Corpus-based Generation of Prosodic Features from Text Based on Generation Process Model

Keikichi Hirose[1], Keiko Ochi[2], & Nobuaki Minematsu[2]

[1] Dept. of Inf. and Commu. Engineering, School of Inf. Science and Tech.
[2] Dept. of Frontier Informatics, School of Frontier Sciences
University of Tokyo, Tokyo, Japan
{hirose, ochi, mine}@gavo.t.u-tokyo.ac.jp

Abstract

A total scheme of generating prosodic features from a text input was constructed. The method consists of corpus-based prediction of pauses, phone durations and fundamental frequencies (F_0's), in this order, and information predicted in an earlier process is utilized in the following processes. Since prediction of F_0's is done on the command values of F_0 contour generation process model instead of direct F_0 values, a stable and flexible control of F_0 contours is possible. By adding constraints on the accent command timings as a post processing, a better quality was realized when speech was synthesized using prosodic features generated by the method. Validity of the developed method was confirmed through the listening test of the synthetic speech.

Index Terms: speech synthesis, prosodic features, F_0 contour

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 at least a whole sentence. This situation comes more serious when we try to realize various styles in synthetic speech, where large variations may observable in prosodic features.

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 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 includes some problems. Frame-by-frame processing 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 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 [3-5]. 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 [6]. 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. Also, it is rather easy to analyze the prosodic controls obtained by statistical methods and to modify generated F_0 contours according to our knowledge obtained through observations of natural utterances.

Although a rather good quality was realized in the synthetic speech with generated F_0 contours by the method, other prosodic features, such as pauses and phone durations, were copied from the target speech. In the current paper, a total method of generating prosodic features in a corpus-based way from text was realized. In the method, pauses, phone durations, and F_0 contours were generated in this order, and the generated features were used for the prediction of other features.

A constraint was applied when predicting command timings of the F_0 model. A better quality of speech due to this constraint indicates the above-mentioned advantage of using the F_0 model: problems in the predicted results can be analyzed and be modified according to our knowledge.

The following sections are organized as follows: After describing the total flow of the generation of prosodic features in section 2, prediction of pauses, phone durations, F_0 contours are explained in sections 3, 4, 5, respectively. Effects of adding timing constraints during F_0 model command prediction process is shown in section 6. The developed method is evaluated through listening test of synthetic speech in section 7. Section 8 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 consisting of content word(s) followed or not followed by particles. It also serves as a basic unit of utterance: a *bunsetsu* is mostly uttered in an accent component. 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. This is because an insight is possible for the constructed trees. The CART (Classification And Regression Tree) included in the Edinburgh Speech Tools Library [7] 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 [8].

3. Prediction of pauses

Pauses play an important role in conveying utterance structure and are tightly related to syntactic structure of the sentence. Also occurrence of a pause is a function of the distance from the preceding pause. Taking these into account, input parameters shown in Table 1 are selected as input parameters of BDT's for pause prediction. As for the "relation between preceding and current *bunsetsu*'s," it is one of the following three: parallel noun phrases, parallel predicates, or non-parallel. Two methods are compared; method 1 first predicts pause durations for all the *bunsetsu* boundaries and assumes as no pauses when the duration is less than a threshold, while method 2 first predicts pause locations and then predict durations for the predicted pauses. The threshold was set to 0.13 sec for the method 1 through observations of the distribution of pause durations.

Table 1. *Input parameters for the pause prediction. Bunsetsu immediately after the bunsetsu boundary in question is denoted as current bunsetsu.*

Input parameter	Category
Number of *morae* from preceding pause	32
Number of *morae* of sentence	49
Number of *bunsetsu*'s from sentence initial to *bunsetsu* boundary in question	12
Number of *morae* of preceding *bunsetsu*	16
Part-of-speech of morpheme at the end of preceding *bunsetsu*	12
Case of particle at the end of preceding *bunsetsu*	13
Relation between preceding and current *bunsetsu*'s	3
Type of morpheme at the end of preceding *bunsetsu*	33
Number of *morae* of current *bunsetsu*	21
Part-of-speech of morpheme at the top of current *bunsetsu*	11
Conjugation form of morpheme at the top of current *bunsetsu*	14
Boundary depth code (BDC)	9

Speech material used for the experiment is utterances of a male narrator for the ATR 503 sentences recorded in 10 kHz sampling rate and 16-bit accuracy. Utterances of 50 sentences are reserved for the testing and the rest are used for training the BDT predictors. Total numbers of *bunsetsu*'s are 2,193 and 226 for training and testing data, respectively. They include 542 and 40 pauses, respectively. The speech material is also used for experiments on prediction of phone durations and F_0 model parameters, and speech synthesis.

Table 2 shows the result of pause prediction. The detection rate is defined as: (number of *bunsetsu* boundaries with correct pause/non-pause judgment)/(total number of *bunsetsu* boundaries). Since a better detection rate is obtained by the method 2, in the following experiments on phone duration and F_0 model parameter prediction, results of method 2 are adopted.

Table 2. *Result of pause prediction.*

	Training		Testing	
	Method		Method	
	1	2	1	2
Detection rate (%)	85.5	85.5	86.8	87.7
Number of pauses correctly detected	394	371	29	28
Average length of pauses correctly detected (sec)	0.37	0.37	0.30	0.31
Root mean square error of predicted pause length (sec)	0.20	0.17	0.19	0.17

4. Prediction of phone durations

Table 3. *Input parameters for the prediction of phone durations.*

Input parameter	Category
Category of current (preceding/following) phones	7 (9)
Part-of-speech of current morpheme	13
Part-of-speech of morpheme at the top (end) of current *bunsetsu*	12 (12)
Conjugation form of the initial (last) word in current *bunsetsu*	14 (9)
Accent type of current (preceding) *bunsetsu*	18 (15)
Position in sentence of current *bunsetsu*	13
Belongs to sentence initial, final, or other breath groups.	3
Number of *morae* of current (preceding) *bunsetsu*	22 (17)
Conjugation form of current morpheme	20
Boundary depth code (BDC) of the boundary immediately after current *bunsetsu*	10
Number of words of current (preceding) *bunsetsu*	12 (8)
Length of preceding (following) pause	Continuous
Number of *morae* from preceding (to following) pause	39 (39)
Number of *morae* between preceding and following pauses	33

A rather large number of works have already been reported on corpus-based prediction of phone durations, where the possible correlates on phone duration were well discussed [9]. Following to these works, input parameters for phone durations are selected as shown in Table 3. Taking the limitation on the amount of training corpus, phones are categorized into 7 groups according to their manner of articulation. For preceding/following phones, the category number of phone group is 9 (larger by 2 than that for the

current phone to represent sentence initial/end and pause). Since pause position is tightly related phone durations, such as longer durations before pauses, information on predicted pauses is also included in the input parameters.

Assuming phone boundaries detected by the forced alignment as correct ones, the binary decision trees were trained and predicted durations were checked. Table 4 shows the evaluation result.

Table 4. *Result of prediction of phone durations.*

	Training	Testing
Average "correct" phone duration (sec)	0.070	0.069
Root mean square error of predicted phone duration (sec)	0.026	0.027

5. Prediction of F_0 model parameters

As for F_0 model parameters, phrase command parameters are first predicted, followed by accent command parameter prediction [5].

Table 5. *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 5). 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:

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

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.

6. Constraints on F_0 model parameters

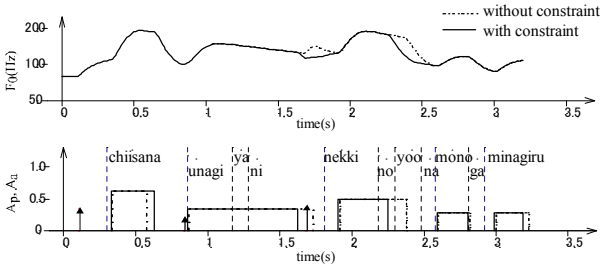

Figure 1. *F_0 contours and F_0 model commands predicted without constraints (dashed line) and with constraints (solid line) on accent command timings. The content of the utterance is: "chiisana unagiyani nekkinoyoonamonoga minagiru (Something like hot moisture spreads in a small eel restaurant.)."*

Table 6. *Distances (msec) of accent command onsets and resets from their corresponding segmental boundaries.*

	Type 1	Type 2	Type 0	Others
Onset	-15.8	-78.3	-75.8	-74.5
Reset	55.8	19.1	-30.2	6.7

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, as shown in Fig. 1 (dashed line). These 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 as shown in Fig. 1 (solid line). A close observation was already conducted for a male announcer's speech from a radio program on the positions of accent command onsets and resets [10]. Table 6 shows the result as distances from the corresponding segmental boundaries (reference points) as averages for each accent type. Here, the reference point for onset is the initial of the second *mora* for accent types other than type 1, and is the initial of the first *mora* for type 1. For reset, it is the end of *mora* with accent nucleus (for type 0 accent without accent

nucleus, the end of *mora* at the end of prosodic word). In the current experiments, accent command timings were constrained in ± 50 msec range centered with values listed in Table 6. The average F_0MSE values of generated F_0 contours in section 5 were reduced to 0.037 and 0.041 for training and testing utterances, respectively.

7. 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 7. 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 [11]. Tri-phone models were trained using the 453 sentence utterances used for the training of the prosodic feature predictors. The segmental features were 75^{th} order vectors consisting of 0^{th} to 24^{th} cepstrum coefficients and their Δ and Δ^2 values.

Table 7. 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

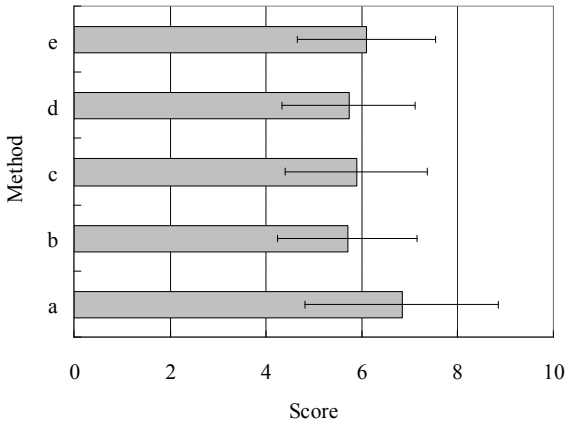

Figure 2. *Result of listening experiment.*

Result of listening test is shown in Fig. 2 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, though considerations are still necessary for the content of constriction. This kind of "empirical" correction comes possible only when the method is based on a quantitative modeling with clear relations with linguistic information.

8. Conclusion

A total scheme of generating prosodic features was developed. All the prosodic features are predicted in a corpus-based way. Since prediction of F_0 contours was conducted in the framework of F_0 model, the method has a merit that the generated F_0 contour can be manipulated to realize a better quality and/or a wider style variation. Result of the listening test of synthetic speech showed the validity of the developed method.

For the future work, we are planning to apply the method for the generation of prosodic features for various styles of speech including emotional ones.

9. 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*, Geneva, 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*, Phoenix, 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*, Philadelphia, Vol.1, pp.378-381 (1996).

[4] Sakurai, A., Hirose, K., and Minematsu, N., "Data-driven generation of F_0 contours using a superpositional model," *Speech Communication*, Vol.40, No.4, pp.535-549 (2003).

[5] 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).

[6] 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).

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

[8] 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 (2002-5).

[9] Sagisaka, Y., "Corpus based speech synthesis," *J. of Signal Processing*, Vol.2, No.6, pp.407-414 (1998-11).

[10] 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. Oriental COCOSDA*, Taipei, pp.38-45 (2003-10).

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