Ann.Bull.RILP(1988)No.22, 183-191

REALIZATION OF LINGUISTIC INFORMATION IN THE VOICE FUNDAMENTAL FREQUENCY CONTOUR OF THE SPOKEN JAPANESE*

Hiroya Fujisaki** and Hisashi Kawai**

1. INTRODUCTION

The prosody of the spoken Japanese has two major factors, i. e., the word accent and the intonation, which are both manifested by the contour of the voice fundamental frequency (henceforth F_0 contour). The former reflects the lexical information and appears as local rise/fall patterns of the F_0 contour, while the latter reflects the syntactic information and appears as more or less global ondulations of the F_0 contour. Previous studies[1)-10)] by Fujisaki and his coworkers have shown that the F_0 contour of an utterance can be decomposed into two types of components (i. e., the accent components and the phrase components) which are closely related to these factors. These studies, however, have also made clear that these components are not straightforward manifestations of the lexical word accent and the syntactic structure, but have structures of their own and are also influenced by the discourse structure. In the present paper, we define prosodic units of spoken Japanese on the basis of F_0 contour characteristics, and describe some of our experimental findings on how various informations are manifested, or even fail to be manifested in certain situations, by the characteristics of the F_0 contour.

2. PROSODIC UNITS OF SPOKEN JAPANESE

Conventional textbooks on Japanese grammar posit a syntactic unit which consists of a content word with or without following function words. From the point of view of prosody, however, this is not necessarily the minimal unit. As the minimal prosodic unit of spoken Japanese, we introduce the "prosodic word", which is defined as a part or the whole of an utterance that forms an accent type. As will be discussed later, a string of prosodic words, under certain conditions, can form a larger prosodic word due to "accent sandhi". On the other hand, a phrase component

---

* A version of this paper was presented at the 1988 IEEE International Conference on Acostics, Speech, and Signal Processing, New York, April 1988.

** Dept. of Electronic Engineering, Faculty of Engineering, The University of Tokyo

of the F_0 contour of an utterance defines a larger prosodic unit, i. e. a "prosodic phrase", which may contain one or more prosodic words. Generally, a prosodic word never extends over two prosodic phrases. Furthermore, in longer sentences, several prosodic phrases may form a section delimited by pauses. Such a section is defined as a "prosodic clause". On the other hand, we adopt word, phrase, clause and sentence as syntactic units. Roughly speaking, the following parallelism exists between the hierarchy of syntactic units and the hierarchy of prosodic units.

syntactic units:			
word ———	phrase ——	clause ——	sentence
prosodic units:			
prosodic word	prosodic phrase	prosodic clause	spoken sentence.

3. METHOD OF ANALYSIS

The F_0 contour of an utterance can be regarded as the response of the mechanism of vocal cord vibration to a set of commands which carry information concerning lexical accent, syntactic and discourse structures of the utterance. Two different kinds of command have been found to be necessary; one is an impulse-like command for the onset of a prosodic phrase while the other is a stepwise command for the accented mora or morae of a prosodic word. Consequences of these two types of commands have been shown to appear as the phrase components and the accent components, each being approximated by the response of a second-order linear system to the respective commands. If we represent an F_0 contour as a pattern of the logarithm of the fundamental frequency along the time axis, it can be approximated by the sum of these components. The entire process of generating an F_0 contour of a sentence can thus be modeled by the block diagram of Fig. 1.

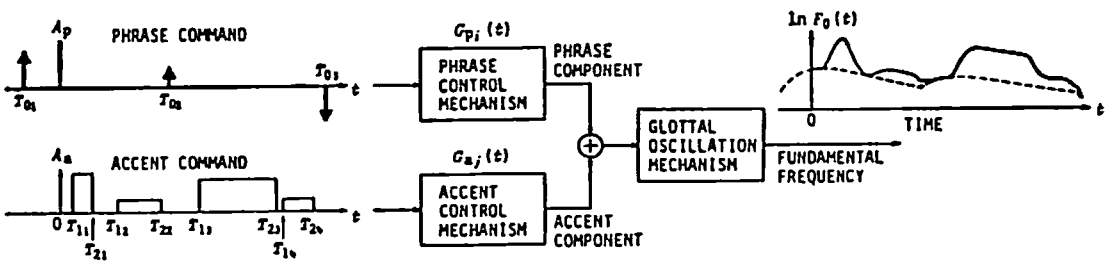

Fig. 1. A functional model for the process of generating sentence F_0 contours.

In the following analysis, the model is used to decompose a given F_0 contour into its constituents, i. e., the phrase components and the accent components. This is accomplished by finding, by the method of Analysis-by-Synthesis, the optimum set of model parameters that gives minimum mean squared error between the measured F_0 contour and the model-generated contour. The results of such decomposition are then used to examine the influences of various linguistic factors upon the accent and the phrase components.

4. SPEECH MATERIALS

The speech materials for the present study consisted of four sets of utterances. Sets 1, 2 and 3 contained simple phrases embedded in a sentence while Set 4 contained longer sentences.

Phrases of Set 1 consist of two prosodic words. The first prosodic word (W_1) is an adjective phrase, and the second prosodic word (W_2) is a noun phrase. For each of W_1 and W_2, two prosodic words of different accent types are selected, i. e., the accent type with a rapid downfall (to be denoted by D) and the type without such a downfall (to be denoted by F). Phrases of Set 2 consist of three prosodic words with two different syntactic structures, i. e., the left-branching and the right-branching structures. Phrases of Set 3 consist of four or five prosodic words with a left-branching structure. These phrases are uttered in the context of "___kaimasu." Furthermore, focal condition is controlled by having the sentence uttered as an answer to a preceding question. These sets of utterances are recorded by two speakers of the Tokyo dialect, and are used mainly to investigate the influences of syntactic and discourse information on the accent components.

Utterances of Set 4, on the other hand, are spoken sentences of a weather forecast recorded by two radio announcers, and are used mainly to investigate the influences of syntactic and discourse information on the phrase components.

The recorded materials were digitized at 10 kHz with 12 bit accuracy, and the fundamental frequency contours were extracted for further analysis by the method described in the preceding section.

5. SYNTACTIC AND DISCOURSE INFORMATION IN ACCENT COMPONENTS

Figure 2 shows the results of analysis for materials from Set 1 uttered by one speaker. The panels on the left refer to utterances of the DD type, while those on the right refer to utterances of the FD type. The uppermost panels show utterances without intended focus (i. e. default condition), while the middle and the lowermost panels show utterances with focus on W_1 and W_2, respectively. In each panel, the "+" symbols indicate a measured F_0 contour, the curve displayed by a solid line indicates the best approximation given by the model-based analysis, and the curve displayed by a dashed line indicates the phrase component.

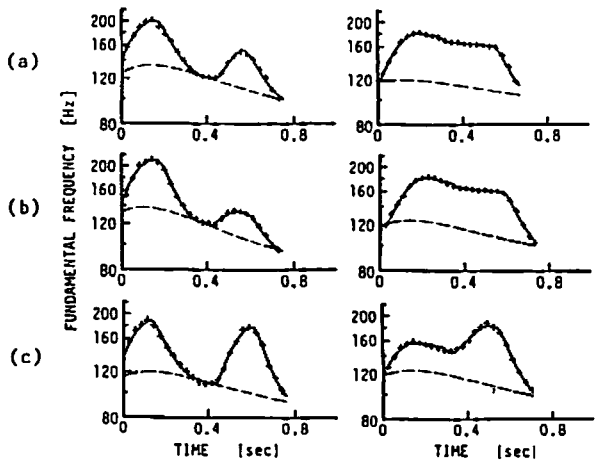

Fig. 2. Results of F_0 contour analysis for W_1W_2.
Left: "anino(W_1) amaguo(W_2)" (DD type),
Right: "aneno(W_1) amaguo(W_2)" (FD type),
(a) without intended focus,
(b) with focus on W_1,
(c) with focus on W_2.

In utterances of the DD type, the amplitude of accent component of W_1 remains unchanged against changes in the focal condition, and is always as large as its value when W_1 is uttered in isolation. On the other hand, the amplitude of accent component of W_2 varies with the focal condition. The results for utterances of the DF type are not shown in the figure, but display similar characteristics.

In utterances of the FD type, the amplitude of accent component of W_2 remains unchanged and is always as large as its value when W_2 is uttered in isolation. On the other hand, the accent component of W_1 varies with the focal condition and is elevated to the same amplitude as that of W_2 both in the default and in the W_1-focused conditions. In these cases, the accent components for W_1 and W_2 are merged into one and form a larger prosodic word ("accent sandhi"), and the two cases are not differentiated. When focus is placed on W_2, the accent sandhi does not occur and the accent component of W_1 assumes a lower value which is usually found when the word is uttered in isolation.

Figure 3 shows the results of analysis for materials from Set 2, only for utterances of the DFD type. The panels on the left refer to the left-branching, and those on the right refer to the right-branching structure. From top to bottom, these panels show utterances without intended focus, then with focus on W_1, W_2, and W_3.

The results shown in Fig. 3 indicate that the amplitude of the accent component of W_1 remains unchanged against changes in the focal condition and in the syntactic structure, and is similar to its value when the word is uttered in isolation. The difference in the syntactic structure is indicated by the absence/presence of an additional phrase component for W_2W_3, but the amplitude of the second phrase component (found in the right-branching syntax) varies rather widely depending on the focal condition. Thus, for example, the difference in the syntactic structure is not well expressed in the W_1-focused condition. In the case of left-branching syntax, differences in the focal condition are fairly well expressed by the F_0 contour (except for the default condition). In the case of right-branching syntax, however, expressions of focal differences are quite limited because of the accent sandhi between W_2 and W_3. These results show that the information on accent types, syntactic structure, and discourse structure can not always be manifested by the F_0 contour.

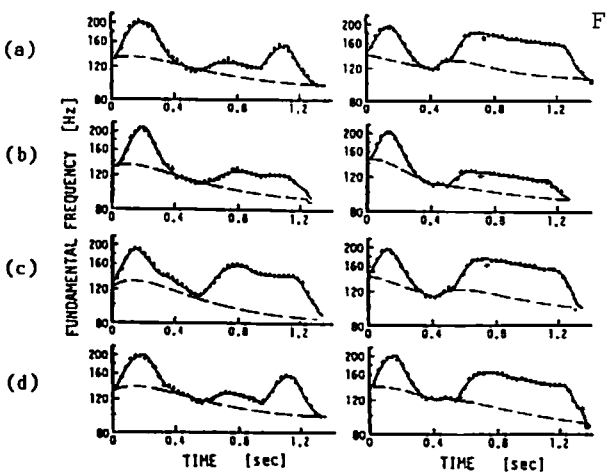

Fig. 3. Results of F_0 contour analysis for $W_1W_2W_3$ (DFD type).
Left: "aomorino aneno amaguo"(left-branching structure),
Right: "anino mizuirono amaguo" (right-branching structure),
(a) without intended focus,
(b) with focus on W_1,
(c) with focus on W_2,
(d) with focus on W_3.

Figure 4 shows results of F_0 contour analysis of two cases of three prosodic words where syntactic and discourse structures present conflicting requirements. In panel (a) for the left-branching structure and focus on W_2, accent sandhi is seen to occur between W_2 and W_3, indicating that the speaker opted to give priority to the discourse requirement over the syntactic requirement. In panel (b) for the right-branching structure and focus on W_1, on the other hand, the accent component of W_1 is not prominent in spite of the discourse requirement, indicating that the speaker opted to give priority to the syntactic requirement. In these situations, it is generally unpredictable which of the two requirements are met by the speaker.

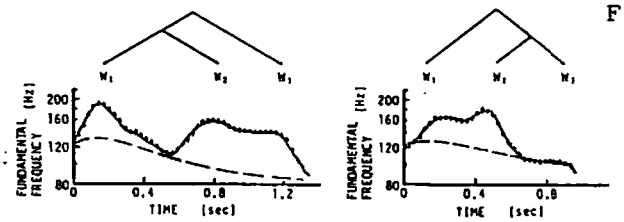

Fig. 4. Two examples in which syntax and discourse present conflicting requirements.
(a) "aomorino aneno amaguo" (DFD, focus on W_2), priority is given to discourse requirement,
(b) "aneno aono amaguo" (FDD, focus on W_1), priority is given to syntactic requirement.

6. SYNTACTIC INFORMATION IN PHRASE COMPONENTS

While the accent component defines a prosodic word, larger prosodic units are defined by the phrase component. Analysis of speech materials of Set 4 indicates that there are three different ways of starting a new phrase component: (1) a phrase component may be preceded by a pause during which the immediately preceding phrase component is completely reset, (2) a phrase component may be preceded by a brief pause but the immediately preceding phrase component is not reset so that the new phrase component is superposed on the old one, (3) a phrase component may be simply added to the old one without pause. Prosodic boundaries marked by these different ways shall be named Type I, Type II, and Type III boundaries. Sentences and prosodic clauses are marked by prosodic boundaries of Type I and Type II, while prosodic phrases are marked by Type III boundaries.

The occurrence of these prosodic boundaries is primarily determined by the syntactic structure, and most probably coincides with syntactic boundaries. It is, however, also subject to other factors such as style of speaking, respiration,

etc. As for the syntactic boundaries at which prosodic boundaries may occur, we may distinguish four different boundary types: (1) between sentences, (2) between clauses, (3) between immediate constituents with a recursively left-branching structure, (4) within an immediate constituent with a recursively left-branching structure. An immediate constituent with a recursively left-branching structure (to be abbreviated by ICRLB) is a syntactic phrase which is delimited by right-branching boundaries and contains only left-branching boundaries. The relationships between these syntactic boundaries and the three types of prosodic boundary are listed in Table 1. The correspondence between the syntactic boundaries and the prosodic boundaries is not one-to-one but is rather stochastic, and the probability that a prosodic boundary occurs at a given syntactic boundary is influenced also by the depth of the syntactic boundary as defined elsewhere[8].

Table 1. Classification of syntactic boundaries and their manifestations as prosdic boundaries.

Prosodic boundaries	Syntactic boundaries			
	Between sentences	Within a sentence		
		Between clauses	Within a clause	
			Between ICRLB's*	Within an ICRLB*
Type I (Phrase resetting with pause)	100%	20%		
Type II (Phrase addition with pause)		30%	5%	
Type III (Phrase addition without pause)		50%	85%	15%

*ICRLB: An immediate constituent with a recursively left-branching structure.

7. CONCLUSIONS

Prosodic units of spoken Japanese have been defined on the basis of components of observed F_0 contours. The influences of various linguistic factors such as lexical word accent, syntactic structure and discourse structure upon the accent components of prosodic words have been described on the basis of analysis of F_0 contours of a number of utterances. The relationships between prosodic boundaries defined by the phrase components and the syntactic boundaries have also been discussed. Furthermore, our analysis has indicated that there are cases where prosody fails to meet all the requirements presented by word accent, syntax and discourse. These results serve as a basis for a high quality synthesis of speech from text.

REFERENCES

1) Fujisaki, H. and Nagashima, S.: "A Model for Synthesis of Pitch Contours of Connected Speech," Annu. Rep. Eng. Res. Inst., Univ. Tokyo, 28, pp. 53-60, 1969.
2) Fujisaki, H. and Sudo, H.: "Synthesis by Rule of Prosodic Features of Connected Japanese," Proc. 7th ICA, 23C2, 1971.
3) Fujisaki, H. and Sudo, H.: "A Generative Model for the Prosody of Connected Speech in Japanese," Conference Record, 1972 IEEE-AFCRL Conference on Speech Communication and Processing, pp. 140-143, 1972.
4) Fujisaki H. and Hirose, K.: "Modeling the Dynamic Characteristics of Voice Fundamental Frequency with Applications to Analysis and Synthesis of Intonation," Preprints of Papers, Working Group on Intonation, XIIIth International Congress of Linguists, Tokyo, pp. 109-119, 1982.
5) Fujisaki H. and Hirose, K.: "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese," J. Acoust. Soc. Jpn. (E), Vol. 5, No. 4, pp. 233-242, 1984.
6) Hirose, K., Fujisaki, H. and Yamaguchi, M.: "Synthesis by Rule of Voice Fundamental Frequency Contours of Complex Sentences," Proc. ICASSP 84, 2.13, 1984.
7) Fujisaki, H., Hirose, K., Takahashi, N. and Yoko'o, M.: "Realization of Accent Components in Connected Speech," Trans. of the Committee on Speech Research, Acoust. Soc. Jpn., S84-36, 1984.
8) Hirose, K., Fujisaki, H. and Kawai, H.: "Generation of Prosodic Symbols for Rule-synthesis of Connected Speech of Japanese ," Proc. ICASSP 86, 45.4, 1986.
9) Kawai, H., Hirose, K. and Fujisaki, H.: "Quantiza-tion of Accent Command Amplitude for Rule Synthesis of Fundamental Frequency Contours," Trans. of the Committee on Speech Research, Acoust. Soc. Jpn., SP86-93, 1987.
10) Fujisaki H. and Kawai, H.: "Realization of Word Accent in Connected Speech of Japanese," Session Paper Se 15.5, 11th ICPhS, 1987.