Analysis and Perception of Intonation Expressing Paralinguistic Information in Spoken Japanese

Hiroya Fujisaki
Dept. of Applied Electronics, Science University of Tokyo
2641 Yamazaki, Noda, 278 Japan

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

ABSTRACT
In addition to the linguistic information, prosody conveys para- and non-linguistic information. The present paper deals specifically with the role of intonation in transmitting the speaker's attitude/intention in spoken Japanese. Short declarative sentences, uttered with various attitudes/intentions, were analyzed to find their acoustic correlates. Perceptual experiments were also conducted using the same utterances as stimuli to find out the accuracy/reliability of transmission.

1. INTRODUCTION

Recent studies by the present authors and others on the prosodic features of the spoken Japanese have contributed much to the elucidation of the role of intonation in conveying linguistic information concerning such factors as lexical word accent, syntactic structure and discourse focus (Fujisaki and Kawai 1988). In the present paper, we define the linguistic information as the information that is explicit in or almost uniquely inferable from the written message.

In addition to linguistic information, however, intonation also conveys para- and non-linguistic information. Here we define paralinguistic information as the information that is not inferable from the written message but is added by the speaker to modify or complement the linguistic information. For instance, a written message can be uttered with various intonational patterns to express different intentions, attitudes and speaking styles which can be controlled by the speaker. On the other hand, nonlinguistic information concerns such factors as the age, gender, idiosyncrasy, physical and emotional conditions of the speakers which are not directly related to the linguistic or paralinguistic contents of the utterance and generally cannot be controlled by the speaker. It is, however, possible for the speaker to control the prosody of utterance to convey his/her emotion.

In comparison with the studies on the linguistic aspects of intonation, studies on the para- and non-linguistic aspects of intonation have been rather limited. However, studies on these aspects are no less important both for the basic understanding of human communication and for the realization of a high-quality man-machine communication through the spoken language. The present paper describes our initial effort toward the elucidation of these aspects of intonation and deals only with the paralinguistic aspects.

2. EXPRESSION OF SPEAKER'S ATTITUDE/INTENTION IN SPOKEN JAPANESE

In Japanese, as in many other languages, a sentence can be uttered with at least several different intonational patterns to express differences in the attitude/intention of the speaker. Let us take, for example, a positive declarative sentence such as "gakkō-e iku," meaning "go to the school," without specifying the subject. The sentence can be uttered at least with the following five different intonational patterns.

(1) Default intonation, indicating that the speaker is merely reporting the fact that someone (most commonly the speaker himself) goes to the school, without further attitudinal/intentional commitment.
(2) Assertive intonation, indicating that the speaker is definitely committed to the fact (i.e., "I am determined to go to the school.").
(3) Interrogative intonation, indicating that the speaker is addressing a question to a

second person (i.e., "Do you go to the school (now)?").
(4) Exhortative intonation, indicating that the speaker is addressing an invitation to a second person (i.e., "Shall we (now) go to the school?").
(5) Hesitative intonation, which indicates an interrogation to which the speaker is reluctant to accept a positive response, or expecting a negative response (e.g., "Do you (still) go to the school (in spite of this bad snowstorm, etc.)?").

On the other hand, some of these attitudes/intentions can also be expressed linguistically by adding a particle (or particles) to the verb which comes at the end of the sentence. For instance, the original sentence can be made into an assertion by adding the particle "-yo" to the end of the verb. Thus "gakkō-e iku-yo" is an assertion, while "gakkō-e iku-ka" is an interrogation. Some (but not all) of these final particles (or particle strings) and their functions are given below.

The addition of the particle(s), however, does not replace the role of intonation. Thus the sentence with the interrogative particle "-ka" is uttered still with an interrogative intonation. Furthermore, various modifications of their default linguistic meaning can be introduced by intonation, producing a variety of finer 'nuances.' For instance, "gakkō-e ika-nai-ka," with a falling intonation indicates a strong suggestion rather than a negative interrogation, and thus can be considered to represent a directive attitude/intention.

Table 1. *Some of the final particles and their default functions.*

particle(s)	function
-ka (-kai)	interrogation
-ne	confirmation
-yo	assertion
-ka-ne	interrogation
-nai	negation
-nai-ka	negative interrogation
-nai-ka-ne	negative interrogation

The purpose of the present study is to find out the objective features that represent these differences on the other hand, and to find out to what extent these differences are perceived.

3. ANALYSIS OF INTONATION EXPRESSING PARALINGUISTIC INFORMATION

3.1 The speech material and the method of analysis

The sentences used for the current study has a very simple syntactic structure consisting of an object phrase and a verb phrase. The object phrase consists of a noun plus an accusative particle "o", while the verb phrase consists of a verb with or without being followed by a particle or a string of particles shown in Table 1, introducing various linguistic modifications of the original verb. Thus a total of eight sentence types (the original and its seven variants) are selected.

Since there exists an interaction between the intonation and the lexical accent of the constituent words, both 'accented' and 'unaccented' words were chosen for the noun and the verb; i.e., the accented "mame]" (bean(s)) and the unaccented "ame" (candy) for the noun, and the accented "mi]ru" (to look at) and the unaccented "niru" (to cook). Combination of these sentence types and word accent types produces a total of 32 sentences. The number is further doubled by adding the polite form of verb ending (e.g., 'mimasu' instead of 'miru'), resulting in a total of 64 sentences.

Each of these sentences was uttered with four or five different attitudes/intentions by four informants who were adult speakers of the common Japanese (i.e., the Tokyo dialect). At least three utterances were produced by a speaker for each sentence.

The speech material was digitized at 10 kHz with 12-bit precision. Fundamental frequencies were extracted by a modified autocorrelation method, and the F_0 contour was further analyzed using a model of the process of F_0 contour generation (Fujisaki and Hirose 1982). In addition to the phrase and accent components which constitute the F_0 contour of a declarative sentence with a default intonation, paralinguistic modification is often found to be expressed by another positive component which occurs toward the end of an utterance. Although this component may involve a mechanism other than that for the accent component, it was assumed in the present study that this component is generated by the accent control mechanism.

3.2 Results of F_0 contour analysis

Figure 1 shows the waveform, the F_0 contour and its closest approximation obtained by Analysis-by-Synthesis together with the extracted phrase component, and the underlying accent commands. The panels on the left-hand side are for the sentence "mame]-o mi]ru," and those on the right-hand side are for the sentence "mame]-o niru." The numbers on each panel indicate ① default, ② assertive, ③ interrogative, ④ exhortative, and ⑤ hesitative intonations, respectively. Comparison of these and other analysis results can be summarized as follows.

(1) Of all the five different intonation patterns analyzed, ③, ④ and ⑤ are commonly characterized by a large utterance-final rising component, but its timing and magnitude are different among the three cases and differ also depending on the accent type of the verb. Significant differences also exist in the local and global tempo, especially for ② and ⑤.

(2) Compared with the default intonation ①, the assertive intonation ② is accompanied by a slightly faster overall speech rate and a longer accent command for the verb.

(3) Compared with the default intonation ①, the interrogative intonation ③ has the similar overall speech rate except for a longer final mora, accompanied by a significantly larger final rise command. When the verb is accented at the initial mora, this rise command appears as a separate command which is much larger than the accent commands of the 'accented' morae of the verb as well as for the object (noun), starting approximately at the segmental onset of the vowel of the final mora. When the verb is unaccented, however, this rise command coincides with the accent command for the second mora of the 'unaccented' verb, and starts approximately 30 msec prior to the segmental onset of the vowel of the final mora. Thus the timing of the final rise is siginificantly different depending on the accent type of the verb.

(4) Compared with the interrogative intonation ③, the exhortative intonation ④ is quite similar except that the magnitude of the final rise command is lower.

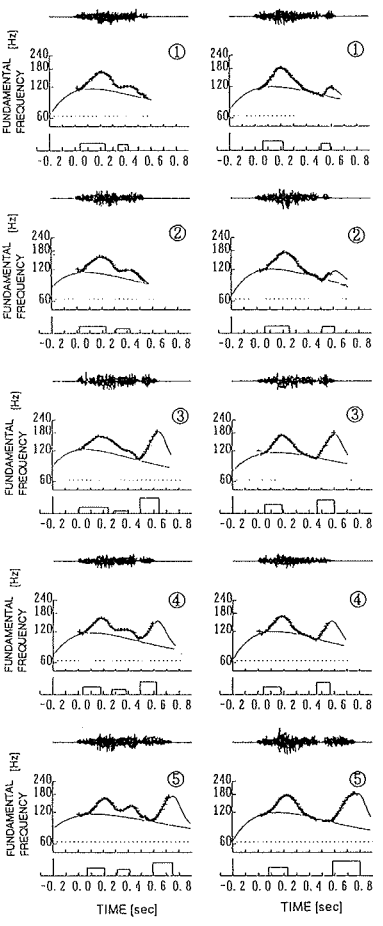

Figure 1. *Results of analysis of F_0 contours for the sentences "mame]-o mi]ru" (left) and "mame]-o niru" (right), with ① default, ② assertive, ③ interrogative, ④ exhortative, and ⑤ hesitative intonations.*

(5) Compared with the interrogative intonation ③, the hesitative intonation ⑤ is characterized by a marked change in the local tempo, i.e., the elongation of the final mora by a factor of two or more. This elongation is also accompanied by a delayed onset of the final rise command, starting at approximately 40 msec after the segmental onset of the vowel of the final mora. The magnitude of this rise command is similar to that for the interrogative intonation, but its duration is increased.

4. PERCEPTION OF SPEAKER'S ATTITUDE/INTENSION IN SPOKEN JAPANESE

The results of analysis mentioned in the foregoing section have indicated that differences in the attitude/intention are reflected in the acoustic characteristics, mainly in the F_0 contours and segmental durations. It remains to be investigated, however, whether or not the intended paralinguistic information is perceived as accurately as in the case of linguistic information such as the word accent type, etc. Perceptual experiments were thus conducted using both natural and synthetic speech sound stimuli.

4.1 Experiments using natural speech

One utterance each of the five variants of the sentence "mame^{l}-o mi-mase^{l}-n-ka," uttered by one male informant of the common Japanese, with five different attitudes/intentions was selected from the recorded speech samples. In this case, a default (neutral) intonation indicates a directive attitude and a falling intonation indicates a confirmative attitude, while interrogation, exhortation, and hesitation are expressed by more or less similar intonational patterns as in the case of "mame^{l}-o mi^{l}ru." They were arranged in random order with an inter-stimuli interval of 4 seconds and presented through headphones to the subjects, whose task was to identify the five attitudes/intentions.

Two normal-hearing subjects, who were both speakers of the common Japanese, took part in the experiment. Table 2 shows the averaged results of the two subjects in the form of a confusion matrix. The numbers indicate the percentages. It can be seen that the attitudes/intentions of the speaker are identified fairly accurately except for the confusion between interrogation and exhortation.

Table 2. *Results of a perceptual experiment on the accuracy of recognition of speaker's attitude/intention. The stimuli are natural utterances of "mame-o mima-sen-ka" uttered with five different attitudes/intentions. The numbers indicate the percentage.*

Stimulus	Response				
	① directive	② confirmative	③ interrogative	④ exhortative	⑤ hesitative
① directive	90			10	
② confirmative		97			3
③ interrogative	3		65	32	
④ exhortative	12		30	58	
⑤ hesitative					100

4.2 Experiments using synthetic speech

While perceptual experiments using natural speech confirmed that the paralinguistic information concerning the speaker's attitude/intention can be transmitted with a fair degree of accuracy in natural speech, further investigation was necessary to obtain a guideline for speech synthesis by rule. Thus another perceptual experiment was conducted using synthetic speech stimuli for the sentences "mame^{l}-o mi^{l}ru" and "ame-o niru."

Although the details cannot be given because of space limitations, the results of this perceptual experiment showed the range of parameter values for each intonational category and also indicated the influence of the word accent type of the verb on the timing of the commands for the terminal F_0 rise for interrogation, exhortation and hesitation.

REFERENCES

H. Fujisaki and H. Kawai (1988), "Realization of linguistic information in the voice fundamental frequency contour of the spoken Japanese," *Proc. 1988 Intl. Conf. on Acoust., Speech, and Signal Processing*, New York, 11–14 April 1988, vol. 1, pp.663–666.

H. Fujisaki and K. Hirose (1982), "Modeling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation," *Preprints of the Working Group on Intonation, the 13th Intl. Congress of Linguists*, Tokyo, 31 August 1982, pp.57–70.