A Method for Automatically Estimating F0 Model Parameters and A Speech Re-Synthesis Tool Using F0 Model and STRAIGHT

Shota Sato, Taro Kimura, Yasuo Horiuchi, Masafumi Nishida, Shingo Kuroiwa, and Akira Ichikawa

Chiba University, Japan, Nintendo Co., Ltd, Japan, Waseda University, Japan

Abstract

In this paper, we describe a speech re-synthesis tool using the fundamental frequency (F0) generation model proposed by Fujisaki et al. and STRAIGHT, designed by Kawahara, which can be used for listening experiments by modifying F0 model parameters. To create the tool, we first established a method for automatically estimating F0 model parameters by using genetic algorithms. Next, we combined the proposed method and STRAIGHT. We can change the prosody of input speech by manually modifying the F0 model parameters with the tool and evaluate the relation between human perception and F0 model parameters. We confirmed the ability of this tool to make natural speech data that have various prosodic parameters.

Index Terms: fundamental frequency, F0 generation model, genetic algorithm, STRAIGHT

1. Introduction

Prosody conveys a lot of information, e.g., emotions, individual characteristics, syntactic structure, and so on. By using prosody, we can understand the semantic and emotional information from an utterance in real time communication [1][2]. However, in the dialogue between humans and machines, prosody is not used effectively. Hence, we have been studying its effective use, especially that of fundamental frequency (F0), for spoken dialogue systems [3]. F0 plays an important role in spoken Japanese, such as intonation. We also found that F0 includes useful information for detecting speaker change [4]. We think that these facts can be used for improving the response time of a system, and for naturally prompting users to utter. Thus, we developed a speech re-synthesis tool that can change an F0 contour according to the F0 generation model.

Fujisaki et al. [5], proposed an F0 generation model (F0 model) and showed that the F0 contour of an utterance can be generated from two types of components: phrase and accent. This model has high approximation accuracy for the F0 contour of Japanese speech. However, in conventional methods, F0 model parameters were estimated manually by experts using linguistic information. We propose an automatic estimation method, which does not need linguistic information, of increasing and decreasing the number of parameters and repeating the general algorithm (GA) process.

STRAIGHT [6], designed by Kawahara to investigate human speech perception, is a speech analysis, modification, and synthesis system. Even when the prosody is changed, this system can make high-quality re-synthesized speech. We developed a tool combining STRAIGHT and our method for automatically estimating F0 model parameters to make the speech data used for listening experiments. This tool can freely change F0 model parameters and can re-synthesize speech data with changed F0 model parameters in a GUI environment.

In section 2, we introduce the F0 model. In section 3, we describe our proposed F0 model parameter estimation method and show the results for the ATR phonetically-balanced sentence speech database. The developed tool and experimental results are shown in section 4.

2. F0 model

An F0 model generates an F0 contour from two main components: phrase and accent. Phrase components are described by an impulse response function generated from the control mechanism of equation (1), and accent components are described by a step response function generated from the control mechanism of equation (2). The F0 contour is then calculated as the sum of these components by equation (3).

\[
G_{\text{p}}(t) = \begin{cases} 
\alpha \cdot t e^{-\alpha t} & : t \geq 0 \\
0 & : t < 0 
\end{cases} 
\]

(1)

\[
G_{\text{a}}(t) = \begin{cases} 
\min\left[1 - \left(1 + \beta \right) e^{-\beta t} \right] & : t \geq 0 \\
0 & : t < 0 
\end{cases} 
\]

(2)

\[
\ln F_0(t) = \ln F_0 + \sum_{i=1}^{j} A_{\text{p}} G_{\text{p}}(t - T_{i0}) + \sum_{j=1}^{j} A_{\text{a}} \left[ G_{\text{a}}(t - T_{j1}) - G_{\text{a}}(t - T_{j2}) \right] 
\]

(3)

Fig. 1 F0 model

Copyright © 2008 ISCA
Accepted after peer review of full paper

September 22 – 26, Brisbane Australia
The symbols in equations (1), (2), and (3) are defined as follows: \( i \) the number of phrase commands; \( j \) the number of accent commands; \( a \) the natural angular frequency of the phrase mechanism; \( \beta \) the natural angular frequency of the accent mechanism; \( \gamma \), a parameter to indicate the ceiling level of the accent components; \( F_0 \) the baseline value of fundamental frequency; \( A_n \) the magnitude of the phrase command; \( T_0 \) the time of the phrase command; \( T_s \), the onset time of the phrase command; and \( T_2 \), the termination time of the phrase command.

3. Estimation of F0 model parameters

Determining the number of phrase and accent components is important in the F0 model estimation. In conventional methods [7][8][9], F0 model experts manually determined the number of components by using linguistic information and comparing the original F0 contour and synthesized F0 contours. This meant that dealing with large amounts of data was difficult. To circumvent this limitation, we propose an automatic estimation method that obtains the optimal parameter set (the optimal number of optimal parameters) by increasing and decreasing the number of parameters and repeating the GA process. The flow of the proposed method is shown in Fig. 2.

(1) The utterance is divided into speech segments at places where there are long pauses, and then the following steps (2) to (6) are applied to each segment.

(2) The initial number of phrase components is determined according to the time length of the divided section, and the number of accent components is set to twice the number of phrase components.

(3) The parameters are estimated by GA.

(4) Unnecessary phrase and accent components are determined by the method described in subsection 3.3. If they are found, they are deleted, and the process returns to step (3) When no more unnecessary components are found, the estimated parameters are saved as the final candidate.

(5) One phrase component and two accent components are increased, and the parameters are estimated again.

(6) If the increased components are determined as unnecessary, the optimal number of components is fixed by the final candidate and the process progresses to step (7). If not, the estimated parameters are saved as the final candidate, and the process returns to step (5).

(7) The F0 model parameters for the whole utterance are constructed by concatenating the final candidate of each speech segment. After that, a final estimation is made by GA.

3.1. Initialization

In the proposed method, the input utterance is divided into speech segments at places where there are long pauses to reduce the computation time for the estimation by GA. In our experiments, we divided the input speech at pauses of 250 ms or longer. If a speech segment is shorter than one second, the number of phrase components is initialized to one and the number of accent components is initialized to two. When a speech segment is more than one second, the number of phrase components is increased for every second and each phrase component is accompanied by two accent components.

3.2. Estimation of parameters by GA

Estimation of F0 model parameters from an F0 contour of natural speech is an optimization problem where many parameters must be determined. Optimization should be performed by changing all the parameters simultaneously because the F0 model parameters are interdependent.

For this, we propose a novel method of estimating F0 model parameters using GA. In GA, solutions are represented in bit strings. We quantized each F0 model parameter and set all of them as genetic representations. A fitness function is defined by the mean square error between the observed F0 contour and the synthesized F0 contour by the F0 model parameters, as shown in equation (4).

\[
\text{adp} = \left[ \frac{1}{N} \sum_{t}^{N} (F_0(t) - F_m(t))^2 + 1 \right]^{-1} \tag{4}
\]

3.3. Determination of necessity of components

If the number of components is determined to be more than is suitable, the GA process tries to decrease the influence of extra phrases and accents through evolution. In the proposed method, the components that satisfy the following conditions are judged to be unnecessary as a result of evolution.

The conditions for phrase components follow:
- The components are moved to a silent period.
• The $A_p$ (magnitude of phrase command) is smaller than the predetermined threshold. In this paper we use 0.05 as the threshold.

The conditions for accent components follow:
• The components are moved to a silent period.
• The $A_a$ (magnitude of accent command) is smaller than the predetermined threshold. In this paper we use 0.1 as the threshold.

The length from $T_1$ to $T_2$ becomes shorter than the predetermined threshold. In this paper we use 0.1 [s] as the threshold.

3.4. Estimation results of F0 model parameters
The estimation results of F0 model parameters by the proposed method are shown in Figs. 3 and 4. Male utterances in the ATR phonetically-balanced sentence speech database were used for input. The symbols in the figures are defined as follows: dotted line, original speech of F0; solid line, estimated F0 contour; up-pointing arrow, phrase command; and rectangular pulse, accent command.

Although errors are seen in transition sections between the voiced and unvoiced areas in the figures, the estimated F0 contours have high approximate accuracy for the F0 contour of the original speech.

Fig. 3 Estimation result of F0 model for Japanese sentence /arayuru geNjitsuwo subete jibuNhouhe nejimagetanoda/ spoken by MHT

Fig. 4 Estimation result of F0 model for Japanese sentence /isshuukaNbakari nyuuyookuwo suzaishita/ spoken by MHT

4. Development of speech re-synthesis tool
We are planning subjective experiments to compare speech data that have various prosodic parameters. To make such speech data, the following functions were necessary for the developed tool.

The tool can
• estimate F0 model parameters from input speech automatically,
• change prosodic parameters freely, and
• re-synthesize speech with changed prosodic parameters.

In the first stage of development, we can change F0 model parameters and speech speed with this tool.

4.1. Structure of tool
Combining STRAIGHT and the automatic estimation method of F0 model parameters achieved the tool’s functions.

A flowchart of the process is shown in Fig. 5. The processes from input to re-synthesis are executed in steps (1) to (5).

(1) The F0 of the input speech is extracted using the F0 extract function of STRAIGHT.
(2) The number of components and F0 model parameters are estimated using the GA automatic estimate method.
(3) The F0 model parameters are changed, and the F0 of the speech is replaced by the F0 contour calculated from the changed parameters.
(4) The speech speed is changed by expanding and contracting the time axis.
(5) The speech is re-synthesized with the changed F0 model parameters and speech speed using STRAIGHT.

Fig. 5 Flowchart of speech re-synthesis tool
4.2. GUI

The developed tool has a GUI (Fig. 6) for users to make speech data easily.

![GUI Screenshot](image)

Fig. 6 Screenshot of GUI

4.3. Results of changed F0 model parameters

Examples of F0 contours obtained by the developed tool by changing the original F0 model parameters are shown in Figs. 7 and 8. In these examples, the input data are male utterances from the same ATR phonetically-balanced sentence speech database as in subsection 3.4. In Figs. 7 and 8, the solid line indicates the original F0 contour and the dotted line indicates the modified F0 contour.

- Change of phrase command (Fig. 7)
  For the first phrase command, the magnitude \( A_{p1} \) increased from 0.5 to 0.8 and the time \( T_{p1} \) slowed from 0.05 to 0.2. For the second phrase command, the natural angular frequency \( \omega_2 \) changed from 3.51 to 2.

![Phrase Command](image)

Fig. 7 Change of phrase command for a Japanese sentence /isshuuka/Nbakari nuuyookuwo shuzaishita/ spoken by MHT

- Change of accent command (Fig. 8)
  The magnitude of the first accent command \( A_{a1} \) changed from 0.5 to 0.8, and the natural angular frequency of the second accent command \( \beta_2 \) changed from 3.41 to 10. The onset time of the third phrase command \( T_{p3} \) changed from 2.38 to 2.2.

![Accent Command](image)

Fig. 8 Change of accent command for a Japanese sentence /isshuuka/Nbakari nuuyookuwo shuzaishita/ spoken by MHT

Figures 7 and 8 show that the F0 contour is modified in accordance with changes of the F0 model parameters. As shown in these examples, we can easily make natural speech data that have various prosodic parameters.

5. Conclusions

We developed a novel F0 modification and speech re-synthesis tool using an F0 model. To estimate the F0 model parameters automatically, we proposed an estimation method of increasing and decreasing the number of parameters and repeating the GA process. Experimental results using the proposed method showed that the estimated F0 contours had high approximate accuracy for the F0 contour of the original speech. We developed a re-synthesis tool by combining the proposed method and STRAIGHT. The tool can change the prosody of input speech by manually modifying the F0 model parameters. We confirmed the ability of this tool to make natural speech data that have various prosodic parameters.

We plan to improve the tool by adding more functions, such as a change of pause and power, and performing listening experiments.

6. Acknowledgements

This research has been partially supported by the Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research(B),17300065,18300045, and the Okawa Foundation for Information and Telecommunications.

7. References