AUTOMATIC EXTRACTION OF MODEL PARAMETERS FROM FUNDAMENTAL FREQUENCY CONTOURS OF ENGLISH UTTERANCES

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 generation process model of the fundamental frequency contours (F_0 contours) of speech is known to be capable of generating F_0 contours quite close to observed ones. The extraction of model parameters from an observed contour, however, requires an iterative process starting from a set of initial parameter values. In order to guarantee a rapid convergence to an optimum solution, the values should be appropriate ones. We already have developed a method of automatically extracting these from given F_0 contours, and applied it to Japanese sentences with good results. The method is based on approximating an observed contour by a continuous curve differentiable everywhere. In the present paper, it was applied to English utterances. Experiments were conducted for 4 native speakers' utterances with 14.5% and 17.5% of average miss and false alarm rates for the accent commands, and 35.7% and 15.5% for 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 generation process model of F_0 contours (henceforth, the model) proposed by Fujisaki and his coworkers [1] can formulate the relationship between these components and the underlying linguistic information quite well [2]. 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 parameter values from a given F_0 contour, cannot be solved analytically, but can be solved only by the method of successive approximation. This process should start from a good first-order approximation. Otherwise, it tends to be quite inefficient, and may not guarantee convergence to a true solution. The first-order approximation is usually given manually, making it difficult to obtain a large corpus with the model parameters. Although several methods have already been developed to automatically estimate the first-order approximation, their performance was not high enough as mentioned in section 3. From this point of view, we have developed a method based on approximating an observed F_0 contour by a set of piecewise third-order polynomials, which is continuous and differentiable everywhere. The method was already applied to Japanese read speech with good results [3], but its performance was not clear for other cases. In the present paper, the method is checked for English, whose prosodic features are rather different from those of Japanese.

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

Figure 1 shows the model outlines. 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

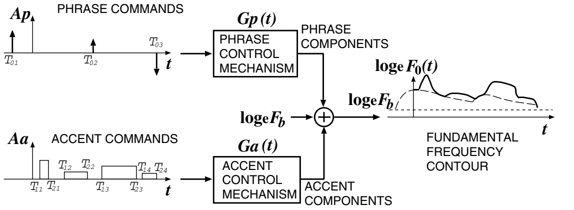

Fig. 1. A functional model for the process of generating F_0 contours.

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 $\log_e F_0(t)$. 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

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

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

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

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 [2].

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 was shown that $\alpha = 3.0[1/s]$ and $\beta = 20.0[1/s]$ 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 [4, 5]. 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 a 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 [6] is based upon the combination of approaches adopted in [4], 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. OUTLINE OF THE CURRENT APPROACH

The proposed procedure consists of the following four stages:

(1) Pre-processing of a measured F_0 contour

After correcting gross errors and removing microprosodic disturbances, the measured F_0 contour is approximated by a piecewise polynomial function of time that is continuous and differentiable everywhere.

(2) Derivation of the first-order approximations 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 onsets 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 of the absolute amplitudes of maximum and minimum of a pair 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.

(3) Derivation of the first-order approximations of phrase command parameters

After removing the accent components estimated in stage (2) 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.

(4) Derivation of the optimum parameters by Analysis-by-Synthesis

Parameters of accent and phrase commands are refined by successive approximation to obtain a set of parameters that are optimum by a pre-determined error criterion, say the least mean squared error in the domain of $\log_e F_0(t)$.

5. EXPERIMENT

5.1. Speech Material

The speech material for the present study was 30 English sentences uttered by 4 native speakers (2 male and 2 female speakers). The speakers shall be denoted as male 1, male 2, female 1 and female 2 from now on. The speech signal was digitized at 10 kHz with 16-bit precision, and F_0 was extracted by a modified autocorrelation analysis of the LPC residual signal [7].

5.2. Results

Figure 2 shows the process of estimating the first-order approximation for English sentence "He grasped the rope with his free hand and swung his legs round to the side." uttered by male 1. From the top to the bottom, panels of the figure respectively show: (1) the speech waveform, (2) the measured F_0 contour, (3) the contour after gross error correction and microprosody removal, (4) the contour after smoothing, (5) the derivative of the smoothed contour, (6) first-order approximations of accent commands, (7) the residual contour, (8) integration of the residual contour up to the point at which it reaches the threshold θ_1, (9) the first-order approximations of phrase commands, (10) the first-order approximation to the smoothed F_0 contour, (11) the final approximation to the original F_0 contour, and (12) the optimum set of commands. The last two panels are the results of successive approximation, usually called analysis-by-synthesis. As indicated in panel (11), a contour quite

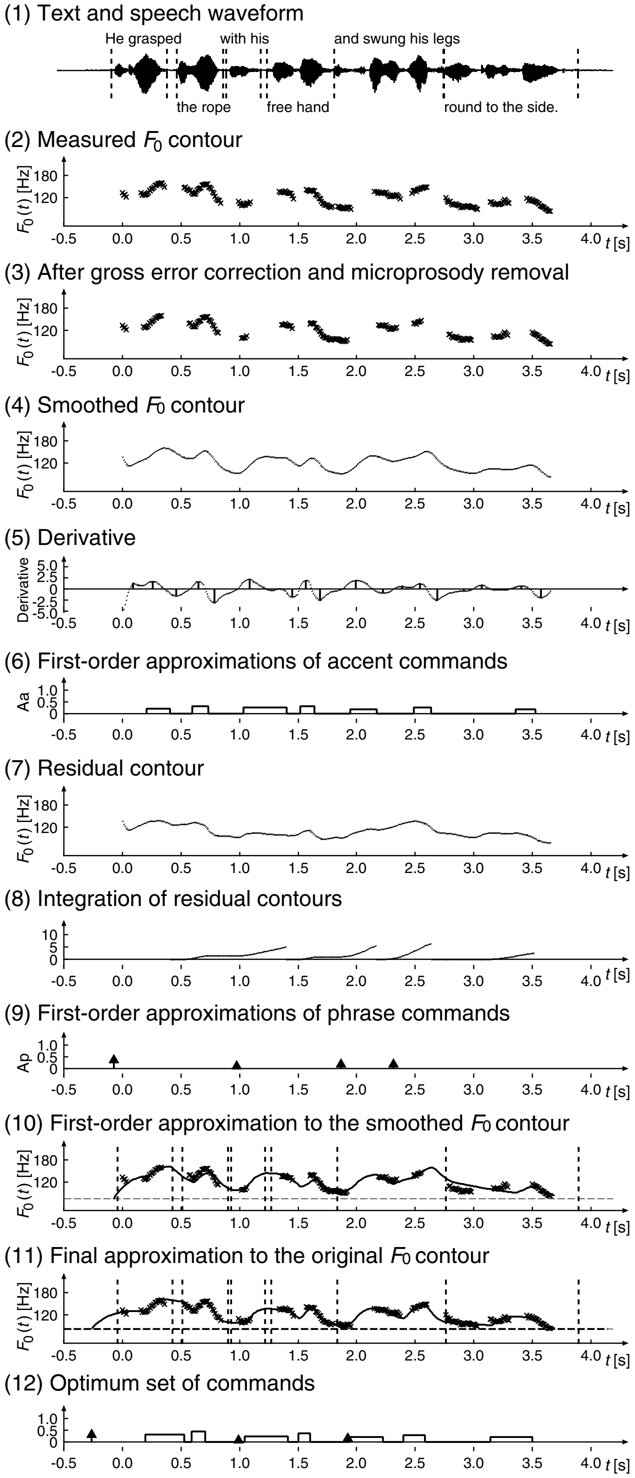

Fig. 2. An example of pre-processing and estimation of commands from an F_0 contour.

similar to that of the corrected contour (in panel (3)) of the observed one is obtained after the successive approximation. The mean square error between these two contours is 0.0004, which is almost the same with 0.0034 between the corrected contour and the contour obtained by the successive approximation starting from manually assigned first-order approximation.

Then, in order to totally evaluate the method, numbers of accent and phrase commands correctly detected by the method are summed up for 30 sentences, and the results are summarized in Table 1 for each speaker. The table also includes numbers of deletion and insertion errors and rates of miss and false alarm, which are respectively defined as:

$$miss = \left(1 - \frac{correct}{correct + deletion}\right) \times 100, \tag{4}$$

$$false\,alarm = \left(1 - \frac{correct}{correct + insertion}\right) \times 100. \tag{5}$$

Commands detected manually by an expert deeply involved in the prosody research are used as the correct answer for the command detection, which is necessary as the reference for the correct, deletion and insertion judgments. Therefore, $correct + deletion$ coincides with the command number manually detected, and $correct + insertion$ coincides with that detected by the method. The reason why only the command detection is counted is that, through the successive approximation process, magnitudes/amplitudes of the commands are adjusted to the correct values and small errors in the command timings are eliminated.

When the miss and false alarm rates are averaged over 4 speakers, they are respectively 14.5% and 17.5% for accent command detection, and are 35.7% and 15.5% for phrase command detection. Although a rather large number of phrase commands fail to be detected, in many cases, they are corresponding to minor syntactic boundaries and little influence to the synthesized speech quality [3]. In the current method, linguistic information is not counted when detecting the model commands. By taking it into account, deletion errors will be reduced.

Table 1. Results of estimation of commands.

	male 1	male 2	female 1	female 2
Accent				
Manual	176	167	163	173
Automatic	179	169	162	195
correct	151	145	132	153
deletion	25	22	31	20
insertion	28	24	30	42
miss [%]	14.2	13.2	19.0	11.6
false alarm [%]	15.6	14.2	18.5	21.5
Phrase				
Manual	81	80	88	88
Automatic	68	58	63	68
correct	54	49	56	58
deletion	27	31	32	30
insertion	14	9	7	10
miss [%]	33.3	38.8	36.4	34.1
false alarm [%]	20.6	15.5	11.1	14.7

Recently, a method of automatic extraction of F_0 model parameters was developed as a freeware [5]. This method is based on assuming a sentence F_0 contour as a waveform and applying a high-pass filtering for the better estimation of accent commands. When F_0 model command detection was done using this method for the same data, the miss and false alarm rates were respectively 36.9% and 11.9% for accent commands, and 42.0% and 38.2% for phrase commands. These values are rather high as compared to those by our method.

6. CONCLUSIONS

The method of automatically extracting the model parameters from observed F_0 contours, formerly developed for Japanese utterances, was successfully applied to read English. The method is similar to that formerly developed by authors [4] in that it looks derivative of the F_0 contour, but is unique in the process of obtaining a smooth curve from the observed quasi-continuous F_0 contours; a piecewise 3rd order polynomial curve which is differentiable everywhere. The method extracts the model parameters only from the observed F_0 contours and does not utilize linguistic information of the sentences. When building up speech corpuses, this information is usually available. We are now developing a way to utilize the information especially for the better estimation of phrase command parameters.

7. 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 and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Jpn (E)*, vol. 5, no. 4, pp. 233–242 (1984).

[3] H. Fujisaki and S. Narusawa, "Automatic extraction of model parameters from fundamental frequency contours of speech," *Proc. 2001 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, pp. 133–138 (2002).

[4] 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).

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

[6] 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).

[7] K. Hirose, H. Fujisaki and S. Seto, "A scheme for pitch extraction of speech using autocorrelation function with frame length proportional to the time lag," *Proc. ICASSP '92*, vol. 1, pp. 149–152 (1992).