A METHOD FOR AUTOMATIC EXTRACTION OF MODEL PARAMETERS FROM FUNDAMENTAL FREQUENCY CONTOURS OF SPEECH

Shuichi Narusawa¹, Nobuaki Minematsu¹, Keikichi Hirose² and Hiroya Fujisaki³

¹ Graduate School of Information Science and Technology, University of Tokyo
² Graduate School of Frontier Sciences, University of Tokyo
³ Prof. Emeritus, University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, JAPAN
{narusawa, mine, hirose, fujisaki}@gavo.t.u-tokyo.ac.jp

ABSTRACT
The process of generating the F₀ contour of speech has been modeled quite accurately in mathematical terms by Fujisaki and his coworkers, but the extraction of parameters of the underlying commands from an observed F₀ contour is an inverse problem that can be solved only by successive approximation. In order to guarantee an efficient and accurate search for the solution, one needs to start with a set of initial values that are close enough to the optimum. This paper presents a method for pre-processing a measured F₀ contour to obtain its approximation consisting of third-order polynomial segments that are continuous and differentiable everywhere. It is shown that the proposed method allows one to obtain first-order approximations to the parameters of accent commands for about 90% of all the accent commands, and of phrase commands for about 84% of all the phrase commands.

1. INTRODUCTION
The contour of the voice fundamental frequency (henceforth F₀ contour) plays an important role in expressing information on the prosody of an utterance, i.e., the information concerning the lexical tone/accents, syntactic structure, and discourse focus. As it is well known, an F₀ contour generally consists of slowly-varying components corresponding to phrases and clauses and rapidly-varying components corresponding to word accents or syllable tones. The exact relationships between these components of an F₀ contour and the underlying linguistic information have been formulated by Fujisaki and his coworkers [1], and expressed as a model for the process of F₀ contour generation. It has been widely shown that the model can generate very close approximations to observed F₀ contours from a relatively small number of parameters representing the linguistic information, and is therefore quite useful in speech synthesis.

While it is quite straightforward to derive an F₀ contour from a set of model parameters, the inverse problem, i.e., the derivation of model parameters from a given F₀ contour, cannot be solved analytically, but can be solved only by the method of successive approximation. Unless one does not start with a good first-order approximation, however, successive approximations tend to be quite inefficient, and may not guarantee convergence to a true solution. The present paper describes a method for finding a good first-order approximation for the set of model parameters from an observed F₀ contour. 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. A MODEL FOR THE GENERATION PROCESS OF F₀ CONTOURS OF JAPANESE UTTERANCES
Figure 1 shows the model for the process of generation of F₀ contours of Japanese utterances. The mechanism that produces changes in log(e) 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(e) 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(e) F₀ to produce the final log(e) 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(e) F₀(t). For the rest of the paper, we shall use the word ‘F₀-contour’ to indicate log(e) F₀(t).

In this model, the F₀ contour is expressed by

\[ \log(e) F₀(t) = \log(e) F₀ + \sum_{i=1}^{l} A_p(t - T_i) \]

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.

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

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

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.

Fig. 1. A functional model for the process of generating F₀ contours.
4. PRE-PROCESSING OF MEASURED F₀ CONTOURS

Pre-processing of an actual F₀ contour consists of four stages: (1) gross error correction, (2) microprosody removal, (3) interpolation, and (4) smoothing.

4.1. 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:

4.1.1. Correction of gross errors in voiced intervals

If the total number of frames with non-zero F₀ values is larger than 2m + 1, and if

\[
\left| \frac{\log_{10} F_0(i)}{\log_{10} F_0(M|j, 2m + 1)} - 1 \right| > S,
\]

where \(F_0(M|j, 2m + 1)\) indicates the median value of F₀ over the \((2m + 1)\) frames centered at the jth 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 \(\log_{10} F_0\) domain. For a frame step of 10ms, \(m = 2\) and \(S = 0.01\) were found to be appropriate on the basis of preliminary experiments.

4.1.2. Correction of gross errors in silent or unvoiced intervals

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

4.2. Removal of Microprosody

The influence of consonantal articulation on \(F_0\) contours, called ‘microprosody’, 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 \(G_0(j)\)) 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 \((i - n_1, i - n_1 + 1, \ldots, i - 1, i)\). 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 \((j, j + 1, \ldots, j + n_2 - 1, j + n_2)\).
4.3. 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

$$
\log_{10} F_0(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3, \quad (5)
$$

whose coefficients \([a_0, a_1, a_2, a_3]\) are obtained by solving the following simultaneous equations:

$$
\begin{align*}
F_0(i) &= a_0 + a_1 i + a_2 i^2 + a_3 i^3, \\
G_0(i) &= a_1 + 2a_2 i + 3a_3 i^2, \\
F_0(j) &= a_0 + a_1 j + a_2 j^2 + a_3 j^3, \\
G_0(j) &= a_1 + 2a_2 j + 3a_3 j^2. \\
\end{align*}
$$

(6)

This interpolation is performed for intervals whose lengths are less than 333 (= 1/α)ms (i.e., for \(j - i + 1 < 34\) at a frame interval of 10ms). Longer intervals are considered as pauses and are not interpolated. The length of the adjacent ‘voiced’ interval at each end is selected to be 50ms (i.e., \(p = 4\) at a frame interval of 10ms). This procedure assigns a continuous contour for the expanded ‘voiceless’ interval, but does not guarantee its continuity with the adjacent $F_0$ data.

4.4. Smoothing

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

(1) For the first 150ms, the coefficients \([a_0, a_1, a_2, a_3]\) of the best third-order polynomial approximation (in the sense of the least mean squared error) are obtained by solving the following set of linear equations:

$$
\begin{align*}
\sum_{i=1}^{N_1} F_0(i) &= N_1 a_0 + \sum_{i=1}^{N_1} a_1 t(i) + \sum_{i=1}^{N_1} a_2 t(i)^2 + \sum_{i=1}^{N_1} a_3 t(i)^3, \\
\sum_{i=1}^{N_1} F_0(i) t(i) &= N_1 a_1 + \sum_{i=1}^{N_1} a_2 t(i) + \sum_{i=1}^{N_1} a_3 t(i)^2, \\
\sum_{i=1}^{N_1} F_0(i) t(i)^2 &= \sum_{i=1}^{N_1} a_2 t(i) + \sum_{i=1}^{N_1} a_3 t(i)^3, \\
\sum_{i=1}^{N_1} F_0(i) t(i)^3 &= \sum_{i=1}^{N_1} a_2 t(i)^2 + \sum_{i=1}^{N_1} a_3 t(i)^4, \\
\sum_{i=1}^{N_1} F_0(i) t(i)^4 &= \sum_{i=1}^{N_1} a_2 t(i)^3 + \sum_{i=1}^{N_1} a_3 t(i)^5, \\
\sum_{i=1}^{N_1} F_0(i) t(i)^5 &= \sum_{i=1}^{N_1} a_2 t(i)^4 + \sum_{i=1}^{N_1} a_3 t(i)^6, \\
\sum_{i=1}^{N_1} F_0(i) t(i)^6 &= \sum_{i=1}^{N_1} a_2 t(i)^5 + \sum_{i=1}^{N_1} a_3 t(i)^7, \\
\end{align*}
$$

(7)

where \(N_1\) indicates the number of frames within the initial interval of 200ms (\(N_1\) is equal to 20 at a frame interval of 10ms, but is smaller if the initial interval is shorter than 200ms).

(2) For the subsequent 150ms, the coefficients \([a_0, a_1, a_2, a_3]\) of the best third-order polynomial approximation are obtained with the additional constraint that the third-order polynomial should be continuous with the immediately preceding one both in amplitude and derivative, and are given by solving the following set of linear equations:

$$
\begin{align*}
F_0(j) &= a_0 + a_1 t(j) + a_2 t(j)^2 + a_3 t(j)^3, \\
G_0(j) &= a_1 + 2a_2 t + 3a_3 t^2, \\
\sum_{j=1}^{N_2} F_0(j) &= N_2 a_0 + \sum_{j=1}^{N_2} a_1 t(j), \\
\sum_{j=1}^{N_2} F_0(j) t(j) &= \sum_{j=1}^{N_2} a_2 t(j)^2 + \sum_{j=1}^{N_2} a_3 t(j)^3, \\
\sum_{j=1}^{N_2} F_0(j) t(j)^2 &= \sum_{j=1}^{N_2} a_2 t(j)^2 + \sum_{j=1}^{N_2} a_3 t(j)^3, \\
\sum_{j=1}^{N_2} F_0(j) t(j)^3 &= \sum_{j=1}^{N_2} a_2 t(j)^3 + \sum_{j=1}^{N_2} a_3 t(j)^4, \\
\sum_{j=1}^{N_2} F_0(j) t(j)^4 &= \sum_{j=1}^{N_2} a_2 t(j)^4 + \sum_{j=1}^{N_2} a_3 t(j)^5, \\
\sum_{j=1}^{N_2} F_0(j) t(j)^5 &= \sum_{j=1}^{N_2} a_2 t(j)^5 + \sum_{j=1}^{N_2} a_3 t(j)^6, \\
\sum_{j=1}^{N_2} F_0(j) t(j)^6 &= \sum_{j=1}^{N_2} a_2 t(j)^6 + \sum_{j=1}^{N_2} a_3 t(j)^7, \\
\end{align*}
$$

(8)

where \(j = 1\) corresponds to \(i = 3N_1/4\) and \(N_2\) indicates the number of frames within the current interval of 200ms. This procedure is repeated within a segment of utterance delimited by two adjacent pauses.

These procedures can give an approximation to the original $F_0$ contour consisting of piecewise third-order polynomial segments that are continuous and differentiable everywhere except for pause intervals.

5. DERIVATION OF THE FIRST-ORDER APPROXIMATIONS OF COMMAND PARAMETERS

5.1. Extraction of accent command parameters

Since the final outcome of smoothing is continuous and differentiable everywhere, it is quite straightforward to compute its derivative and find its maxima and minima analytically. If we neglect the effects of phrase components, the maxima and the minima of the first derivative of the contour should correspond to the onset and the offsets of accent commands with a constant delay of 1/β. The actual procedure is to detect the largest maximum and the smallest minimum for each interval where the sign of the derivative remains the same. A pair of maximum and minimum thus corresponds to the onset and the offset of an accent command. The mean absolute amplitudes of such a pair of a maximum and a minimum can be adopted as the first-order approximation to the amplitude of the corresponding accent command. If the initial part of the first-order derivative is negative and gives a minimum, then it can be regarded as the offset of the utterance-initial accent command, in which case one has to assume the existence of onset of the accent command before the start of an utterance.

5.2. Extraction of phrase command parameters

After approximately removing the accent components from the smoothed $F_0$ contour, one obtains a residual contour which consists mainly of phrase components. Since the influence of each phrase command is essentially a semi-infinite function of time starting from the onset of the command, each phrase command is detected successively by a left-to-right procedure from the residual contour.
6. EXPERIMENT

6.1. The Speech Material

The speech material for the present study was a 15-minute recording of a male announcer’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.

6.2. Results

Figure 2 illustrates an example of the speech waveform and the results of successive processing for the Japanese utterance: "Ikutsukano otodake sokokara shakuyooshite, raibuno fun’ikio sokonawazuni henshuusuru kotoga dekiru." (The recording of the concert 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 present paper has described our on-going work toward fully automatic extraction of $F_0$ contour parameters. We have shown that the inverse problem of deriving the commands from a measured $F_0$ contour can be converted into an analytically solvable problem, by approximating a measured $F_0$ contour by a smooth curve consisting of third-order polynomial segments that are continuous and differentiable everywhere except for pause intervals.

Experimental results have shown the validity of the approach, but have also indicated the need for further work.

7. CONCLUSIONS

The present paper has described our on-going work toward fully automatic extraction of $F_0$ contour parameters. We have shown that the inverse problem of deriving the commands from a measured $F_0$ contour can be converted into an analytically solvable problem, by approximating a measured $F_0$ contour by a smooth curve consisting of third-order polynomial segments that are continuous and differentiable everywhere except for pause intervals.

Experimental results have shown the validity of the approach, but have also indicated the need for further work.

8. REFERENCES