A METHOD FOR AUTOMATIC EXTRACTION OF PARAMETERS OF THE FUNDAMENTAL FREQUENCY CONTOUR

Shuichi Narusawa 1, Hiroya Fujisaki 1 and Sumio Ohno 2

1 Science University of Tokyo, 2641 Yamazaki, Noda, 278-8510, Japan
2 Tokyo University of Technology, 1404-1 Katakura, Hachioji, 192-0982, Japan

ABSTRACT
The process of generating an F0 contour has been modeled quite accurately in mathematical terms by Fujisaki and his coworkers, but the derivation of the underlying commands from an observed F0 contour is an inverse problem that cannot be solved analytically. Although it can be solved by successive approximation, a good first-order approximation is necessary to guarantee an efficient and accurate search for the optimum solution. The present paper describes a method for pre-processing an observed F0 contour to obtain a smooth contour, from which a good first-order approximation can be analytically obtained. Experimental results show that correct extraction rates for the accent and the phrase commands are about 90% and 79%, respectively.

1. INTRODUCTION
In many languages of the world, the contour of the fundamental frequency of voice (henceforth F0 contour) is used to convey linguistic information concerning lexical tone/accent, syntax, and focus, as well as para- and non-linguistic information concerning speaker’s intention, attitude, speaking style, gender, physical and emotional states, etc. Fujisaki and his co-workers have shown that the process of F0 contour generation can be accurately represented by a mathematical model, which generates an F0 contour in response to a set of commands whose parameters correspond well to the linguistic and paralinguistic information of the utterance [1]. The model was first developed for F0 contours of Japanese, but has since been shown to apply to other languages [2,3].

While the generation of an F0 contour from a set of input commands is quite straightforward if we use the model, the derivation of the underlying commands from a given F0 contour is an inverse problem which cannot be solved analytically. Although it can be solved by successive approximation, a good first-order approximation is necessary to guarantee an efficient and accurate search for the optimum solution. The present paper first describes a method of pre-processing to remove gross errors and other disturbances from an observed F0 contour, and to approximate it by a mathematically simpler curve. This approximation reduces the derivation of the first-order approximation to an analytically solvable problem. The first-order approximation is then refined by a two-stage Analysis-by-Synthesis to obtain a more accurate solution. Although the method is applicable, with certain language-specific modifications, to F0 contours of various languages, the present paper deals with F0 contours of Japanese.

2. FUNCTIONAL MODEL FOR THE PROCESS OF GENERATION OF F0 CONTOURS OF AN UTTERANCES OF JAPANESE
Figure 1 shows the model for the process of generation of F0 contours of Japanese utterances. The mechanism that produces changes in log F0(t) from the phrase commands is named ‘phrase control mechanism’ and its outputs are named ‘phrase components.’ Likewise, the mechanism that produces changes in log F0(t) from the accent commands is named ‘accent control mechanism’ and its outputs are named ‘accent components.’ The outputs of these two mechanisms are added to a constant component log F0 to produce the final log F0(t). Although a further mechanism (‘glottal oscillation mechanism’) is required to obtain the glottal source waveform, this final stage can be disregarded in the discussion of log F0(t). For the rest of the paper, we shall use the word ‘F0-contour’ to indicate log F0(t).

In this model, the F0 contour is expressed by

$$\log F_0(t) = \log F_b + \sum_{i=1}^{I} A_p_i G_p(t - T_{1i}) + \sum_{j=1}^{J} A_a_j \left( G_a(t - T_{1j}) - G_a(t - T_{2j}) \right)$$

(1)

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

(2)

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

(3)

where Gp(t) represents the impulse response function of the phrase control mechanism and Ga(t) represents the step response function of the accent control mechanism.

Figure 1: A functional model for the process of generating F0 contours.
The symbols in these equations indicate:

- $F_b$: baseline value of fundamental frequency.
- $I$: number of phrase commands.
- $J$: number of accent commands.
- $A_p$: magnitude of the $i$th phrase command.
- $A_o$: amplitude of the $j$th accent command.
- $T_o$: timing of the $i$th phrase command.
- $T_a$: onset of the $j$th accent command.
- $\alpha$: natural angular frequency of the phrase control mechanism.
- $\beta$: natural angular frequency of the accent control mechanism.
- $\gamma$: relative ceiling level of accent components.

Parameters $\alpha$ and $\beta$ are known to be almost constant within an utterance as well as across utterances of a particular speaker. Although certain individual differences exist across speakers, it has been shown that $\alpha = 3.0$ and $\beta = 20$ can be used as default values. Parameter $\gamma$ may be variable across utterances and speakers, but it has also been shown that $\gamma = 0.9$ can be used as a default value.

3. PRE-PROCESSING OF MEASURED $F_0$ CONTOURS

Pre-processing [4] of an actual $F_0$ contour consists of four stages: (1) gross error correction, (2) microprosody removal, (3) interpolation, and (4) smoothing.

3.1. Correction of Gross Errors

Due to irregularities inherent in the mechanism of vocal fold vibration, no existing algorithm is completely free from gross errors. Gross errors can be classified into two types: (1) assignment of a false value (including zero) to a frame corresponding to a voiced interval, and (2) assignment of non-zero frequency value to a frame corresponding to a voiceless interval. The algorithm for correction of these two types of gross errors consists of the following two stages:

1. Removal of gross errors in voiced intervals

If the total number of frames with non-zero $F_0$ values is larger than $m$, and if

$$\frac{\log F_0(i)}{\log F_0(M[i, 2m + 1])} - 1 > S,$$

where $F_0(M[i, 2m + 1])$ indicates the median value of $F_0$ over the $(2m + 1)$ frames centered at the $i$th frame, then mark $F_0(i)$ to be a gross error. In all other cases, $F_0(i)$ is not regarded as a gross error.

After $F_0$ values of all the frames are thus examined, $F_0$ values judged to be gross errors are replaced by linear interpolation in the $\log F_0$ domain. For a frame step of 10 ms, $m = 2$ and $S = 0.01$ were found to be appropriate on the basis of preliminary experiments.

2. Removal of gross errors in silent or voiceless intervals

Since gross errors due to false detection of $F_0$ in silent or voiceless intervals seldom occur in successive frames, they can be removed by median smoothing over $(2n + 1)$ frames. For a frame step of 10 ms, $n = 2$ was found to be appropriate, allowing the removal of gross errors in at most two consecutive unvoiced frames.

3.2. Removal of Microprosody

The influence of consonantal articulation on $F_0$ contours, called 'microprosody' [4], is often quite large especially in voiceless consonants, and thus has to be removed, since it is not included in the model. It appears as $F_0$ transitions at boundaries between adjacent vowels. The procedure for removing the consonantal disturbances can be stated as follows.

Let $i$ be the frame number that immediately precedes the voiceless consonant and $j$ be the frame number that immediately follows it. Calculate the gradients of the $F_0$ contour (to be denoted by $G_0(i)$) in the vicinity of these boundaries.

If, for $n_1 \leq 10$, $G_0(i - n_1)$ has the same sign as $G_0(i - 1)$ and $|G_0(i - n_1)| > |G_0(i - 1)|/2$, remove all $F_0$ data at frames $(i - n_1, i - n_1 + 1, \ldots, i - 1)$. Likewise, for $n_2 \leq 10$, if $G_0(j + n_2)$ has the same sign as $G_0(j + 1)$ and $|G_0(j + n_2)| > |G_0(j + 1)|/2$, remove all $F_0$ data at frames $(j, j + 1, \ldots, j + n_2 - 1, j + n_2)$.

3.3. Interpolation of Intervals of Voiceless Consonants

After removal of $F_0$ data perturbed by microprosody, the $F_0$ contour for the interval including the original voiceless consonant and the microprosodic sections is interpolated by the following procedure.

By re-defining the starting and ending frame numbers of the interval to be interpolated as $[i + 1, j - 1]$, the $F_0$ contour for an expanded interval $[i - p, j + p]$ is approximated by a third order polynomial equation

$$F_0(t) = a_0 + a_1t + a_2t^2 + a_3t^3,$$

whose coefficients $[a_0, a_1, a_2, a_3]$ are obtained by the method of least mean squared error. This interpolation is performed for intervals whose lengths are less than 50 ms (i.e., for $j - i + 1 \leq 50$ at a frame interval of 10 ms). Longer intervals are considered as pauses and are not interpolated. The length of the adjacent ‘voiced’ interval at each end is selected to be 50 ms (i.e., $p = 4$ at a frame interval of 10 ms). This procedure assigns a continuous contour for the expanded ‘voiceless’ interval, but does not guarantee its continuity with the adjacent $F_0$ data.

3.4. Smoothing

The interpolated $F_0$ contour is further smoothed by the following procedure to obtain an approximation that is continuous and differentiable everywhere.

Take the first 200 ms of data and calculate the best third-order polynomial approximation in the sense of least mean squared error, and repeat this procedure at a frame step of 150 ms, with the additional constraint that the new third-order approximation and its derivative should both be continuous with the preceding ones at $t = 150$ ms. The procedure is to be repeated until the end of the utterance, and gives an approximation to the original $F_0$ contour consisting of piecewise third-order polynomial segments that are continuous and differentiable everywhere. We shall henceforth refer to it simply as ‘the smoothed $F_0$ contour’ throughout the rest of the paper.
4. FIRST-ORDER APPROXIMATIONS OF ACCENT COMMAND PARAMETERS

4.1. Points of Inflection of Smoothed $F_0$ Contours

Equation (3) for the accent component shows that it has a point of inflection at $t = 1/\beta$. Since temporal changes of phrase components are generally much more gradual than those of accent components, the inflection points of the $F_0$ contour will roughly correspond to those of the accent components, and hence to the onsets and offsets of the corresponding accent commands except for a delay of $1/\beta$ [s]. Since the smoothed $F_0$ contour as obtained above is differentiable everywhere, its points of inflection can be obtained by taking the second derivative of each third-order polynomial segment and putting it equal to zero. Thus the problem is reduced to a trivial one of solving a linear equation, and the piecewise third-order polynomial will have an inflection point only when the line crosses the abscissa within the interval for which the approximation is valid.

4.2. Estimation of Onset and Offset of Accent Commands

Since the smoothed $F_0$ contour generally has additional points of inflection due to noise and other disturbances, only those corresponding to onset/offset of accent commands have to be selected. This is done first by detecting only larger maxima and smaller minima of the first derivative of the smoothed $F_0$ contour, and then by selecting the point of the largest derivative within the interval over which the derivative is positive, and the point of the smallest derivative within the following interval over which the derivative is negative. On the basis of preliminary experiments, the thresholds for positive and negative values of the slope were set equal to 0.6 and 0.4, respectively. Thus the onset and offset of an accent command are estimated to be $1/\beta$ [s] before the these points.

4.3. Estimation of Accent Command Amplitude

Once the onset and offset of an accent command are estimated, the amplitude is estimated to be

$$A_a = \frac{A_{a1} - A_{a2}}{2},$$

(6)

where $A_{a1}$ and $A_{a2}$ respectively indicate the slope of the corresponding inflection points.

5. FIRST-ORDER APPROXIMATIONS OF PHRASE COMMAND PARAMETERS

5.1. Estimation of Phrase Command Parameters

Once the accent command parameters are determined, they are used to subtract the corresponding accent components from the smoothed $F_0$ contour. The first-order approximations of phrase command parameters are then estimated using this residue.

5.2. Estimation of Timing of Phrase Commands

(1) Utterance-initial phrase command

Equation (2) for the phrase component shows that it attains the maximum value at $t = 1/\alpha$. Thus the timing of the utterance-initial phrase command is estimated to be $1/\alpha$ [s] before the point of maximum of the above-mentioned residue.

(2) Utterance-medial phrase command

The residue signal is divided into consecutive intervals of 200 ms duration each. The point of the minimum within each 200 ms interval is adopted as the first-order approximation to the timing of utterance-medial phrase commands.

5.3. Estimation of Magnitude of Phrase Commands

Once the timing of the $J$th phrase command is estimated to be $T_{0J}$, the first-order approximation of its magnitude $A_{pJ}$ can be obtained to fit model-generated $F_0$ contour to the smoothed $F_0$ contour at the center of the accent command which occurs immediately after $T_{0J}$. If the accent command in question is the $J$th command in the utterance with onset and offset timing respectively equal to $T_{1J}$ and $T_{2J}$, the first-order approximation of the magnitude of the $J$th phrase command can be given by

$$A_{pJ} = \frac{\log F_0(T) - \log F_0' - \log PHR - ACC}{\alpha^2 t_0 \exp(-\alpha T)}$$

(7)

where

$$t_0 = T - T_{0J}, \quad T = (T_{1J} + T_{2J})/2,$$

$$PHR = \sum_{i=1}^{t-1} A_{p1}(T - T_{0i}),$$

$$ACC = \sum_{j=1}^{J} A_{p1}(G_{a1}(T - T_{1j}) - G_{a1}(T - T_{2j})),$$

(8)

and $F_0'$ is the lowest value of $F_0$ in the $F_0$ contour adopted as the first-order approximation to the baseline frequency $F_0$, $PHR$ and $ACC$ respectively represent the contributions at $T$ of all the phrase commands and the accent commands that have contributions to the $F_0$ contour at $T$.

6. OPTIMIZATION OF PARAMETERS

The optimization of command parameters is conducted by the following two-stage Analysis-by-Synthesis.

(1) Optimization of parameters on the smoothed $F_0$ contour

Starting from the set of first-order approximation parameters, the optimum parameter values are obtained by Analysis-by-Synthesis to minimize the mean squared error between the smoothed $F_0$ contour and the model-generated $F_0$ contour. The set of parameters obtained shall be referred to as the second-order approximation.

(2) Optimization of parameters on the original $F_0$ contour after removal of gross errors

Starting from the second-order approximation parameters, parameter values are further optimized by Analysis-by-Synthesis to minimize the mean squared error between the original $F_0$ contour after removal of gross errors and the model-generated $F_0$ contour. The set of parameters obtained shall be referred to as the third-order approximation.
7. EXPERIMENT

7.1. The Speech Material

The speech material for the present study was 15-minute recording of a male annoucer’s speech from a radio program “From My Bookshelf.” It is a reading of a book. The speech signal was digitized at 10 kHz with 16-bit precision, and the fundamental frequency was extracted by a modified autocorrelation analysis of the LPC residual signal. The measured \( F_0 \) contour, which contains occasional gross errors, local disturbances due to microprosody, and voiceless intervals, went through the pre-processing to remove gross errors and microprosody and to interpolate the gaps due to voiceless consonants.

7.2. Results

Figure 2 illustrates an example of the speech waveform and the results of successive stages of processing for the utterance: “Sōieba, bokutachiga kurutoki, kōsokudōra jikode tsūkōdomeni natteitakara makikomaretakana, to watasi.” It shows that the pre-processing can remove gross errors and other disturbances in the measured \( F_0 \) contour (2), and converts it into a piecwise third-order polynomial in \( t \) that is continuous and differentiable everywhere (5), from which appropriate first-order approximations of the commands can be obtained automatically (7), and the approximation can be further improved automatically by the two-stage Analysis-by-Synthesis. Results of the experiments on a total of 85 utterances indicate that out of 823 gross errors 815 were corrected, and out of 563 microprosodic disturbances 509 were removed. Thus the automatic correction was successfully performed on more than 95% of all the gross errors and disturbances.

Table 1 compares the results of automatic extraction of the phrase and accent commands with the results of manual analysis by an experienced researcher. Assuming that the results of manual analysis are 100% correct, the correct rate of extraction was about 90% for the accent commands and about 79% for the phrase commands. Most of the failures of accent command extraction occur either toward the end of utterances where the first derivatives of smoothed \( F_0 \) contours are generally very small or at places where voiced intervals are very short. Therefore the failures are limited only to very small accent commands. On the other hand, false extraction rate was about 4%. These false alarms, however, do not present problems since the command amplitudes due to false alarms are generally reduced to a very low values during the course of successive approximation. Likewise, failures of extraction of phrase commands are also limited to those of smaller magnitude. The false alarm rate for the phrase commands is about 11%.

![Figure 2](image-url) An example of pre-processing and estimation of commands from an \( F_0 \) contour.

Table 1: Results of estimation of commands.

<table>
<thead>
<tr>
<th>Commands</th>
<th>Accent</th>
<th>Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>678</td>
<td>406</td>
</tr>
<tr>
<td>Automatic correct</td>
<td>611 (90%)</td>
<td>322 (79%)</td>
</tr>
<tr>
<td></td>
<td>67</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>40</td>
</tr>
</tbody>
</table>

REFERENCES