Corpus-based Generation of F_0 contours of Japanese based on the Generation Process Model and its Control for Prosodic Focus

Keikichi Hirose[1], Keiko Ochi[1], and Nobuaki Minematsu[2]
[1] *Department of Information and Communication Engineering, the University of Tokyo, Tokyo*
[2] *Department of Electrical Engineering and Information Systems, the University of Tokyo, Tokyo*
hirose, ochi, minematsu@gavo.t.u-tokyo.ac.jp

Abstract

A total corpus-based process of generating prosodic features form text is developed. The process first predicts pauses and phone durations, and then generates F_0 contours. Since F_0 contour generation is based on the generation process model, it is rather easy to manipulate the generated F_0 contours in command level. A method was developed for generating sentence F_0 contours, when a focus is placed in one of "bunsetsu" of an utterance. The method is to predict differences in the F_0 model commands between with and without focus utterances, and applies them to the F_0 model commands predicted beforehand by the baseline method. The validity of the method was proved by the experiment on F_0 contour generation and speech synthesis.

1. Introduction

Introduction of corpus-based concatenative scheme largely improved the quality of synthetic speech to a "close to human" level. However, the improvement is mostly on the segmental features of speech, and, if we view from the aspect of prosodic features, there still remain problems to be solved. Since prosodic features cover a range longer than phonemes, concatenation of prosodic features in such units may cause unnatural speech sounds; prosodic features need to be generated by viewing a whole sentence or longer units.

Recently, in speech synthesis community, an attention is paid to works on HMM-based speech synthesis, where a flexible control in speech styles is possible by adapting phone HMMs to a new style [1]. In the method, both of segmental and prosodic features of speech are processed together in a frame-by-frame manner, and, therefore, it has an advantage that synchronization of both features is kept automatically [2]. Although various styles such as attitudes and emotions were realized with rather high quality by the method, frame-by-frame processing of prosodic features, however, includes some problems. It has a merit that fundamental frequency (F_0) of each frame can be used directly as the training data, but, in turn, it sometimes causes sudden F_0 undulations (not observable in human speech) especially when the training data are limited. As mentioned already, prosodic features cover a wider time span than segmental features, and should be treated differently.

From these considerations, we have developed a corpus-based method of synthesizing F_0 contours in the framework of the generation process model (F_0 model) and realized speech synthesis in reading and dialogue styles with various emotions [3, 4]. The model represents a sentence F_0 contour as a superposition of accent components on phrase ones; each type of components assumed to be responses to step-wise accent commands and impulse-like phrase commands, respectively [5]. By predicting the model commands instead of frame-by-frame F_0 values, a good constraint is automatically applied on the generated F_0 contours; still keeping acceptable speech quality even if the prediction is done incorrectly.

When synthesizing F_0 contours, phone and syllable boundary information is necessary. A corpus-based method was developed also for predicting pauses and phone durations from text input. By combining the method with that for F_0 contour synthesis, a total scheme was constructed to generate prosodic features for speech synthesis from a text [6].

By handling F_0 contours in the F_0 model framework, a clear relationship is obtainable between generated F_0 contours and their background linguistic (and para-/non-linguistic) information, enabling "flexible" control of prosodic features. It is rather easy to analyze the prosodic controls obtained by statistical methods and to modify generated F_0 contours in another corpus-based way, which is trained using a small speech corpus. As an example for the flexible control, we have developed a method of focus control [7]. Given

978-1-4244-2179-4/08/$25.00 ©2008 IEEE

a speech synthesis system without specific focus control, it is not efficient to prepare a large speech corpus with focus control and train the speech synthesis system from the beginning. The proposed method realizes prosodic focus as a supplemental process to our corpus-based method of F_0 contour generation; train binary decision trees for differences in phrase command magnitudes and accent command amplitudes between utterances with and without focuses. The command values predicted by our baseline method (for utterances without specific focuses) are modified using the differences. By concentrating to the differences, a better training for F_0 change due to focal position comes possible only with a limited speech corpus. Moreover, speakers for the training need not be the same for those of the baseline.

The following sections are organized as follows: After a brief explanation on our total corpus-based scheme of generating prosodic features from text input in section 2, prediction of F_0 contours are explained in section 3. Effects of adding timing constraints during F_0 model command prediction process is shown in section 4. The method of prosodic feature generation is evaluated through a listening test on synthetic speech in section 5. The method of realizing prosodic focus is proposed in Section 6. Section 7 concludes the paper.

2. Generation of Prosodic Features

Each sentence of the input text is first parsed into a morpheme sequence using a freeware CHASEN. Parsing using another freeware JUMAN+KNP is also conducted to obtain syntactic structures. The syntactic structure is given as a boundary depth code (BDC) of each *bunsetsu* boundaries, which indicates the distance from the *bunsetsu* immediately before the boundary to the *bunsetsu* directly modified. Here, *bunsetsu* is defined as a basic unit of Japanese syntax and pronunciation consisting of content word(s) followed or not followed by particles. Then the linguistic information thus obtained is used to predict position of pauses and their lengths. Similar processes of predicting phone durations and F_0 model parameters follow. Since all the timing structures need to be decided before the F_0 contour generation, the prediction of F_0 model parameters is conducted as the last process of prosodic feature generation. Binary decision trees (BDT's) are adopted for the prediction. The CART (Classification And Regression Tree) included in the Edinburgh Speech Tools Library [8] was utilized to construct BDT's. Training corpus (with necessary annotations) is prepared automatically using the above parsers, an HMM-based segmentation scheme, and an F_0 model command extractor [9]. Due to the page limitation, prediction process is shown only for F_0 model parameters in the next section.

3. Prediction of F_0 model parameters

Table 1. *Input parameters for the phrase command prediction.*

Input parameter	Category
Position in sentence of current *bunsetsu*	13
Number of *morae*	18 (18)
Accent type (location of accent nucleus)	14 (15)
Number of words	8 (8)
Part-of-speech of the first word	12 (13)
Conjugation form of the first word	14 (15)
Part-of-speech of the last word	12 (13)
Conjugation form of the last word	9 (10)
BDC at the boundary immediately before current *bunsetsu*	10
Pause immediately before current *bunsetsu*	2
Length of pause immediately before current *bunsetsu*	Continuous
Phrase command for the preceding *bunsetsu*	2
Number of *morae* between preceding phrase command and head of current *bunsetsu*	26
Magnitude of preceding phrase command	Continuous

It is known that the information of preceding units has a larger influence on the prosodic features of the current unit than that of following units. Taking these into consideration, information of the directly preceding *bunsetsu* is included in the input parameters for the phrase command predictor as well as that for the current *bunsetsu* in question (Table 1). The category numbers in the parentheses for the preceding *bunsetsu* are larger than those of the corresponding parameters of the current *bunsetsu* by one to represent "no preceding *bunsetsu*." Since pauses have a tight relation with phrase commands, information of predicted pauses was included also, while it was not used for the prediction of accent command parameters.

Similar to the case of phrase commands, the parameters on accent commands (position and amplitude) are tightly related to the information of the current and preceding units (prosodic words), such as position in sentence, length, grammatical information of the first and last words of the units, and syntactic boundary between the units. They also change according to the accent types of the units. Taking these into consideration, the input parameters for accent command predictor were selected (not shown here, due to space limitation).

As an objective measure to evaluate the F_0 contour generated using the predicted F_0 model parameters, the mean square error between the generated contour and the target contour is defined as:

$$F_0MSE = \frac{\sum_t (\Delta \ln F_0(t))^2}{T} \qquad (1)$$

where $\Delta \ln F_0(t)$ is the F_0 distance in logarithmic scale at frame t between the two F_0 contours. The summation is done only for voiced frames and T denotes their total number in the sentence. When timing information (pause lengths and phone durations) of the target speech were utilized, average F_0MSE values of generated F_0 contours were 0.039 and 0.042 for training and testing utterances, respectively.

4. Constraints on F_0 model parameters

A preliminary listening test was conducted for the speech synthesized using the generated prosodic features. Although the synthetic speech sounded natural for many cases, accent types were occasionally perceived incorrectly. They are caused mostly by the inaccurate prediction of accent command location. This inaccurate prediction may be due to inaccurate F_0 model command extraction for the training corpus. By applying a certain constraint on the accent command timing, this type of errors can be corrected. Although detail of the constraint is not shown here, the average F_0MSE values of generated F_0 contours in section 3 were reduced to 0.037 and 0.041 for training and testing utterances, respectively.

5. Speech synthesis and evaluation

Since errors in pauses and phone durations cause mismatch in timings for generated F_0 contours, it is not appropriate to evaluate the developed method only from F_0MSE values. Also large F_0MSE may not directly causes degradation in synthetic speech. From this point of view, a listening experiment was conducted for speech synthesized using prosodic features generated from predicted parameters. Ten sentences were randomly selected from the 50 testing sentences, and each one was synthesized with prosodic features with five variations (methods) shown in Table 2. They are randomly presented to 12 native speakers of Japanese, who were asked to conduct ten-scale scoring from the viewpoint of the naturalness of synthetic speech (10: Sounds like natural speech, 1: Sounds quite poor.). Speech synthesis was conducted using the HMM-based speech synthesis toolkit [10]. Tri-phone models were trained using the 453 sentence utterances used for the training of the prosodic feature predictors. The segmental features were 75th order vectors consisting of 0th to 24th cepstrum coefficients and their Δ and Δ^2 values.

Table 2. Combinations of prosodic features for speech synthesis. *Methods "d" and "e" denote accent command timing prediction without constraints and with constraints, respectively.*

Method	Pause	Phone duration	F_0 contour
a	Target	Target	Target
b	Target	Target	Generated
c	Target	Generated	Generated
d, e	Generated	Generated	Generated

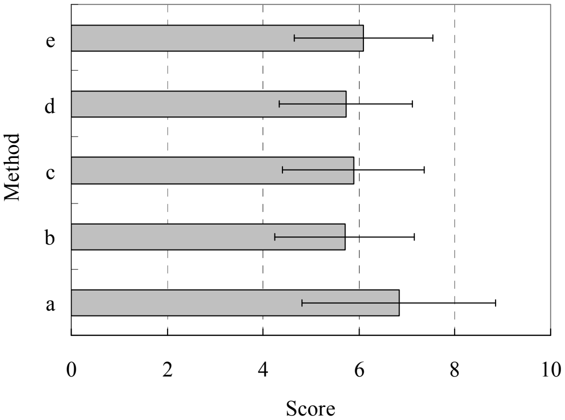

Figure 1. *Result of listening experiment.*

Result of listening test is shown in Fig. 1 as averages and standard deviations. It is noted that no degradation is observable when predicted pause lengths and phone durations are used. This is considered to be because of information on predicted pause lengths and phone durations being used for the prediction of F_0 model commands. Better score for method "e" as compared to method "d" indicates that the constriction on accent command timing works as expected. This kind of "empirical" correction comes possible only when the method is based on a quantitative modeling with clear relations with linguistic information.

6. Focus control

Observations of F_0 contours imply that we can represent F_0 change due to focal positions as changes in F_0 model command magnitudes/amplitudes. The differences in phrase command magnitudes are predicted using CART with inputs of distance from focused *bunsetsu*, position of current *bunsetsu* in phrase/sentence, *mora* numbers and accent types, distance in *mora* from the preceding phrase command, depth of syntactic boundary, and pause between preceding and current *bunsetsu*s. The differences in accent command amplitudes are also predicted in a similar way. 172 utterances by a female speaker were used for the prediction. Figure 2 shows an example of generated F_0 contour when the predicted changes are applied to F_0 model parameters predicted by the baseline method. Result of preliminary perceptual experiment confirmed that the focus was perceived on the right place by the method.

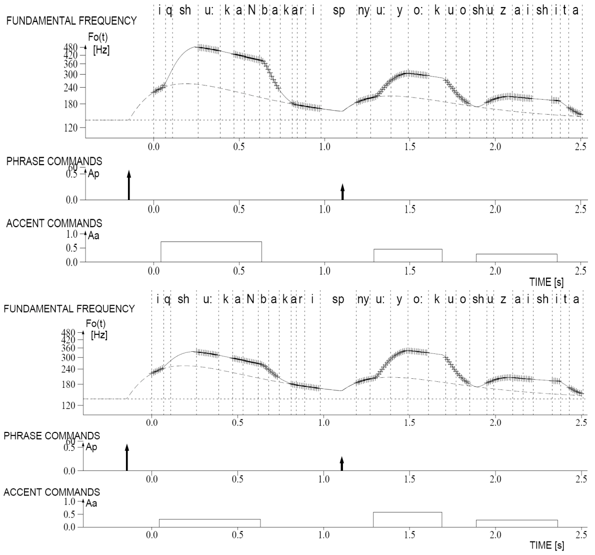

Figure 2. *Generated F_0 contours and F_0 model parameters of Japanese sentence "isshukanbakari nyuuyookuo shuzai shita (I collected news data in New York for about a week)". The first and second panels show when "isshukanbakari" and "nyuuyookuo" are focused, respectively.*

7. Conclusion

A total corpus-based method for generating prosodic features from text is presented. The key point of the method is that F_0 contours are predicted based on the F_0 model. As an example of "flexibility" of the developed method, realization of prosodic focus is addressed. The developed method is to predict differences in command magnitudes/amplitudes with and without focuses. The validity of the method was confirmed by a preliminary experiment.

8. References

[1] Yamagishi, J., Onishi, K., Masuko, T., and Kobayashi, T., "Modeling of various speaking styles and emotions for HMM-based speech synthesis," Proc. *EUROSPEECH*, pp.2461-2464 (2003).

[2] Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T., "Hidden Markov models based on multispace probability distribution for pitch pattern modeling," *Proc. IEEE ICASSP*, pp.229-232 (1999).

[3] Hirose, K., Sakata, M., Kawanami, H., "Synthesizing dialogue speech of Japanese based on the quantitative analysis of prosodic features," *Proc. ICSLP*, Vol.1, pp.378-381 (1996).

[4] Hirose, K., Sato, K., Asano, Y., and Minematsu, N., "Synthesis of F_0 contours using generation process model parameters predicted from unlabeled corpora: Application to emotional speech synthesis," *Speech Communication*, Vol.46, Nos.3-4, pp.385-404 (2005-7).

[5] Fujisaki, 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 (1984-10).

[6] Hirose, K., Ochi, K., and Minematsu, N., "Corpus-based generation of prosodic features from text based on generation process model," *Proc. Interspeech*, pp.1274-1277 (2007).

[7] Ochi, K., Hirose, K., and Minematsu, N., "Control of prosodic focus in corpus-based generation of fundamental frequency based on the generation process model," *Proc. Interspeech*, to be published (2008).

[8] The Edinburgh Speech Tools Library, http://www.cstr.ed.ac.uk/projects/speech_tools/

[9] 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).

[10] Galatea Project, http://hil.t.u-tokyo.ac.jp/~galatea/regist-jp.html (in Japanese).