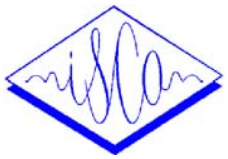

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Evaluation of an Improved Method for Automatic Extraction of Model Parameters from Fundamental Frequency Contours of Speech

Shuichi Narusawa[1], *Nobuaki Minematsu*[1], *Keikichi Hirose*[2] *& 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}@gavo.t.u-tokyo.ac.jp fujisaki@alum.mit.edu

10.21437/SpeechProsody.2004-101

Abstract

The authors have already presented a method for automatic extraction of accent and phrase commands of a model from a given F_0 contour of speech. This paper describes improvements introduced to cope with difficulties encountered by the previous method, especially in connection with the extraction of accent commands, and reports the results of experiments conducted for the evaluation of the current method using two sets of speech materials differing in sentence length and syntactic complexity. It is shown that the method works quite well for the majority of utterances tested. Analysis of performance in terms of misses and false insertions of commands indicates that the performance is slightly better for shorter utterances, and that most of the errors are related to commands of smaller magnitude/amplitude, suggesting that their effects on the perception of naturalness of prosody are of minor importance.

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 or clauses and rapidly-varying components corresponding to word accents or syllable tones. Fujisaki and his co-workers have shown that the process of F_0 contour generation can be accurately represented by a mathematical model. The model generates an F_0 contour in response to a set of commands [1, 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 [3].

While the generation of an F_0 contour from a set of input commands is quite straightforward if we use the model, the derivation of the underlying commands from a given F_0 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. We have developed a method that converts the problem of finding a good first-order approximation into an analytically solvable problem [4]. The method has already been applied successfully to F_0 contour analysis of read speech of Japanese [5]. In this paper, we introduce further improvements of the method, and show the results of experiments conducted for the evaluation of its performance on two different sets of material of spoken Japanese.

2. A model for the generation process of F_0 contours of Japanese utterances

Figure 1 shows the outlines of the model. 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 $\log_e F_0(t)$. For the rest of the paper, we shall use the word 'F_0-contour' to indicate $\log_e F_0(t)$.

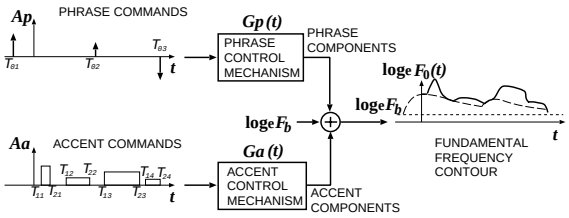

Figure 1: *A functional model for the process of generating F_0 contours.*

In this model, the ith phrase command is characterized by its magnitude Ap_i and its temporal location T_{0i}, while the jth accent command is characterized by its amplitude Aa_j, its onset T_{1j} and offset T_{2j}. The control mechanisms are characterized by the parameters α, β and γ, which can be regarded as constants with their respective default values of 3.0 [1/s], 20.0 [1/s], and 0.9.

3. Basic principles [4]

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 [6, 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 a smooth (*i.e.*, continuous and differentiable everywhere) 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.

4. Outline of the current approach

4.1. Pre-processing of measured F_0 contours

4.1.1. 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 non-zero frequency value to a frame corresponding to a voiceless interval, and (2) assignment of a false value (including zero) to a frame corresponding to a voiced interval. In the present study, we adopt the following algorithm utilizing the correlation between F_0 and short-term power in order to cope with these two types of gross errors.

Gross errors are detected on the basis of correlation between F_0 and the short-term power. The short-term power of speech is calculated for each 20-ms window, and is normalized by its maximum within an utterance. Frames with a normalized short-term power below -30dB are considered to be voiceless, and the values of F_0 detected within these frames are removed as gross errors. For frames with relative short-term pauses above -30dB, if the value of $\log_e F_0$ at a frame is deviated from those of adjacent frames beyond a certain threshold but the deviations of the short-term power are below a threshold, then it is regarded as a candidate for a gross error, and the F_0 value is replaced by interpolation or extrapolation using values of $\log_e F_0$ of adjacent frames. The new algorithm is especially effective in removing bursts of gross errors.

4.1.2. 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 the consonants and their adjacent vowels. The procedure for removing microprosody is described elsewhere [5].

4.1.3. Interpolation of voiceless consonants and short pauses

In order to realize smoother interpolation than that was possible by our previous method, we introduce the following algorithm utilizing a larger number of data points for interpolation.

A voiceless or silent interval $[t_{i+1}, t_{j-1}]$ is interpolated by the following third order polynomial, using values of $\log_e F_0$ at $(N+1)$ frames on each side of the interval in question.

$$\log_e F_0(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3, \tag{1}$$

where the coefficients a_0, a_1, a_2, and a_3 are determined by minimizing the mean squared error over the two intervals of $(N+1)$ frames each adjacent to the interval $[t_{i+1}, t_{j-1}]$, and can be obtained by solving the following set of linear equations:

$$\begin{cases} \sum_{n=i-N}^{i} \log_e F_0(t_n) = N a_0 + \sum_{n=i-N}^{i} a_1 t_n + \sum_{n=i-N}^{i} a_2 t_n^2 + \sum_{n=i-N}^{i} a_3 t_n^3, \\ \sum_{n=i-N}^{i} \log_e F_0(t_n) t_n = \sum_{n=i-N}^{i} a_0 t_n + \sum_{n=i-N}^{i} a_1 t_n^2 + \sum_{n=i-N}^{i} a_2 t_n^3 + \sum_{n=i-N}^{i} a_3 t_n^4, \\ \sum_{m=j}^{j+N} \log_e F_0(t_m) = N a_0 + \sum_{m=j}^{j+N} a_1 t_m + \sum_{m=j}^{j+N} a_2 t_m^2 + \sum_{m=j}^{j+N} a_3 t_m^3, \\ \sum_{m=j}^{j+N} \log_e F_0(t_m) t_m = \sum_{m=j}^{j+N} a_0 t_m + \sum_{m=j}^{j+N} a_1 t_m^2 + \sum_{m=j}^{j+N} a_2 t_m^3 + \sum_{m=j}^{j+N} a_3 t_m^4. \end{cases} \tag{2}$$

The optimum value of N was found experimentally to be equal to 10.

If the maximum absolute value of the gradient of the interpolating curve exceeds a certain threshold, Equation (1) is not used, but the voiceless interval is linearly interpolated.

4.1.4. Smoothing

The interpolated F_0 contour is further smoothed to obtain an approximation in terms of piecewise third-order polynomial segments that are continuous and differentiable everywhere [5].

4.2. Derivation of the first-order approximations of command parameters

4.2.1. Accent commands

Since the final outcome of the 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, and if all the accent commands are well separated, the maxima and the minima of the first derivative of the smoothed contour should generally appear as pairs, and should respectively correspond to the onsets and the offsets of accent commands with a constant delay of $1/\beta$.

There are cases, however, that this does not occur. (1) The maxima and minima do not necessarily alternate with each other, either due to noise or due to an increase or decrease of accent command amplitude without apparent resetting, which can occur in prosodic words consisting of a word of the so-called 'unaccented' type followed by a word of the 'accented' type. (2) The onset or the offset of an accent command cannot be detected from the smoothed contour when the onset of an accent command occurs before the initial mora of an utterance,

or when the utterance-final word belongs to the so-called 'unaccented type' which do not have the offset of the accent command within the utterance. In order to cope with these cases, we introduce the following algorithm.

Maxima or minima due to noise are removed first by introducing a threshold for the absolute amplitude. After this threshold operation, up to two largest maxima or two smallest minima are detected for each interval where the sign of the derivative remains the same. If the number of maxima or minima within such an interval is one, a maximum and an immediately following minimum is regarded to correspond to the onset and offset of an accent command, and the mean absolute amplitude of the pair is adopted as the first-order approximation to the amplitude of the accent command. If the number of maxima (or minima) within such an interval is two, and if their separation is smaller than 50% of the mean length of a mora, then the two are replaced by one maximum (or minimum) whose absolute amplitude is the sum of the two. If, on the other hand, their separation is larger, they are regarded as belonging to two consecutive commands. The first maximum is used as the first-order approximation to the amplitude of the first command. Its offset is assumed to coincide with the onset of the second accent command, and the sum of the two original maxima is used as the corresponding maximum (of the first derivative) of the second accent command. The first-order approximation to the amplitude of the second accent command is then given by the mean absolute amplitude of the new maximum and the immediately following minimum.

If the first derivative is negative and gives a minimum at the utterance-initial interval, then it is 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. Likewise, if the first derivative is positive at the utterance-final interval, it is regarded as the onset of the utterance-final accent command.

4.2.2. phrase commands

After removing the accent components estimated in 4.2.1 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 [5].

4.3. Derivation of the optimum set of 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, namely the least mean squared error in the domain of $\log_e F_0(t)$ [5].

5. Experiment

5.1. Speech material

The speech material for the present study consists of the following two sets of utterances.

A: Readings of 100 short sentences by a male speaker, randomly extracted from the corpus of continuous speech (consisting of 513 isolated sentences) provided by the Advanced Telecommunications Research International (ATR).

B: Reading of a chapter of a book by another male speaker (consisting of 85 sentences that are longer on the average than those of Speech Material A) recorded from a radio program "From My Bookshelf" by the Japan Broadcasting Corporation (NHK).

These speech signals were digitized at 10 kHz with 16-bit precision, and the fundamental frequency was extracted by a modified autocorrelation analysis of the LPC residual signal [8].

5.2. Illustration of successive processing stages and their results

Figure 2 illustrates an example of the speech waveform for Speech Material B and the results of successive processing for

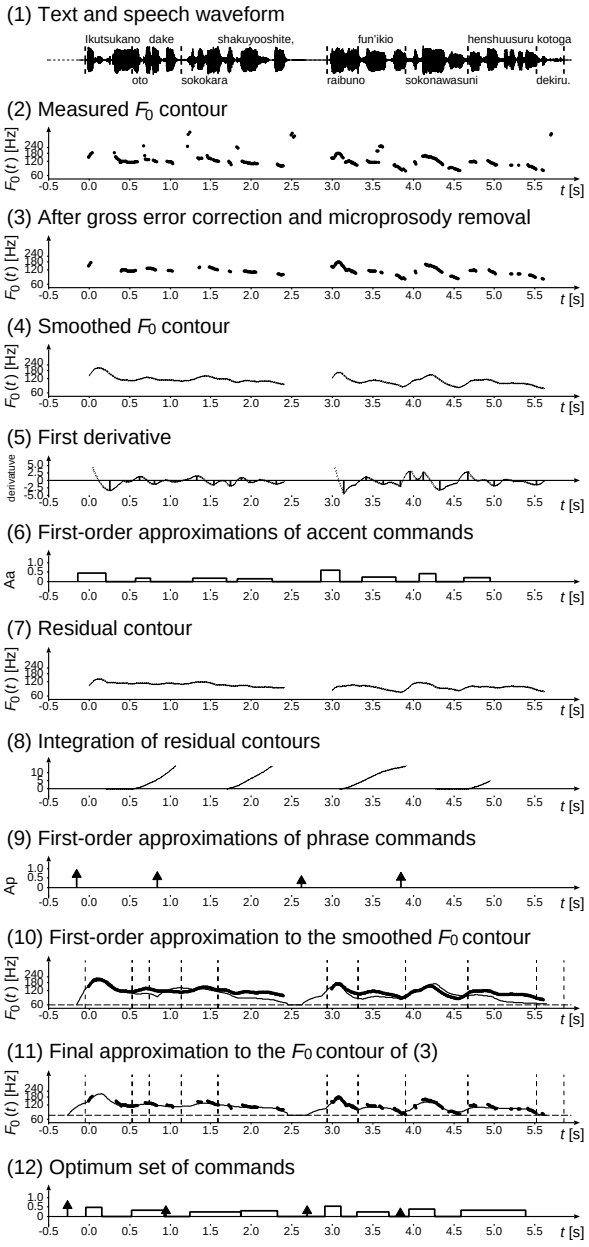

Figure 2: *An example of pre-processing and estimation of commands from an F_0 contour.*

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.). From the top to the bottom, the panels of the figure indicate: (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 to the accent commands, (7) the residual contour, (8) integration of the residual contour up to the point at which it reaches a threshold, (9) first-order approximations to the phrase commands, (10) the first-order approximation to the smoothed F_0 contour, (11) the final approximation to the F_0 contour after gross error correction and microprosody removal, and (12) the optimum set of commands. The last two panels indicate the results of successive approximation. The final approximation shown in (11) is quite close to the observed and corrected contour shown in (3). The root mean squared error is 0.038 (3.8%), which is quite close to 0.030 (3.0%) obtained by successive approximation starting from manually assigned first-order approximations to the commands.

5.3. Evaluation in terms of misses and false insertions

While the present method can give parameter values that are capable of generating very close approximations to the original F_0 contours in the majority of utterances, there are cases in which certain commands are missed or incorrectly inserted. Table 1 shows the total numbers of these errors for the two sets of speech material. Also shown in the table are the percentages of these errors against the total number of commands manually extracted by experts in an interactive analysis of F_0 contours.

Table 1: *Performance evaluation.*

No. of Phrase Commands	Material A	Material B
Manually extracted	453	393
Automatically extracted	453	365
missed	29 (6.4%)	59 (15.0%)
falsely-inserted	20 (4.4%)	31 (7.9%)
No. of Accent Commands	**Material A**	**Material B**
Manually extracted	662	604
Automatically extracted	618	565
missed	99 (15.0%)	130 (21.5%)
falsely-inserted	55 (8.3%)	91 (15.1%)

* The percentages are against the total number of manually extracted commands, which are considered to be correct.

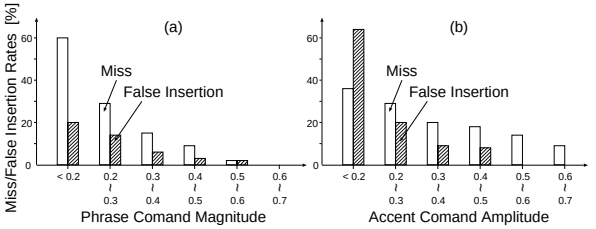

Figure 3: *Distributions of magnitude/amplitude of missed and falsely-inserted commands for Speech Material B.*
(a) Phrase commands (b) Accent commands.

Table 1 shows that all the error rates are higher for Material B than for Material A, reflecting the fact that sentences in Material B are on the average longer and more complex than those of Material A. It also shows that miss rates are higher than false insertion rates. This is due to the fact that pre-processing of observed F_0 contours tends to disregard minor inflections.

Figure 3 shows the distributions of magnitude/amplitude of missed and falsely-inserted commands in the automatic extraction of phrase/accent commands from Speech Material B. It is clear that the errors decrease rapidly with the increase in the magnitude/amplitude of commands.

6. Conclusions

Based on the principle already proposed by the authors for the automatic extraction of accent and phrase commands from observed F_0 contours of speech [4, 5], an improved method has been presented in order to cover cases which presented difficulties to the method previously reported. The new method has been tested on two sets of speech materials. The results have indicated that the new method works quite well for the majority of utterances. Analysis of errors in detecting phrase and accent commands in terms of misses and false insertions has indicated that miss rates are consistently higher than false insertion rates. The results have also shown that the errors are mainly related to commands of relatively small magnitude/amplitude so that their effects on the naturalness of prosody are considered to be of minor importance. Work is under way on the perceptual evaluation of prosody of speech re-synthesized using the extracted parameters, and on the methods and tools for constructing a database of prosody of spoken Japanese containing information on the phrase and accent commands.

7. Acknowledgment

This work was supported by the Grant-in-Aid for Scientific research on Priority Areas (B) No.12132102 "Analysis, Formulation, and Modeling of Prosody" (Principal investigator: Hiroya Fujisaki) from the Ministry of Education, Culture, Science and Technology of Japan.

8. References

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

[2] Fujisaki, H.; Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *J. Acoust. Soc. Jpn (E)*, 5(4), 233–242.

[3] Fujisaki, H, 1999. The fundamental frequency contour — Its modeling, the underlying mechanisms, and application to multilingual speech synthesis —. *Proc. ICSP '99*, 1, 19–26.

[4] Narusawa, S.; Fujisaki, H.; Ohno, S., 2000. A method for automatic extraction of parameters of the fundamental frequency contour. *Proc. ICSLP 2000*, 1, 649–652.

[5] Narusawa, S.; Minematsu, N.; Hirose, K.; Fujisaki, H., 2002. A method for automatic extraction of model parameters from fundamental frequency contours of speech. *Proc. ICASSP 2002*, 1, 13–17.

[6] Fujisaki, H.; Hirose, K.; Seto, S., 1990. A study on automatic extraction of characteristic parameters of fundamental frequency contours. *Proc. Fall Meeting, Acoust. Soc. Jpn.*, 1, 255–256. (In Japanese)

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

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