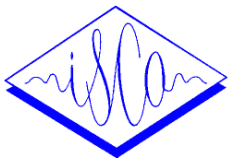

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

MANIFESTATION OF LINGUISTIC AND PARA-LINGUISTIC INFORMATION IN THE VOICE FUNDAMENTAL FREQUENCY CONTOURS OF SPOKEN JAPANESE

Hiroya Fujisaki, Keikichi Hirose and Noboru Takahashi

Dept. of Electronic Engineering, Faculty of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

ABSTRACT

Prosodic features of spoken language play an important role in the transmission of linguistic information concerning word meaning, sentence structure and discourse structure but also in the transmission of para- and non-linguistic information such as speaker's intention/emotion, idiosyncrasy, and naturalness.

In this paper, we first define the units of prosody of the spoken Japanese on the basis of analysis of voice fundamental frequency contours of a large number of spoken sentences, and then clarify the relationship between linguistic and para-linguistic information and the components of the F_0 contour. While the phrase components convey mainly the syntactic information the accent components convey the information on accent type, syntactic structure, and discourse structure. Para-linguistic information is mainly conveyed by an additional accent component at the sentence end.

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 mainly reflects the lexical information and appears as local rise/fall patterns of the F_0 contour, while the latter mainly reflects the syntactic information and appears as more or less global undulations of the F_0 contour. Previous studies 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 [1-3]. These studies, however, have also made clear that these components are not straightforward manifestations of the lexical word accent and the syntactic structure, and are also influenced by the discourse structure and the para-linguistic information [4-10]. In the present paper, we first define prosodic units of spoken Japanese on the basis of F_0 contour characteristics, and then describe some of our experimental findings on how various information is manifested, or even fail to be manifested in certain situations, by the characteristics of the F_0 contour.

2. PROSODIC UNITS OF SPOKEN JAPANESE

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, and is usually composed of a content word followed by a sequence of function words. Under certain conditions, a string of prosodic words can form a larger prosodic word due to "accent sandhi." On the other hand, a phrase component 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, ICRLB, clause and sentence as syntactic units. "ICRLB" is an abbreviation for "immediate constituent with a recursively left-branching structure" which is a syntactic phrase delimited by right-branching boundaries and contains only left-branching boundaries. Roughly speaking, the parallelism shown in Fig. 1 exists between the hierarchy of syntactic units and the hierarchy of prosodic units [8,9].

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

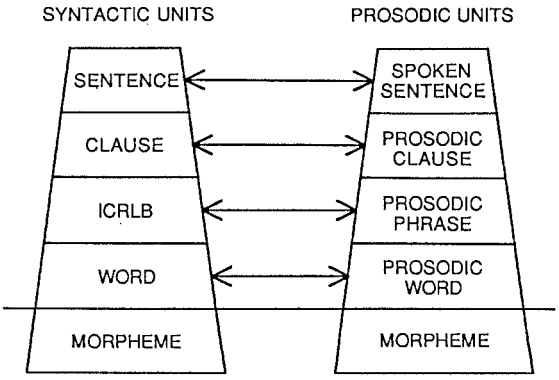

Fig. 1. Hierarchy of the syntactic units and that of the prosodic units, and their relationship.

command have been found to be necessary to account for the formation of an F_0 contour of an utterance of the common Japanese; 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. 2 [3].

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 three sets of utterances. Set 1 contains sentences of weather forecasts and news uttered by professional announcers. These utterances were used mainly for investigating the influences of syntactic information on the phrase components. Set 2 contains simple phrases uttered by two male speakers of the Tokyo dialect in the context of "___kaimasu. (I will buy___.)" The phrases consist of two or three prosodic words, and, for the latter case, the first and the second prosodic words (W1 and W2) are adjective phrases while the third prosodic word (W3) is a noun phrase [5]. Utterances for all the possible combinations for accent types, syntactic structures, and focal conditions were recorded and used mainly to investigate the influences of syntactic and discourse information on the accent components. Set 3 contains short sentences consisting of a noun phrase and a verb phrase like "mameo miru. (I see beans.)" For each phrases, two words of different accent types were selected. These sentences were uttered by a male speaker of the Tokyo dialect to express a variety of attitudes or intentions, such as neutrality, decisiveness, interrogation, exhortation, and disbelief. Utterances were also recorded with various sentence-final particles such as "ka," "ne," and "yo," that are commonly used to express some of the above-mentioned intentions. These utterances were used to investigate the influences of para-linguistic information on the F_0 contours.

The recorded materials were digitized at 10 kHz with 12 bit accuracy, and the fundamental frequency contours were extracted for further analysis of F_0 contours.

5. SYNTACTIC INFORMATION IN PHRASE COMPONENTS

A pause is always accompanied with a phrase command, while a small phrase command usually occurs without a pause. Three different ways are possible for starting a new phrase component: (1) preceded by a pause during which the immediately preceding phrase component is completely reset, (2) 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, and (3) simply added to the old one without pause. Let us denote prosodic boundaries marked by these ways by Type I, Type II, and Type III boundaries. As already mentioned, spoken sentences and prosodic clauses are marked by Type I or Type II boundaries, while prosodic phrases are marked by Type III boundaries.

The occurrence of these prosodic boundaries is primarily determined by the syntactic structure, and most often coincides with syntactic boundaries. It is, however, also subject to other factors such as style of speaking, respiration, etc. As already shown schematically in Fig. 1, Type I, II and III boundaries respectively occur mainly at syntactic boundaries between sentences, clauses, and ICRLB's, but this correspondence is not exactly one-to-one but is rather stochastic.

The probability that a prosodic boundary occurs at a given syntactic boundary is influenced by the depth of the syntactic boundary [6,7,10]. Figure 3 shows an example of the result of analysis along with the syntactic tree of the sentence. Each leaf of this syntactic tree is a prosodic word. The number on each leaf of the syntactic tree denotes the number of generations from the primal predicate of the sentence, i. e., the number of right-branches contained in the pass from the root to the leaf. Let us denote the "depth of a boundary" by j-i+1(=k), where i and j respectively denote the numbers on leaves at the left-hand side and at the right-hand side of the boundary. Using this notation, we can define the "left-branching boundary" as a boundary at which k=0 and the "right-branching boundary" as a boundary at which k>0. The results of analysis indicate that longer pauses and larger phrase commands tend to occur at boundaries with larger k.

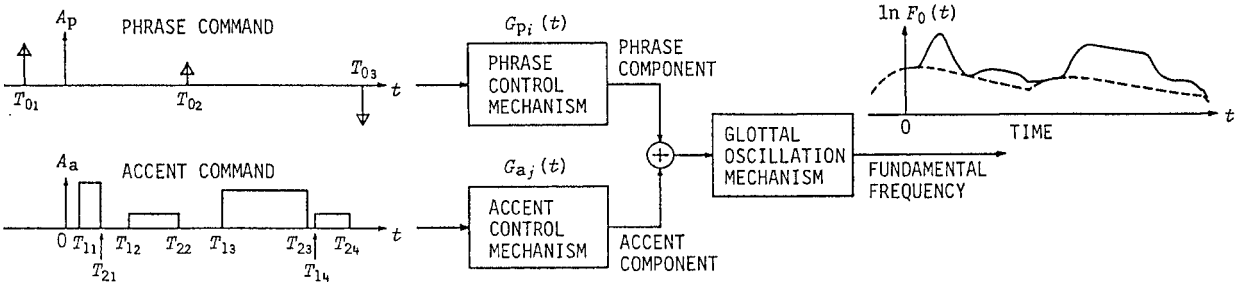

Fig. 2. Block diagram of a functional model for the process of generating sentence F_0 contours.

6. SYNTACTIC AND DISCOURSE INFORMATION IN ACCENT COMPONENTS

For the ease of representation, let us henceforth denote a prosodic word with a rapid downfall in the F_0 contour by "D" and that without any downfall by "F". When the prosodic words are uttered in isolation, there is a significant difference between the accent commands for D and F, being higher for D and lower for F. When more than two prosodic words are uttered in connected speech, however, they often interact with each other and their accent commands change both in amplitude and in shape. The manner of interaction is dependent on the accent types, the syntactic structure of the phrase and the focal condition.

Analysis of materials from Set 2 was conducted to clarify the rules underlying these changes [5,7,8]. The following characteristics were found for phrases without right-branching boundaries (ICRLB phrases).

1) In a phrase consisting only of D's, the accent command for the initial word is essentially of the same amplitude as that when the word is uttered in isolation, while the accent commands for all the subsequent words are suppressed. An example for three prosodic words is shown in Panel (a) of Fig. 4. When both the second and third prosodic words are of the F-type, they are concatenated to form one prosodic word with a low accent command.

2) In a phrase of DFD, each of the two prosodic words of the D-type has a high command, while the word of the F-type has low one.

3) In a phrase of FD or FFD, the F's have accent commands as high as that of the succeeding D. A new prosodic word is formed by the concatenation between F and D.

4) In a phrase consisting only of F's, they are usually concatenated and form a larger prosodic word.

5) When focus is placed on a prosodic word with a low accent command, its amplitude becomes higher (Panels (b), (c), (d) of Fig. 4). If a phrase contains F-type prosodic words, the manner of concatenation is affected by the focus. For phrases with right-branching boundaries, the characteristics of accent component are somewhat different. For instance, a low phrase command often occurs at the right branching boundary, and it usually prevents the mutual interactions between the accent components of the prosodic words on both sides of the boundary (Fig. 5). Figure 6 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 W2, accent sandhi is seen to

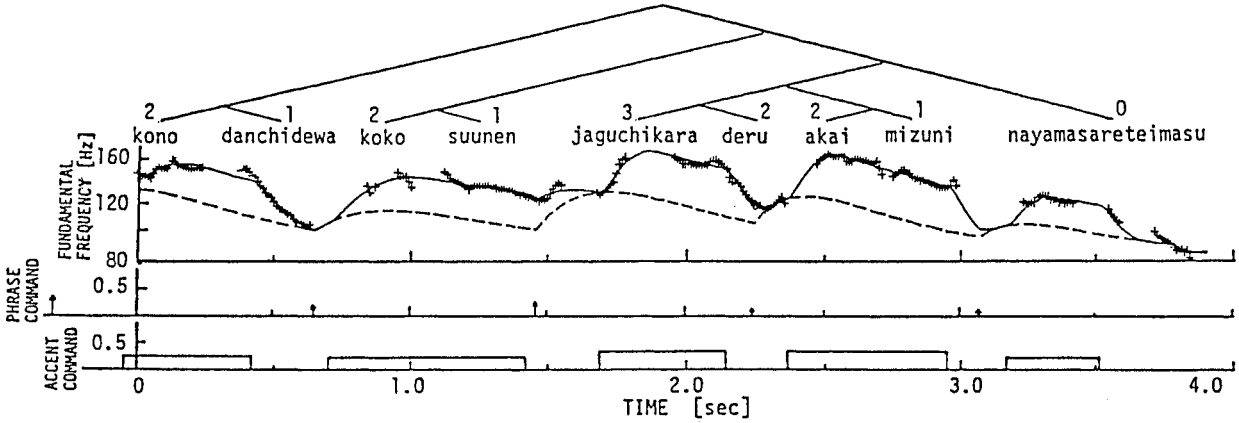

Fig. 3. The F_0 contour and the syntactic tree of a Japanese sentence. The sentence can be translated into English as "In this apartment house complex, for the past several years, residents have been annoyed by the stained water coming out of the taps."

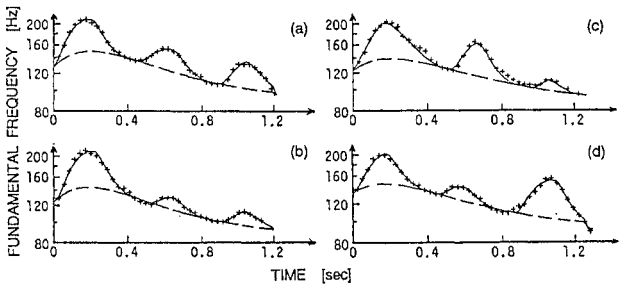

Fig. 4. Results of the F_0 contour analysis for a Japanese noun phrase "aomorino(W1) anino(W2) amaguo(W3)" uttered in four different manners: (a) without any obvious focus, (b) with focus on W1, (c) with focus on W2, (d) with focus on W3. The phrase may be translated as "a raincoat of my brother in Aomori."

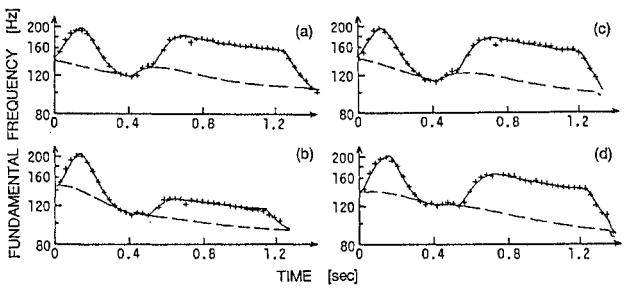

Fig. 5. Results of F_0 contour analysis for a Japanese noun phrase "anino(W1) mizuirono(W2) amaguo(W3)" uttered in the same manners as those in Fig. 4. The phrase may be translated as "my brother's light-blue raincoat."

occur between W2 and W3, indicating that the speaker opted to give priority to the discourse requirement. In Panel (b) for the right-branching structure and focus on W1, on the other hand, the accent component of W1 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.

7. REPRESENTATION OF PARA-LINGUISTIC INFORMATION

Compared with the neutral utterance in Panel (a), a decisive utterance in Panel (b) is characterized by a higher accent component for the second prosodic word. The remaining three utterances are all characterized by the so-called "sentence-final intonation," whose acoustic correlate is found to be a rather steep rise in the F_0 contour at the final mora. Although the actual mechanism for this sentence final rise may not necessarily be identical to that of the accent component, it is more appropriately interpreted as the consequence of a high accent command occurring at the final mora.

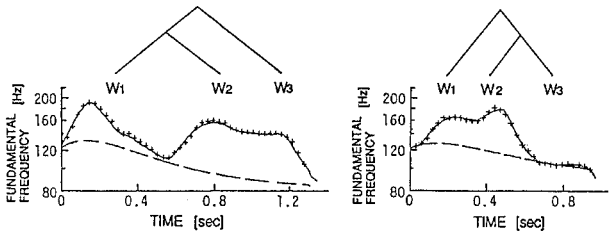

Fig. 6. Two examples in which syntax and discourse present conflicting requirements on F_0 contours. In the case of (a) "aomorino aneno amaguo" (DFD, focus on W2), priority is given to discourse requirement, while in the case of (b) "aneno aono amaguo" (FDD, focus on W1), priority is given to syntactic requirement.

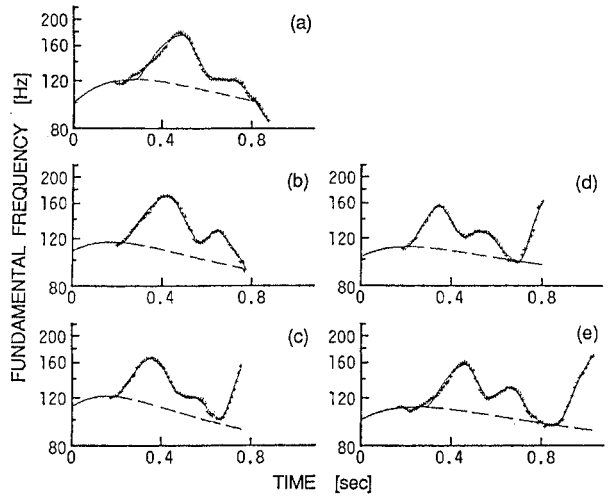

Fig. 7. Results of the F_0 contour analysis for a Japanese sentence "mameo niru" uttered with various intentions: (a) neutrality, (b) decisiveness, (c) interrogation, (d) exhortation, and (e) disbelief.

Interrogation in Panel (c) and exhortation in Panel (d) are quite similar in the timing and the height of the final "accent" component, while disbelief in Panel (e) is expressed by an elongated final mora with a delayed "accent" component. Perceptual tests were also carried our to investigate the accuracy at which these intonations are transmitted, and the results generally confirmed fairly accurate interpretation on the part of the listener, except that simple interrogation and exhortation are rather confusable.

8. CONCLUSIONS

After defining prosodic units of spoken Japanese on the basis of components of observed F_0 contours, the relationship between the prosodic boundaries and the syntactic boundaries have been described. The influences of various linguistic factors such as lexical word accent, syntactic structure and discourse structure upon the accent components of prosodic words have also been discussed on based on the analysis of F_0 contours of a number of utterances. Our analysis has indicated that there are cases where prosody fails to meet all the requirements presented by word accent, syntax and discourse. Investigations were further conducted on how the para-linguistic information is conveyed by the F_0 contours.

This work is partly supported by a Grant-in-Aid for Scientific Research (No. 02224106) from the Ministry of Education.

REFERENCES

[1] H. Fujisaki and H. Sudo, "Synthesis by Rule of Prosodic Features of Connected Japanese," Proc. 7th ICA, 23C2, 1971.

[2] H. Fujisaki and K. Hirose, "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.

[3] H. Fujisaki and K. Hirose, "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese," J. Acoust. Soc. Jpn. (E), 5, 4, pp. 233-242, 1984.

[4] K. Hirose, H. Fujisaki and M. Yamaguchi, "Synthesis by Rule of Voice Fundamental Frequency Contours of Complex Sentences," Proc. IEEE ICASSP 84, 2.13, 1984.

[5] H. Fujisaki, K. Hirose, N. Takahashi and M. Yoko'o, "Realization of Accent Components in Connected Speech," Trans. of the Committee on Speech Research, Acoust. Soc. Jpn., S84-36, 1984.

[6] K. Hirose, H. Fujisaki and H. Kawai "Generation of Prosodic Symbols for Rule-synthesis of Connected Speech of Japanese," Proc. IEEE ICASSP 86, 45.4, 1986.

[7] K. Hirose and H. Fujisaki, "Accent and Intonation in Speech Synthesis," Journal of IEICE, 70, 4, pp. 378-385, 1987.

[8] H. Fujisaki and H. Kawai, "Realization of Linguistic Information in the Voice Fundamental Frequency Contour of the Spoken Japanese," Proc. IEEE ICASSP 88, S14.3, pp. 663-666, 1988.

[9] K. Hirose, H. Kawai and H. Fujisaki, "Synthesis of Prosodic features of Japanese Sentences," Preprints Second Symposium on Advanced Man-Machine Interface through Spoken Language, Hawaii, pp. 3.1-3.13, 1988.

[10] K. Hirose, H. Fujisaki, H. Kawai and M. Yamaguchi, "Speech Synthesis of Sentences Based on a Model of Fundamental Frequency Contour Generation," Trans. IEICE, J72-A, 1, pp. 32-40, 1989.