# Use of Linguistic Information for Automatic Extraction of $F_0$ Contour Generation Process Model Parameters

*Keikichi Hirose\*, Yusuke Furuyama\*, Shuichi Narusawa\*\*, Nobuaki Minematsu\*\*, and Hiroya Fujisaki\*\*\**

\*Graduate School of Frontier Sciences, \*\*Graduate School of Information Science and Technology, \*\*\*Prof. Emeritus
University of Tokyo, Japan
{hirose, furuyama, narusawa, mine}@gavo.t.u-tokyo.ac.jp, fujisaki@alum.mit.edu

## Abstract

A method was developed to utilize linguistic information (lexical accent types and syntactic boundaries) to improve the performance of the automatic extraction of the $F_0$ contour generation process model commands. The extraction scheme is first to smooth the observed $F_0$ contour by a piecewise 3rd order polynomial function and to locate accent command positions by taking the derivative of the function. If the results of automatic extraction differ from those estimated from the linguistic information, they are modified according to the several rules. The results showed that some errors could be corrected by the use of linguistic information, especially when the initial word of an accent phrase is type 0 (flat) accent. As a whole, the correct extraction rate (recall rate) was increased from 79.8 % to 82.3 % for phrase commands and from 81.6 % to 85.9 % for accent commands.

## 1.  Introduction

Recent technology on speech information processing largely relies on the speech corpora. Several rather large speech corpses have already been developed. However, most of them only include phonemic labels with no prosodic labels. The major reason for this situation is that the corpora are mostly arranged for speech recognition study, where prosodic features are not so much utilized yet. However, control of prosodic features is an important issue in speech synthesis already and use of prosodic features will become mandatory for the future development of speech recognition.

The major problem on developing prosodic speech corpus (speech corpus with prosodic labels) is that we have no good system for prosody annotation for Japanese. The well-known Tone and Break Indices (ToBI) system is a good candidate, and J-ToBI has already been developed with its extended version X-JToBI [1]. Several speech corpora have already been developed under the ToBI labeling framework, 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 labelers. Also, labeling by hand is time consuming work. To cope with this situation, there were several attempts 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 with 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 research works were conducted on the automatic extraction of model commands. Recently, by smoothing the observed $F_0$ contours using 3rd order polynomial functions and searching accent command location from their derivatives, we succeeded extracting model commands with a rather high accuracy [3]. However, there were still a number of cases the extraction was done incorrectly. The erroneous results can be corrected by utilizing the linguistic information of the utterances, which is not taken into account in the current method. We have already reported a method using linguistic information, but it only looked at the timing relationship between accent command timings and corresponding segmental boundaries [4]. This was because we had no good system available for obtaining the accent types of accent phrases. Recently, we have developed such a system under the project on "Development of Fundamental Software of Anthropomorphic Agents," and using the system, succeeded to include the linguistic information in the scheme of automatic extraction of model commands. In the current paper, the developed method is explained with some experimental results.

## 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. The phrase component is generated by a second-order, critically-damped linear filter in response to an impulse called phrase command;

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

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

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

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 $F_0$ contour analysis results, they were fixed at 3.0 s$^{-1}$ and 20.0 s$^{-1}$, respectively. The ceiling parameter $\gamma$ was also fixed at 0.9.

An $F_0$ contour is given by the following equation:

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

(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 $i$th phrase command, $A_{aj}$ is the amplitude of the $j$th accent command, $T_{0i}$ is the time of the $i$th phrase command, $T_{1j}$ is the onset time of the $j$th accent command, and $T_{2j}$ is the reset time of the $j$th accent command. A schematic view of the model is given in Fig. 1.
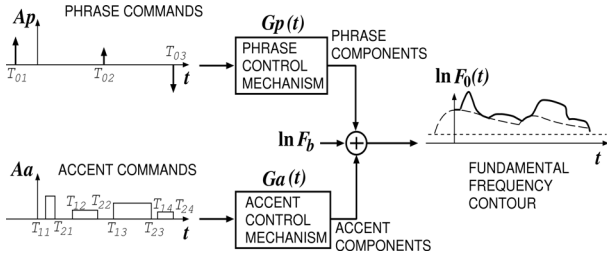


Figure 1: *Generation process model for sentence $F_0$ contours.*

## 3. Original method without linguistic information

The observed $F_0$ contour may include pitch extraction errors, sharp $F_0$ movements due to articulation of speech sounds, and voiceless parts with no $F_0$ values. These portions may cause errors in the $F_0$ model parameter extraction. Therefore, the pre-processing of the method includes operations for these portions: correction of gross errors, removal of microprosody, and interpolation of voiceless consonants.

The gross errors are defined as sudden and large $F_0$ shifts due to pitch extraction errors and irregularity in vocal folds vibration, often observable at the end of utterances. First, frames with the power lower than a threshold were assumed to be voiceless, even if they were judged voiced and had $F_0$ values by the pitch extraction algorithm. Then, the isolated $F_0$ clusters, whose $F_0$ values were different from those of other parts than a threshold, were assumed as gross errors and substituted by the interpolated values.

Articulation of some consonants, such as plosives, influences the vocal folds vibration and causes sharp $F_0$ movements around the phoneme boundaries. These so-called microprosodic $F_0$ movements are not included in the $F_0$ model and, therefore, should be removed before the model parameter extraction. Since the microprosody for voiceless plosives appears as the sharp down-drift in $F_0$ contour into the succeeding vowels, it can be removed by a rather simple logic.

A voiceless portion without $F_0$ was interpolated by the 3rd order polynomial, whose coefficients were decided from $F_0$ values immediately before and after the portion. If a portion without $F_0$ had duration larger than $1/\alpha$ (0.33 s), it was assumed as a pause and no interpolation was conducted.

After these processes, the $F_0$ contours were smoothed out by a piecewise 3rd order polynomial function. 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. Basically, an accent command appears in the derivative as a set of peak and valley. However, when an accent component is immediately followed by another component with similar or larger amplitude, the valley will not appear clearly. An algorithm of accent command detection was developed taking these features into account. Since the timings of derivative peak and valley correspond to the middle of the $F_0$ upward and downward movements, respectively, and delay from the command onset and reset times by $1/\beta$, 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 by shifting it backward by $1/\alpha$. 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.

The commands extracted by the above process are used as the initial parameter values for the analysis-by-synthesis process to obtain the optimized command values.

## 4. Extraction of linguistic information from the text

The text analysis was conducted using Japanese parsers *Chasen* [5] and *KNP* [6]. *Chasen* divides the input sentence into morphemes with their readings, part-of-speeches, accent types, accent attributes, and so on. *KNP* detects "bunsetsu" boundaries using information obtained from another parser *Juman* (*Juman* is used, because KNP cannot directly handle the outputs from *Chasen*). It also tells 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), and its accent type is somewhat different from those of the words in it. While the accent phrase coincides with "bunsetsu" in many cases, it is defined in a different way: a unit with one accent component (sometimes, accompanied by secondary accent components). Two or more "bunsetsu" can be an accent phrase and vice versa.

The large vocabulary speech recognition system Julius [7] was utilized to obtain phoneme boundary locations through the forced alignment process.

The accent type of each accent phrase was decided applying the accent sandhi rules to the morphemes of the phrase. The rules are those originally developed by Sagisaka et. al. [8] and modified though perceptual experiment by the authors [9]. They include rules for a content word and a function word, two content words, an affix and a content

word, and so on. The accent type of an accent phrase can be decided by referring to the accent types and attributes of the morphemes.

## 5. Relation of $F_0$ model commands and linguistic information

The speech corpus used for the experiment is the 15-minute recordings of a male announcer's speech from a radio program "From My Bookshelf." The relation between "bunsetsu" boundary depths and phrase commands was analyzed for the corpus. The $F_0$ model commands are those manually extracted, which are also used later as correct commands to evaluate the method. Table 1 summarizes the result according to the existence/absence of preceding pauses and boundary depth codes. The boundary depth codes are obtained by KNP as shown in Fig. 1, which shows an example of parsing for the sentence "arayuru geNjitsuo subete jibuNno houe nejimagetanoda ([He] twisted all the reality to his side.)."

Table 1: *Relation between manually extracted phrase commands and linguistic information. The magnitudes and distances are shown as averages of extracted phrase commands. There are many cases without phrase commands at boundaries, especially when Boundary Depth Code is 1 and without pauses.*

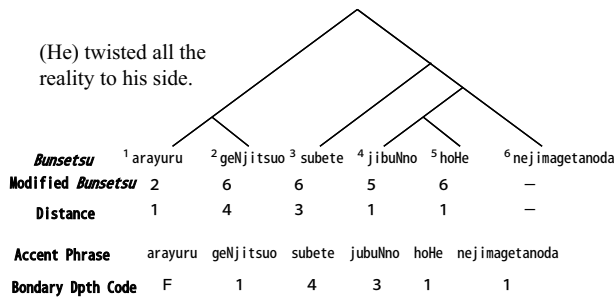| Pause at boundary | Boundary depth code | Magnitude | Distance from corresponding *bunsetsu* beginning (ms) |
|---|---|---|---|
| | Sentence beginning (F) | 0.616 | -251 |
| No | 1 | 0.293 | -106 |
| | 2 | 0.303 | -81 |
| | 3 | 0.363 | -39 |
| | 4 or larger | 0.389 | -76 |
| Yes | 1 | 0.415 | -200 |
| | 2 | 0.434 | -170 |
| | 3 | 0.522 | -192 |
| | 4 or larger | 0.582 | -193 |



Figure 2: *Result of syntactic analysis by KNP and codes to represent "bunsetsu" boundary depth.*

Also analysis on accent commands was conducted on the same corpus. Table 2 shows relation of the correct (manually-extracted) accent commands and those estimated from the linguistic information.

Table 2: *Correspondence between manually extracted accent commands and those estimated from the linguistic information.*

| | | Type 1 | Type 2 | Type 0 | Others |
|---|---|---|---|---|---|
| Number of commands | | 151 | 109 | 206 | 188 |
| Onset | Number of corresponded commands | 121 | 68 | 168 | 168 |
| | Average distance from corresponding mora initial (ms) | -15.8 | -78.3 | -75.8 | -74.5 |
| Reset | Number of corresponded commands | 124 | 68 | 130 | 153 |
| | Average distance from corresponding mora end (ms) | 55.8 | 19.1 | -30.2 | 6.7 |

## 6. Modification of $F_0$ model commands using linguistic information (Developed method)

Based on the results in Section 5 and other knowledge on $F_0$ contours obtained in the former experiments, the following rules were developed for the modification of the extracted $F_0$ model commands:

1) When the extracted accent command onset (or reset) position coincide with that generated from linguistic information, and when the extracted accent command reset (or onset) position does not coincide, the reset (or onset) position is changed to the generated position. The generated position is obtainable from Table 2 by referring to the accent type.

2) When no accent command is extracted where the linguistic information requires an accent command, add a command with amplitude "average value minus a standard deviation."

3) When an accent command is extracted where the linguistic information requires no accent command, and when its amplitude is below a threshold, it is deleted.

4) When no phrase command is extracted at a "bunsetsu" boundary with a short pause or with the boundary depth code equal to or larger than 2, a new phrase command is added. The magnitude of the phrase command is that listed in Table 1. When adding a new phrase command, the amplitude of the first accent command after the boundary is halved.

5) When a large and long-period accent command is extracted without phrase command immediately after a "bunsetsu" boundary, a phrase command is added. The magnitude of the phrase command is that listed in Table 1. When adding a new phrase command, the amplitude of the first accent command after the boundary is halved.

6) When a phrase command is extracted around the middle of a "bunsetsu," it is deleted.

Using the commands modified by the above rules as initial parameter values for analysis-by-synthesis process, the final command values are obtained.

Figures 3 and 4 show the results of the automatic $F_0$ model command extraction by the original and developed methods, respectively. It is clearly shown that the extraction error at the sentence initial, by the original method, is corrected by the developed method using accent type information. In the case of Japanese, when a command of type 0 accent is located at the beginning part of an accent phrase, the original method often cannot extract it correctly.
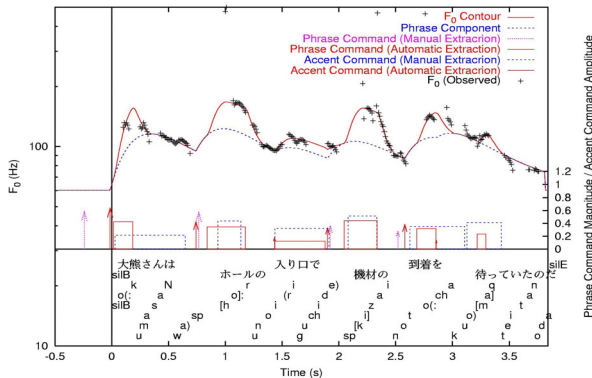


Figure 3: *$F_0$ command extraction by the original method. (English translation of the sentence: Mr. Ohkuma had been waiting for the arrival of the machine parts at the hall entrance.)*
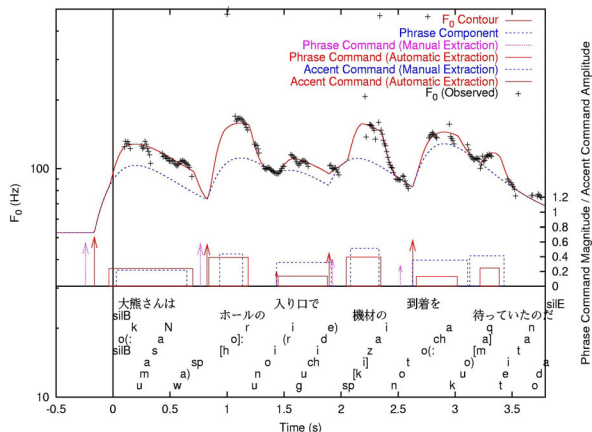


Figure 4: *$F_0$ command extraction by the developed method. (English translation of the sentence: Mr. Ohkuma had been waiting for the arrival of the machine parts at the hall entrance.)*

As the quantitative measures for command extraction performance, we adopted recall rate $R$ and precision rate $P$, which are respectively defined as follows:

$$R = \frac{C}{C+D}, \qquad P = \frac{C}{C+I}$$

where $C$, $D$, $I$ respectively mean number of commands correctly extracted, number of deletion errors, and number of insertion errors. In this evaluation, the magnitude/amplitude of the command is not taken into account, because it can be adjusted through the analysis-by-synthesis process. Table 3

summarizes the result. It clearly shows improvements in the performance for the developed method.

Table 3: *Comparison of $F_0$ model command extraction by the original and developed methods. The developed method utilizes linguistic information.*

|  | Phrase command (396 commands) | | Accent command (624 commands) | |
|---|---|---|---|---|
|  | Original method | Developed method | Original method | Developed method |
| Recall rate $R$ (%) | 79.8 | 82.3 | 81.6 | 85.9 |
| Precision rate $P$ (%) | 80.6 | 85.3 | 83.7 | 85.5 |

## 7. Conclusion

Information on the $F_0$ model command obtainable from linguistic information was successfully introduced to our method of automatic extraction of $F_0$ model commands. Further study is planned to introduce a statistical method, which becomes possible when the speech data with $F_0$ model commands increase.

## 8. References

[1] Kikuchi, H. and Maekawa, K., "Accuracy of prosodic labeling of spontaneous speech by X-JToBI," *Record of Acoustical Society of Japan Fall meeting*, pp.259-260, 2002. (in Japanese)

[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] 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*, pp.509-512, 2002.

[4] Sakurai, A. and Hirose, K., "Designing a parameter-based prosodic speech database," *Proc. Oriental COCOSDA Workshop*, Taipei, pp.5-8, 1999.

[5] Nara Institute of Science and Technology, Morphological Analyzer ChaSen, http://chasen.aist-nara.ac.jp/.

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

[7] Kyoto University, Large vocabulary continuous speech recognition decoder Julius, http://winnie.kuis.kyoto-u.ac.jp/pub/julius/.

[8] Sagisaka. Y. and Sato, H., "Accentuation rules for Japanese word concatenation," *Trans. IECE Jpn.*, Vol.66D, No.7, pp.849-856, 1983. (in Japanese)

[9] Minematsu, N., Kita, R. and Hirose, K., "Automatic estimation of accentual attribute values of words for accent sandhi rules of Japanese text-to-speech conversion," *IEICE Trans. Information and Systems*, Vol.E86-D, No.1, pp.550-557, 2003.