ON THE EFFECTS OF SPEECH RATE UPON PARAMETERS OF THE COMMAND-RESPONSE MODEL FOR THE FUNDAMENTAL FREQUENCY CONTOURS OF SPEECH

Sumio Ohno *Hiroya Fujisaki* *Yoshikazu Hara*

Science University of Tokyo
2641 Yamazaki, Noda, 278-8510, Japan

ABSTRACT

A command-response model for the process of F_0 contour generation has been presented by Fujisaki and his coworkers. The present paper describes the results of a study on the variabilty and speech rate dependency of the model's parameters in utterances of a speaker of Japanese. It was found that parameters α and β can be considered to be practically constant at a given speech rate, while Fb may vary slightly from utterance to utterance. Among these three parameters, only α was found to have a small but systematic tendency to increase with the speech rate.

1. Introduction

Needless to say, the contour of the voice fundamental frequency (henceforth F_0 contour) is the consequence of control of vocal fold vibration by neuromotor commands carrying information concerning lexical accent, syntactic structure, etc. The whole process of F_0 contour generation has been quantitatively modeled by a set of commands that provide the input information and by the mechanism that produces the F_0 contour as the response [1]. The model's validity was shown initially for F_0 contours of spoken Japanese [2,3], but has since been demonstrated, with certain language-specific modifications, to apply to F_0 contours of a number of languages [4].

Our previous studies have also shown that the baseline frequency (Fb) remains almost constant across utterances of one speaker, while the time constants (α and β) of the phrase control mechanism and the accent control mechanism are almost the same across different speakers. On the other hand, our pilot study suggested that these parameters may vary with the speech rate. The present paper describes our recent study on the speech rate dependency of these parameters in the utterances of one native speaker of Japanese.

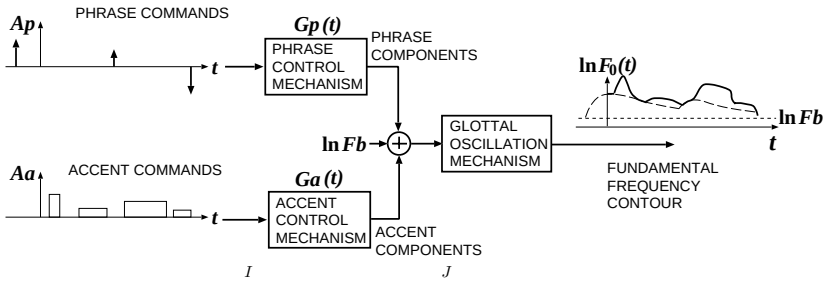
2. SPEECH MATERIAL AND METHOD OF ANALYSIS

The speech material was recordings of a short story consisting of 14 sentences read by a male speaker of the common Japanese at three speech rates: 'slow', 'normal', and 'fast'. The average speech rates were 6.6, 7.6, and 8.8 *morae* per second, respectively. The speech material was digitized at 10 kHz with 16 bit precision, and the fundamental frequency was extracted at 10 ms intervals by a modified autocorrelation analysis.

By the method of Analysis-by-Synthesis based on the quantitative model for the process of F_0 contour generation shown in Fig. 1, the parameters of the input commands, *viz.*, the magnitudes and timings of the phrase and accent commands, the time constants (α and β) of the phrase and accent control mechanisms, and the baseline frequency (Fb) can be determined for a given utterance or a set of utterances.

3. ESTIMATION OF α, β AND Fb FOR INDIVIDUAL UTTERANCES

Assuming that parameters α, β, and Fb can vary from utterance to utterance even within a speaker and at a given speech rate, analysis was made individually on each sentence utterance. Since estimation of these parameters



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

$$Gp(t) \begin{cases} = \alpha^2 t \exp(-\alpha t), & t \geq 0 \\ = 0, & t < 0 \end{cases}, \quad Ga(t) \begin{cases} = \min[1 - (1+\beta t)\exp(-\beta t), \gamma], & t \geq 0 \\ = 0, & t < 0 \end{cases}$$

Figure 1: A command-response model for F_0 contour generation.

are affected by the presence of microprosodic variations due mainly to voiceless consonants, these variations have been removed by pre-processing based both on threshold of normalized autocorrelation function of the LPC residual and on median smoothing of measured F_0 values.

Figure 2 illustrates one example each of the results of analysis of a sentence uttered at three speech rates. The sentence is *"Yuugureno hayashio nukeruto, atariwa ameni kemutte nanimo mienai."* (When I came out from the woods in the twilight of evening, there was almost no sight of the scenery because of the misty rain.) The three panels correspond to the three speech rate: (a) slow, (b) normal, and (c) fast. Each panel displays, from top to bottom, the speech waveform, approximate positions of word boundaries (vertical dotted lines), measured F_0 values (+ symbols), the best approximation by the model (solid line), the baseline frequency (horizontal dotted line), the phrase commands (impulses), and the accent commands (stepwise functions). The broken lines indicate the phrase components wherever they differ from the model's approximation, and the differences between the solid line and the broken lines indicate the accent components. The patterns of phrase commands are essentially similar except for timing, but the patterns of accent commands are seen to vary to some extent at different speech rates, *viz.*, concatenation of accent commands are seen to occur more often at higher speech rates. The means and standard deviations of α, β, and Fb of all the 14 sentence utterances are shown in Table 1 for each of the three speech rates, and are also plotted in Fig. 3. Parameter α tends to increase with the speech rate. The tendency is also found, though to a lesser degree, in parameter β. On the other hand, the Fb does not seem to vary systematically with the speech rate. Table 2 shows the results of a t-test on significance of differences in the mean values of these parameters for the three speech rate conditions.

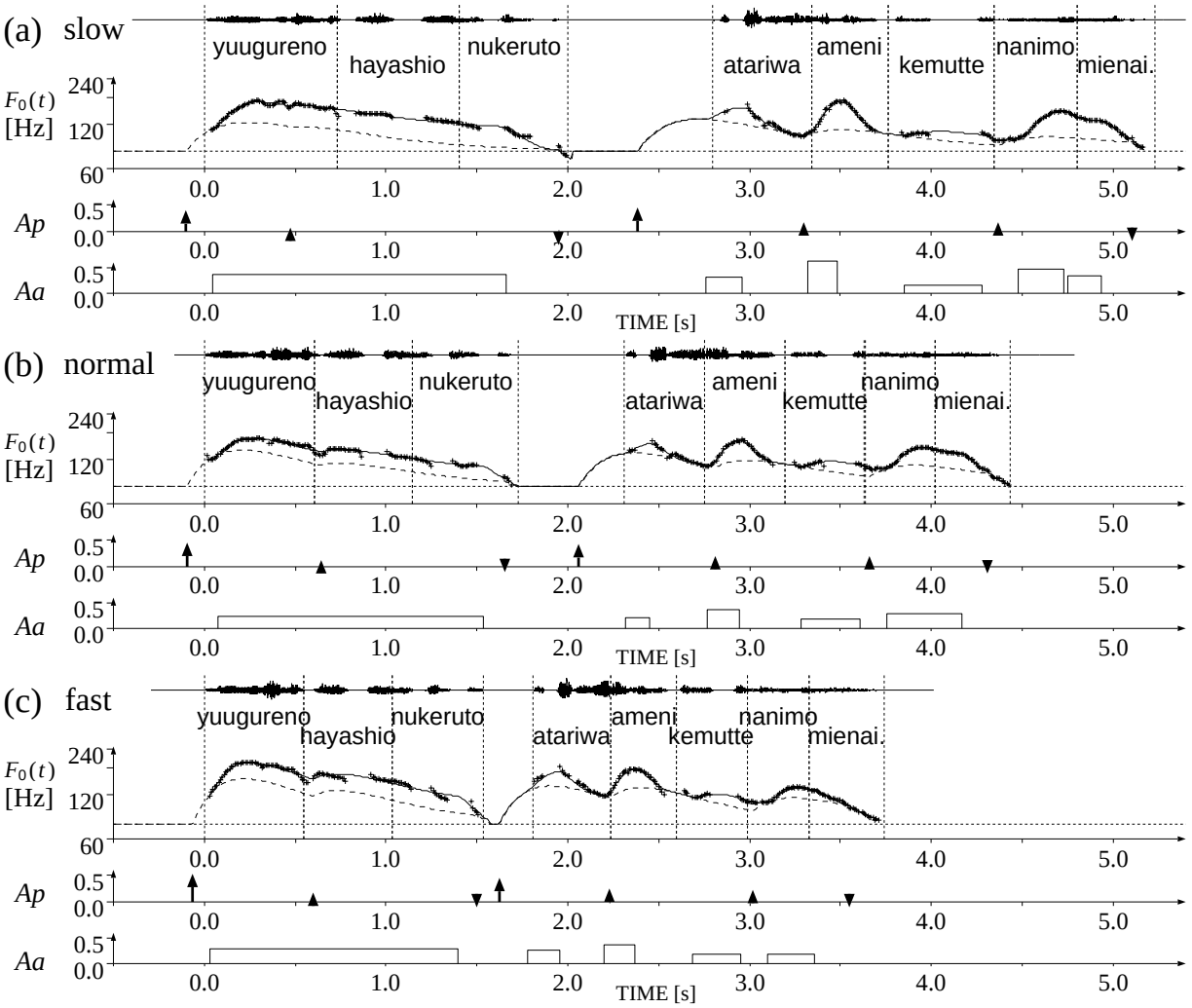


Figure 2: Examples of Analysis-by-Synthesis of F_0 contours of the same text uttered at three speech rates: (a) slow, (b) normal, and (c) fast.

Table 1: Means (μ) and standard deviations (σ) of α, β, and Fb of all the 14 sentence utterances for each of the three speech rates.

	α		β		Fb [Hz]	
	μ	σ	μ	σ	μ	σ
slow	3.18	0.08	18.3	1.49	77.3	4.0
normal	3.31	0.04	20.0	0.94	76.5	2.4
fast	3.62	0.10	20.5	1.03	77.3	2.3

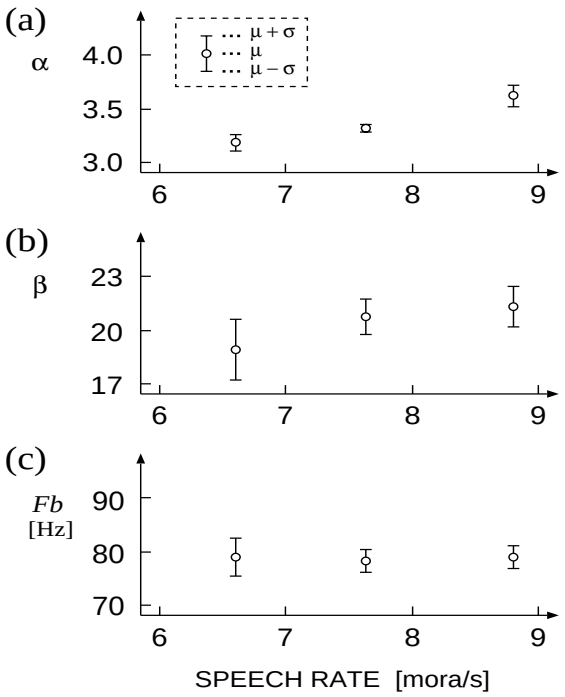


Figure 3: Means (μ) and standard deviations (σ) of α, β, and Fb of all the 14 sentence utterances for each of the three speech rates.

Table 2: Results of a t-test on significance of difference in mean values of α, β, and Fb of all the 14 sentence utterances for each of the three speech rates.

α	normal	fast
slow	$+++$	$+++$
normal		$+++$

β	normal	fast
slow	$+++$	$+++$
normal		$-$

Fb	normal	fast
slow	$-$	$-$
normal		$-$

$+++$	at 1 % level	$+$	at 10 % level
$++$	at 5 % level	$-$	not significant

4. ESTIMATION OF α AND β FOR A SET OF UTTERANCES PRODUCED AT A GIVEN SPEECH RATE

Although most of the variations in the F_0 contour such as those due to segmental microprosody are removed by pre-processing, F_0 contours of individual utterances are not completely free from local variations due to factors that are not taken into account by the F_0 contour model. These variations are considered to be responsible, to a certain extent, for the fluctuations of the estimated values of α and β obtained in the foregoing analysis. The effect of these F_0 contour variations will be generally reduced in longer utterances. In fact, if we assume that the values of α and β remain constant for the entire set of utterances produced at a given speech rate, we can estimate optimum values of α and β that are common to all the utterances within the set.

The following procedure was adopted to find the optimum values of α and β for a set of utterances produced at the same speech rate.

1) Starting from a fixed set of values of α and β, find the optimum set of values of all other parameters (Fb as well as magnitudes and timings of both phrase and accent commands) for each utterance within the set, minimizing the mean squared error between the actual F_0 contour and the model-generated F_0 contour on the logarithmic scale of F_0.
2) Take the inverse of the mean squared error for each utterance as a measure of goodness of fit, obtain the sum of their values for all the utterances of the set, and use it as the measure of overall goodness of the current set of α and β. This goodness measure is introduced in order to give greater weight for utterances with F_0 contours that have closer agreement with model-generated F_0 contours, thereby reducing the weights for F_0 contours that have larger deviations from model-generated contours due to local variations.
3) Search for the overall optimum set of values of α and β in the α-β plane.

Figure 4 shows the behavior of the overall measure of goodness on the α-β plane for the three speech rates: (a) slow, (b) normal and (c) fast. Panel (a) shows a three-dimensional display of the measure as a function of α and β. The behavior of the overall goodness measure is also displayed on the α-β plane as equi-potential contours. The results of all three cases indicate that the overall goodness measure shows a broad but global maximum for each speech rate. The optimum set of α and β are shown in Table 3, together with the mean value of Fb for each speech rate, obtained by using the overall goodness measure for each utterance as the weighting factor. These results are also plotted in Fig. 5.

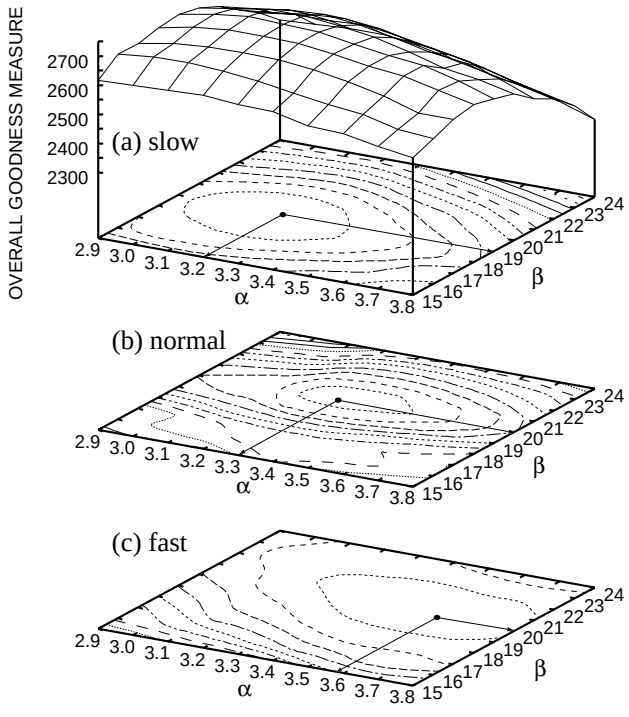


Figure 4: The behavior of the overall measure of goodness on the α-β plane as equi-potential contours for the three speech rates: (a) slow, (b) normal and (c) fast.

Comparison of these results with those obtained in the previous section indicates that the estimated values of α and β are very close and show essentially the same tendency against changes in the speech rate.

5. DISCUSSION AND SUMMARY

The results shown in Figs. 3 and 5 indicate that parameters α and β can be considered to be practically constant at a given speech rate, while the baseline frequency Fb may vary slightly from utterance to utterance. Among these three parameters, only α has a small but systematic tendency to increase with the speech rate. Further studies are being conducted on the effects of the speaking style, individual differences, and language differences.

6. REFERENCES

1. Fujisaki, H. and Nagashima, S. "A model for the synthesis of pitch contours of connected speech," *Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo*, Vol. 28, pp. 52–60, 1969.

Table 3: The optimum set of α and β and the mean and standard deviation of Fb.

	α	β	Fb [Hz] μ	σ
slow	3.22	18.8	77.1	4.0
normal	3.32	20.2	77.5	3.1
fast	3.56	20.2	77.3	2.1

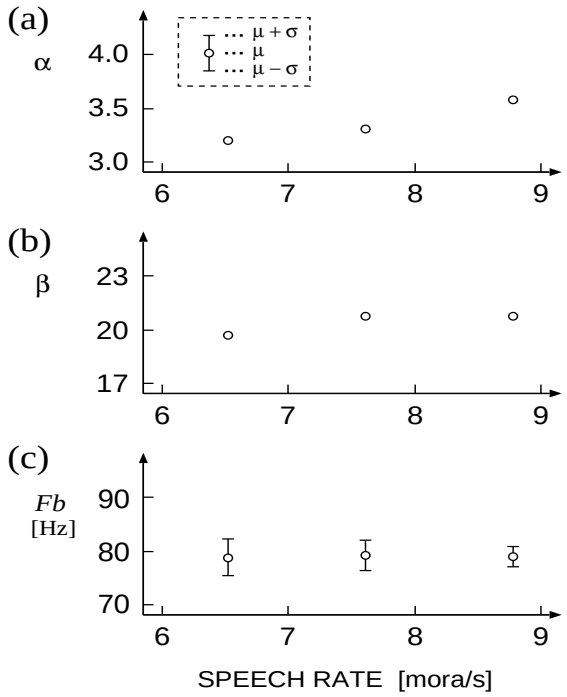


Figure 5: Means (μ) and standard deviations (σ) of α, β and Fb for each set of speech rates.

2. Fujisaki, H. and Sudo, H. "A model for the generation of fundamental frequency contours of Japanese word accent," *Journal of the Acoustical Society of Japan*, Vol. 27, pp. 445–453, 1971.
3. Fujisaki, H. and Hirose, K. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *The Journal of the Acoustical Society of Japan (E)*, Vol. 5, no. 4, pp. 233–242, 1984.
4. Fujisaki, H. "Modeling the process of fundamental frequency control of speech for synthesis of tonal features of various languages," *Proceedings of 1997 China-Japan Symposium on Advanced Information Technology*, pp. 1–12, 1997.