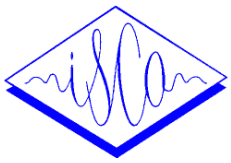

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

PROSODIC CHARACTERISTICS OF A SPOKEN DIALOGUE FOR INFORMATION QUERY

*Hiroya Fujisaki**, *Sumio Ohno**, *Masafumi Osame**, *Mayumi Sakata*** and *Keikichi Hirose***

* Department of Applied Electronics, Science University of Tokyo
2641 Yamazaki, Noda, 278 Japan

** Department of Electronic Engineering, University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

ABSTRACT

Prosodic characteristics of a simulated dialogue have been analyzed and compared with those of read speech. The simulated dialogue was produced by referring to a written text, while read speech was produced by reading individual sentences of the same text in isolation. Analysis of fundamental frequencies and speech rates indicated that dialogue-style utterances have larger mean value and standard deviation of fundamental frequencies as well as higher mean value of speech rates than reading-style utterances. Further analysis of the F_0 contour parameters extracted by using a quantitative model has revealed that the differences in the F_0 characteristics are caused by a higher baseline value of the F_0 contours as well as an expanded range of variation in the amplitude of accent commands in the dialogue-style utterances.

1. INTRODUCTION

The prosodic characteristics of a spoken dialogue exhibit marked differences from those of text reading in two aspects. Firstly, they provide focal information to assist rapid and proper comprehension of brief but correct utterances. Secondly, they provide information to facilitate reconstruction of correct messages from ill-formed utterances. Elucidation of the first aspect is especially important for generating a high-quality speech output, while clarification of the second aspect is indispensable for realizing reliable speech input in a spoken dialogue system.

The present paper is primarily concerned with the analysis of the first aspect. The prosodic characteristics represented by the fundamental frequency contours (henceforth the F_0 contours) and the speech rate have been analyzed using speech material consisting of a simulated dialogue and readings of individual sentences of the same dialogue by the same speakers. Characteristics of the F_0 contours were parameterized by the method of Analysis-by-Synthesis using a quantitative model for the generation process, and the results are interpreted in relation to their roles in the dialogue.

2. THE SPEECH MATERIAL AND THE METHOD OF ANALYSIS

The speech material for the present study consists of recordings of a simulated dialogue in an information query/answer situation concerning ski resorts. The simulated dialogue was produced by a speaker or pair of speakers by referring to a written text. The dialogue consists of 34 alternating utterances by a client and an agent, and most of the utterances contain only one sentence. Table 1 shows the overall structure of the dialogue. Odd-numbered utterances, mostly questions and requests, are by the client, while even-numbered utterances, mostly answers and explanations, are by the agent. In a few cases, the agent asks a question to the client for clarification. For the sake of comparison, individual sentences of the text were also read by the same speakers in a normal reading style, without paying attention to the context of the dialogue in which they occur. Utterances in the simulated dialogue and reading shall henceforth be referred to as the 'dialogue-style (or simply DS)' utterances and 'reading-style (or simply RS)' utterances, respectively.

The recorded material was digitized at 10 kHz with 12 bit precision for further analysis. The durations of utterances and pauses were measured from the speech waveform, while fundamental frequencies were extracted by a modified autocorrelation analysis of the LPC prediction residual. The F_0 contours were further analyzed by the method of Analysis-by-Synthesis using a quantitative model for the process of F_0 contour generation [1].

Table 1. Topical structure of the dialogue on ski resort information.

Topic	Sub-topic	Utterance No.
A. Facilities	1. number of lifts	1 , 2
	2. night-skiing facilities	3 , 4
B. Traveling time from Tokyo	1. general	5 , 6
	2. by car	7 ~ 10
	3. by train	11 , 12
	4. confirmation	13 , 14
C. Current condition of one of the resorts	1. snow	15 , 16
	2. waiting time for the lift	17 , 18
	3. accommodation	19 , 20
D. Further information on accommodation	1. availability	21 , 22
	2. rates, room types, etc.	23 ~ 26
	3. availability of specific room types	27 ~ 30
E. Request for reservation		31 , 32
F. Closing the dialogue		33 , 34

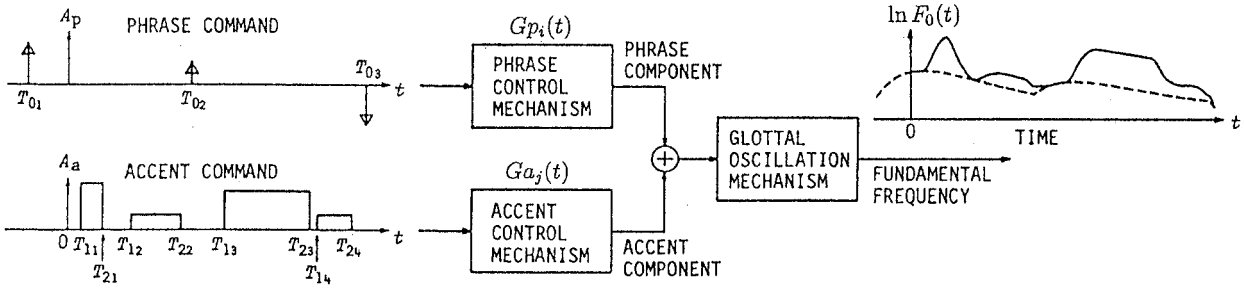

Fig. 1. A quantitative model for the process of F_0 contour generation.

As shown in Fig. 1, the model is based on the assumption that the F_0 contour of a declarative sentence in the logarithmic frequency scale can be considered as the sum of two kinds of components, i.e., the phrase components and the accent components, and that the phrase component is the response of a critically-damped second-order linear system to an impulse-like phrase command, while the accent component is the response of another critically-damped second-order linear system to a step-like accent command. Thus an F_0 contour of a spoken sentence is represented by the following equation:

$$\ln F_0(t) = \ln Fb + \sum_{i=1}^{I} Ap_i Gp_i(t - T_{0i}) + \sum_{j=1}^{J} Aa_j \{Ga_j(t - T_{1j}) - Ga_j(t - T_{2j})\}, \quad (1)$$

$$Gp_i(t) \begin{cases} = \alpha_i^2 t \exp(-\alpha_i t), & \text{for } t \geq 0, \\ = 0, & \text{for } t < 0, \end{cases} \quad (2)$$

$$Ga_j(t) \begin{cases} = \min[1 - (1 + \beta_j t)\exp(-\beta_j t),\ \gamma], & \text{for } t \geq 0, \\ = 0, & \text{for } t < 0, \end{cases} \quad (3)$$

where $Gp_i(t)$ represents the impulse response function of the phrase control mechanism and $Ga_j(t)$ represents the step response function of the accent control mechanism. The symbols in these equations indicate

- Fb : baseline value of an F_0 contour,
- I : number of phrase commands,
- J : number of accent commands,
- Ap_i : magnitude of the i-th phrase command,
- Aa_j : amplitude of the j-th accent command,
- T_{0i} : timing of the i-th phrase command,
- T_{1j} : onset of the j-th accent command,
- T_{2j} : end of the j-th accent command,
- α_i : natural angular frequency of the phrase control mechanism to the i-th phrase command,
- β_j : natural angular frequency of the accent control mechanism to the j-th accent command,
- γ : a parameter to indicate the ceiling level of the accent component (generally set equal to 0.9).

α_i and β_j are parameters characterizing glottal control mechanisms and are allowed to vary only within limited ranges of values. The parameter Fb represents the approximate lower limit of the voice register for a particular utterance and thus do not vary appreciably across utterances of the same speaker, but may be varied to indicate changes in the speaking style or to express certain kinds of non-linguistic information.

In addition to the phrase and accent components which constitute the F_0 contour of a declarative sentence uttered as a statement, an interrogative intonation is commonly characterized by the existence of another positive component occurring toward the end of an utterance, with or without a particle '/ka/'. Other particles such as '/ne/' can be used also with this type of intonation to indicate the intention of confirmation, etc. [2]. Since the rise time of this component has a similar value to that of an accent component, we assume in this paper that this component is also generated by an accent command applied to the accent control mechanism, though it may actually involve a third mechanism other than those for phrase and accent.

The speech rate was obtained for each sentence on the basis of the number of morae contained and its duration measured by visual inspection of the speech waveform.

3. RESULTS OF ANALYSIS

3.1 Illustrative Examples

Figure 2 shows a typical result of analysis each of the sentence /djaH naitaHwa arimasu ka/ ("Then, is there night-skiing (facility)?") in the RS utterance (a) and in the DS utterance (b) produced by the same speaker. Each panel shows, from top to bottom, the speech waveform, the observed F_0 contour as a sequence of + symbols, the best approximation generated by the model as a curve in a solid line, the estimated phrase components as a curve in a dashed line, the estimated baseline value Fb in a dotted horizontal line, and the estimated accent commands. The difference between the solid line and the dashed line corresponds to the estimated accent components.

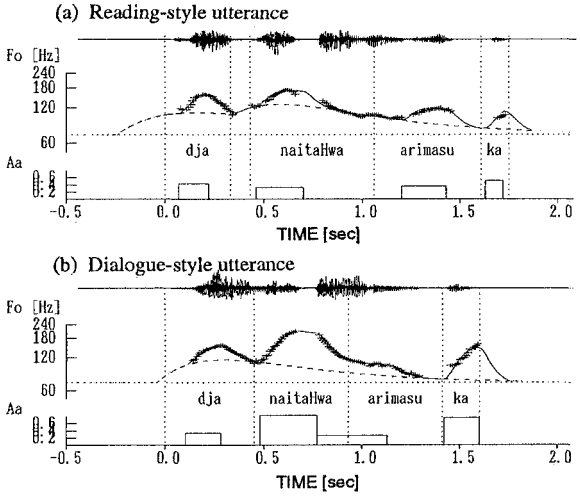

Fig. 2. F_0 contour analysis of the utterances of the sentence /djaH naitaHwa arimasu ka/. (a) reading-style, (b) dialogue-style.

Comparison of the two panels indicate characteristic differences between the two utterance styles. Compared with the RS utterance, the DS utterance is characterized by a wider range of F_0 variation and a shorter total duration. A closer examination of the result of F_0 contour analysis reveals that the wider F_0 range in the latter is partly due to a smaller number of phrase components (i.e., one for the DS utterance against two for the RS utterance), a larger initial phrase component, and a larger accent component assigned to the prosodic word /naitaHwa/ which is the focus of the utterance.

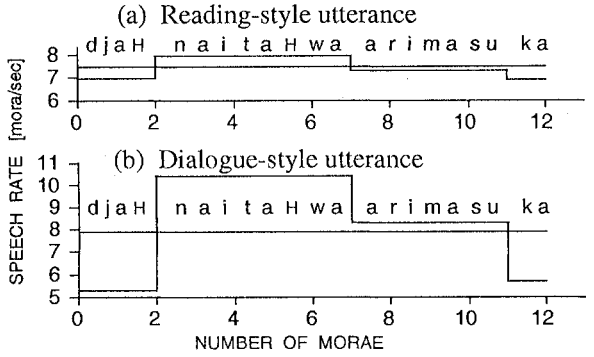

Fig. 3. Local speech rate versus position and length of prosodic words in (a) reading-style and (b) dialogue-style utterances of /djaH naitaHwa arimasu ka/.

Figure 3 shows the analysis of local speech rates expressed in terms of number of morae per second for each of the prosodic word that constitutes the utterances shown in Fig. 2. The abscissa indicates the position and length of each prosodic word in unit of mora, and the local speech rates are plotted as deviations from the mean speech rate of the whole utterance. The utterance-final particle /ka/ is shown here as a separate unit. There is a general tendency that the speech rate starts at a low value, increases at utterance-medial positions, and drops at the end. In comparison with the RS utterance shown in (a), this tendency is much more marked in the DS utterance in (b).

Figure 4 shows another example of analysis of the DS utterance /eH naebato kusatsuno dotʃiraga rihutoga oHi desu ka/ ("Eh... which of (the two resorts) Naeba and Kusatsu has a larger number of ski lifts?"). This is an opening utterance of the dialogue by the client and starts with a filler sound, often used to begin a question or a request. The speaker is the same as for the utterances shown in Fig. 2. Compared with the F_0 contours of Fig. 2, the F_0 contour in Fig. 4 is characterized by a much smaller rate of overall declination and a higher baseline value Fb. This is apparently the result of production of the initial filler sound at an approximately constant F_0 value, by raising Fb and using very low amplitude for the utterance-initial phrase component, which is seen to have an effect on the shape of F_0 contour of the entire utterance.

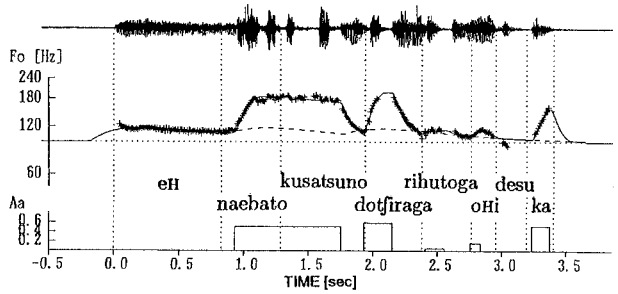

Fig. 4. F_0 contour analysis of a dialogue-style utterance of /eH naebato kusatsuno dotʃiraga rihutoga oHi desu ka/.

In order to separate the effects of characteristic differences between the RS and DS utterances from those of individual differences, the following sections will show results obtained from utterances by one typical speaker playing the role of client and agent.

3.2 Statistical Properties of Prosodic Parameters

In order to quantify some of the characteristic differences between the two styles, we will show the statistical distributions of the relevant parameters obtained from the analysis of all the utterances by one of the typical speakers. Since the filler sound /eH.../ and certain adverbs such as /hai/ (yes) and /dja/ (then), often used in isolation or at the beginning of an utterance, have characteristics different from the rest of the utterances, they are excluded in the calculation of the overall distributions. The commands for the utterance-final intonation are also excluded but their characteristics will be discussed in a separate section.

Table 2 shows the mean value (μ) and the standard deviation (σ) of the following prosodic parameters for both RS and DS utterances:

(1) Fundamental frequency (F_0),
(2) Speech rate in mora/sec of each sentence (Sr),
(3) Number of phrase commands per sentence (Np),
(4) Number of accent commands per sentence (Na),
(5) Magnitude of phrase command (Ap),
(6) Amplitude of accent command (Aa),
(7) Baseline value of F_0 contour of each sentence (Fb).

The first column of the table indicates that DS utterances possess higher mean value and wider range of variation in fundamental frequency than RS utterances. Column (2) shows that they have higher mean speech rate and wider range of its variation across sentences. Columns (3) and (4) indicate that they have smaller number of phrase and accent commands per sentence. This is equivalent to saying that the speaker tends to put more words into a prosodic phrase or clause, and also to produce longer prosodic words by *accent sandhi*. These features may be a consequence of increased speech rates. Columns (5) and (6) respectively show the results for the phrase command magnitude and the accent command amplitude. While the phrase command magnitudes of the DS utterances have smaller mean values than those of the RS utterances, their accent command amplitudes have larger mean and standard deviation, indicating an expanded range of variation of the degree of accentuation of prosodic words. Finally, column (7) indicates that there is a significant difference in the baseline value of F_0 contours, being higher for the DS utterances by more than 10%.

Table 2. Comparison of mean (μ) and standard deviation (σ) of prosodic parameters for the reading-style utterances and dialogue-style utterances. The unit is Hz for F_0 and Fb and mora/sec for Sr.

		F_0	Sr	Np	Na	Ap	Aa	Fb
Reading-	μ	131.3	7.72	3.00	4.40	.379	.333	74.3
style	σ	22.9	.44	1.61	2.12	.162	.106	3.1
Dialogue-	μ	140.4	8.94	2.35	3.92	.331	.362	83.0
style	σ	33.7	1.05	1.19	1.98	.145	.180	6.3

3.3 Prominence and Accent Command Amplitude

The wide range of variation of the accent command amplitude stems from various factors such as the difference in the lexical word accent types, the degree of accentuation placed on a prosodic word which is in turn influenced by linguistic (syntactic/semantic/pragmatic), paralinguistic (intentional), as well as non-linguistic (e.g. emotional) factors [3]. Here we will discuss only the effect of contrastive accentuation on the distribution of accent command amplitudes.

In Japanese, prosodic words (henceforth to be referred to as words for the sake of brevity) in a spoken sentence are pronounced with various degrees of accentuation. In general, listeners can easily differentiate 'prominent' words from others, since there exist considerable differences in the accent command amplitudes between 'prominent' words and 'non-prominent' words. It should be noted that the level of accentuation varies from sentence to sentence so that the perception of 'prominence' is based on a relative judgment within a sentence.

In general, prominence is placed on words that are semantically important in a sentence. In a discourse, it is placed also on words that are pragmatically important, i.e., words that convey key information on the course of discourse. However, *not all* the important words receive prominence in spoken Japanese. For instance, if two successive words within a prosodic phrase are both important, it is generally the first one that becomes prominent. In terms of accent command amplitude, the amplitude for the first word is elevated and that for the second word is usually suppressed by contrast. The word that is given 'prominence' becomes a focus of the utterance. Certain words such as interrogatives almost invariably become prominent in questions. When a sentence consists of two or more prosodic phrases, each containing important words, prominence may be placed on one of the words in each prosodic phrase, but the accent command amplitude for the second prominent word is usually lower than the one for the first prominent word, and words following it receive still lower accent command amplitudes.

The results of the foregoing analysis show that accent command amplitudes of prominent words are generally higher in DS utterances than in RS utterances. Accent command amplitudes of non-prominent words have similar values in both styles except at utterance-final positions, where they are often extremely suppressed and become almost zero in DS utterances. These words shall be referred to as 'de-accented' in this paper. Figure 5 (a) shows the distribution of accent command amplitude of prominent and non-prominent words in RS utterances, while Fig. 5 (b) shows those of prominent, non-prominent and de-accented words in DS utterances. While prominent and non-prominent words have overlapping distributions, they can be clearly distinguished within an utterance because of immediate contrast.

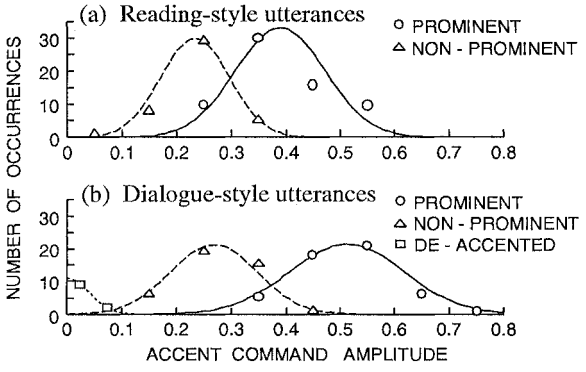

Fig. 5. Distributions of accent command amplitudes for prominent, non-prominent and de-accented prosodic words in (a) reading-style and (b) dialogue-style.

3.4 Commands for Utterance-Final Intonation

Figure 6 shows the distributions of command amplitudes for the utterance-final intonation for both utterance styles. The mean value is significantly larger almost by a factor of two in DS utterances, indicating that the speaker tries to be more explicit in showing his intention in the simulated dialogues.

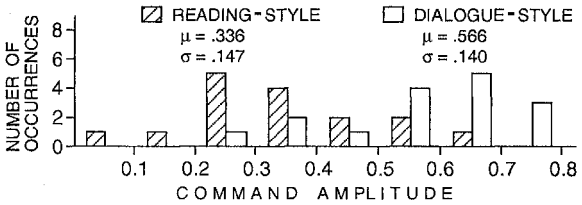

Fig. 6. Distributions of utterance-final command amplitudes.

4. SUMMARY AND CONCLUSION

Prosodic characteristics of a simulated dialogue have been analyzed and compared with those of read speech. The simulated dialogue was produced by referring to a written text, while read speech was produced by reading individual sentences of the same text in isolation. Analysis of fundamental frequencies and speech rates indicated that dialogue-style utterances have larger mean value and standard deviation of fundamental frequencies as well as higher mean value of speech rates than reading-style utterances. Further analysis of the F_0 contour parameters extracted by using a quantitative model has revealed that the differences in the F_0 characteristics are caused by a higher baseline value of the F_0 contours as well as an expanded range of variation in the amplitude of accent commands in the dialogue-style utterances. The latter phenomenon has been interpreted in terms of differences in the number of levels of accentuation of prosodic words. Analysis of the temporal characteristics of the same speech material and their relationships to the F_0 contour characteristics is being undertaken, but will be discussed elsewhere.

REFERENCES

[1] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan (E)*, Vol. 5, No. 4, pp. 233–242 (1984).

[2] H. Fujisaki and K. Hirose, "Analysis and perception of intonation expressing paralinguistic information in spoken Japanese," *Proceedings of the ESCA Workshop on Prosody, Working Papers 41, Lund University, Department of Linguistics*, pp. 254–257 (1993).

[3] H. Fujisaki, "From information to intonation," *Proceedings of the 1993 International Symposium on Spoken Dialogue, Tokyo*, pp. 7–18 (1993).

[4] K. Hirose, M. Sakata, M. Osame and H. Fujisaki, "Analysis and synthesis of fundamental frequency contours for the spoken dialogue in Japanese," Paper published in *Proceedings of the 2nd ESCA/IEEE Workshop on Speech Synthesis, New Paltz, New York* (1994).