Ann. Bull. RILP, No. 13, 163-173 (1979)

ACOUSTIC FEATURES OF THE FUNDAMENTAL FREQUENCY CONTOURS OF DECLARATIVE SENTENCES IN JAPANESE

Hiroya FUJISAKI*, Keikichi HIROSE*, and Kazuhiko OHTA*

ABSTRACT

For the purpose of elucidating the relationship between the sentence F_0-contour and the linguistic and non-linguistic information, an attempt was made to extend the model for the word F_0-contour already published by the authors to the sentence F_0-contour. The extended model is based on the formulation of the process whereby the logarithmic fundamental frequency is controlled in proportion to the sum of two components corresponding respectively to the effects of voicing and accent. Fundamental frequencies were extracted from utterances of various declarative sentences, and the F_0-contours were used to test the validity of the model. The model's parameters were determined to give the best approximation to an observed F_0-contour on the basis of the mean squared error. The model's validity was demonstrated by its ability of closely approximating observed F_0-contours. The extracted parameters were found to be closely related to linguistic factors and factors constituting the naturalness of speech, and thus provide a means for generating natural F_0-contours from a small set of parameters and rules for synthesis.

1. Introduction

The contour of the voice fundamental frequency (henceforth F_0-contour) plays an important role not only in the transmission of linguistic information concerning word meaning and sentence structure, but also in the transmission of non-linguistic information such as naturalness, emotion, and speaker idiosyncrasy.

Because of the difficulty in the accurate analysis and the quantitative description of the characteristics of the F_0-contour, which is usually a quasi-continuous curve with rather complex undulations, the relationships between the acoustic characteristics and the linguistic and non-linguistic information are not always clear. The elucidation of these relationships requires firstly the selection of characteristic parameters that are capable of describing the essential features of an F_0-contour, and secondly a method for extracting these parameters from an observed F_0-contour. In other words, an analytical formulation (i.e., a model) of the control process of the fundamental frequency is indispensable for the quantitative analysis and linguistic interpretation of the F_0-contour characteristics.

It is of course widely recognized that F_0-contours of words and sentences are generally characterized by a gradual declination from the onset toward the end of an utterance, superposed by local humps corresponding to word accent as well as those corresponding to such intonational

* Department of Electrical Engineering, Faculty of Engineering, University of Tokyo.

factors as interrogation and emphasis. Careful observation of these contours reveals that F_0-contours of the same linguistic expression uttered by a number of speakers with different voice pitches are almost similar when plotted on the logarithmic scale of the fundamental frequency as functions of time [1], and that F_0-contours of different linguistic expressions uttered by the same speaker possess more or less the same baseline which drops rather steeply at the beginning and then gradually tends to an asymptotic value. The humps of the F_0-contour are also found to rise and fall smoothly, and the magnitude of a hump corresponding to a prosodic element, say a word with a certain accent type in Japanese, expressed by the percentage increase in F_0 from its baseline value, remains nearly the same regardless of its position within a sentence.

Most of the models that have been proposed for the analysis and interpretation of F_0-contour characteristics, however, are based on a rather crude piecewise-linear approximation of an F_0-contour on the linear scale of the fundamental frequency, i.e., on a combination of a straight, declining base line and humps of rectangular, trapezoidal, or triangular shapes corresponding to various prosodic elements [2-7].

In view of the fact that a piecewise-linear function can yield a reasonable approximation to any continuous curve, it is not surprising that a fair degree of naturalness can be obtained in speech synthesis by specifying only one F_0 value for each mora and by constructing a piecewise-linear contour connecting these values [5-6]. Although such a simplification may seem to provide an apparent advantage, the advantage is more than offset by the complexity of the rules for generating close approximations to observed F_0-contours, and by the lack of clear and simple correspondence between the characteristics of such crude approximations and the underlying prosodic elements. For example, it is often claimed under the assumption of the existence of such linear declination, that the rate of declination varies with the sentence length [6-8], and therefore the speaker must be provided with a "look-ahead" mechanism to control the rate in accordance with the length of an utterance yet to be produced [8]. Likewise, the magnitude of the hump corresponding to the same presodic element does not remain constant when it is expressed by the absolute value, but varies with its position in a sentence. The simplicity and the generality of rules for F_0-contour synthesis thus depend strongly on the accuracy of the model in describing the essential characteristics.

It was from this point of view that a functional model was proposed for generating F_0-contours of various word-accent types in the *Tokyo* and the *Osaka* dialects of Japanese, and was demonstrated to be capable of producing close approximations to these contours from binary commands for voicing and accent [9-11]. Some of the model's parameters were found to correspond to linguistic information on the word accent types, while the other parameters were found to be characteristic of the F_0 control mechanism of the individual speakers. The model also proved to be valid for disyllabic words of English [12, 13]. A preliminary effort was made to extend the model to F_0-contours of spoken Japanese sentences, showing that at least two additional intonational factors, i.e., those of interrogation and emphasis, had to be incorporated into the model [14].

Based on the results of these preliminary studies, the present report describes an improved formulation of F_0-contours of simple declarative

sentences of Japanese, and presents a quantitative analysis of the relationship between the sentence length and the declination of the F_0-contours, as well as an analysis of the variations in the characteristics of humps corresponding to various prosodic elements found within an utterance.

2. Speech Materials

Considering the aforementioned factors, the selection of the utterances for this study observed the following conditions.

(1) The utterances to be analyzed present a continuous F_0-contours and the fundamental frequencies are easily extracted.

(2) The sample sentences possess various lengths.

(3) The same words appear in different positions of the sentences.

Table 1 shows the list of ten sentences selected for the present analysis. The sample sentences contain only the voiced segments, ranging in length from 8 to 24 morae. A male Japanese speaker of the *Tokyo* dialect read the list naturally without respiratory pauses for each sentence and the readings were recorded on an audio tape. The list of 10 sentences were read five times, but the first reading was excluded from the analysis. The speech materials were sampled at 10 kHz, quantized with 10 bit accuracy and stored in a magnetic tape memory. A detailed analysis was made on the readings of sentences 1, 2 and 8 of Table 1.

Table 1. List of sentences adopted for the analysis of F_0-contours.

Number	Phonemic transcription	Number of morae
1	/aoinoewaaru/	8
2	/aoiaoinoewaaru/	11
3	/aoinoewaieniaru/	11
4	/anoaoiaoinoewaaru/	13
5	/aoiaoinoewaieniaru/	14
6	/aoinoewajamanouenoieniaru/	17
7	/aoiaoinoewauenoieniaru/	17
8	/aoiaoinoewajamanouenoieniaru/	20
9	/anoaoiaoinoewajamanouenoieniaru/	22
10	/anoaoiaoinoewaanojamanouenoieniaru/	24

3. Analysis of Sentence F_0-contour

3.1 Model of the sentence F_0-contour generation

Through the preliminary analysis of the extracted sentence F_0-contours, it became clear to us that, similar to the word F_0-contour, the sentence F_0-contours can be decomposed into two types of components, viz., into the voicing and accent components. They correspond to the gradual decay and the local humps of the F_0-contour, respectively. In

order to formulate the process of sentence F_0-contour generation, we made three assumptions.

(1) Commands for voicing and accent take the form of step function.

(2) Separate mechanisms exist for voicing and accent, which can be approximated by linear systems that convert these commands into their respective control signals.

(3) These control signals are combined and applied to the mechanism of glottal oscillation, whose fundamental frequency is an exponential function of the control signal.

The model of sentence F_0-contour generation based on the above assumptions is illustrated schematically in Fig. 1. Note that the number of the voicing commands and the number of the accent commands are not limited to one, which differs from the case of word F_0-contours. By representing the amplitudes of the i-th voicing and the j-th accent command as Av_i and Aa_j, respectively, the output fundamental frequency $F_0(t)$ as a function of time t is given by

$$\ln[F_0(t)/F_{min}] = \sum_{i=1}^{I} Av_i \{Gv_i(t - T_{0i}) - Gv_i(t - T_{3i})\} + \sum_{j=1}^{J} Aa_j \{Ga_j(t - T_{1j}) - Ga_j(t - T_{2j})\} . \quad (1)$$

where F_{min} is the lowest frequency of the vocal chord vibration; I and J are the number of voicing and accent commands, respectively; T_{0i} is the onset time (in sec) and T_{3i} is the offset time of the i-th voicing command; T_{1j} is the onset time and T_{2j} is the offset time of the j-th accent command. In accordance with the word F_0-contour, Gv_i and Ga_j in equation (1) can be approximated by the following step response functions of the critically damped linear system of the second order:

$$Gv_i(t) = \alpha_i\, t \exp(-\alpha_i\, t)\; u(t) , \quad (2)$$

$$Ga_j(t) = \{1 - (1 + \beta_j\, t) \exp(-\beta_j\, t)\}\, u(t) , \quad (3)$$

where $u(t)$ is the unit step function. In the absence of respiratory pauses within a spoken sentence, the offset time T_{3i} for all the voicing commands

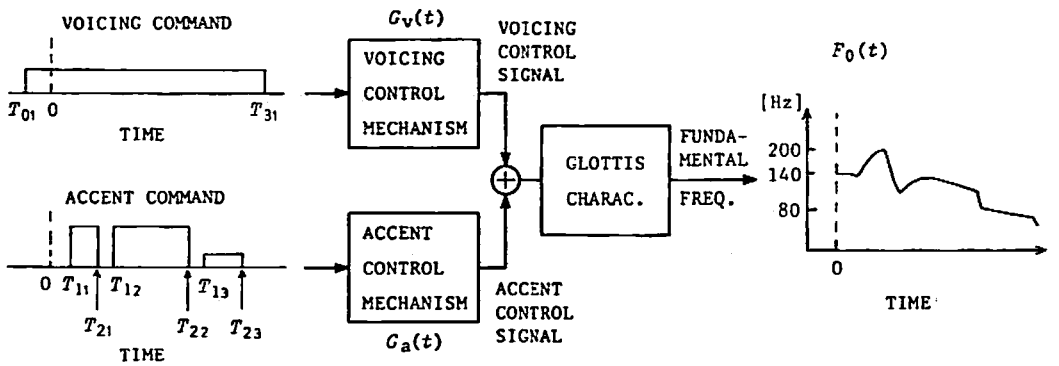

Fig. 1. A functional model for the process of generating a sentence F_0-contour.

are assumed to be identical for all i's within an utterance. On the other hand, the following restriction is introduced on the timing of the accent commands,

$$T_{2j} < T_{1j+1}, \tag{4}$$

in order to avoid the overlap between accent components.

3.2 Method of analysis

The F_0-contour analysis involves the extraction of fundamental frequency followed by the feature extraction of F_0-contour. The fundamental periods are detected pitch synchronously by the method based on the short-term autocorrelation analysis and the peak detection.[15]. These fundamental periods are converted to fundamental frequencies, which are further interpolated to produce F_0-contours uniformly sampled at intervals of 12.8 msec. The parameters of the above-mentioned model are determined by minimizing the mean squared error between the extracted F_0-contour and that of the model in logarithmic scale. The minimization process utilizes the method of Analysis-by-Synthesis, and is conducted as follows[10]. First, the number of voicing commands I and that of the accent commands J for the sample are determined by the preliminary observation of the extracted F_0-contour. Second, the point of minimun mean squared error is sought using the computer in the $4(I+J)+1$ dimensional space, composed of the parameters for the voicing, viz., T_{0i}, T_{3i}, A_{vi}, α_i, the parameters for the accent, viz., T_{1j}, T_{2j}, A_{aj}, β_j, and the lowest frequency of vocal chord vibration F_{min}. The optimal values of these parameters are defined as the characteristic values of the observed F_0-contour. The initial values and the intervals of the parameters for the search are given manually to the computer. An appropriate selection of the initial values is important for an efficient search. The intervals used in the present analysis are 2 Hz for F_{min}, 0.01 sec for T_0, T_3, T_1, T_2, 0.05 for A_v, 0.02 for A_a, 0.2 sec^{-1} for α, and 0.5 sec^{-1} for β.

3.3 Results of analysis

In Fig. 2 the extracted logarithmic fundamental frequencies (indicated by "o") and the best approximations by the model (solid lines) are shown for the samples: (a) sentence 1 (8 morae), (b) sentence 2 (11 morae), and (c), (d) sentence 8 (20 morae) listed in Table 1. For ease of exposition, the extracted fundamental frequencies are shown with intervals of 38.4 msec in the figure. The number of voicing commands I and that of accent commands J for Fig. 2 (a) through (d) are (I=1, J=2), (I=1, J=3), (I=1, J=4), and (I=2, J=4), respectively.

When the number of voicing commands I is assumed to be one, the discrepancy between the observed F_0-contour and its best approximation by this model is very small in the case of sentences 1 and 2 as shown in Fig. 2 (a) and (b), but is large especially in the latter half of sentence 8 as shown in Fig. 2(c). On the other hand, if I is assumed to be 2, the matching in sentence 8 between the observed contour and the approximation becomes as close as those in sentences 1 and 2 as shown in Fig. 2(d). Consequently, the number of voicing commands for the samples of sentence 8 must be two, although the sentence was read in one breath. This fact implies that each

voicing command does not necessarily correspond to a breath group but, instead, closely correlates with the syntactic structure of the sentence. The seemingly slow declinations observed in long sentences can be ascribed to the existence of such "re-voicng" within a sentence as exemplified above, in addition to the exponential decay in the voicing component.

The close agreement between the abserved F_0-contours and the best approximations, shown in Fig. 2(a), (b) and (d) confirms the validity of the model, and indicates that the parameters of the model determined by the

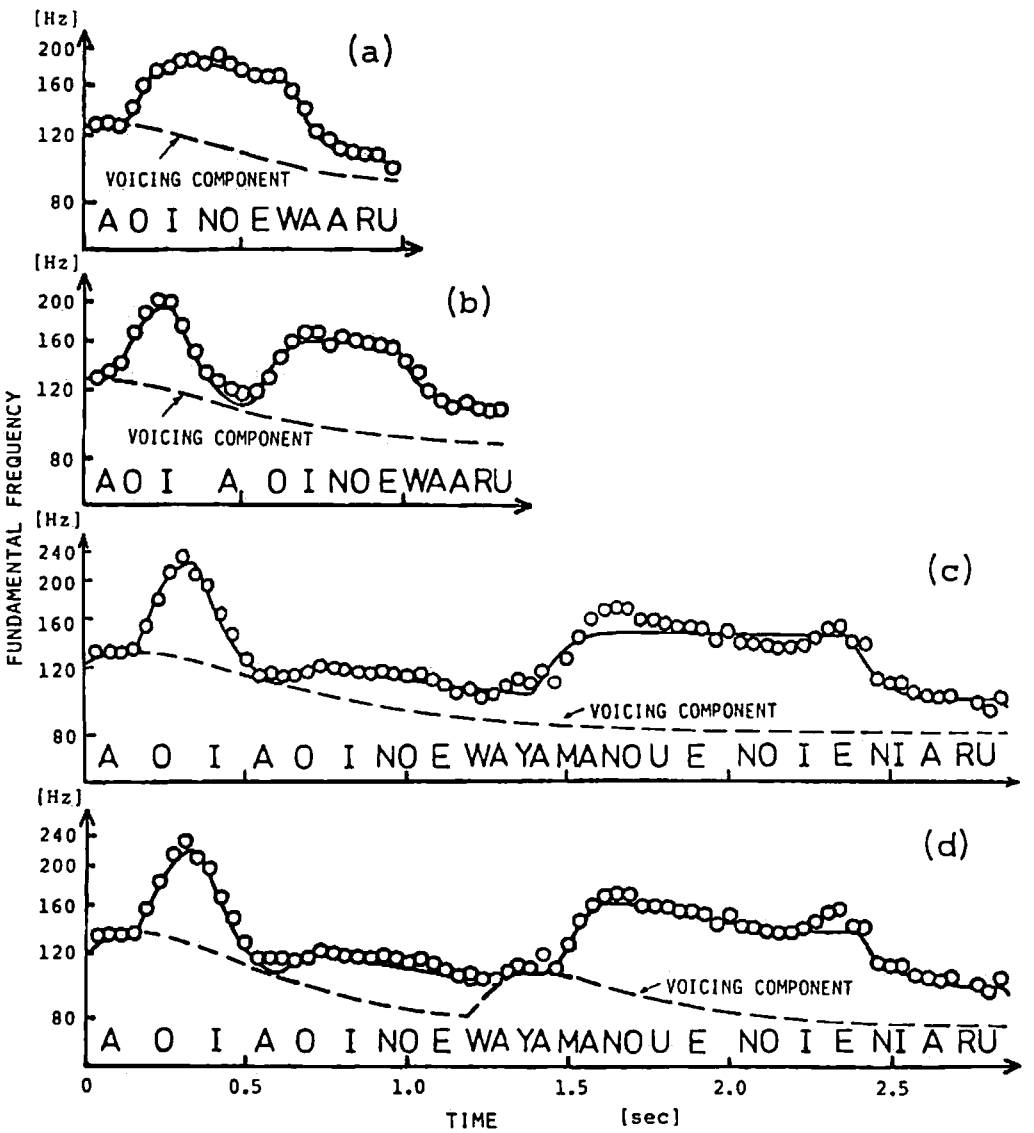

Fig. 2. Analysis-by-Synthesis of sentnece F_0-contours: (a)/aoinoewaaru/, (b)/aoiaoinoewaaru/, (c) and (d)/aoiaoinoewajamanouenoieniaru/. The number of voicing commands is one for (a), (b) and (c), and is two for (d).

method mentioned in section 3.2 can represent the precise characteristics of the sentence F_0-contour. The values of the parameters of the best approximations are listed in Table 2.

Table 2. Parameters of sentence F_0-contours extracted by Analysis-by-Synthesis without constraints.

Sentence	F_{min} (Hz)	i	T_0 (sec)	T_3 (sec)	Av	α (sec^{-1})	j	T_1 (sec)	T_2 (sec)	Aa	β (sec^{-1})
Ex.(1)	83	1	-0.21	1.01	1.18	3.3	1	0.13	0.64	0.46	20.5
8 morae							2	0.71	0.99	0.13	20.8
Ex.(2)	84	1	-0.21	1.42	1.20	3.6	1	0.08	0.27	0.54	21.2
11 morae							2	0.51	1.00	0.49	20.5
							3	1.02	1.34	0.20	20.0
	76	1	-0.13	2.90	1.55	3.4	1	0.15	0.34	0.64	20.0
Ex.(8)		2	1.19	2.90	0.80	4.2	2	0.59	1.15	0.24	23.5
20 morae							3	1.44	2.40	0.54	21.5
							4	2.44	2.84	0.24	30.5

4. Characteristics of Sentence F_0-contour

The analysis of F_0-contours in the preceding section was conducted without any constraints on the values of various parameters characterizing the voicing and accent components of a sentence. It is, however, natural to assume that some of the parameters, i.e., such as those characterizing the speaker's physiological mechanism, may not vary over a wide range but remain rather constant within a sentence or even across sentences. On the other hand, parameters conveying linguistic information are expected to vary over a wide range depending on the information content of each element. It is also expected that the distribution of values of the latter parameters may not be unimodal but may rather be multimodal, reflecting the discrete nature of the linguistic code.

Among the parameters characterizing the accent component, β_j represents the response rate of the glottal control mechanism to an accent command, so that its value remains rather constant, as shown in Table 1. On the other hand, the magnitude Aa_j of the accent command is seen to be distributed over two distinct ranges of values: one from 0.46 to 0.62, and the other from 0.15 to 0.24. These results suggest that the approximation by the present model to an F_0-contour may not be seriously impaired by constraining all the β_j's to be equal, and by allowing only two values for the Aa_j's. As for the voicing components, their overall characteristics are naturally expected to be similar, if not equal in magnitude, when there are two or more voicing components in a sentence. Hence the value of α_i may also be constrained to be equal without seriously impairing the approximation by the model.

Figure 3 compares the results of Analysis-by-Synthesis of the F_0-contour of one utterance sample of sentence 8 "/aoiaoinoewajamanoueno-

ieniaru/, " obtained with various constraints. Figure 3(a) shows the result obtained without any constraints, Fig. 3(b) shows the result obtained when all the β_j's are constrained to be equal, Fig. 3(c) shows the results obtained with an additional constrant on Aa_j, viz., all the Aa_j's being constrained to take either one of two values, and Fig. 3(d) shows the results obtained with still another constraint on α_i, viz., the two α_i's being constrained to be equal. These results indicate the validity of the present

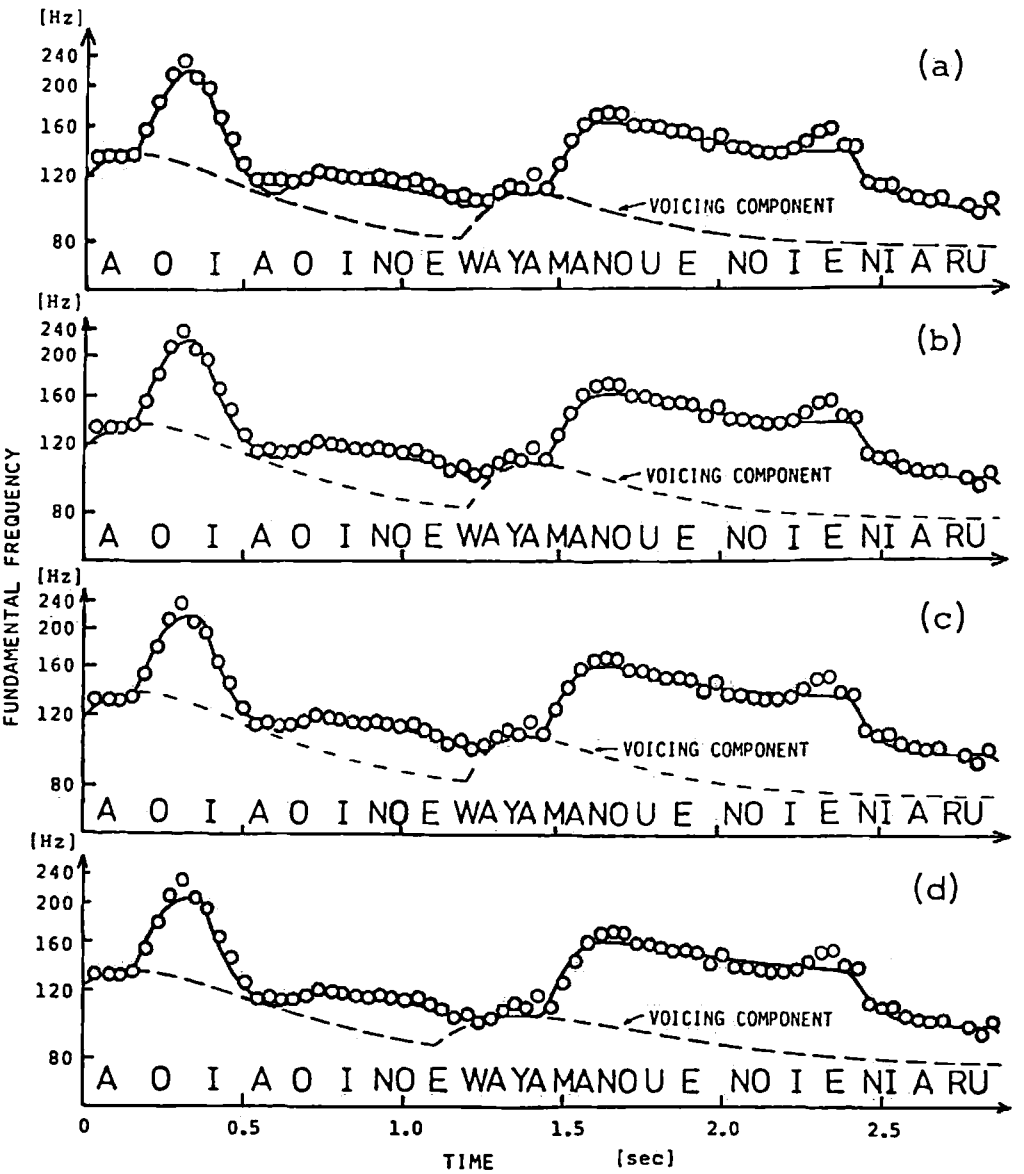

Fig. 3. Analysis-by-Synthesis of the sentence F_0-contour of /aoiaoinoewajamanouenoieniaru/; (a) without constraints, (b) with a constraint on β, (c) with constraints on both β and Aa, and (d) with constraints on α, β and Aa.

formulation in obtaining simple yet accurate rules for generating realistic F_0-contours of sentences. The values of the parameters for sentences 1, 2 and 8 obtained under the same condition as in Fig. 3 (d) are listed in Table 3.

Table 3. Parameters of sentence F_0-contours extracted by Analysis-by-Synthesis with constraints on α, β and Aa.

Sentence	F_{min} (Hz)	i	T_0 (sec)	T_3 (sec)	Av	α (sec^{-1})	j	T_1 (sec)	T_2 (sec)	Aa	β (sec^{-1})
Ex.(1)	83	1	-0.22	1.01	1.18	3.3	1	0.12	0.64	0.46	20.3
8 morae							2	0.72	0.99	0.15	20.3
Ex.(2)	84	1	-0.21	1.42	1.20	3.6	1	0.07	0.28	0.50	20.8
11 morae							2	0.51	1.00	0.50	20.8
							3	1.02	1.34	0.20	20.8
	76	1	-0.17	2.90	1.55	3.0	1	0.15	0.36	0.52	24.0
Ex.(8)		2	1.10	2.90	0.65	3.0	2	0.61	1.08	0.22	24.0
20 morae							3	1.43	2.39	0.52	24.0
							4	2.39	2.84	0.22	24.0

The relationships between the sentence length and the parameters α, β are shown in Fig. 4, where the ordinate indicates the values of α and β obtained under the same condition as in Fig. 3(d), and the abscissa indicates the sentence length represented in terms of the number of morae n within a sentence. The influence of n upon α and β are seen to be minimal. The constancy of α indicates that the shape of the natural declination in the fundamental frequency is almost independent of the sentence length within our formulation. As mentioned in Section 3.3, the apparent slow declination in longer sentences can be well explained within our hypothesis of "re-voicing." The relationships between the sentece length and the

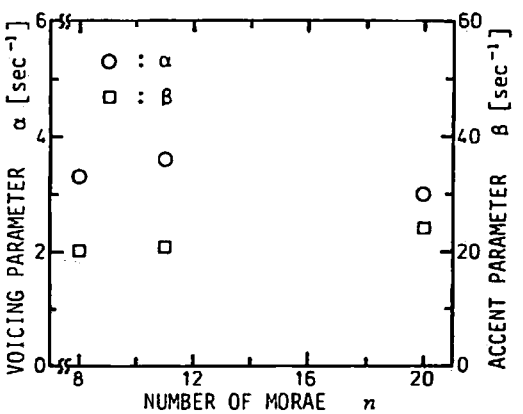

Fig. 4. Time parameters of voicing and accent components versus number of morae in a sentence.

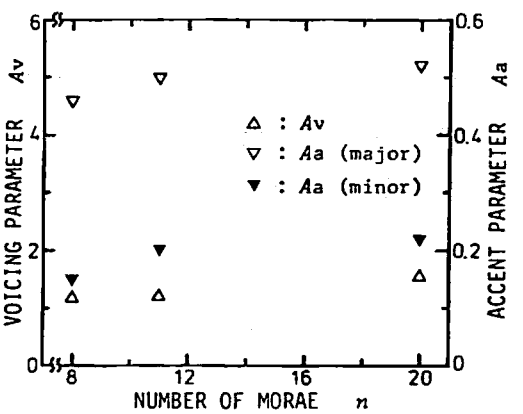

Fig. 5. Amplitude parameters of voicing and accent components versus number of morae in a sentence.

amplitude parameters of voicing A_V or of accent A_G are shown in Fig. 5, indicating that the influence of n upon these parameters is also negligible. The ratio of the two values determined for the amplitude of the accent commands also does not vary appreciably with n, suggesting that the amplitude is controlled in a binary way to show the presence/absence of emphasis, at least as far as the speech samples for the present study are concerned.

5. Conclusion

In order to elucidate the essential characteristics of the sentence F_0-contour in Japanese, and to establish simple and universal rules for the synthesis of F_0-contour of speech, an attempt was made to extend the model of the word F_0-contour already proposed and validated in formulating the characteristics of F_0-contours of both Japanese and English words. The present model is based on the formulation of the F_0-contour on the logarithmic frequency scale, and allows for the existence of one or more voicing and accent commands within a sentnece. The model's validity was tested by an analysis of F_0-contours of various declarative sentences of Japanese. It was shown that the characteristics of the sentence F_0-contour can be represented by a rather small number of parameters, and, conversely, that the sentence F_0-contour can be reproduced with high accuracy from these parameters. Within our formulation, it was shown that the shape of the voicing component, responsible for the natural declination of the F_0-contour, does not fluctuate appreciably with the sentence length, and that the amplitude of the accent command can be constrained to have only one of two values, corresponding to the presence/absence of emphasis. These results lead to simple and accurate rules for describing as well as for reproducing the essential characteristics of F_0-contours of declarative sentences in Japanese. Further investigations are necessary, however, on the effects of syntax, speech rate, and speaker idiosyncracies, using a larger number of utterance samples than in the present study.

References

1. Öhman, S. (1967); "Word and Sentence Intonation: A Quantitative Model," Speech Transmission Laboratory Quarterly Progress and Status Report, STL-QPSR 2-3/1967, 20-54.

2. Isačenko, A. V. and H. J. Schädlich (1966); "Untersuchungen über die deutsche Satzintonation," Studia Grammatica 7, 7-67.

3. 't Hart, J. (1966); "Perceptual Analysis of Dutch Intonation Features," I. P. O. Annual Progress Report 1, 47-51.

4. Hashimoto, S. (1968); "A Study of Intonation and Its Application to Speech Synthesis Rule," Rec. Fall Meeting, Acoust. Soc. Japan, 2-3-11, 155-156.

5. Hashimoto, S. and S. Miyahara (1974); "A Method of Linear Approximation Model of Pitch Contour," Trans. of the Committee on Speech Research, Acoust. Soc. Japan. S74-11.

6. Hakota, K. and H. Sato (1975); An Experimental Study on Characteristics of Pitch Patterns in Spoken Sentences," Record of 1975 National Convention of the Institute of Electronics and Communication Engineers of Japan, 1236, 1143.

7. Maeda, S. (1974); "A Characterization of Fundamental Frequency Contours of Speech," Quarterly Progress Report, No. 114, Research Laboratory of Electronics, M. I. T., 193-211.

8. Breckenridge, J. (1977); "The Declination Effect," J. Acoust. Soc. Am. 61, Suppl. 1, S90.

9. Fujisaki, H. and S. Nagashima (1969); "A Model for Synthesis of Pitch Contours of Connected Speech," Annual Report, Eng. Res. Inst., University of Tokyo 28, 53-60.

10. Fujisaki, H. and H. Sudo (1971); "A Model for the Generation of Fundamental Frequency Contours of Japanese Word Accent, J. Acoust. Soc. Japan 27, 445-453.

11. Fujisaki, H. and M. Sugito (1978); "Analysis and Perception of Two-Mora Word Accent Types in the Kinki Dialect," J. Acoust. Soc. Japan 34, 167-176.

12. Hirose, K., H. Fujisaki and M. Sugito (1978); "Acoustic Correlates of Word Accent in English and Japanese," Trnas. of the Committee on Speech Research, Acoust. Soc. Japan, S78-41.

13. Hirose, K., H. Fujisaki and M. Sugito (1978); "Word Accent in Japanese and English: A Comparative Study of Acoustic Characterisitcs in Disyllabic Words," J. Acoust. Soc. Am. 64, Suppl. 1, S114.

14. Fujisaki, H. and H. Sudo (1973); "Prosodic Rules of Speech — A Model for the Synthesis of Intonation in Standard Japanese," in Speech Information Processing, S. Hiki, ed., University of Tokyo Press, Tokyo, 123-142.

15. Fujisaki, H. and Y. Tanabe (1973); "A Time-Domain Technique for Pitch Extraction of Speech," J. Acoust. Soc. Japan 29, 418-419.