Automatic Extraction of F_0 Control Parameters Using Utterance Information

Hiromasa Ogawa & Yoshinori Sagisaka

Global Information and Telecommunication Institute
hiro@suou.waseda.jp; sagisaka@giti.waseda.ac.jp

Abstract

Aiming at automatic extraction of F_0 control parameters based on a generation model, we proposed an automatic extraction method using utterance information. In the proposed method, constituent phrase information is used for the prediction of initial value of F_0 control parameters in the optimal value search. Extraction experiments using short and long sentence sets showed higher extraction accuracy than the extraction without utterance information. The extraction accuracy difference between single sentence samples and major phrases in long sentences showed the need of further information on neighboring phrases for further improvements of extraction accuracy in long sentences.

1. Introduction

In speech synthesis, recent developments of large-scale speech corpora have enabled unit selection approach and corpus based speech synthesis become popular [1]. They have also been used for prosody processing and quite interesting F_0 control characteristics have been found [2][3]. However, as most of these corpora only have phonetic labels, we needed quite laborious hand annotations of prosodic events for these analyses, As easily seen in these studies, not only simple phonetics-based prosody event tags such as ToBI labels, but also much finer transcriptions have been quite useful for quantitative analysis. In particular, by analyzing F_0 phenomena from generation viewpoints, it is widely known that the superimposition model proposed by Fujisaki can account for many superficially varying F_0 manifestations in very reasonable ways [4]. For efficient studies of F_0 control, we needed a large amount of F_0 control parameters based on this generation model. In this paper, we report experimental results for their automatic extraction.

The automatic extraction of F_0 control parameters based on this generation model have been studied by many researchers. Edouard Geoffois proposed dynamic analysis of F_0 decomposition for speech recognition purposes [5]. Hansjörg Mixdorf has tried an extraction using Low-pass filtering technique to remove phrase components [6]. Narusawa, Hirose and Fujisaki have applied the derivative of the piecewise 3rd order polynomial function to assign initial parameter values of accent and phrase components to search optimal command values using analysis-by-synthesis technique [7]. In these attempts, they have been trying to extract F_0 control parameters simply using contour by itself. As the extraction of F_0 control parameters is very difficult even for experienced experts if additional utterance information is not provided. We decided to extract them with full use of side information obtained for an input utterance.

In the following sections, after explaining the description of F_0 control parameters based on this generation model, we discuss the difficulties of automatic extraction in section two. In section three, we propose an extraction scheme that we have tried with experimental results for two databases. Finally we summarize the current results and show our future perspectives.

2. Parametric description of an F_0 contour and their automatic extraction

Parametric description of an F_0 contour has been proposed by Fujisaki based on a generation model [8]. This model decomposes an F_0 contour into phrase components and accent components. These components are obtained as an output of second order critical damping system for impulsive inputs referred as a phrase command and an accent command.

A phrase component $G_p(t)$ is given by the following equation;

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

A accent component $G_a(t)$ is given by the following equation;

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

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

An F_0 contour is consisting of multiple combination of these components. It can be described as a linear sum of these components as follows.

$$lnF_0(t) = lnF_{min} + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}\{G_a(t - T_{1j}) - G_a(t - T_{2j})\} \quad (3)$$

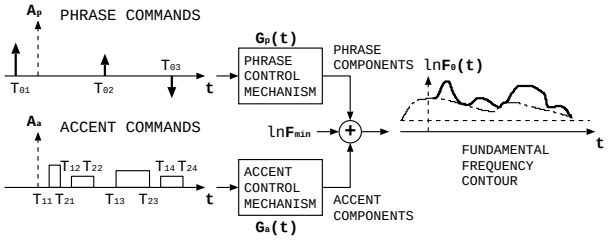

Figure 1: *Generation process model for sentence F_0 contours [8].*

In the equation, F_{min} 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 time of the jth accent command, T_{2j} is the reset time of the jth accent command. The visual outline of this model is shown in Figure 1.

As well known, this model can consistently describe a superficially quite varying F_0 contour by the small number of its control parameters, such as time and magnitude of phrase commands and onset time, reset time and amplitude of accent commands, for speech samples with various prosodic variations in many languages. The extraction of these parameters from a speech waveform is quite important as an efficient parametric representation of F_0. As there is no analytic solution to this inverse problem, an iterative procedure has been proposed for their extraction using analysis by synthesis method [8]. However, automatic extraction of these parameters from unrestricted samples is difficult and many studies have been started.

In the automatic extraction of F_0 control parameters, reasonable amount of research efforts have been spent for the problems of discontinuities in F_0 contours resulting from unvoiced portions and micro-prosody variations caused by neighboring consonants. To enable fully automatic extraction of F_0 control parameters we need systematic approach by using further information useful for extraction. Most of previous research aimed at an extraction of F_0 control parameters directly from F_0 contour only. Simply judging from the fact that the extraction is extremely difficult even for experienced experts without knowing its content, side information obtained from its content is quite helpful. In this paper, we have tried to extract F_0 control parameters by using a corresponding utterance content as much as possible. We do not only use the information directly obtained from its context, but also statistical characteristics of F_0 control parameters obtained other samples in the database.

3. Automatic extraction using utterance information

3.1. Localized extraction using constituent phrase information

As the first step of F_0 control parameter extraction using a corresponding utterance content, we started the extraction where constituent phrase information is given. We localize the parameter extraction to a portion of Japanese read speech consisting of only one or two phrase component with multiple accent components. This restriction is natural both from difficulty reduction of F_0 control parameter extraction and from the specification of corresponding portion in an utterance using their utterance content.

As it is known that a skeleton F_0 contour of Japanese read speech can be roughly determined by its constituent accent phrase numbers, their accent phrase length and accent types, F_0 control parameter extraction is much easier than when we do not have any information. At the same time, the specification of corresponding portion can be done using speech synthesis and recognition technology, though we are not sticking to full automatic extraction. Possible phrase boundary candidates can be restricted in the corresponding text using a phrasing module in a Japanese text-to-speech system and their constituent accent phrase information should also be available. The allocation of phrase position can be done by a phone aligner using input text and rough boundary search using an F_0 contour.

3.2. Extraction procedures

To carry out the search of timing points and magnitudes of phrase and accent command T_{0i}, A_{pi}, T_{1j}, T_{2j} and A_{aj}, we need their initial values. We use linear regression models for their prediction. These prediction models are trained before the first step using a set of manually analyzed F_0 control parameters and their corresponding constituent phrase data. In the regression modeling, time and magnitude of phrase command and onset time, reset time and amplitude of accent command are estimated by linear combinations of constituting accent phrase length (in mora) and their accent-type categories. For this calculation Quantification method I [9] is employed since input values are categories. The coefficients of these linear combinations are determined by minimizing the errors of F_0 control parameters same as conventional linear regression.

Automatic extraction is carried out as follows. First, initial values of F_0 control parameters for the test data are predicted using the regression models where the input parameters are mora counts and accent types of constituent accent phrases in the test data. Then, search the optimal parameter values until the F_0 contour difference between the generated one and the observed does not decrease. Finally let the convergent parameter value be an extraction result. The search method is as same as described in Fujisaki & Hirose [8]. In the search, minimum mean squared error is obtained through a hill-climbing search in the (2I+3J) dimensional space constructed by the parameters for the phrase and accent commands, where the number of phrase commands is I and accent commands is J. The timing of the accent commands are constrained to avoid overlap of neighboring accent commands. The step sizes used in the present analysis are $0.01s$ for T_{0i}, T_{1j}, T_{2j}, 0.01 for A_{pi}, A_{aj}.

Table 1: *Phrase based extraction accuracy for single sentences.*

test sentence set	with utterance information	without utterance information
two accent phrase sentences	91.5%	77.0%
three accent phrase sentences	91.7%	80.0%

Table 2: *Phrase based extraction accuracy for long sentences.*

test sentence set	with utterance information	without utterance information
long sentences	85.9%	72.5%

4. Experiment of automatic parameter extraction

4.1. Extraction for short sentences with a simple phrase structure

4.1.1. Extraction experiment

As the first experiment, we chose single sentences speech data with one or two phrase components and two or three accent components. The speech data set consist of 800 single sentences with two accent phrases and 100 single sentences with three accent phrases in ATR Japanese speech database for speech synthesis [10]. For the training of the regression models with two and three accent phrases, 700 and 80 sentences were used as training data respectively. The rest 100 sentences with two accent phrases and 20 sentences with three accent phrases were used as test data. F_0 control parameter extraction was manually performed to all training and test data.

Before the automatic extraction, the regression models were trained using training data to estimate the initial value of F_0 control parameters T_{0i}, A_{pi}, T_{1j}, T_{2j} and A_{aj}. We could have confirmed that the initial value prediction had been carried out reasonably through the comparison between observed F_0 contours and generated ones.

To confirm the usefulness of utterance information needed in initial value prediction, we carried out search without initial value prediction as a control experiment. For the search without initial value prediction, we adopted the F_0 control parameter average of the training data as initial values.

4.1.2. Extraction results

As the correctness of F_0 control parameter extraction can be evaluated by comparing the timing points of phrase command T_{0i}, the onset time of accent command T_{1j} and the reset time of accent command T_{2j}, we considered the automatic extraction is correct when the onset time and reset timing points of the accent command are within half mora distance from the manually extracted timing points.

Table 1 shows the extraction accuracies calculated for phrases contained in the sentence sets consisting of either two or three accent components. As shown in the table, the extraction using utterance information attained around 92% accuracy in both cases. The comparison with the extraction without utterance information shows that the extraction errors are reduced to almost one third. Although about 8% extraction error is still not small, the error analysis showed that more than half of errors are resulting from insufficiencies of F_0 contour caused by long voiceless portions, devocalization and miss-extractions. Furthermore, the most of other errors are observed in restricted accent phrase combinations that experts need additional phrase information. In Japanese, when an accented phrase follows an unaccented phrase, it is often quite difficult to find the first phrase boundary with F_0 contour only. These analysis results support reasonable extraction improvement using context information and revealed further possibilities to reduce extraction errors.

4.2. Extraction for a major phrase interval in long sentences

4.2.1. Application of automatic extraction to long sentences

The extraction experiment in the previous section showed that the proposed automatic extraction can use utterance information effectively in a single sentence. To enable automatic extraction of F_0 control parameters, not only we have to cope with extraction errors analyzed in the previous section but also to confirm this extraction method can be applicable to ordinary long sentences. As so many different combinations of phrase components and accent components are possible, it is not efficient to simply make the same kind of models one by one for all combinations both from exhaustive combinatorial search efforts and the accurate estimation of initial values needed for the search.

We restricted the extraction of F_0 control parameters to speech intervals consisting of one phrase component. As we can automatically segment a long sentence into accent components quite accurately using F_0 contour only [11], we can expect an automatic segmentation into speech intervals consisting of one phrase component using both F_0 contour and its utterance content.

For speech intervals consisting of one phrase component, we apply the same extraction procedures applied to single sentences. For the training of the linear regression model for initial value prediction, we classified training data extracted from long sentence speech according to the number of constituent accent phrases. They are used to make multiple regression models as illustrated in Figure 2.

4.2.2. Extraction experiment

In this experiment, we used 150 sentences including 528 speech intervals consisting of one phrase components and multiple accent components. These sentences were selected from ATR Japanese speech synthesis database

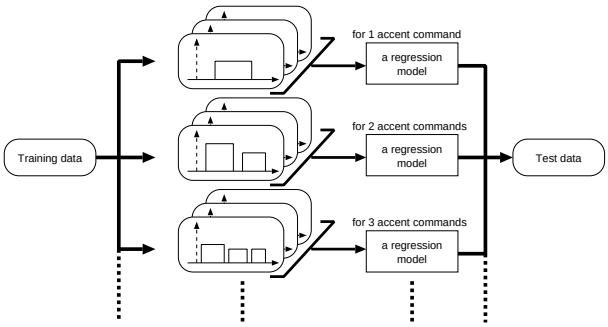

Figure 2: *Multiple models for initial value prediction of F_0 control parameters*

[12]. For the training of the regression models, 130 sentences were used and the rest 20 sentences were used as test data.

Automatic extraction is carried out as follows. First, set the initial values for all speech intervals consisting of one phrase components in the long sentence by using the regression models where the input parameters are mora counts and accent types of constituent accent phrases in the test data. At this time, the regression model corresponding to the number of accent components contained in each interval is used. Then, set a parameter search regions as an interval delimited by accent phrase boundaries. For the first speech interval, we carry out the same search as the case of short sentences. For the 2nd and the latter intervals, we take into the consideration of F_0 bias given by the preceding phrase when we apply the same search algorithm.

The 2nd or later speech interval carries out the same search as the case of short sentences, also taking the fixed parameter till then into consideration.

4.2.3. Extraction results

The extraction experiment results are shown in Table 2. As shown in this table, the extraction accuracy of using utterance information is higher than the one without utterance information. However, the extraction accuracy is lower than the one for single sentences although we pre-assigned the phrase command timing points. Furthermore, the improvement from the extraction without utterance information is smaller than the one for single sentences. The error analysis showed that these lower performances for long sentences are resulting from the neighboring phrase control effects on the F_0 contour in the current interval. Further utterance information on the neighboring phrases is needed for the improvement of extraction accuracy.

5. Conclusion

In this paper, we proposed an automatic extraction of F_0 control parameters based on generation model using utterance information. Extraction experiments showed the improvement of extraction accuracy using constituent phrase information for the prediction of initial values in the optimal value search. The extraction accuracy difference between single sentence samples and major phrases in long sentences showed the need of further information on neighboring phrases. For a full automatic extraction of F_0 control parameters, we have to not only improve the extraction method proposed in this paper but also put together other procedures such as automatic segmentation into major phrase and phrase command position estimation for long sentences. However, if semi-automatic extraction can be possible for some kind of speech data, there are so many applications. We would like to approach this extraction problem to cope with urgent current needs.

6. References

[1] Y. Sagisaka, "Speech synthesis : Overview" in Survey of the State of the Art in Human Language Technology, R. Cole et al (Eds.), 165-170, *Linguistica Computazionale* Vol. XII XIII, 1997.

[2] Y. Sagisaka, N. Campbell, N, Higuchi(Editors), *Computing Prosody*, Springer, 1997.

[3] T. Hirai, N. Iwahashi, N. Higuchi, Y. Sagisaka, "Automatic Extraction of F_0 Control Rules Using Statisical Analysis," in *Progress in Speech Synthesis* van Santen et al (Eds.), 333-346, Springer, 1997.

[4] H. Fujisaki, "Modeling the process of fundamental frequency contour generation," 313-326, in *Speech, perception, production and linguistic structure*, Y. Tohkura et al (Eds.) Ohmusha IOS Press, 1992.

[5] E. Geoffrois, "A pitch contour analysis guided by prosodic event detection," *Eurospeech'93*, 1993.

[6] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," *ICASSP 2000*, vol. 3(1281-1284), 2000.

[7] S. Narusawa, N. Minematsu, K. Hirose, H. Fujisaki, "Automatic Extraction of Model Parameters from Fundamental Frequency Contours of Speech and Its Evaluation," *Technical report of IEICE*, 19-24, 2002.

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

[9] C. Hayashi, "On the Quantification of Qualitative Data from the Mathematico-Statistical Point of view," *Annals of the Institute of Statistical Mathematics*, Vol. 2, 1950.

[10] M. Miyatake, Y. Sagisaka, "Japanese speech database for Synthesis," *ATR Technical Report* TR-I-0056, Nov. 1988.

[11] M. Nakai, H. Singer, Y. Sagisaka, H. Shimodaira, "Accent Phrase Segmentation by F_0 Clustering Using Superpositional Modelling," in Y. Sagisaka, N. Campbell, N. Higuchi (Editors), *Computing Prosody* Springer, 343-359, 1997.

[12] M. Abe, Y. Sagisaka, T. Umeda, H. Kuwabara, "Speech Database," *ATR Technical Report* TR-I-0166, Sep. 1990.