Corpus-based Extraction of F_0 Contour Generation Process Model Parameters

Keikichi Hirose, Yusuke Furuyama**, and Nobuaki Minematsu****

*Graduate School of Information Science and Technology, **Graduate School of Engineering,
***Graduate School of Frontier Sciences
University of Tokyo, Japan
{hirose, furuyama, mine}@gavo.t.u-tokyo.ac.jp

Abstract

A corpus-based method was developed for automatic extraction of the F_0 contour generation process model parameters (phrase and accent commands). The method first smoothes the observed F_0 contour by a piecewise 3rd order polynomial function and finds points of inflection. Then several parameters related to the points are used as input parameters for the predictor of the model commands. Finally the predicted commands are tuned to the observed contour by the analysis-by-synthesis. An experiment was conducted using ATR 504 sentence speech corpus, and the performance close to a rule-based method, also developed by the authors, was obtained. An experiment was further conducted by adding linguistic information of the content of the utterance (such as accent type, depth of *bunsetsu* boundary) to input parameters. The performance was largely improved; the extraction rates reached around 90 % for phrase commands and 84 % for accent commands.

1. Introduction

Due to an increasing need on speech corpora in spoken language processing, several large ones have already been developed for major languages. However, most of them only include phonemic labels without labels on prosodic features. The major reason for this situation is that the corpora are mostly arranged for speech recognition study, where prosodic features are little used. However, control of prosodic features is quite important in realizing high quality in speech synthesis and use of prosodic features will become mandatory for the future development of speech recognition.

The major problem on developing prosodic corpus (speech corpus with prosodic labels) is that we have no established system for prosody annotation yet. The well-known Tone and Break Indices (ToBI) system is a good candidate. It was originally developed for English and adapted to several languages. As for Japanese, J-ToBI has already been developed with its extended version X-JToBI [1]. Several speech corpora have already been developed in the framework of ToBI labeling, but the ToBI system has an unavoidable defect that it is not based on the quantitative definition of prosodic features. It requires "intuition" of the human labelers, and, therefore, the labeling results may fluctuate between them. Also, labeling by hand is a time consuming work. To cope with this situation, several attempts were made for the automatic ToBI labeling, though the results were still far from satisfaction and needed manual correction.

The generation process model for fundamental frequency contour (henceforth F_0 model) will offer us a prosody labeling system with quantitative definition. The model assumes two types of commands, phrase and accent commands, as model inputs, and these commands have been proven to have a good correspondence with linguistic and para-/non linguistic information of speech [2]. For instance, we can easily define the prosodic boundary depth between accent phrases using the absence/existence and magnitude of phrase command between the phrases. The major problem when using the F_0 model for prosodic labeling is that we have no good method for automatically extracting the model commands from the observed F_0 contours.

From this point of view, several works have been conducted on the automatic extraction of the F_0 model commands [3-5]. Recently, by smoothing the observed F_0 contours using 3rd order polynomial functions and searching accent command location from their derivatives, a rather high accuracy was realized [6]. The performance can be improved by adding several constraints, which come from linguistic information of the speech samples [7]. However, the method consists of ad hoc rules and further improvements are rather difficult.

In view of the problem for the rule-based method, we newly developed a corpus-based method for F_0 model command extraction. The method predicts model commands using binary decision trees with input parameters obtainable from observed F_0 contours. Constraints can easily be applied by adding parameters on linguistic information to the input parameter set.

The rest of the paper is constructed as follows: Section 2 describes the F_0 model. Section 3 explains our rule-based method, which is used as the baseline of the corpus-based method. The developed corpus-based method is explained in sections 4 and 5 for the versions without and with linguistic information, respectively. Section 6 shows experimental results. Section 7 concludes the paper.

2. F_0 model

The F_0 model is a command-response model that describes F_0 contours in logarithmic scale as the superposition of phrase and accent components [2]. The phrase component is generated by a second-order, critically-damped linear filter in response to an impulse called phrase command;

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (1)$$

and the accent component is generated by another second-order, critically-damped linear filter in response to a step function called accent command;

$$G_a(t) = \begin{cases} \min\left[1-(1+\beta t)\exp(-\beta t), \gamma\right], & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (2)$$

where α and β are the time constants for the phrase and accent control mechanisms, respectively. Since these parameters are tightly related to the mechanical system of larynx, they are considered to be similar for all the utterances. Based on the former analyses on F_0 contours, they are fixed at 3.0 s^{-1} and 20.0 s^{-1}, respectively. The ceiling parameter γ is also fixed at 0.9.

An F_0 contour is given by the following equation:

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}\left\{G_a(t - T_{1j}) - G_a(t - T_{2j})\right\} \quad (3)$$

In the equation, F_b is the bias level, i is the number of phrase commands, j is the number of accent commands, A_{pi} is the magnitude of the ith phrase command, A_{aj} is the amplitude of the jth accent command, T_{0i} is the time of the ith phrase command, T_{1j} is the onset of the jth accent command, and T_{2j} is the end of the jth accent command.

3. Rule-based method

The input parameters for the corpus-based method (developed method) are calculated from observed F_0 contours through a process similar to that for the rule-based method [6]. Also, the rule-based method is used as the reference for evaluating the corpus-based method in section 6. Considering these situations, the rule-based method is first explained briefly in this section.

The observed F_0 contour may include pitch extraction errors, sharp F_0 movements due to articulation of speech sounds, and voiceless parts without F_0 values. These may cause errors in the F_0 model parameter extraction. Therefore, the observed contour is gone through the processes of correction of gross errors, removal of microprosody, interpolation of intervals of voiceless consonants, and smoothing by a piecewise 3rd order polynomial function [6]. A stable extraction of the F_0 model commands is possible from the smoothed F_0 contour thus obtained. The advantage of using 3rd order polynomial over using other curves is that the derivative can be given straightforwardly through a mathematical calculation.

The accent commands are obtained from the 1st order derivative of the smoothed F_0 contour. The command location can be decided from the derivative peak and valley points easily. The command amplitude is estimated from the peak and valley values.

The extraction of phrase commands were conducted for the F_0 contour residual, which was obtained by subtracting accent components generated from the extracted accent commands from the smoothed contour. The onset of the first phrase component is estimated from the timing of the first peak of the residual. The magnitude of the component was estimated through a successive approximation process. Then the estimated phrase component was subtracted from the residual F_0 contour to obtain a new residual F_0 contour, which was used to extract the second phrase command in a similar way. The process starts from the utterance initial and ends when it reaches the utterance final.

4. Corpus-based method without linguistic information

The corpus-based method extracts F_0 model commands from the smoothed F_0 contours obtained by the same process as the rule-based method. Similar to the rule-based method, it first extracts the accent commands and then extracts the phrase commands from the F_0 contour residuals. The extraction process consists of the following steps.

1. Search points of inflection (peak and valley points of 1st derivative) for the smoothed F_0 contour.
2. Classify the points into three categories: points corresponding to accent command onsets (A1), points corresponding to accent command ends (A2), and points not related to accent commands (A0). This classification is conducted by the corpus-based method (using a binary decision tree) with the input parameters listed in Table 1.
3. For each point classified to A1 or A2, distance from the corresponding onset/end of accent command is predicted using other binary decision trees (one for onset and another for end). The input parameters are again those listed in Table 1.
4. Calculate onsets and ends of accent commands from the result of the above process. Since an accent command should have a pair of onset and end points, a special process is necessary to recover a missing end/onset point between two succeeding onset/end points. When two onset points succeed, an F_0 contour peak immediately following to the first onset point is searched. If it is found, an end point is assumed at *T*peak+*T*dis, where *T*peak and *T*dis mean the instance of the peak and the distance of the onset point and the peak, respectively. If the peak is not found, an end point is assumed at one mora behind the first onset point. When two end points succeed, an onset point is inserted at one mora before the second end point. When a pause is found between onset and end points, insert an end point at the beginning of the pause and an onset point at the end of the pause. The process above reflects the accent system of the Tokyo dialect, and is basically the same with that developed for the rule-based method.
5. Calculate the amplitude of accent command from the 1st derivative of smoothed F_0 contour at the corresponding inflection points.
6. Classify the accent command onset points into two categories: points corresponding to phrase commands (P1) and points not related (P0). This classification is conducted using a binary decision tree with the inputs listed in Table 2. The residual F_0 is calculated by subtracting accent components obtained in the preceding steps (and the bias level F_b) from the smoothed F_0 contour.
7. For each point classified to P1, distance from the corresponding phrase command is predicted using another binary decision tree. The input parameters are again those listed in Table 2. The magnitude of phrase command is calculated from the F_0 residual at the extracted command point.
8. Refine the extracted parameters of accent and phrase commands through a recursive process (analysis-by-synthesis).

For the speech samples used to train binary decision trees, classification of inflection points into A0, A1, A2, and classification of accent command onsets into P0 and P1 are necessary. This was done simply by the minimum distance

rule: for instance, an inflection point closest to an accent command onset is classified to A1.

Table 1: Input parameters of the accent command predictor. PI denotes "point of inflection."

Input parameter	Category
Current PI Position, First derivative of F_0 contour, F_0	Continuous
Directly preceding/following PI First derivative of F_0 contour, F_0	Continuous
Distance between current and directly preceding/following PI's	Continuous
Existence of voiceless period between current and directly preceding/following PI's	2
Distance between end/start of directly preceding/following voiceless period and current PI	Continuous

Table 2: Input parameters of the phrase command predictor. PACO denotes "predicted accent command onset."

Input parameter	Category
Current PACO Position, Residual F_0, F_0, Cumulative value of residual F_0	Continuous
Directly preceding/following PACO Residual F_0, F_0, Cumulative value of residual F_0	Continuous
Distance between current and directly preceding/following PACO's	Continuous
Difference in cumulative values of residual F_0 at current and directly preceding/following PACO's	Continuous
Existence of voiceless period between current and directly preceding/following PACO's	2
Distance between end/start of directly preceding/following voiceless period and current PACO	Continuous

5. Corpus-based method with linguistic information

The F_0 model commands are tightly related to linguistic information. Therefore, the performance of command extraction can be improved by checking the consistency between extracted commands and linguistic information of the utterance. Minor components, which have no clear relation with linguistic information, may be ignored through the process. This ignorance may cause errors from the viewpoint of "precisely copying observed F_0 contour by the model." However, it increases the matching between extracted commands and linguistic information, leading to a better prosodic corpus for speech synthesis and recognition. The major problem, when using linguistic information for rule-based command extraction, is how combine it with information obtained from F_0 contours. In corpus-based method, linguistic information can be easily combined by including it to input parameters for the predictors.

As for the accent command extraction, besides the points of inflection of smoothed F_0 contours, mora boundaries corresponding to the beginning of accent initial morae and to the end of accent nucleus morae (henceforth, accent related mora boundaries) are added to candidates for the accent command onsets/ends, and classified into three categories (A1, A2 and A0) by the 3rd step in section 4. Here, accent initial and nucleus morae are respectively defined as the first and last high-F_0 morae, which coincide with the accent types of the accent phrases. The addition is done only for mora boundaries without point of inflection in the period from one mora before and to one mora after. The parameters shown in Table 3 are added to the inputs for accent command extraction. This process may increase the extraction accuracy where accent command onset/end points do not clearly appear in the F_0 contour, as in the cases before and after pauses.

As for the phrase command extraction, the parameters shown in Table 4 are added to the input parameters shown in Table 2. The depth of bunsetsu boundary is added, because it has a tight relation with the phrase command occurrence.

Table 3: Input parameters added to the accent command predictor. A dummy category is added for the first to the sixth input parameters to represent the candidate point being a point of inflection, for which no accent type information is obtained from the utterance content.

Input parameter	Category
Accent type at current candidate point	11
Accent type at directly preceding/following candidate point	11
Current candidate point being mora boundary corresponding to the beginning of accent initial mora	3
Directly preceding/following candidate point being mora boundary corresponding to the beginning of accent initial mora	3
Distance between directly preceding/following accent related mora boundary and current candidate point	Continuous
Existence of voiceless period between directly preceding/following accent related mora boundary and current candidate point	2
Distance between directly preceding/following *bunsetsu* boundary and current candidate point	Continuous

Table 4: Input parameters added to the phrase command predictor. PACO denotes "predicted accent command onset."

Input parameter	Category
Existence of *bunsetsu* boundary between current and preceding PACO's	2
Bunsetsu boundary depth between current and preceding PACO's	18
Distance between directly preceding/following *bunsetsu* boundary and current PACO	Continuous

The linguistic information of utterance content is obtained through text analysis using Japanese parsers [7]. The *KNP* parser [8] tells us which *bunsetsu* directly modifies another *bunsetsu*, and thus gives us *bunsetsu* boundary depth

information. The *bunsetsu* is defined as a basic unit of Japanese grammar and pronunciation consisting of a content word (or content words) followed or not followed by a function word (or function words). It mostly coincides with the accent phrase.

6. Experiments on F_0 model command extraction

Experiments on the F_0 model command extraction were conducted using the ATR continuous speech corpus (503 sentence utterances by speaker MHT). First the F_0 model commands were extracted for all of 503 utterances through the analysis-by-synthesis with initial parameters manually assigned. The extracted commands were assumed as the correct ones and used as the references to evaluate the rule-based and corpus-based methods. The results of command extraction for 503 utterances are given at:

http://www.gavo.t.u-tokyo.ac.jp/parameters.html.

The 503 utterances were divided into 403 and 100. The 403 utterances were used for the training of binary decision trees for the corpus-based method. The 100 utterances were used to evaluate the rule-based and corpus-based methods. Results are summarized in Table 5 for the phrase commands and Table 6 for the accent commands. In the tables, *C*, *S*, *D*, *I* respectively mean number of commands correctly extracted, number of substitution errors, number of deletion errors, and number of insertion errors. The substitution error indicates that more than 2/3 of an extracted command period coincides with that of the corresponding correct command, though onset and/or end points are erroneous. Discrepancy in timing smaller than 1mora is ignored, because it can be corrected though the analysis-by-synthesis process. Also, the magnitude/amplitude of the command is not included in the evaluation because of the same reason.

The tables clearly show that the use of linguistic information improves the performance of command extraction. It is rather difficult to say which is better between the rule-based and corpus-based methods. (Different answers when viewed from correct extraction rates or total number of errors.) Roughly speaking, we should say the performance is similar for the both methods. Here, we should note that the rule-based method was developed by an expert on Japanese prosody through a careful inspection on a number of F_0 contours.

Table 4: Comparison of phrase command extraction by the rule-based and corpus-based methods. The corpus-based method has two versions: without and with linguistic information of the speech sample. The number of phrase commands manually extracted is 294.

	Rule-based method	Corpus-based method	
		Without linguistic information	With linguistic information
C	278 (94.6 %)	248 (84.4 %)	264 (89.8 %)
D	16	46	30
I	50	16	20

Table 5: Comparison of accent command extraction by the rule-based and corpus-based methods. The corpus-based method has two versions: without and with linguistic information of the speech sample. The number of accent commands manually extracted is 563.

	Rule-based method	Corpus-based method	
		Without linguistic information	With linguistic information
C	425 (75.4 %)	432 (76.7 %)	473 (84.0 %)
S	39	47	30
D	82	84	69
I	20	43	45

7. Conclusion

Corpus-based method was developed for automatically extracting the F_0 model commands from observed F_0 contours. Better performance was realized by using linguistic information of the utterance. Further experiments are planned for other speakers' utterances. Some extraction errors may be fatal for synthesized speech quality and some not. Speech synthesis experiments are planned using the developed prosodic corpora.

The work was partly supported by Grant in Aid for Scientific Research of Priority Areas (#746).

8. References

[1] Maekawa, K., Kikuchi, H., Igarashi, Y., and Venditti, J., "X-JToBI: An extended J_ToBI for spontaneous speech," *Proc. ICSLP*, Denver, pp.1545-1548, 2002.

[2] Fujiaski, H. and Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Japan (E)*, Vol.5, No.4, pp.233-242, 1984.

[3] Hirose, K., Fujisaki, H., Yamaguchi, M., and Watanabe, Y., "Automatic estimation of feature parameters for fundamental frequency contours," *Record of Acoustical Society of Japan Spring meeting*, pp.93-94, 1983. (in Japanese)

[4] Mixdorf, H., "A novel approach to the fully automatic extraction of Fujisaki model parameters," *Proc. IEEE ICASSP*, Istanbul, pp.1281-1284, 2000.

[5] Ogawa, H. and Sagisaka, Y., "Automatic extraction of F_0 control parameters using utterance information," *Proc. Speech Prosody*, Nara, pp.447-450, 2004.

[6] Narusawa, S., Minematsu, N., Hirose, K. and Fujiaski, H., "A method for automatic extraction of model parameters from fundamental frequency contours of speech," *Proc. IEEE ICASSP*, Orlando, pp.509-512, 2002.

[7] Hirose, K., Furuyama, Y., Narusawa, S., Minematsu, N., and Fujisaki, H., "Use of linguistic information for automatic extraction of F_0 contour generation process model parameters," *Proc. EUROSPEECH*, Geneva, pp.141-144, 2003.

[8] Kyoto University, Japanese Syntactic Analysis System KNP http://www-nagao.kuee.kyoto-u.ac.jp/projects/nl-resource/.