhoven, Larry Hyman, Anita Liang, John Ohala, Manjari Ohala, Natasha Warner, and Raymond

Weitzman.. This project was supported in part by a grant from the Center for Japanese Studies at the University of California, Berkeley.

[1] Mattingly (1966) reports that male-female differences in formant frequencies are also chiefly due to social conventions, not to physiological differences.

[2] The phenomenon of *ososagari* was first reported by Neustupný (1966).

[3] We suspect that nasals have some influence on the F0-peak delay.

[4] We have chosen two different F0-peak frequencies for the non-delayed and delayed tokens in order to balance the perceived overall pitch ranges. Because the non-delayed tokens have more gradual F0 change, if the peak frequency were 300 Hz, as is the case for the delayed tokens, each non-delayed token as a whole would sound much higher than the corresponding delayed token. Comparing several non-delayed tokens with varying F0 peak frequencies, we have determined that a 270-Hz peak renders a voice range most compatible with the delayed tokens. Other characteristics (e.g. formant frequencies, amplitude, speech rate) are identical in both types of tokens.

REFERENCES

Bartels, Christine, and John Kingston. 1994. Salient pitch cues in the perception of contrastive focus. *Journal of the Acoustical Society of America* 95, 2973.

Bolinger, Dwight. 1958. A theory of pitch accent in English. *Word* 14, 109-49.

Brend, Ruth. 1975. Male-female intonation patterns in American English. In B. Thorne and N. Henley (eds.), *Language and Sex: Difference and Dominance*. Rowley, MA: Newbury House.

Brown, Penelope, and Stephen Levinson. 1978. *Politeness: Some Universals in Language Usage.* Cambridge: Cambridge University Press.

Chiba, Tsutomu, and Masato Kajiyama. 1941. *The Vowel - Its Nature and Structure*. Tokyo: Kaiseikan.

Coleman, Ralph. 1976. A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice. *Journal of Speech and Hearing Research* 19, 168-80.

Edelsky, Carole. 1979. Question intonation and sex roles. *Language in Society* 8, 15-32.

Fry, Dennis. 1958. Experiments in the perception of stress. *Language and Speech* 1, 126-52.

Gussenhoven, Carlos. 1984. *On the Grammar and Semantics of Sentence Accent.* Dordrecht: Foris Publications.

Hasegawa, Yoko, and Kazue Hata. 1988. Delayed pitch fall in Japanese. *Journal of the Acoustical Society of America* Suppl. 1.83, S29.

Hata, Kazue, and Yoko Hasegawa. 1988. Delayed pitch fall in Japanese: a perceptual experiment. *Journal of the Acoustical Society of America* Suppl. 1.84, S156.

Hata, Kazue, and Yoko Hasegawa. 1991. The effect of F0 fall rate on accent perception in English. *BLS* 17, 121-29.

Hata, Kazue, and Yoko Hasegawa. 1992. A study of F0 reset in naturally-read utterances in Japanese. *Proceedings of the International Conference of Spoken Language Processing*, 1239-42.

Jugaku, Akiko. 1979. *Japanese Language and Women* (In Japanese). Tokyo: Iwanami Shoten.

Ladd, Robert. 1983. Phonological features of intonational peaks. *Language* 59, 721-59.

Lakoff, Robin. 1975. *Language and Woman's Place*. New York: Harper and Row.

Lakoff, Robin. 1990. *Talking Power: The Politics of Language in Our Lives*. New York: Basic Books.

Loveday, Leo. 1981. Pitch, politeness and sexual role: an exploratory investigation into the pitch correlates of English and Japanese politeness formulae. *Language and Speech* 24, 71-89.

Mattingly, Ignatius. 1966. Speaker variation and vocal-tract size. *Journal of the Acoustical Society of America* 39, 1219.

Neustupný, J. V. 1966. Is the Japanese accent a pitch accent? *Onsei Gakkai Kaihoo* 121. Reprinted in M. Tokugawa (1980), *Akusento*, 230-39. Tokyo: Yuuseido.

Nordström, P.-E. 1977. Female and infant vocal tracts simulated from male area functions. *Journal of Phonetics* 5, 81-92.

O'Connor, J. D., and G. F. Arnold. 1961. *Intonation of Colloquial English*. London: Longman.

Ohala, John. 1983. Cross-language use of pitch: an ethological view. *Phonetica* 40, 1-18.

Peterson, Gordon, and Harold Barney. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24, 175-184

Pickett, James. 1980. *The Sounds of Speech Communication.* Baltimore: University Park Press.

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

Pierrehumbert, Janet, and Shirley Steele. 1989. Categories of tonal alignment in English. *Phonetica* 46, 181-96.

Sachs, Jacqueline, Phillip Lieberman, and Donna Erickson. 1973. Anatomical and cultural determinants of male and female speech. In R. Shuy and R. Fasold (eds.), *Language Attitude: Current Trends and Prospects*, 74-84. Washington, D.C.: Georgetown University Press.

Sugito, Miyoko. 1968. Dootai sokutei ni yoru nihongo akusento no kaimei. *Gengo Kenkyuu* 55. Reprinted in M. Sugito (1982), *Nihongo akusento no kenkyuu*, 49-75. Tokyo: Sanseidoo.

Sundberg, Johan. 1987. *The Science of the Singing Voice.* Dekalb, Illinois: Northern Illinois University Press.

Tannen, Deborah. 1990. *You Just Don't Understand: Women and Men in Conversation.* New York: William Morrow.

Trudgill, Peter. 1974. *Sociolinguistics: An Introduction.* Harmondsworth: Penguin.