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

Shuichi Narusawa[1], *Nobuaki Minematsu*[1], *Keikichi Hirose*[2] *and Hiroya Fujisaki*[3]

[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 F_0 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 F_0 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 F_0 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 F_0 contour) plays an important role in expressing information on the prosody of an utterance, *i.e.*, the information concerning the lexical tone/accent, syntactic structure, and discourse focus. As it is well known, an F_0 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 F_0 contour and the underlying linguistic information have been formulated by Fujisaki and his coworkers [1], and expressed as a model for the process of F_0 contour generation. It has been widely shown that the model can generate very close approximations to observed F_0 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 F_0 contour from a set of model parameters, the inverse problem, *i.e.*, the derivation of model parameters from a given F_0 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 F_0 contour. Although the method is applicable, with certain language-specific modifications, to F_0 contours of various languages, the present paper deals with F_0 contours of Japanese.

2. A MODEL FOR THE GENERATION PROCESS OF F_0 CONTOURS OF JAPANESE UTTERANCES

Figure 1 shows the model for the process of generation of F_0 contours of Japanese utterances. The mechanism that produces changes in $\log_e F_0(t)$ from the phrase commands is named 'phrase control mechanism' and its outputs are named 'phrase components.' Likewise, the mechanism that produces changes in $\log_e F_0(t)$ 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 $\log_e F_b$ to produce the final $\log_e F_0(t)$. 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 $\log_e F_0(t)$. For the rest of the paper, we shall use the word 'F_0-contour' to indicate $\log_e F_0(t)$.

In this model, the F_0 contour is expressed by

$$\begin{aligned}\log_e F_0(t) &= \log_e F_b + \sum_{i=1}^{I} Ap_i Gp(t - T_{0i}) \\ &+ \sum_{j=1}^{J} Aa_j \{Ga(t - T_{1j}) - Ga(t - T_{2j})\},\end{aligned} \quad (1)$$

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

$$Ga(t) = \begin{cases} \min[1 - (1 + \beta t)\exp(-\beta t), \gamma], & \text{for } t \geq 0, \\ 0, & \text{for } t < 0, \end{cases} \quad (3)$$

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.

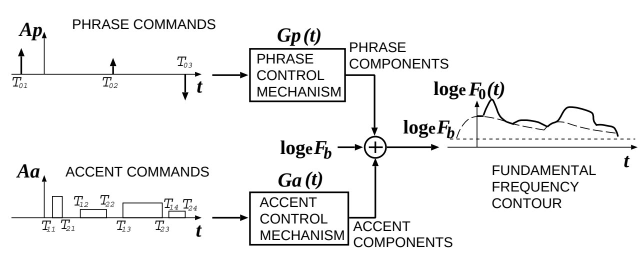

Fig. 1. A functional model for the process of generating F_0 contours.

The symbols in these equations indicate

F_b : baseline value of fundamental frequency,
I : number of phrase commands,
J : number of accent commands,
Ap_i : magnitude of the ith phrase command,
Aa_j : amplitude of the jth accent command,
T_{0i} : timing of the ith phrase command,
T_{1j} : onset of the jth accent command,
T_{2j} : offset of the jth accent command,
α : natural angular frequency of the phrase control mechanism,
β : natural angular frequency of the accent control mechanism,
γ : relative ceiling level of accent components.

Parameters α and β 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 $\beta = 20.0$ can be used as default values. Parameter γ 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. NECESSITY OF PIECEWISE SMOOTHING OF MEASURED F_0 CONTOURS

Since it is possible to use default values for α, β, and γ, the inverse problem is reduced to finding good first-order approximations to the number, temporal locations (henceforth 'timing'), and magnitudes/amplitudes of the phrase/accent commands. The baseline frequency F_b can be obtained automatically by minimizing the mean squared error between the measured F_0 contour and the model-generated F_0 contour.

Several attempts have already been reported toward automatic extraction of F_0 contour parameters using the above-mentioned model [2] - [7]. These approaches, however, have made only limited success. The major reason is that the actual F_0 contour contains a number of factors that are not covered by the model, such as (1) gross errors in the measurement of F_0, (2) local deviations due to microprosody caused by certain consonants, (3) discontinuities due to the presence of voiceless consonants and utterance-medial pauses, and (4) lack of smoothness (*i.e.*, non-differentiability). For the reliable estimation of the first-order approximations of model parameters, therefore, it is necessary to cope with these factors.

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]. If the measured F_0 contour is approximated by smooth curve consisting of third-order polynomial segments, 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.

Since the current approach [8] is based upon the combination of approaches adopted in [2] and [3], it is necessary to convert the measured F_0 contour of an utterance into a continuous curve consisting of third-order polynomial segments, in such a way that the resulting curve is differentiable everywhere. Once this is done, its points of inflection (*i.e.*, points where the first derivative is at a maximum or a minimum) should indicate points that are closely related to the onset and offset of the accent commands with an approximately constant delay.

4. PRE-PROCESSING OF MEASURED F_0 CONTOURS

Pre-processing of an actual F_0 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 F_0 values is larger than $2m+1$, and if

$$\left| \frac{\log_e F_0(i)}{\log_e F_0(M|i, 2m+1)} - 1 \right| > S, \tag{4}$$

where $F_0(M|i, 2m+1)$ indicates the median value of F_0 over the $(2m+1)$ frames centered at the ith 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 $\log_e F_0$ 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 $G_0(\cdot)$) 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 $(i-n_1, i-n_1+1, \cdots, i-1, i)$. 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 $(j, j+1, \cdots, j+n_2-1, j+n_2)$.

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

$$\log_e F_0(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3, \qquad (5)$$

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

$$\begin{cases} F_0(i) = a_0 + a_1 i + a_2 i^2 + a_3 i^3, \\ G_0(i) = a_1 + 2a_2 i + 3a_3 i^2, \\ F_0(j) = a_0 + a_1 j + a_2 j^2 + a_3 j^3, \\ G_0(j) = a_1 + 2a_2 j + 3a_3 j^2. \end{cases} \qquad (6)$$

This interpolation is performed for intervals whose lengths are less than 333$(=1/\alpha)$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:

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

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:

$$\begin{cases} F_0(j) = a_0 + a_1 t(j) + a_2 t(j)^2 + a_3 t(j)^3, \\ G_0(j) = a_1 + 2a_2 t(j) + 3a_3 t(j)^2, \\ \sum_{j=1}^{N_2} F_0(j) = N_2 a_0 + \sum_{j=1}^{N_2} a_1 t(j) \\ \qquad\qquad + \sum_{j=1}^{N_2} a_2 t(j)^2 + \sum_{j=1}^{N_2} a_3 t(j)^3, \\ \sum_{j=1}^{N_2} F_0(j)t(j) = \sum_{j=1}^{N_2} a_0 t(j) + \sum_{j=1}^{N_2} a_1 t(j)^2 \\ \qquad\qquad + \sum_{j=1}^{N_2} a_2 t(j)^3 + \sum_{j=1}^{N_2} a_3 t(j)^4, \end{cases} \qquad (8)$$

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 $1/\beta$. 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 can be edited without destroying the impression of a live performance, by borrowing only a few sounds from the rehearsal.)

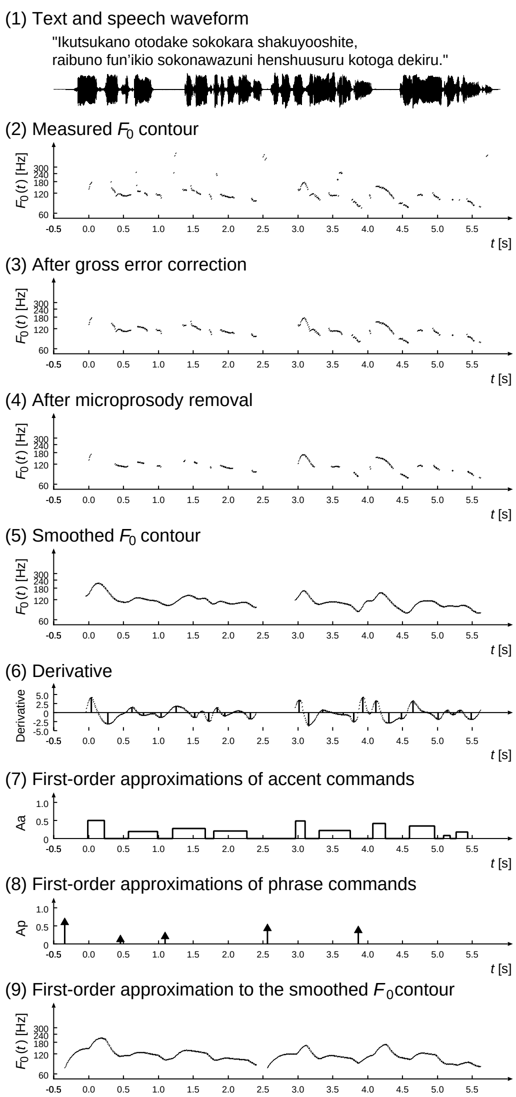

Fig. 2. An example of pre-processing and estimation of commands from an F_0 contour.

The figure shows, from top to bottom, (1) the speech waveform, (2) the measured F_0 contour, (3) the contour after gross error correction, (4) the contour after microprosody removal, (5) the contour after smoothing, (6) the derivative of the smoothed contour, (7) first-order approximations of accent commands, (8) first-order approximations of phrase commands, and (9) the first-order approximation to the smoothed F_0 contour. These panels clearly indicate that the method can give very good first-order approximations for the accent commands, and is therefore quite useful as the initial stage of fully automatic estimation of F_0 contour parameters.

Results of 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 microprosodic disturbances.

The results of automatic extraction of the phrase and accent components were then compared with those of manual analysis by an experienced researcher. Assuming that the results of manual analysis are 100% correct, the rate of correct extraction was about 90% for the accent commands, and was about 84% for the phrase commands.

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

[1] H. Fujisaki and S. Nagashima, "A model for synthesis of pitch contours of connected speech," *Annual Report, Engg. Res. Inst., University of Tokyo*, vol. 28, pp. 53–60 (1969).

[2] H. Fujisaki, K. Hirose and S. Seto, "A study on automatic extraction of characteristic parameters of fundamental frequency contours," *Proc. Fall Meeting, Acoust. Soc. Jpn.*, vol. 1, pp. 255–256 (1990).

[3] E. Geoffrois, "A pitch contour analysis guided by prosodic event detection," *Proc. Eurospeech '93, Berlin*, vol. 2, pp. 793–796 (1993).

[4] H. Fujisaki, S. Ohno and Y. Wada, "A method for automatic estimation of parameters of a model for the generation process of fundamental frequency contours of speech," *Proc. Spring Meeting, Acoust. Soc. Jpn.*, vol. 1, pp. 17–18 (1995).

[5] H. Fujisaki, S. Ohno and O. Tomita, "Automatic parameter extraction of fundamental frequency contours of speech based on a generative model," *Proc. ICSP'96, Beijing*, vol. 1, pp. 729–732 (1996).

[6] J. Mersdorf, A. Rinscheid, M. Brűggen and K. W. Schmidt, "Coding of large intonational units by linear prediction," *Proc. ESCA Workshop on Intonation, Athens*, pp. 235–238 (1997).

[7] H. Mixdorff, "A novel approach to the fully automatic extraction of fujisaki model parameters," *Proc. ICASSP 2000. Istanbul*, vol. 3, pp. 1281–1284 (2000).

[8] S. Narusawa, N. Minematsu, K. Hirose and H. Fujisaki, "Automatic extraction of parameters from fundamental frequency contours of speech," *Proc. ICSP 2001, Seoul*, vol. 2, pp. 833–838 (2001).