CONTROL OF PROSODIC FOCUS IN CORPUS-BASED GENERATION OF FUNDAMENTAL FREQUENCY CONTOURS OF JAPANESE BASED ON THE GENERATION PROCESS MODEL

Keiko Ochi[1], *Keikichi Hirose*[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

ABSTRACT

A total corpus-based process of generating prosodic features from 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 the "*bunsetsu*" of an utterance. The method is to predict differences in the F_0 model commands between with and without focus utterances, and apply 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.

***Index Terms*—** Generation process model, F_0 contour, Corpus-based method, Speech synthesis, Prosodic focus

1. INTRODUCTION

Recently, in the speech synthesis community, attention has been focused on works on HMM-based speech synthesis, where a flexible control in speech styles is possible by adapting phone HMMs to a new style. In the method, both 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 [1]. 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. Prosodic features cover a wider time span than segmental features, and should be treated differently.

From this consideration, 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 [2]. 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 [3]. 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 [4].

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 [5]. Given 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, prediction of F_0 contours are explained in section 2. The method of prosodic feature generation is evaluated through a listening test on synthetic speech in the same section. The method of realizing prosodic focus is proposed and tested in Section 3. Section 4 concludes the paper.

2. GENERATION OF PROSODIC FEATURES

Each sentence of the input text is first parsed into a morpheme sequence using the open source software CHASEN. Parsing using another freeware JUMAN+KNP is also conducted to obtain syntactic structures. The syntactic structure is given as a

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

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 [6] 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 [7]. Due to the page limitation, prediction process is shown only for F_0 model parameters in this paragraph.

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[2]. 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.

Table 1. Input parameters for the phrase command prediction.

Input parameter	Category
Position in sentence of current *bunsetsu*	13
Number of *morae*	18 (19)
Accent type (location of accent nucleus)	14 (15)
Number of words	8 (9)
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

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). All the trainings were conducted using 453 utterances out of 503 utterances of ATR continuous speech corpus by a female narrator.

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.

To investigate the validity of the proposed method of F_0 contour generation when applied in a TTS system, a full speech synthesis system was constructed using the HMM-based speech synthesis as shown in Figure 1. 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 Mel-cepstrum coefficients and their Δ and Δ^2 values.

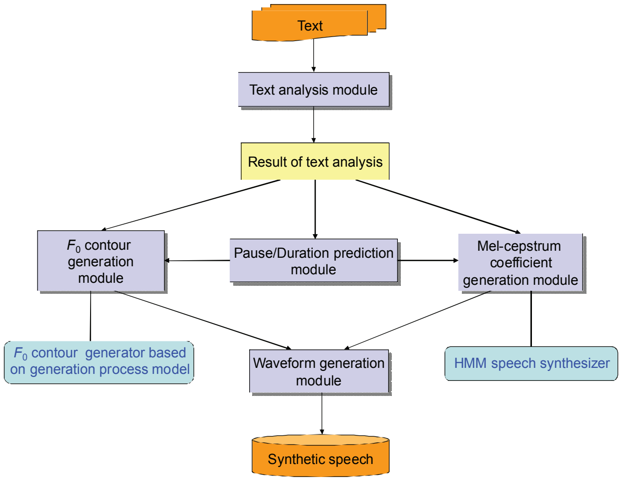
Text
Text analysis module
Result of text analysis
F_0 contour generation module
Pause/Duration prediction module
Mel-cepstrum coefficient generation module
F_0 contour generator based on generation process model
Waveform generation module
HMM speech synthesizer
Synthetic speech

Figure 1. Total configuration of developed speech synthesizer.

A listening experiment was conducted for speech synthesized using prosodic features generated from predicted parameters. Ten sentences not included in the training corpus were selected from the 503 sentences and were synthesized with prosodic features with five variations shown in Table 2. They are randomly presented to 12 native speakers of Japanese, who were asked to conduct ten-point scoring from the viewpoint of the naturalness of synthetic speech (10: Sounds like natural speech, 1: Sounds quite poor.).

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

The average score is 6.8, when original prosodic features are used (method “a”). It reduces to 5.7 when all the prosodic features are predicted (method “d”). It increases to 6.1 when a constraint is applied on the accent command timing (method “e”). 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 becomes possible only when the method is based on a quantitative modeling with clear relations with linguistic information.

3. FOCUS CONTROL

Although emphasis of word(s) is not handled explicitly in most of current speech synthesis systems, its control comes important in many situations, such as when the systems are used for generating reply speech in spoken dialogue systems: words conveying key information to the user’s question need to be emphasized. Emphasis associated with narrow focus in speech can be achieved by contrasting the F_0’s of the word(s) to be focused from those of neighboring words.

This contrast can be achieved by placing a phrase command (or increasing phrase command magnitude, when a command already exists) at the beginning of the word(s), by increasing the accent command amplitudes of the word(s), and by decreasing the accent command amplitudes of the neighboring words. The way of using these three controls maybe different from language to language. In order to investigate the situation for Japanese, we selected 50 sentences from the 503 sentences of the ATR continuous speech corpus, and asked a female speaker to utter each sentence without (specific) focus and with focus in one of assigned words (*bunsetsus*). For each sentence, 2 to 4 *bunsetsu*s were assigned depending on the sentence length. Figure 2 shows F_0 contours together with results of F_0 model approximations for utterances of the same sentence in different focal conditions. From the figure it is clear that the above three controls occur in the case of Japanese. It is also clear that there are one-to-one correspondences in phrase and accent commands for different focal conditions. (Although “jibuNnohooe” has one accent command when focus is placed on “subete,” it can be processed to have two commands with the same amplitude.) This one-to-one correspondence inspires us to realize focuses by controlling command magnitudes/amplitudes.

The proposed method for focus control is to modify command magnitudes/amplitudes predicted by the BDT's trained for utterances without specific focuses (baseline method) depending on the differences in command magnitudes/amplitudes between without and with focus utterances. The differences are trained also using BDT's. The modification is first applied to the phrase command magnitudes and then to the accent command amplitudes taking the (modified) phrase command information into account. Tables 3 and 4 show input parameters for the binary decision trees for predicting command magnitude/amplitude differences. Category numbers are reduced from the case of training command magnitudes/amplitudes (of the baseline method), so that training can be done only with a limited speech corpus. The above utterances for investigation on focus control are used to train the trees for the current experiment. They include 50 utterances without focus and 172 utterances with focus on one of noun phrases (*bunsetsu* including a noun).

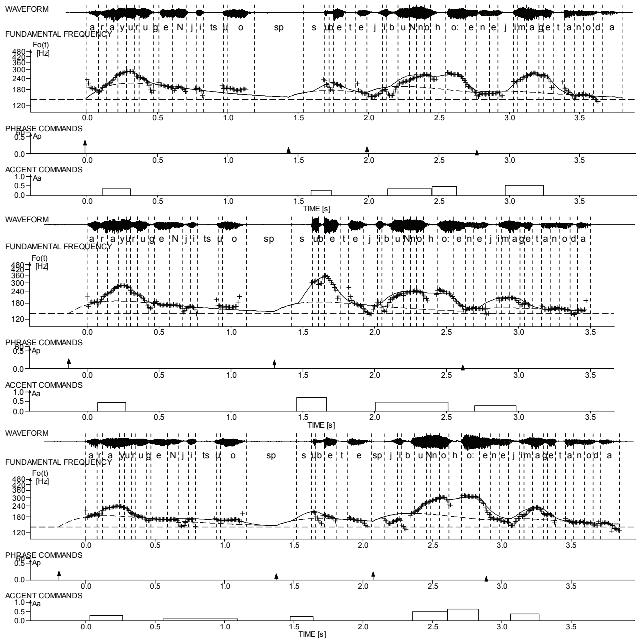

Figure 2. F_0 contours and F_0 model parameters of Japanese sentence “arayuru geNjitsuo subete jibuNnohooe nejimagetanoda ((He) twisted all the reality to his side.)” uttered by a female speaker. From the top to the bottom panels: without specific focus, focus on “subete,” and focus on “jibuNnohooe,” respectively.

Figure 3 shows examples of generated F_0 contours when the predicted changes are applied to F_0 model parameters predicted by the baseline method. The baseline method includes prediction of pauses and phone durations, and no modification is applied to those values. The three controls listed above for focus control can be seen in the figure. Here we should note that the speaker to train the command differences is different from one (the narrator) for training baseline method.

In order to check the effect of the focus control for realizing emphasis, a perceptual experiment was conducted for the synthetic speech. Speech synthesis was conducted using the system shown in Figure 1. Twenty six sentences not included in the 50 sentences for training command magnitude/amplitude differences are selected from the 503 sentences of the ATR continuous speech corpus, and one synthetic utterance is selected for each sentence; 19 utterances with focus and 7 utterances without focus. Eleven native speakers of Japanese were asked to listen to these utterances and check *bunsetsu* where they perceived an emphasis. “No emphasis” answer was allowed. On average, in 76.1 % cases, the *bunsetsu*s focused by the proposed method were perceived as “with emphasis.” If “no emphasis” answers are excluded from the statistics, the rate increases to 83.7 %.

Modification of F_0 contours may cause degradation in synthetic speech quality. In order to check this point, the same 11 speakers were also asked to evaluate the synthetic speech from naturalness in prosody in 5-point scoring (5: very natural, 1: very unnatural). No apparent degradation is observed from the result; 3.03 (standard deviation 1.00) for utterances with focus and 3.12 (standard deviation 0.93) for those without.

Table 3. Input parameters for the prediction of differences in phrase command magnitudes. The category numbers in parentheses are those for the directly preceding *bunsetsu*.

Input parameter	Category
Position in prosodic phrase of current *bunsetsu*	3
Position in prosodic clause of current *bunsetsu*	4
Position in sentence of current *bunsetsu*	5
Distance from focal position (in *bunsetsu* number)	5
Number of *morae*	4 (5)
Accent type (location of accent nucleus)	4 (5)
BDC at the boundary immediately before current *bunsetsu*	9
Pause immediately before current *bunsetsu*	2
Length of pause immediately before current *bunsetsu*	Continuous
Existence of phrase command for the preceding *bunsetsu*	2
Number of *morae* between preceding phrase command and head of current *bunsetsu*	4
Magnitude of current phrase command	Continuous
Magnitude of preceding phrase command	Continuous

Table 4. Input parameters for the prediction of differences in accent command amplitudes. The category numbers in parentheses are those for the directly preceding and proceeding prosodic words.

Input parameter	Category
Position in sentence of current prosodic word	3
Position in prosodic phrase of current prosodic word	3
Position of prosodic phrase to which the current prosodic word belongs	2
Distance from focal position (in number of prosodic word)	5
Accent type (location of accent nucleus)	4 (5)
BDC at the boundary immediately before current prosodic word	2
Amplitude of directly preceding accent command	Continuous
Amplitude of current accent command	Continuous
Magnitude of current phrase command	Continuous
Magnitude of preceding phrase command	Continuous

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. Controls of duration and amplitude are for future research. We are planning to apply the similar supplemental control of F_0 model commands for realizing various styles including emotional speech.

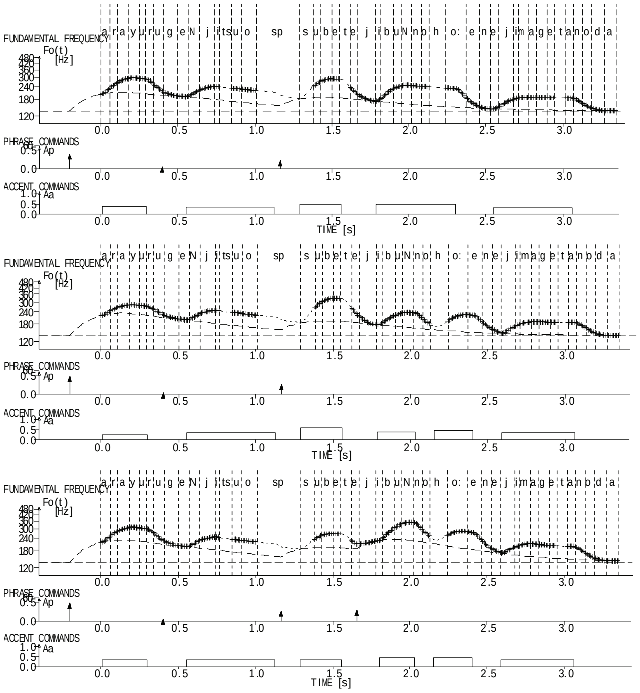

Figure 3. Generated F_0 contours and F_0 model parameters. The sentence and focal conditions are the same with those shown in Figure 2.

8. REFERENCES

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

[2] K. Hirose, K. Sato, Y. Asano, and N. Minematsu, "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).

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

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

[5] K. Ochi, K. Hirose, and N. Minematsu, "Control of prosodic focus in corpus-based generation of fundamental frequency based on the generation process model," *Proc. Interspeech*, p.1216 (2008).

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

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