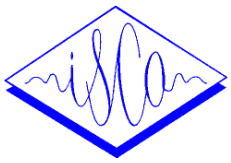

Second ESCA/IEEE Workshop
on Speech Synthesis
Mohonk, New Paltz, NY, USA
September 12-15, 1994

Effect of Speaking Style on Parameters of Fundamental Frequency Contour

Norio Higuchi, Toshio Hirai and Yoshinori Sagisaka
(ATR Interpreting Telecommunications Research Laboratories,
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02, Japan)
higuchi@itl.atr.co.jp, thirai@itl.atr.co.jp, sagisaka@itl.atr.co.jp

Abstract

The authors have analyzed the fundamental frequency (F_0) contours of Japanese sentences spoken in four styles, *e.g.* unmarked, hurried, angry and gentle, for the synthesis of natural sounding speech. Thirty-five sentences in each speaking style spoken by a professional narrator were analyzed. The parameters of the F_0 generation model proposed by Fujisaki, *i.e.* the minimum value of F_0 (F_{min}), the amplitude of the phrase commands (A_p) and the amplitude of the accent commands (A_a), are used here as key factors in the analysis. In the case of the sentences spoken angrily, F_{min} is kept high, and the change due to both the phrase component and the accent component is minimal. Consequently, the F_0 contours of sentences spoken angrily are flat. On the other hand, in the sentences spoken softly, the dynamic range due to the accent component is greater than for the others, and in order to keep it high the amplitude of the phrase component is accordingly suppressed. The F_0 contours of the sentences spoken hurriedly are similar except that the amplitude of the accent commands is slightly smaller than for those spoken normally. It was found that these parameters are useful to express the difference due to the speaking styles.

1. Introduction

The quality of synthetic speech has been improved by recent research [1, 2], but most modifications were designed to simulate natural speech such as uttered by a professional narrator/announcer when he/she reads just one-type of manuscript. As a result, the synthetic speech is still monotonous and the control of the speaking style simulating daily conversations has not been achieved yet. The control of speaking style is one of the most interesting and attractive targets of speech synthesis for the next generation [3] and it will make the synthetic speech more suitable for practical use in *e.g.* a voice Q-A system.

Though the speaking style affects both the spectral and the prosodic features of the utterances, the differences in the prosodic features are more remarkable than those in spectral features. In this paper to characterize the prosody characteristics the fundamental frequency contours are analyzed quantitatively based on the F_0 generation model proposed by Fujisaki [4].

2. Speech material

Sentence utterances spoken by a male professional narrator in simulated conversation were recorded and analyzed. The speaking styles analyzed here are the following four speaking styles: (1) unmarked style (instruction-free style), (2) hurried style, (3) angry style and (4) gentle style. The simulated conversations between the narrator and an interlocutor were performed twice in each speaking style, exchanging roles. The scenarios included (1) a conversation between a tourist and a customs official at an international airport, (2) between a passenger and an employee at a railroad station and (3) between a customer and an official at a post office. Though, each conversation consists of eighteen, twelve and nineteen sentences, respectively, six, three and four sentences from each conversation were excluded because of insufficient length (sentences like "*hai*." meaning "yes"). Consequently, thirty-five sentences in

each speaking style were analyzed, while fourteen were excluded. The total number of sentences in the four different speaking styles analyzed here is one-hundred and forty.

3. Analysis method

The quantitative analysis was performed using parameters of the F_0 generation model proposed by Fujisaki. The observed F_0 contour can be decomposed into three components : (1) the minimum value of F_0, (2) the phrase component and (3) the accent component. The minimum value of F_0 mainly depends on the size of the vocal fold, and up to now it has been considered to be a constant value for each speaker. However, the following analysis suggests that it should be adjusted to speaking style. The phrase component can be deconvoluted to the phrase command. The phrase command indicates the initial position of each phrase, and it is closely correlated to the sentence phrase structure. The accent component is the response of the accent command through a smoothing function which can be approximated by a step-response of a critically-dumped system. It carries the information on the high/low tonal distinction, which is determined lexically, and also on the presence or absence of a prominence for each word.

Parameters used here for the comparison of each speaking style are the following: (1) the minimum value of F_0 (F_{min}), (2) the amplitude of the phrase commands (A_p) and (3) the amplitude of the accent commands (A_a). Each parameter of the model is determined by a semi-automatic Analysis-by-Synthesis method which consists of two stages. In the first stage, the timing parameters are decided based on labeled data, and the amplitude of the phrase commands chosen so that the phrase component does not exceed the observed value of F_0. The difference between the phrase component and the observed F_0 value is calculated. Then, the amplitudes of the accent commands are determined by solving a linear multiple equation so as to minimize the estimation error between the contour generated by the model and the above-mentioned difference. In the second stage, each parameter value is optimized using a hill-climbing method. The approximation of F_0 contours using the Fujisaki model was good in most of the sentences, as shown in Fig. 1.

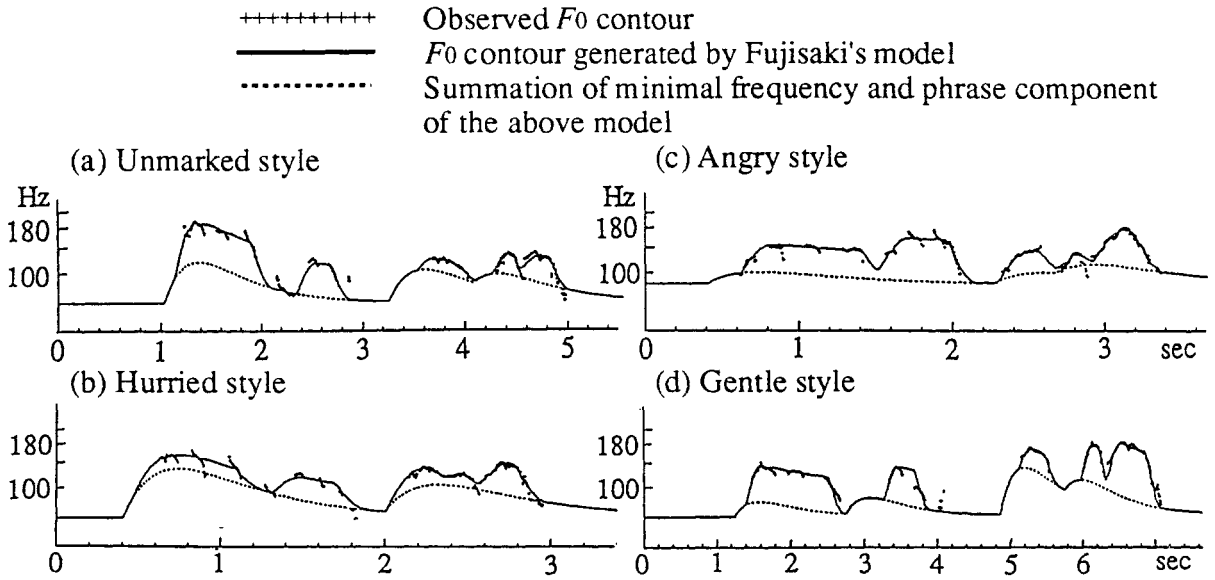

Fig. 1 Example of approximation by Fujisaki's Fo model. Each panel shows the Fo contour and its best approximation using Fujisaki's Fo model. The utterance is /kono suHtsukeHsuto koHkuHkaban soreni ano daNboHrubakodesu/, meaning "(All I have are) this suitcase, some hand baggage and that cardboard box."

4. Results

Figures 2, 3 and 4 show the parameter values of F_{min}, A_p and A_a, respectively. The vertical axis indicates the F_{min}, A_p or A_a values of the utterances spoken in three different speaking styles, while the horizontal axis indicates those of the same part of the same sentences in unmarked style. The points just on the horizontal/vertical axes in Fig. 3 mean that no phrase commands have been found at the corresponding location of the equivalent sentence, while those in Fig. 4 indicate that the accent components were suppressed in the same word of the sentence. The position of the phrase command and the deletion of the accent component depends on the speech rate of the utterance.

The F_{min} value is concentrated in a very narrow range for each speaking style except for two utterances in unmarked style. The A_p value covers a wider range than other parameter values, while that of A_a occupies a relatively narrow range along the line which shows equivalent values in both speaking styles. This shows that the constraints for the accent command are stronger than those for the phrase command, because the accent command is governed by lexical content more than the phrase command. Though the sentences analyzed here are just the same for four different speaking styles, it was found to be difficult to keep the same speaking quality over all the sentence.

The ranking of the average value of F_{min} (from highest to lowest) is angry, hurried, gentle and unmarked. The average value of F_{min} in the sentences spoken angrily is much higher than in the others. The ranking of the amplitude of the phrase command is hurried, unmarked, gentle and angry, while the ranking of the amplitude of the accent component is gentle, unmarked, hurried and angry.

In the case of the sentences spoken angrily, F_{min} is kept high, and the change due to both the phrase component and the accent component is minimal. Consequently, the F_0 contours of sentences spoken angrily are flat. On the other hand, in the sentences spoken softly, the dynamic range due to the accent component is greater than for the others, and in order to keep it high, the amplitude of the phrase component is accordingly suppressed. The F_0 contours of the sentences spoken hurriedly are similar except that the amplitude of the accent commands is slightly smaller than those spoken normally, perhaps because of the difference in speaking rate.

The differences in these parameters among four speaking styles were found to be constant and the feasibility of the synthesis of the utterances spoken in these different speaking styles was also shown by this result.

5. Conclusion

The authors analyzed the F_0 contour of the utterances spoken in four different speaking styles, *e.g.*, unmarked, hurried, angry and gentle styles using the F_0 generation model proposed by Fujisaki, and found differences in F_{min}, A_p and A_a values between four speaking styles. Further research on the analysis of utterances spoken by different speakers, and a subjective evaluation of the synthesized utterances using the typical values of each parameter have already been scheduled. The analysis of the amplitude and the duration of the phonemes is also necessary to create more natural-sounding synthetic speech.

For control of speaking style, we have to consider a more elaborate model, such as a hierarchical model consisting of a higher level related to the politeness, the speech act, the emotion, the social effect and other factors affecting the speaking style, and a lower level related to the strategy which is used for the realization of the speaking style with consideration in physiological constraints of speech organs. This research is only the first step to reveal the variation due to the speaking style. The selection of the strategy depends on individuals. Further research will bring us a good prospect for the realization of a more elaborate model for the control of the speaking style.

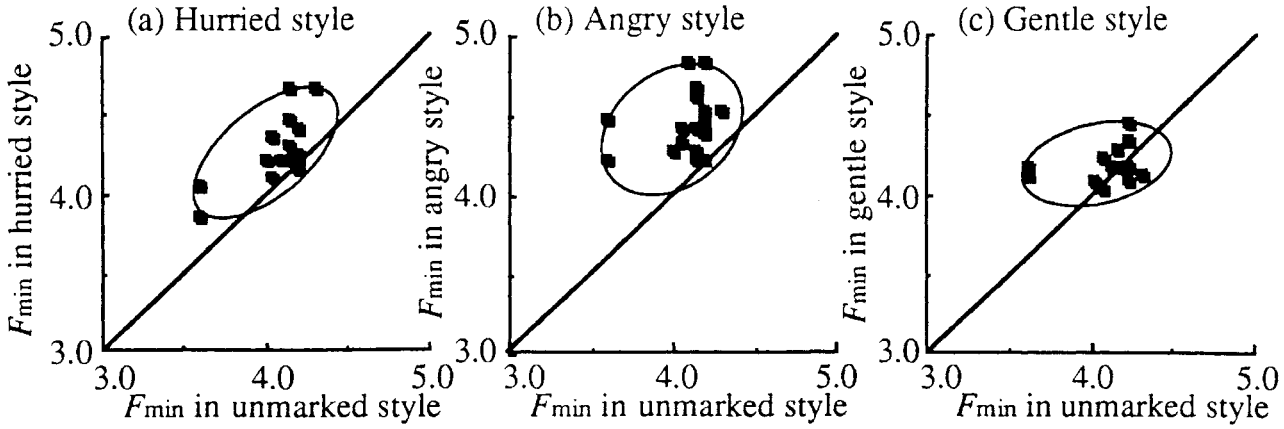

Fig. 2 Extracted values of logarithm of minimal frequency.

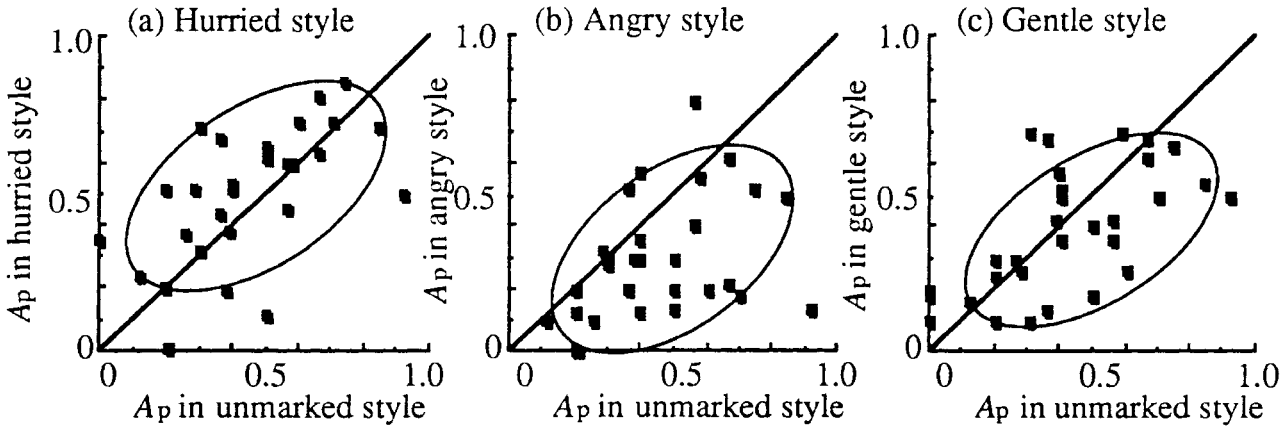

Fig. 3 Extracted values of amplitude of phrase command.

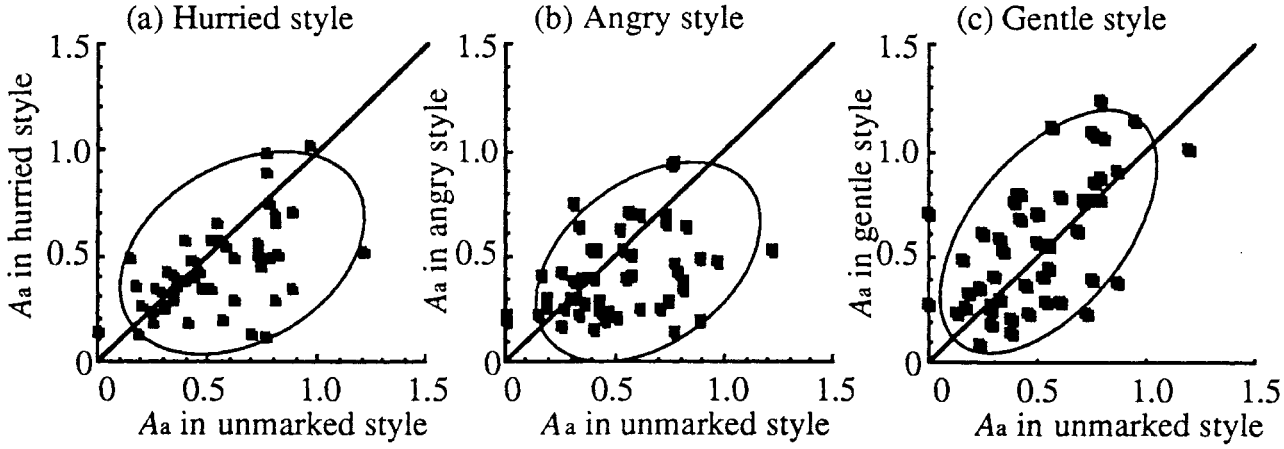

Fig. 4 Extracted values of amplitude of accent command.

References

[1] Y. Sagisaka, N. Kaiki N. Iwahashi and K. Mimura : "ATR ν-Talk speech synthesis system," Proc. ICSLP 92, **2**, Th.fAM.2.2 (1992).

[2] H. Kawai, N. Higuchi, T. Shimizu and S. Yamamoto : "Development of a text-to-speech system for Japanese based on waveform splicing," Proc. ICASSP94, **I**, 569-572 (1994).

[3] D. Lambert, K. Cummings, J. Rutledge and M. Clements : "Synthesizing multistyle speech using the Klatt synthesizer," J. Acoust. Soc. Am., **95**, No. 5, Pt. 2, 2979 (1994).

[4] H. Fujisaki and K. Hirose : "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," J. Acoust. Soc. Jpn (E), **5**, 233-242 (1984).