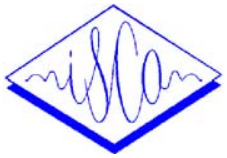

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Evaluation of a Method for Automatic Determination of F_0 Model Parameters

Shehui Bu†, Mikio Yamamoto‡ & Shuichi Itahashi‡

† Graduate School of Systems & Information Engineering
‡ Institute of Information Sciences & Electronics
University of Tsukuba, Japan

†bushehui@milab.is.tsukuba.ac.jp; ‡{myama, itahashi}@is.tsukuba.ac.jp

Abstract

This paper discusses the problems in the automatic method to determine the discrete parameters of the proposed F_0 model from the speech wave. The dynamic programming (henceforth DP method) and the least mean square error (LMSE) methods serve in the two-step algorithm proposed in this paper. Furthermore, in order to automatically detect the optimal number of phrase commands, decrease of LMSE is used. From the experiment results on a set of 11 sentences spoken by four Japanese speakers, we obtained 84.1% correct rate of phrase component extraction.

1. Introduction

The relationship between the F_0 pattern and the prosodic information can be quantitatively analysed if we have a suitable model that describes the process of generation of F_0 pattern in mathematical form. Several models such as Fujisaki model and downstep model, etc., were proposed to cope with this problem [3, 7]. In particular, many past research works have shown that Fujisaki model can describe the intonation mechanism of Japanese and some of other languages very well. The auotmatic method of extracting the parameters of the model is important and needed in the prosodic study. However, good initial values at the starting point is required in order to slove the nonlinear problem; it means that the manual operation is needed in this case. Some attempts have been made to slove this problem such as in references [8, 9]. Thees attempts have beeen made to approximate the input F_0 pattern by spline functions or 3rd-order polynomials and then to approximate the smoothed pattern by the F_0 model. In order to solve this problem, we propose a method which utilizes the DP and LMSE methods based on a revised F_0 model.

2. F_0 Model

Actually, the F_0 pattern of the speech sound can be influenced by various factors such as pause, duration, syntax, etc. and its mechanism is very complex. It is widely recognized that the F_0 patterns of an utterance descend at a fixed rate after the pause accompanied by inhalation of air becasue of the natural reduction of the expiratory pressure at the lungs. So it can be recognized that the F_0 patterns of words and sentences are generally characterized by a gradual declination from the onset toward the end of the utterance, superposed by local humps corresponding to such intonational factors as interrogation and emphasis [5]. In order to approximate the F_0 pattern more precisely, the item F_{min} in Fujisaki Model can be replaced by a slope line: $b_i(t - T_{0i}) + c_i$ as shown in equation (1) [1, 2]. By introducing this line component, we can represent the F_0 declination more precisely. The line equation with negative slope could cause negative F_0 but it does not happen usually within a limited duration of an accentual phrase due to the limited capacity of expiration.

$$\ln\left(\hat{F}_0(t)\right) = b_i(t - T_{0i}) + c_i + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}\left\{G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j})\right\} \quad (1)$$

where

$$G_{pi}(t) = \begin{cases} \alpha_i^2 t exp(-\alpha_i t) & : t \geq 0 \\ 0 & : t < 0 \end{cases} \quad (2)$$

and

$$G_{aj}(t) = \begin{cases} min\left[1 - (1 + \beta_j t)\, exp(-\beta_j t), \gamma\right] & : t \geq 0 \\ 0 & : t < 0 \end{cases} \quad (3)$$

Equation (2) indicates the impulse response function of the phrase-control mechanism and equation (3) indicates the step response function of the accent-control mechanism. The symbols in equations (1), (2) and (3) indicate:

I: number of phrase commands; J number of accent commands; A_{pi}: magnitude of the ith phrase command; A_{aj}: magnitude of the jth accent command; T_{0i}: timing of the ith phrase command; T_{1j}: onset of the jth accent command; T_{2j}: end of the jth accent command; α_i: natural angular frequency of the phrase control mechanism to the ith phrase command; β_j: natural angular frequency of the accent control mechanism to the jth accent command; γ: a parameter to indicate the ceiling level of the accent component [3, 9].

In this stduy, we chose the following values for the three parameters: $\alpha_i = 3.0(rad/sec.)$, $\beta_j = 20.0(rad/sec.)$, and $\gamma = 0.9$ according to the reference [3].

3. Automatic Parameters Determination Methods

3.1. Determination of the Model Parameters

The most important problems of the automatic determination methods are how to determine the optimum values and timing of the model parameters.

First, according to references [1, 6], suppose we approximate the $F_0(t)$ by the function $\hat{F}_0(t)$ as shown in Equation (4),

we can obtain the mean square error by the evaluation function:

$$\varphi(t) = \frac{1}{T}\sum_{t=1}^{T}\left\{\hat{F}_0(t) - F_0(t)\right\}^2 w(t) \qquad (4)$$

The $w(t)$ is 1 when the speech in test is voiced, otherwise $w(t)$ is 0. Let us minimize the Equation (4) in respect of a set of parameters A_{pi}, A_{aj}, b_i and c_i, then we can work out the four parameters. Here, b_i must be negative; howerver A_{pi}, A_{aj}, and c_i must be postive [1, 2].

Second, the problem of the timing of the parameters can be solved by dynamic programming method. The principle of dynamic programming method can be said to reduce a multi-stage problem into a two-stage problem. So the storage tables which keep the optimum values of the previous stages will be required to reduce the cost of the total calculation. First, we divide the speech sentence of the duration T into N interivals. Then, minimizing procedure can be divided into two parts: one is the minimization of kth interval only and another is minimization of $1st$, $2nd$, $\cdots$, $(k-1)th$ intervals. Then we can represent the process so that the interval from 0 to n_k is divided into k stages and the optimization is executed for each segment [1, 2, 6].

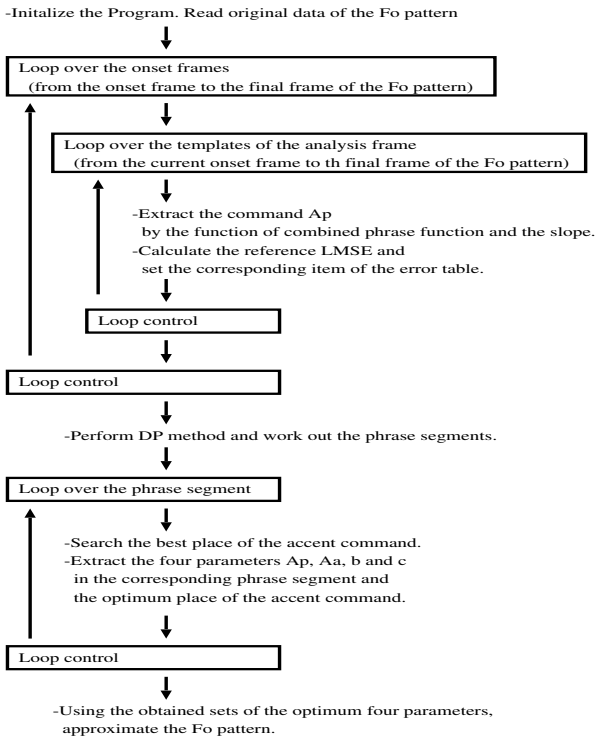

Figure 1: *Schematic diagram of the two-step algorithm*

The simultaneous calculation of all four parameters, A_{pi}, A_{aj}, b_i and c_i, exhausts considerable calculation time in DP method. In order to reduce the computation time exhausted by DP method, a two-step algorithm has been devised as shown in Fig. 1. First step: The timing of the phrase commands can be stably and suitably extracted by only using the combination of the phrase component and the slope function. Second step: The work in this step is to extract the four parameters from each phrase interval. Firstly, according to the result of the first step, we can divide the utterance sentence into phrase segments. Secondly the procedure is composed of two parts in each phrase segment. One is to search the best boundaries of the accent command in the corresponding phrase segment. The other is to work out the four parameters in the current phrase segment by using the least square error method. Repeat this procedure until we reach the final phrase segment. At present, we assume that there is only one accent command within each phrase to save computation time [1, 2].

3.2. Determination of Optimum Number of the Phrase Segments

It is necessary to determine the optimum number of the phrase segments in the automatic algorithm used in the analysis of real speech. In this paper, we tried three indices, least mean square error (LMSE) $E(l)$, decrease of LMSE $c(l)$ and decreasing rate of the LMSE $d(l)$ [1, 2], where l denotes the number of phrase segments.

The decrease of the least mean square error $c(l)$ is defined as

$$c(l) = |E(l) - E(l-1)|\,, \qquad (5)$$

and the decreasing rate of the least mean square error $d(l)$ is defined as

$$d(l) = |E(l) - E(l-1)|\,/E(l), \qquad (6)$$

where, the $E(l)$ is the least mean square error when the number of the phrase segment is l:

$$E(l) = \frac{1}{N}\sum_{n=1}^{N}\left\{\hat{F}_0^{(l)}(n) - F_0(n)\right\} w(n) \qquad (7)$$

Here, $F_0(n)$ is the F_0 extracted from the speech wave, $\hat{F}_0^{(l)}(n)$ is the approximated F_0 when the phrase segment number is l, N denotes the total number of the frames.

4. Experiments

4.1. Speech Material

A set of 11 declarative sentences of Japanese [3], consisting only of voiced segments, were selected as shown in Table 1. In this table, the column item, *P*, is the supposed number of phrase segments which were determined from the real F_0 pattern by visual inspection. Because of the ambiguty of phrase boundaries and individualily of the speaker, some of the sentences in this set can not be uniquely determined. We can consider that these sentences have more than one candidate of the number of phrase segments as shown in Table 1. Each sentence was uttered as naturally as possible, and recorded twice (the first time was an exercise) by four speakers (two males and two females of 20 years of age, named M1, M2, F1 and F2, respectively) of Tokyo dialect in a sound-proof room. The recorded speech materials were sampled at 16kHz, quantized with 16 bit accuracy. The F_0 contours of all the speech material were extracted by the *AMDF* method [4], and the frame interval in the analysis was $10ms$.

4.2. Analysis Example of a Japanese sentence

In order to quantitatively evaluate the proposed method, an experiment was conducted. The material is one of the Japanese sentences $/anoaoiaoinoewaaru/$ (*There is a picture of that*

Table 1: *Speech materials. (P : The number of phrase segments in each sentence)*

No.	Meaning in English	Phonemic transcription	P
1	a blue hollyhock	/aoiaoi/	2
2	a picture of blue hollyhock	/aoiaoinoe/	2,3
3	that picture of blue hollyhock	/anoaoiaoinoe/	3
4	There is a picture of hollyhock.	/aoinoewaaru/	2
5	There is a picture of blue hollyhock.	/aoiaoinoewaaru/	2,3
6	There is a picture of hollyhock picture in a house.	/aoinoewaieniaru/	2,3
7	There is a picture of that blue hollyhock.	/anoaoiaoinoewaaru/	3,4
8	There is a picture of blue hollyhock in a house.	/aoiaoinoewaieniaru/	3,4
9	That picture of blue hollyhock is in a house.	/anoaoiaoinoewaieniaru/	4
10	The picture of a hollyhock is in a house on a mountain.	/aoinoewayamanouenoieniaru/	2,3
11	The picture of a blue hollyhock is in a house on a mountain.	/aoiaoinoewayamanouenoieniaru/	3

blue hollyhock.) listed in Table 1 (Sentence 7) and spoken by the male speaker M1. Figure 2 denotes the least mean square error (LMSE) $E(l)$, the decrease of LMSE $c(l)$, and the decreasing rate of the LMSE $d(l)$, respectively. According to $c(l)$ and $d(l)$, as shown in this figure, we can see that $c(3)$ and $d(3)$ are the maximimum when the number of phrase segments is 3. We can assume that the optimum number of the phrase segments is 3. Figure 3 illustrates the results of the parameter extraction experiment by using the proposed two-step algorithm when the number of phrases is 3. From this figure, it is seen that the approximated F_0 pattern approaches the real F_0 contour.

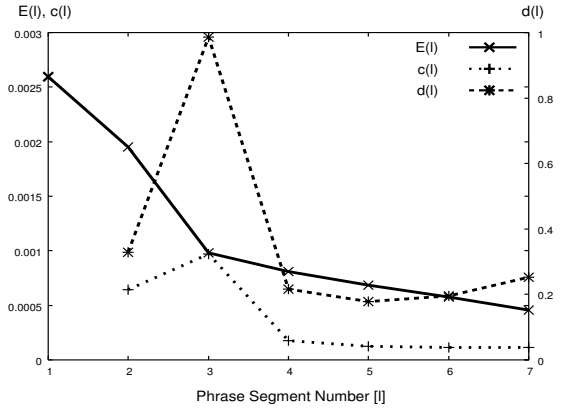

Figure 2: *The least mean square error (LMSE) $E(l)$, decrease of the LMSE $c(l)$, and decreasing rate of the LMSE $d(l)$ of the Japanese sentence $/anoaoiaoinoewaaru/$ (There is a picture of that blue hollyhock.).*

From this result, we can see that the amplitude of the accent command in the second phrase segment is zero, which means there is no accent command in the second phrase segment. The phrase commands, A_{p1}, A_{p2} and A_{p3}, correspond nearly to the actural phrases in the uttered sentence. In addition, the accent commands, A_{a1} and A_{a3}, similarly fit the real accentuation in the utterance. According to these results, we can conclude that the experiment result fits well when the number of phrase segments is 3.

Table 2: *The number of Phrases decided by the decrease of LMSE $c(l)$.* C:*($\surd$: correct; $\times$: incorrect.)*

No.	M1		M2		F1		F2	
	$c(l)$	C	$c(l)$	C	$c(l)$	C	$c(l)$	C
1	2	$\surd$	2	$\surd$	2	$\surd$	2	$\surd$
2	3	$\surd$	2	$\surd$	2	$\surd$	2	$\surd$
3	3	$\surd$	3	$\surd$	3	$\surd$	3	$\surd$
4	3	$\times$	3	$\times$	2	$\surd$	2	$\surd$
5	2	$\surd$	3	$\surd$	2	$\surd$	2	$\surd$
6	2	$\surd$	2	$\surd$	2	$\surd$	2	$\surd$
7	3	$\surd$	3	$\surd$	3	$\surd$	3	$\surd$
8	3	$\surd$	4	$\surd$	2	$\times$	4	$\surd$
9	4	$\surd$	2	$\times$	2	$\times$	2	$\times$
10	2	$\surd$	2	$\surd$	3	$\surd$	2	$\surd$
11	2	$\times$	3	$\surd$	3	$\surd$	3	$\surd$
	81.8%		81.8%		81.8%		90.9%	

Table 3: *The number of Phrases decided by the decreasing rate of LMSE $d(l)$.* C *: ($\surd$: correct; $\times$: incorrect.)*

No.	M1		M2		F1		F2	
	$d(l)$	C	$d(l)$	C	$d(l)$	C	$d(l)$	C
1	3	$\times$	2	$\surd$	2	$\surd$	4	$\times$
2	3	$\surd$	2	$\surd$	3	$\surd$	2	$\surd$
3	3	$\surd$	2	$\times$	3	$\surd$	3	$\surd$
4	3	$\times$	3	$\times$	2	$\surd$	4	$\times$
5	4	$\times$	3	$\surd$	2	$\surd$	2	$\surd$
6	5	$\times$	6	$\times$	2	$\surd$	2	$\surd$
7	3	$\surd$	5	$\times$	2	$\times$	3	$\surd$
8	3	$\surd$	4	$\surd$	5	$\times$	4	$\surd$
9	4	$\surd$	4	$\surd$	2	$\times$	3	$\times$
10	5	$\times$	2	$\surd$	2	$\surd$	3	$\surd$
11	5	$\times$	7	$\times$	3	$\surd$	5	$\times$
	45.5%		54.5%		72.7%		63.6%	

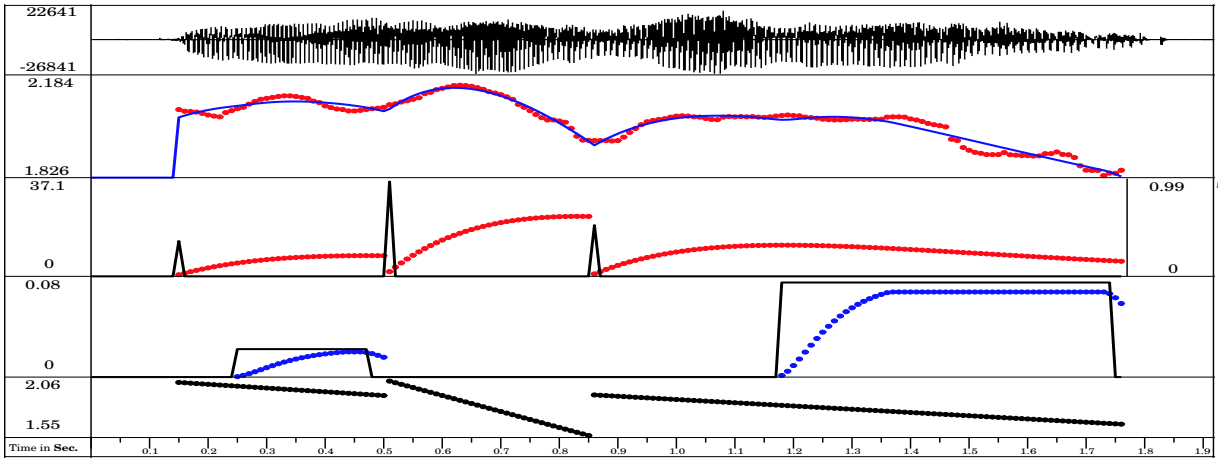

Figure 3: *Results of the analysis in the Japanese sentence /anoaoiaoinoewaaru/ (There is a picture of that blue hollyhock.). Speech wave, F_0 pattern (dots) and its approximation (line) are plotted on the logarithmic scale, phrase commands(triangle, sacle on left side) and phrase segments(curves, scale on right side), accent commands (rectangles) and accent segment(curves), slope lines are illustrated, respectively.*

4.3. Accuracy of the automatic determination methods for the number of phrase segments

A comparative experiment was executed in order to evaluate the accuracy and validity of the indices, $c(l)$ and $d(l)$, which are introduced to determine the number of phrase segments automatically. The procedure of the comparative experiment is as fillows. First we set the maximum number of the phrase segments as 7, and calculate the corresponding $E(l)$, $c(l)$ and $d(l)$ by using the proposed two-step algorithm, and get the best numbers of $c(l)$ and $d(l)$.

Tables 2 and 3 show the results of the experiment. In these tables, two items $c(l)$ and $d(l)$ are the optimum phrase number which corresponds to the maximum values of $c(l)$ and $d(l)$by using the Equations 5 and 6, respectively. From Table 2, we can observe that the accuracy of $c(l)$ corresponding to speakers M1, M2, F1 and F2 is between 81.8% and 90.9%. In contrast, the accuracy of $d(l)$ ranges from 45.5% to 72.7% as shown in Table 3. The average accurate rate of $c(l)$ is 84.1% and that of $d(l)$ is 59.1%. According to these results, we can conclude that the accuracy of $c(l)$ is much higher than $d(l)$.

5. Conclusion and Future Work

In this paper, an automatic method to extract the parameters based on a revised F_0 model was proposed in order to extract the F_0 model parameters automatically. The two-step algorithm can determine the phrase and accent commands without convergence problems. An experiment was conducted by using a set of 11 Japanese sentences spoken by 4 Japanese speakers. From the results of the examples illustrated in this paper, we observed that this method is sufficiently suitable and accurate. And the decrease of LMSE $c(l)$ can automatically determine the number of phrase segments.

For the future work, we will continue to verify the accuracy of this method, and focus on improving this method and seeking a more effective and advanced algorithm with:

- Sentences including consonants,
- More speakers, more sentences,
- Improvement of the two-step algorithm into one-stage DP
- The flexibility of the phrase command start time (so that the phrase command time before a voice start point is accepted)

6. References

[1] S. Bu, M. Yamamoto, S. Itahashi, 2003. A method of automatic extraction of F_0 Model parameters. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition 2003*, Tokyo, Japan, 227-230.

[2] S. Bu, M. Yamamoto, S. Itahashi, 2003. Considerations on automatic parameters estimation of F_0 Model. In *Prep. Autumn Meeting of the Acoust. Soc. of Jpn*, Paper 1-8-23, 227-228. (in Japanese)

[3] H. Fujisaki, K.Hirose, 1984. Analysis of voice fundamental frequency contours for declarative Sentences of Japanese. *Jour. Acoust. Soc. Jpn. (E)Vol.5, No.4,* 233-242.

[4] S. Furui, 2001. *Digital speech processing, synthesis and recognition, second edition, revised and expanded,* Marcel Dekker Inc.

[5] K. Hakoda, H. Sato, 1980. Prosodic Rules in Connected Speech Synthesis, *IEICE Trans. , Vol. J63-D, No. 9,* 715-722. (in Japanese)

[6] S. Itahashi, 1978. Description of speech data pattern by several functions with applications to formant and fundamental frequency trajectories, *STL-QPSR 2-3,* 1-22.

[7] H. Kubozono, 1993. *The organization of Japanese prosody,* Kurosio Publishers, Tokyo Japan.

[8] H. Mixdorff, 2000. A novel approach to the full automatic extraction of Fujisaki model parameters, *ICASSP2000, Vol. 3,* 1281-1284.

[9] S. Narusawa, N. Minematsu, K. Hirose, H. Fujisaki, 2002. A method for automatic extraction of the fundamental frequency contours generation model, *IPSJ Jour. , Vol. 43, No. 7,* 2155-2168. (in Japanese)