TR-I-0025

Prosodic Characteristics and Their Control in Japanese Speech with Various Speaking Styles

種々の発声様式における日本語音声の
韻律の特徴とその制御について

Masanori MIYATAKE　Yoshinori SAGISAKA

宮武 正典　　匂坂 芳典

June 1988

Abstract

Isolated word utterances of Japanese were analyzed to clarify the effects of speaking styles on prosodic parameters. These words were uttered in seven different ways, i.e., normal, slow, fast, strong, weak, high and low by two professional narrators, and their prosodic parameters (fundamental frequency (F_0), power and segmental duration) were compared. The analysis shows a strong correlation between F_0 and power, and a difference in F_0 patterns between speaking styles. Moreover, it was found that the F_0 pattern shows remarkable differences according to speaking styles and that the F_0 pattern changes very systematically. To realize these tendencies in speech synthesis by rule, F_0 control parameter estimation was carried out by applying analysis-by-synthesis technique to all samples at a time. Through this estimation, it was confirmed that F_0 pattern can be systematically controlled.

ATR Interpreting Telephony Research Laboratories

ATR自動翻訳電話研究所

1. Introduction

To produce a much more human-like speech output, natural synthesized speech with various speaking styles is desired. In the current technology of speech synthesis by rule, however, only normal reading speech can be synthesized. There is almost no work[1)2)] which enables the synthesis of speech with various speaking styles.

In this paper, prosodic characteristics and the relationship between prosodic parameters are analyzed using Japanese isolated words uttered in various speaking styles. Furthermore, a new parameter estimation technique for F_0 control is proposed.

2. Prosodic Parameters in Various Speaking Styles

2.1. Speech Utterances

To find the essential characteristics of prosodic parameters, isolated words are analyzed. As shown in Table 1, these data consist of 308 words, including 1,215 vowels. These words were uttered in seven different ways, *high*, *low*, *strong*, *weak*, *fast*, *slow*, and *normal*, by two professional narrators. They were directed to control pitch as *high* and *low*, loudness as *strong* and *weak*, and speed as *fast* and *slow*. To compare the prosodic characteristics between different speaking styles, F_0 and power at vowel centers, and mora segmental duration are measured for each word set and speaking style.

2.2. Correlation Between Prosodic Parameters

Many prosodic control rules and models have been developed. In these rules and models, prosodic parameters are controlled independently. In natural conversation, however, some correlations are found. It is thus desirable that they should be controlled systematically to synthesize natural speech.

To clarify the relationship, the coefficients of correlation between F_0, power and segmental duration are calculated for each set. In each set, power in normal speaking is set to 0 dB as an origin of comparison. This analysis shows a strong

correlation between F_0 and power. (Figure 1) Moreover, the following tendencies are found:

(1) With the exception of the 'high' speaking style, F_0 and power are directly proportional. Correlations of 0.81 - 0.99 were found.

(2) F_0 can, however, be controlled independent of power, when the subject is instructed to do so, e.g., 'high'.

(3) On the other hand, segmental duration is almost independent of F_0 and power.

These correlations seem to be caused by the human speech production mechanism. The pressure in the lung is mainly used to control loudness, and influences subglottal pressure, so that vocal cord vibration is varied. Therefore, in this case F_0 increases or decreases directly proportional to power. On the other hand, both the constriction of the vocal cords and the pressure in the lung can be used to control pitch. As vocal cords which are too loose cannot produce voiced sound, it seems that the pressure in the lung is mainly used to utter in *low* speaking style. Constriction of the vocal cords can, however, cause the speaker to utter in *high* speaking style. Therefore, F_0 can increase independent of power.

From these tendencies, it can be concluded that both F_0 and power should be varied to control pitch and loudness except in the *high* speaking style, and that segmental duration is almost independent of pitch and loudness controls, as shown in Figure 2. (Arrow thickness indicates the degree of correlation)

3. F_0 Patterns in Various Speaking Styles

3.1. SVC F_0 Pattern

As shown in Section 2.2., F_0 patterns show remarkable differences.

As is well known from the previous work[3], the *sampled at vowel centers F_0 pattern* (SVC F_0 pattern) is very stable for each n-mora i-type accent word. To quantify F_0 pattern changes due to speaking styles, the SVC F_0 patterns are precisely analyzed.

3.2. the Characteristics of SVC F_0 Patterns in Various Speaking Styles

SVC F_0 Patterns are calculated for each word utterance. For example, F_0 patterns averaged over ten 5-mora words with accented third mora uttered by a male are shown in Figure 3. In the calculation, the following tendencies are found:

(1) The average value of F_0 patterns differs between speaking styles.

(2) The dynamic range and excursion from the F_0 pattern base line also increase in proportion to F_0 value.

3.3. F_0 Pattern Control

For the control model of the F_0 pattern, the following critical damping model[4] is employed: the F_0 pattern is logarithmically expressed as the sum of voicing and accent factors, i.e.,

$$\ln F_0(m) = \ln F_{min} + A_p G_p(m - M_0) + A_a\{G_a(m - M_1) - G_a(m - M_2)\}$$

where, $G_p(m) = \alpha m \cdot \exp(-\alpha m)$

$G_a(m) = 1 - (1 + \beta m)\exp(-\beta m)$

m: mora position
F_{min}: lower limit
A_p: amplitude of voicing factor
M_0: beginning of voicing control signal
α: coefficient of voicing factor
A_a: amplitude of accent factor
M_1: beginning of accent control signal
M_2: end of accent control signal
β: coefficient of accent factor

3.4. F_0 Control Parameter Estimation

Using SVC F_0 pattern and the critical damping model, the F_0 control parameters are estimated by applying analysis-by-synthesis (AbS). As the SVC F_0 pattern is very stable and contains less data (it is made only at vowel centers), it becomes possible to systematically estimate the parameters at one time. By using some conditions of restriction for each mora and each accent type in the critical damping model, the AbS can be applied to all samples in the above word sets simultaneously.

3.5. F_0 Control Parameter Estimation for Normal Speaking Style

F_0 patterns of actual data are classified according to their mora length and accent types, and the control parameters are estimated by applying AbS to each classified F_0 pattern set. (Table 2) To make the restriction reasonable, the following conditions are fixed for normal speaking style:

(1) As the end of the accent control signal (M_2) differs according to accented mora, even though the estimated values are not precise, M_2 is fixed as follows:

$M_2 = i$ (when the i-th mora is accented)
or $M_2 \geqq n$ (when no mora is accented)
where, n is word mora length.

(2) Voicing control parameters are common to whole words, as they should be independent of accent control parameters.

(3) The lower limit (F_{min}) is fixed, because it allows better control of the other parameters.

(4) As the amplitude of the accent factor (A_a) for F_0 pattern with accented first mora is two or three times as large as A_a for other F_0 patterns, two A_as are used.

The control parameters are estimated for all data in normal speaking style, using the above conditions. (Table 3) The logarithmic estimation error is

0.080, i.e., 7-8 Hz, and the deviation for actual data is 0.065, i.e., 6-7 Hz. This result shows that the control parameters are well estimated.

3.6. Control Parameter Estimation for Various Speaking Styles

There are several possible conditions for estimating the control parameters for various speaking styles:

(1) The lower limit (F_{min}) is fixed and the amplitude parameters of the voicing factor (A_v) are prepared for individual speaking style.

(2) The lower limits (F_{min}) and the beginning of the voicing control signal (M_0) are prepared for individual speaking style.

In either way, the estimate is almost the same, though the conditions in which the F_0 pattern can be more controllable must be clarified. The estimation error is 0.102, i.e., 7 - 20 Hz, and the deviation for actual data is 0.075, i.e., 5-15 Hz. This result shows that the estimation is well done.

Some systematic errors, however, still remain: for example, errors in voicing factor which depend on mora length are shown in Figures 4(a), (b) and (c). For three-mora words, the estimated F_0 pattern is higher than actual F_0 patterns (Figure 4(a)), and for five-mora, it is lower (Figure 4(c)).

4. Conclusion

A strong correlation between F_0 and power, and differences in the average values and the dynamic ranges of F_0 patterns are found under various speaking styles. Furthermore, an F_0 control parameter estimation for various speaking styles is carried out. It is confirmed that the F_0 pattern can be systematically controlled for various speaking styles. It is desired that the above systematic detail errors should be reduced and applied to various conversational speech styles.

References

1) D. Robert Ladd "Evidence for the independent function of intonation contour type, voice quality, and F_0 range in signaling speaker affect" J. Acoust. Soc. Am. 78(2) pp. 435-444 (1985)

2) C. M. Johns-Lewis "Digital analysis of pitch and silence in three speech styles" IEE Conf. Publ. (Inst. Electr. Eng.) No. 258 pp. 281-286 (1986)

3) S. Hashimoto "Several features of Japanese word accent" Transaction of Inst. of Electr. and Communication Engineers vol. 56-D, No.11, pp. 654-661 (1973)

4) H. Fujisaki, H. Sudo "A model for the generation of fundamental frequency contours of Japanese word accent" J. Acoust. Soc. Japan vol.27, No.9 pp. 445-453 (1971)

Table 1
Specification of Isolated Word Sets

Name of Sets	CVb	CVg	CVk	CVs	CVm	Total
Words	62	62	62	62	60	308
Mora	251	252	231	253	228	1,215

Total Utterance Words
Male 2,220 Words(8,758 Mora)
Female 1,848 Words(7,290 Mora)

Table 2 Control Parameters estimated for each classified data (male)

mora-type	F_{min}	A_v	M_0	α	A_a	M_1	M_2	β
3-0	55.7	1.50	0.70	1.13	0.36	-25.7	3.50	1.64
3-1	47.0	1.62	-1.48	0.61	0.84	0.12	1.47	2.87
3-2	55.8	1.23	0.55	0.92	0.48	0.22	1.99	4.35
4-0	58.6	1.54	0.59	0.84	0.30	-22.2	4.50	1.55
4-1	50.0	1.92	-1.43	0.70	1.03	-0.06	1.22	1.35
4-2	57.4	1.40	0.28	0.88	0.52	0.38	2.41	2.60
4-3	60.6	1.22	0.58	0.77	0.41	-0.21	2.93	3.86
5-0	60.5	1.49	0.46	0.67	0.29	-19.3	5.50	1.85
5-3	50.7	1.50	0.71	0.51	0.51	-0.34	3.53	3.05

Table 3 Control Parameters estimated for normal speaking style (male)

F_{min}	A_v	M_0	α	A_a	M_1	M_2	β
37.8	1.95	0.052	0.52	1.32(1-type) 0.55(other)	0.08(1-type) -2.67(other)	n+0.5(0-type) i (other)	2.06

1-type = 1st mora accented, 0-type = no mora accented

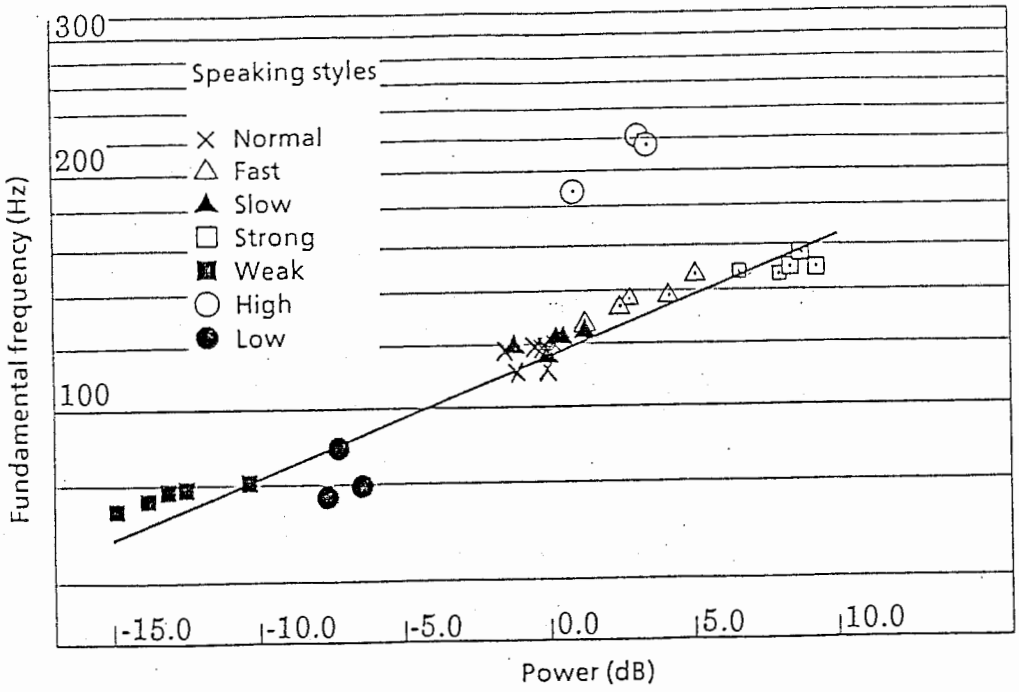


Figure 1. Correlation between power and F_0

(Male, five sets of 62 words for each speaking style. Each dot corresponds to the mean value averaged over each set.)

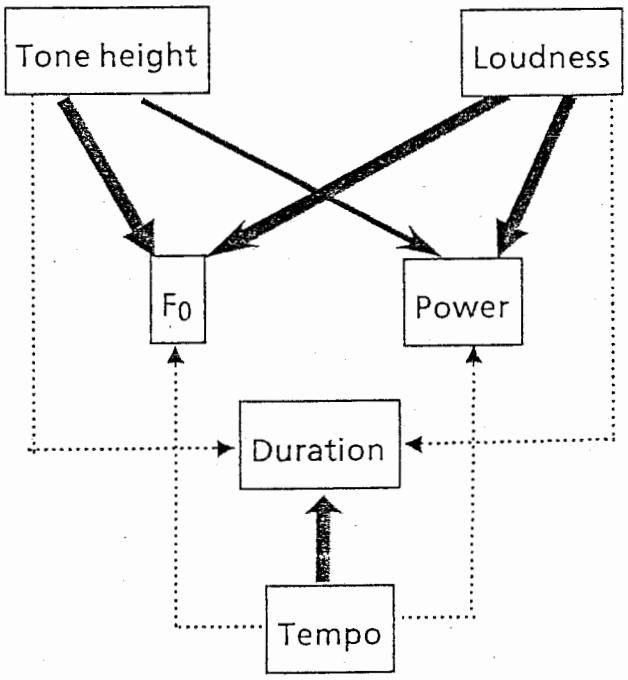


Figure 2. Correlation between prosody controls

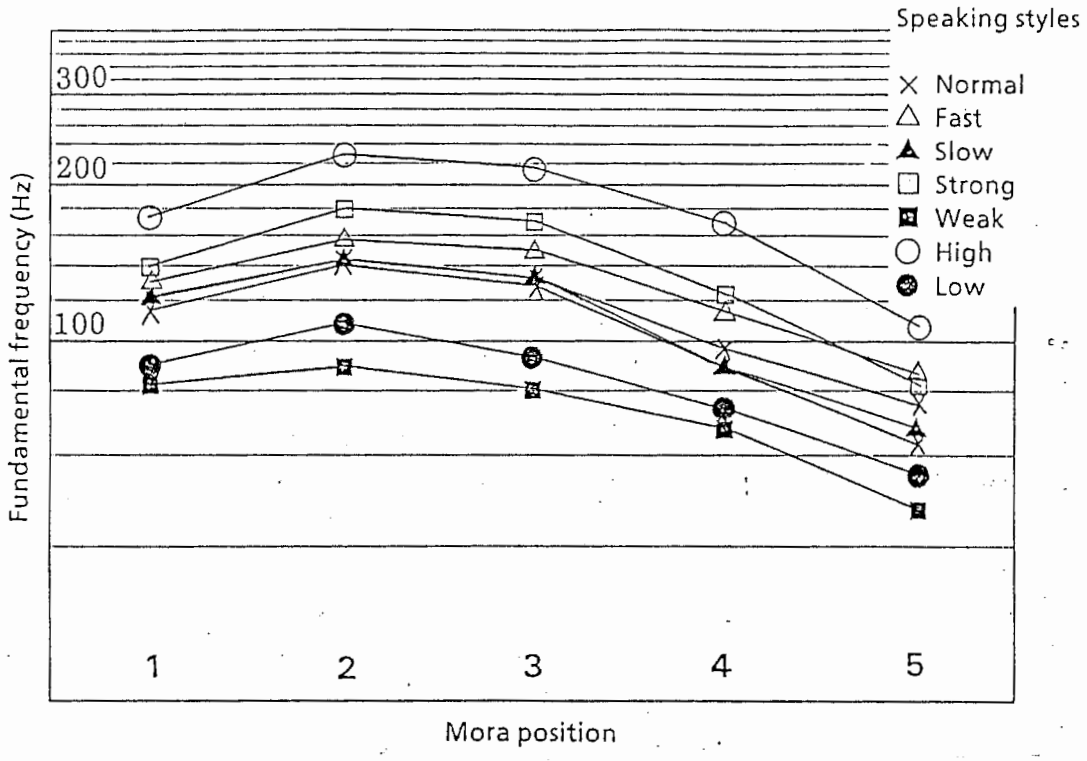
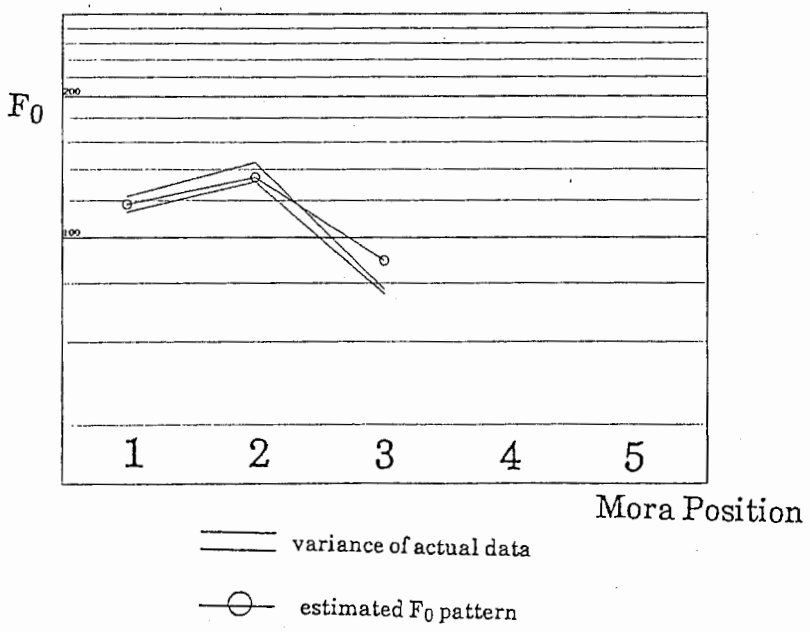
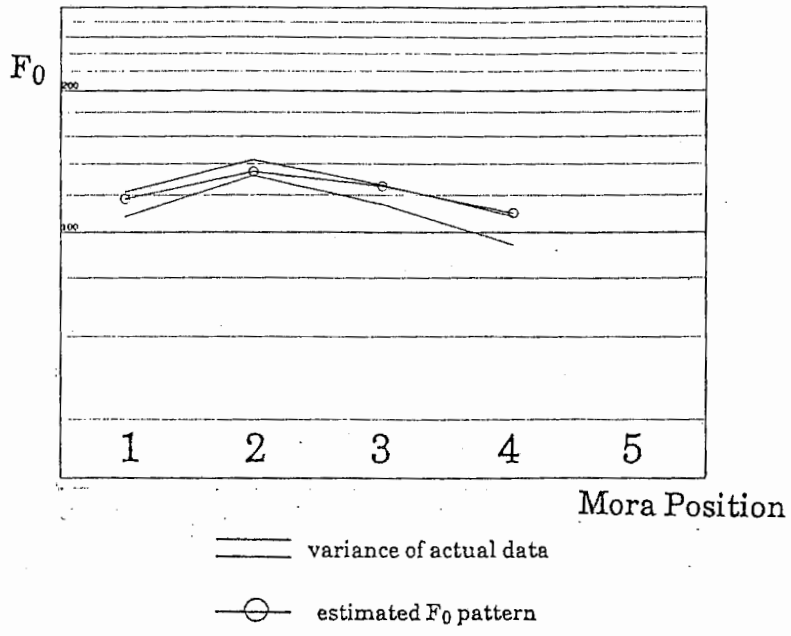


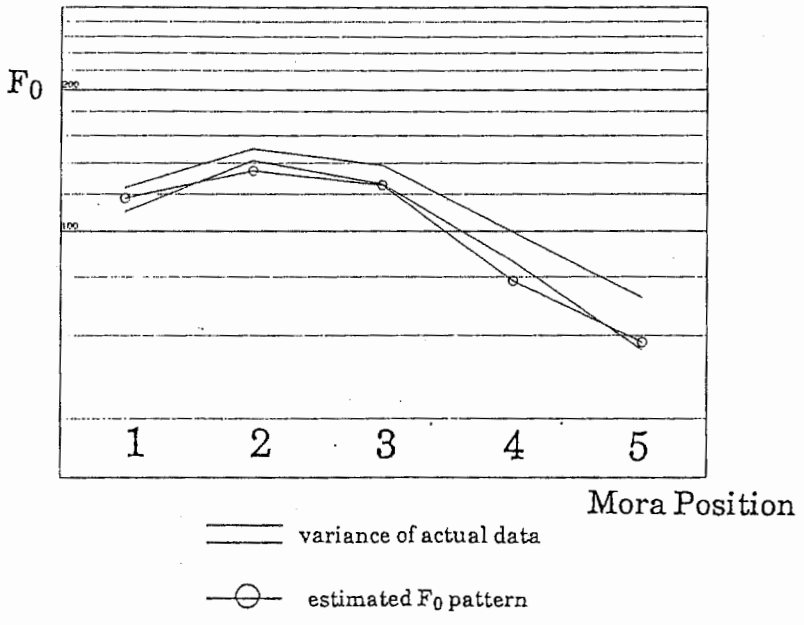
Figure 3. SVC F_0 patterns in various speaking styles
(Male, averaged over ten 5-mora words with accented third mora)



(a) 3-mora word with accented second mora
Figure 4. SVC F_0 patterns in normal speaking style (male)



(b) 4-mora word with no accented mora



(c) 5-mora word with accented third mora

Figure 4. SVC F_0 patterns in normal speaking style (male) (Continued)