NON-PHYSIOLOGICAL DIFFERENCES BETWEEN MALE AND FEMALE SPEECH: EVIDENCE FROM THE DELAYED F0 FALL PHENOMENON IN JAPANESE

Yoko Hasegawa

University of California, Berkeley
Department of East Asian Languages
Berkeley, CA 94720-2230, USA

Kazue Hata

Speech Technology Laboratory
Panasonic Technologies, Inc.
Santa Barbara, CA 93105, USA

ABSTRACT

There is a perceptual difference between male and female speech in fundamental frequency and, less significantly, in formant frequencies. These characteristics are physiologically determined to a great extent, and thus speakers exert little control over these characteristics in their normal utterances. The present study investigates whether some characteristics of fundamental frequency are also manipulated by speakers to make speech quality more feminine. The results of our experiment show that the delay of the F0 fall that signals a lexical accent is associated with femininity in Japanese.

I. INTRODUCTION

1.1 Comparisons of Male and Female Speech

Because male speech and female speech are evidently different, people usually have no difficulties in identification of a speaker's gender. On average, female speakers have higher fundamental frequency (F0) [1] and higher formant frequencies [2], the former being more significantly different than the latter between male and female groups [3]. Male speakers typically use low F0, which is associated not only a large body size but also character traits such as aggressiveness, assertiveness, self-confidence, and so forth [4]. 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 [4].

F0 is certainly influenced by anatomical differences between male and female speakers; however, it has also been reported that observed F0 differences are much greater than those that can be attributed to anatomy. Many researchers have pointed out that F0 differences are more of a social phenomenon, reflecting different social norms laid down for men and women [5-11]. Society assigns different social roles to men and women and expects different behavioral patterns from each group; language simply reflects this social fact [5]. American females, for example, make great use of rising intonation [6], and, when using formal and polite register, Japanese female speakers use higher F0 than female speakers of other languages [10, 11].

The frequent use of higher F0 and rising intonation are non-physiological aspects of female speakers. In this article, we present yet another gender difference observed in Japanese speakers; viz., Japanese female speakers tend to delay the F0 peak that signals a lexical accent. This phenomenon is known as the *delayed F0 fall* (*oso-sagari* in Japanese), which is explained in the next section.

1.2 Delayed F0 Fall

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 [12]. This accentual F0 peak, however, frequently occurs on the post-accent syllable, without listeners detecting any change in accent placement. For example, in /namida/ 'tears' (Noun) the lexical accent falls on the first syllable /na/; however, the actual F0 peak may occur on the second syllable /mi/, and yet /na/ is perceived as accented. This phenomenon, originally called *oso-sagari* (delayed F0 fall), was first reported by Neustupný [13].

Investigating the delayed F0 fall phenomenon, Sugito 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; i.e., native speakers of Japanese perceive an accent on a syllable when it is followed by a falling F0 contour [14]. In the /namida/ example above, if the post-accent syllable /mi/ contains a F0 fall, /na/ is perceived as accented, even if the F0 peak occurs on /mi/.

Although its occurrence varies between speakers as well as between words uttered by the same speaker, delayed F0 falls are more frequently observed in words with an accent in certain positions or segmental environments. In Sugito's data, about 36% of 3-, 4-, 5-syllable words with the accent on the initial syllable had delayed F0 fall [14]. In Hasegawa and Hata, we reported that 37% of the words beginning with /(C)VmV/ were uttered with delayed F0 fall [15]. 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). Our subsequent experiment confirmed the same tendency [16].

Based on our previous experiments, we have hypothesized that the delayed F0 fall is associated with femininity in Japanese, i.e., it makes speech sound more feminine. To test this hypothesis, we conducted a perceptual experiment using a synthetic voice. Subjects were asked to evaluate femaleness for each pair of sentences which was prepared with and without a delayed F0 fall. The result has supported our hypothesis.

II. Experiment

2.1 Materials

Using a MITalk-based system, we synthesized the following sentences. Each sentence contains a target word (shown in Italics), which has the lexical accent on the first syllable.

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

The F0 contour of sentences A, C, and D was a rise-fall shape with the starting F0 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 with a delayed F0 fall (D-token), and the other without it (ND-token). Figures 1 and 2 represent the two types of F0 contour within a target word.

In the D-tokens, the peak (300 Hz) occurred at 30 msec into the second vowel, followed by a 7-semitone F0 fall. Once the fall reached 200 Hz, the F0 was sustained into the third syllable. The ND-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.

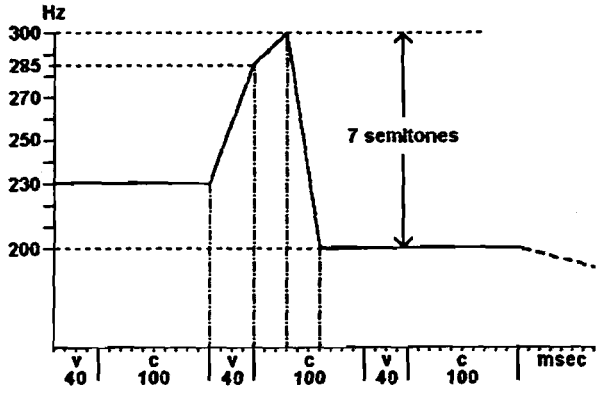

Figure 1: F0 contour of a target word with a delayed F0 fall (D-token)

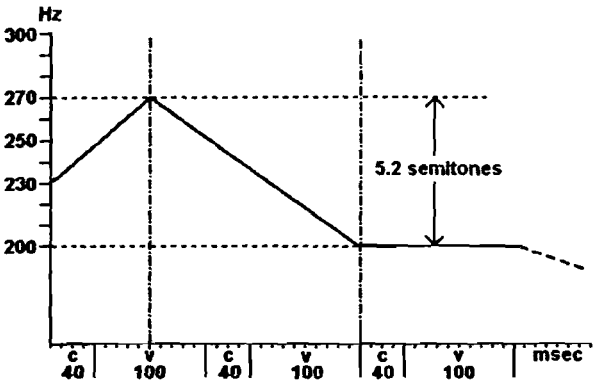

Figure 2: F0 contour of a target word without a delayed F0 fall (ND-token)

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

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

A1.	D-ND	A2.	ND-D	*namida* ga deru.
B1.	D-ND	B2.	ND-D	*tumari* kore desu ne.
C1.	D-ND	C2.	ND-D	*kata* ga itai.
D1.	D-ND	D2.	ND-D	*kanari* omosiroi.

These eight pairs were duplicated and then randomized. We then added four red-herring pairs at the beginning and recorded a total of 20 pairs. 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

32 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 sounded more female-like. In answer sheets they circled 1 if they thought that 1 was likely to be uttered by a female speaker, and circled 2 otherwise.

2.3 Results

Figure 3 summarizes the results. The abscissa shows how many D-tokens each subject identified as uttered by a female speaker, and the ordinate shows how many subjects obtained the indicated score.

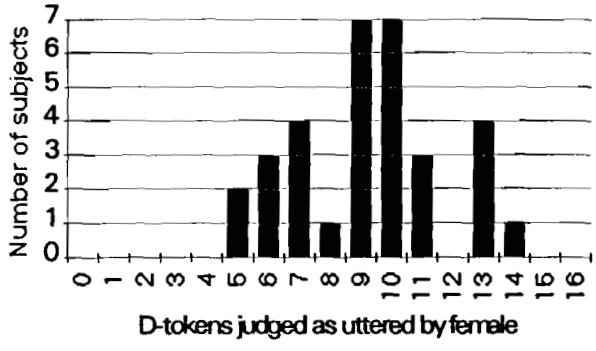

Figure 3: Results of the experiment

There were 16 token-pairs, half of which had the D-ND order and the other half the ND-D order. Therefore, if there is no perceptual difference between the D-tokens and ND-tokens, we expect the average of the subjects' judgments of D-tokens as more feminine to be close to 8 (50%). The result, however, shows that the responses are skewed: the average is 9.47 (59.19%) 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 because all subjects expressed, 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 F0 prominence. For example, among four cues for English accent (F0, duration, amplitude, spectral patterns), F0 is reported to be most significant [17], and yet, other cues being equal, some listeners do not respond to F0 changes alone in determining pitch peaks [18]. Therefore, it is important to separate the subjects according to their strategies in F0-perception experiments [19].

As seen in Figure 3, there is a great variability in the result of our experiment. One subject associated D-tokens with a female speaker 14 times out of 16. It is unlikely that the subject obtained this score by mere guessing. We investigated how consistent the subjects' responses were. Because each order of tokens was duplicated, if subjects unconsciously perceived D-tokens as more feminine, they would give the same responses for the identical pairs.

For this purpose, we counted only those responses, that indicated (A) the D-token more feminine twice (consistent D-feminine responses) or (B) the ND-token more feminine twice (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 4.

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. We conclude, based on this result, that delayed F0 fall is a significant cue to femininity in Japanese, but not all native listeners are sensitive to it.

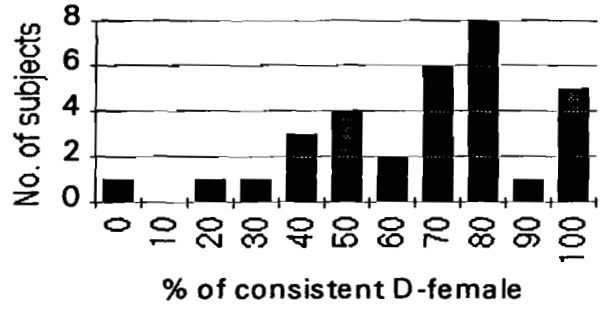

Figure 4: Consistent D-female Responses

III. Discussion

Delayed F0 fall has been discussed in the literature for some time to account for intonational meanings. Ladd, for example, proposes a binary feature [+/-delayed peak] for the relative temporal alignment of the H-tone and the accented syllable [20]. If the H is late in the syllable, i.e. [+delayed peak], we get a rise; if it is early, [-delayed peak], we get a fall. He claims that the difference between Swedish Accents 1 and 2, as well as the accentual patterns in Norwegian and Serbo-Croatian, can be represented with this feature.

Gussenhoven, who proposes the feature [delay] to operate not only on falls but also on rises, claims that the modification delay has some intonational meaning: 'This manipulation is very non-routine, very significant' [21]. He notes that delay in speech to children, where it generally occurs more frequently than in other types of speech, indicates the adult speaker's concern for being understood by the inexperienced and possibly inattentive language user. For example, when the sentence 'Would you rather have your Mummy take you to the hospital?' is addressed to a sobbing child, the adult speaker may utter the word *Mummy* with a rise-fall contour ([+delay]) rather than with an unmodified simple fall .

These works have demonstrated that delay in the alignment between H-tones and accented syllables can be significant in language; however, how this significance actually signals may differ according to language types. Because all languages cited above can be categorized as

intonation languages [22], delays in the alignment of H- or L-tone and accented syllables are less restricted than those in pitch-accent languages like Japanese. For example, the English word *rather*, which has the lexical prominence on the first syllable, can be uttered with a noticeable F0 peak (and a subsequent fall) on the second syllable. This difference was exploited in British English until the early twentieth century to indicate an upper-class pronunciation of the exclamation meaning 'Yes, very much so' [21].

In Japanese, on the other hand, if the F0 peak is detected on the post-accent syllable, the perceived accent will inevitably shift to that syllable, resulting in an anomalous pronunciation. Therefore, when the peak is delayed, the fall rate after the peak must be significantly great to "compensate for" the delay. The later the peak location, the greater the fall rate required to maintain the accentual pattern of the word [23]. Furthermore, the magnitude of delay has an absolute limit beyond which the delayed F0 fall phenomenon does not occur [23].

Based on the result of the present experiment, we conclude that the feature [delay] is a legitimate property of language, which can be exploited in language-specific ways, and that in Japanese, this feature is used to express femininity.

References

[1] Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. 24, 175-184

[2] Chiba, T., and Kajiyama, M. (1941). *The Vowel - Its Nature and Structure* (Kaiseikan, Tokyo).

[3] Coleman, R. O. (1976). "A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice," J. Speech Hear. Res. 19, 168-80.

[4] Ohala, J. J. (1983). "Cross-language use of pitch: an ethnological view," Phonetica 40, 1-18.

[5] Trudgill, P. (1974). *Sociolinguistics: an Introduction* (Penguin, Harmondsworth).

[6] Brend, R. M. (1975). "Male-female intonation patterns in American English," in *Language and Sex: Difference and Dominance*, edited by B. Thorne and N. Henley (Mewbiry, Rewley).

[7] Lakoff, R. (1975). *Language and Woman's Place* (Harper & Row, New York).

[8] Brown, P., and Levinson, S. C. (1978). *Politeness: Some Universals in Language Usage* (Cambridge Univ. Press, Cambridge).

[9] Edelsky, C. (1979). "Question intonation and sex roles," Lg. Sci. 8, 15-32.

[10] Jugaku, A. (1979). *Japanese Language and Women* (In Japanese) (Iwanami, Tokyo).

[11] Loveday, L. (1981). "Pitch, politeness and sexual role: an exploratory investigation into the pitch correlates of English and Japanese politeness formulae," Lg. Speech 24, 71-89.

[12] Pierrehumbert, J. B., and Beckman, M. E. (1988). *Japanese Tone Structure* (MIT, Cambridge, MA).

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

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

[15] Hasegawa, Y., and Hata, K. (1988). "Delayed pitch fall in Japanese," J. Acoust. Soc. Am. Suppl. 1.83, S29.

[16] Hata, K., and Hasegawa, Y. (1992). "A study of F0 reset in naturally-read utterances in Japanese," Proc. ICSLP, 1239-42.

[17] Fry, D. B. (1958). "Experiments in the perception of stress," Lg. Speech 1, 126-52.

[18] Hata, K., and Hasegawa, Y. (1991). "The effect of F0 fall rate on accent perception in English," Proc. Berkeley Linguistics Society, 121-29.

[19] Bartels, C., and Kingston, J. (1994). "Salient pitch cues in the perception of contrastive focus," J. Acoust. Soc. 95, 2973.

[20] Ladd, D. R. (1983). "Phonological features of intonational peaks," Lg. 59, 721-59.

[21] Gussenhoven, C. (1984). *On the Grammar and Semantics of Sentence Accent* (Foris, Dordrecht)

[22] Cruttenden, A. (1986). *Intonation* (Cambridge Univ. Press, Cambridge).

[23] Hata, K., and Hasegawa, Y. (1988). "Delayed pitch fall in Japanese: a perceptual experiment," J. Acoust. Soc. Am. Suppl. 1.84, S156.