CORPUS-BASED GENERATION OF F_0 CONTOURS USING GENERATION PROCESS MODEL FOR EMOTIONAL SPEECH SYNTHESIS

Keikichi Hirose [1], Kentaro Sato [1] & Nobuaki Minematsu [2]

[1] Graduate School of Frontier Sciences, University of Tokyo, Tokyo, Japan
[2] Graduate School of Inf. Science & Tech., University of Tokyo, Tokyo, Japan
hirose@gavo.t.u-tokyo.ac.jp

ABSTRACT

A corpus-based method was developed for generating fundamental frequency contours in emotional speech synthesis. The method assumes the generation process model and predicts its command parameters (positions and amplitudes) using binary regression trees with the input of linguistic information of the sentence to be synthesized. Because of the model constraint, a certain quality is still kept in synthesized speech even if the prediction is done incorrectly. The speech corpus includes three types of emotional speech (anger, joy, sadness) and calm speech uttered by a female narrator. The command parameters necessary for the training and testing of the method were automatically extracted from speech using a program developed by the authors. Since the accuracy of the extraction largely affects the prediction performance, a constraint is applied on the position of phrase commands during the extraction. The method first predicts the phrase command parameters, which are then used for the prediction of accent command parameters. The mismatches between the predicted and target contours for angry speech were similar to those for calm speech. Synthesis of emotional speech was conducted with text inputs. The segmental features were handled by the HMM synthesis method and the phoneme durations are predicted in a similar corpus-based method. Perceptual experiment was conducted using the synthesized speech, and the result indicated that the anger could be well conveyed by the developed method. The result came worse for joy and sadness.

INTRODUCTION

Introduction of corpus-based methods to speech synthesis largely improved the speech quality. However, most speech synthesizers only offer prosodically poor sounds: in monotonous reading style. This is one of the major reasons preventing synthetic speech to be used in human and machine interfaces. To cope with this situation, a scheme enabling to synthesize various speaking styles is necessary. As an example of various speaking styles, emotional speech is selected here as our research target.

A full corpus-based synthesis has already been realized for emotional speech using the ATR selection-based speech synthesis engine, CHATR (Iida et al., 2002). However, in the framework of CHATR, the precise control of prosodic features is rather difficult. HMM-based speech synthesis, where all the acoustic features including F_0 are handled in frame-by-frame basis, was applied to emotional speech synthesis with a certain success (Yamagishi et al., 2003; Tsuduki et al., 2003). However, prosodic features cover a wider time span than segmental features, and, generally speaking, to model frame-by-frame F_0 movement is not a good idea.

From these considerations, we already have developed a corpus-based synthesis of F_0 contours (Hirose et al., 2001) in the framework of the generation process model (Fujisaki & Hirose, 1984; henceforth F_0 model). The model assumes two types of commands, phrase and accent commands, as model inputs. By predicting the model commands instead of F_0 values, a good constraint will automatically applied on the synthesized F_0 contours; still keeping acceptable speech quality even if the prediction is done incorrectly. Although currently no constraints are applied on model commands, they are possible, such as on command timings.

Prediction of the command values is conducted using binary decision trees: one tree for one model parameter. To train the trees, speech corpuses, which contain the model command information, are necessary. In the previous reports for read and emotional speech synthesis (Hirose et al., 2003; Hirose et al., 2004), these corpuses were prepared automatically from speech data using a method of automatic extraction of F_0 model parameters, which was developed by the authors (Narusawa et al., 2002). Although favorable results were obtained, there were often cases where the predicted model commands were not consistent with our knowledge on the commands. For instance, there were cases where phrase commands located inside the accent commands, which were not allowed in the F_0 model.

The major reason of the wrong prediction is that the automatic extraction method of F_0 model commands does not work well for some of the speech samples in the training corpuses. To cope with the wrong prediction, we newly introduced a constraint on the phrase command locations for the better extraction and developed a scheme of text-to-speech conversion, which fitted to the command extraction scheme. We also conducted a speech synthesis experiment and evaluated the synthetic speech through a perceptual experiment. The segmental features for synthetic speech were generated by an HMM-based method.

PROSODIC CORPUS

Emotional speech corpus used for the experiment is the recordings of utterances by a female narrator. It includes 3 types of emotional speech, anger, joy, sadness, and calm speech. She was asked to read several hundreds of sentences, which were prepared for each type of emotion as a written text. The sentences for calm speech are the 503 sentences used for the ATR continuous speech corpus, while those for emotional speech are newly prepared for each emotion type so that the speaker can properly include the emotion in her utterances. An informal listening test was conducted for all the samples to exclude those without designated emotion from the experiment. Then, the remained samples were gone through the following process to obtain a prosodic corpus.

1. Phoneme labels and speech sounds were time-aligned through the forced alignment using the speech recognition software Julius.

2. From the content (text) of each utterance, its morphemes and part-of-speech information were obtained using the Japanese parser Chasen (Matsumoto, 2000). Another parser KNP was used to obtain *bunsetsu* boundaries and their syntactical depths. Here, *bunsetsu* is defined as a basic unit of Japanese grammar and pronunciation, and consists of a content word (or content words) followed or not followed by a function word (or function words). The result of KNP analysis is given as KNP codes, which indicate the *bunsetsu* that the current *bunsetsu* directly modifying.

3. For the F_0 contour extracted from the speech waveform, F_0 model parameters were estimated using the model parameter extraction method developed by the authors (Narusawa et al., 2002). To increase the accuracy of extraction, a constriction was added to the location of the phrase command; a phrase command should locate before a *bunsetsu* boundary. Also, two succeeding accent commands locating close to each other with similar amplitudes were merged.

4. Each *bunsetsu* boundary was checked if it is also a prosodic word boundary according to the accent command information obtained in the above process. If a *bunsetsu* boundary locates between two accent commands, it is also a prosodic word boundary. If no *bunsetsu* boundary locates, the last morpheme boundary between the two commands is assumed to be a prosodic word boundary. Here, prosodic word is defined as a *bunsetsu* or a sequence of *bunsetsu*'s which contains an accent command.

5. For each prosodic word thus obtained, an accent type was assigned by referring to the accent type dictionary. The dictionary has accent type and attribute information, and, using a system developed by the authors (Minematsu et al., 2003), the accent type of each prosodic word can be decided automatically.

After the above processes, around 400 sentence samples with prosodic labeling (F_0 model command information) were obtained for each emotion, which were divided into two groups to be used for the training and testing of the methods (50 sentences for testing and the rest of the sentences for training).

F_0 CONTOUR GENERATION

In our original method, prediction of F_0 model parameters is done for each accent phrase, and a sentence F_0 contour is generated using the F_0 model after the prediction process is completed for all the constituting accent phrases. The method was modified to fit to the processes of corpus preparation (henceforth, the new method). From the text, the following four processes are necessary before the F_0 contour generation:

1. Prediction of phrase command: each *bunsetsu* boundary is judged whether it is accompanied by a phrase command or not. If yes, the magnitude of the command is predicted also.

2. Prediction of prosodic word boundary location: each morpheme boundary is judged whether it is also a prosodic word boundary or not.

3. Decision of accent types: for each prosodic word, an accent type was assigned using the same process as process 5 in the previous section.

4. Prediction of accent command: for each prosodic word, an accent command was predicted.

The processes 1, 2 and 4 are done using a scheme based on binary decision trees (BDT's).

Phrase Command Prediction

The input parameters for BDT of phrase command prediction were selected as shown in Table 1. Besides the features of the current *bunsetsu* in question and those of directly preceding *bunsetsu*, boundary depth code (BDC) between the two *bunsetsu*'s is added. The category

numbers, shown in the parentheses, are those for the preceding *bunsetsu* and are larger than those of the corresponding parameters of the current *bunsetsu* by one to represent "no preceding *bunsetsu*." The *bunsetsu* boundary was obtained by KNP. No manual correction was added.

Table 1. Input parameters for phrase command prediction. The category numbers in the parentheses are those for the directly preceding *bunsetsu*.

Input parameter	Category
Position in sentence	28
Number of *morae*	21 (22)
Accent type (location of accent nucleus)	18 (19)
Number of words	10 (11)
Part-of-speech of the first word	14 (15)
Conjugation form of the first word	19 (20)
Part-of-speech of the last word	14 (15)
Conjugation form of the last word	16 (17)
Boundary depth code (BDC)	20
Phrase command for preceding *bunsetsu*	2
Number of morae between the preceding phrase command and the head of the current *bunsetsu*	25
Magnitude of the preceding phrase command	Continuous

As for the output parameters, besides magnitudes and timings of the phrase commands, a binary flag indicating the existence/absence of a phrase command at the head of the *bunsetsu* is selected, like the case of the original method.

Prediction of Prosodic Word Boundary

Adding to linguistic information of current and directly preceding morphemes, phrase command information predicted in the 1st process is used as the input parameters for the prosodic word boundary prediction. The output parameter is the binary flag indicating whether the boundary between current and preceding morphemes is a prosodic word boundary or not. Around 85 percent of correct prediction was obtained for all emotions.

Prediction of Accent Command

Similar parameters as the phrase command prediction were selected as input parameters for accent command prediction. The output parameters are amplitudes and timings of the accent commands.

Result

Figure 1 shows an F_0 contour for "zeikiNno mudazukaida yametekure (Stop to waste taxes.)" generated using model commands predicted by the proposed method. Although amplitudes for

the second and the third accent commands are slightly under-estimated, a counter close to that of natural utterance (target contour) is obtained.

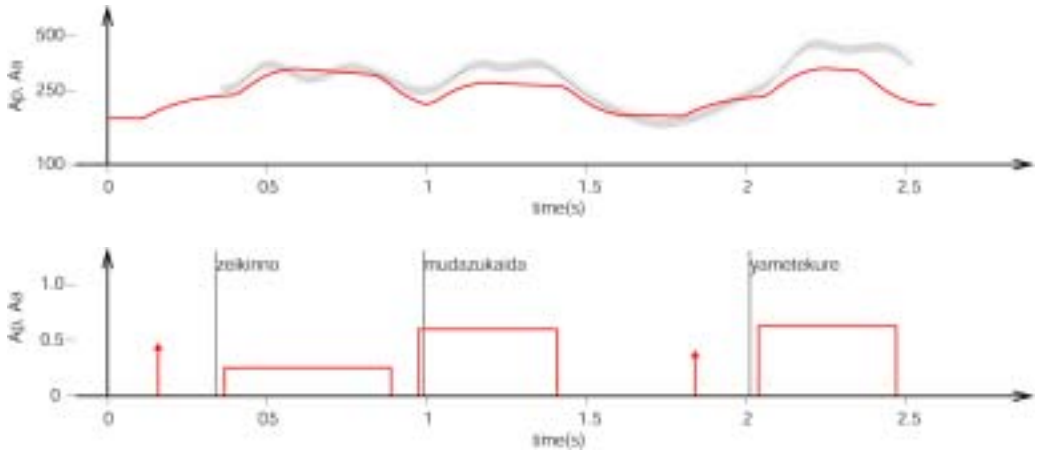

Fig. 1. Model commands (bottom) predicted by the proposed method and F_0 contour generated using them (top).

SPEECH SYNTHESIS AND EVALUATION

Using the new method for F_0 model command prediction, speech synthesis from text was conducted for the 3 types of emotional speech and the calm speech. Segmental duration necessary of the synthesis was predicted in a similar way as the command prediction. Segmental features were generated using the HMM-based speech synthesis toolkit. Tri-phone models were trained for each type of emotion using the same sentences used for the BDT training. The segmental features were 75th order vectors consisting of 0th to 24th cepstrum coefficients and their Δ and Δ^2 values. The sampling frequency, the frame period, and the frame length were set to 16 kHz, 5 ms, and 25 ms, respectively.

Table 2. Scores for the realization emotion and naturalness of prosody.

	Anger		Joy		Sadness	
	Original	New	Original	New	Original	New
Realization	4.01	4.21	3.26	3.36	3.07	3.12
Naturalness	2.06	2.48	1.76	1.90	1.61	2.32

Ten sentences were randomly selected from the test sentences for each type of emotion and were used for the evaluation. For comparison, speech synthesis was also conducted using F_0 contours from model commands predicted by the original method. The synthesized speech was presented to 9 Japanese speakers, who were asked to select one from the four types (calm, angry, joyful, sad) for each sample. Although the better selection of designated emotion was realized in the new method as compared to the original method, the rates were still rather low for joy and sadness. They were also asked to rank the samples, for which type selection were correct; how well they can perceive the emotion designated for each sample (5: quite well, 3:

marginal, 1: poor) and how they evaluate naturalness of prosody (5: natural, 3: somewhat, 1: very synthetic). Table 2 shows the result. As for the realization of designated emotion, a good result was obtained for anger, but the results were slightly worse for joy and sadness. However, for naturalness, the scores were low for all the cases. The HMM synthesis may partly responsible for this.

CONCLUSIONS

A new corpus-based method of generating F_0 contours for emotional speech from text was developed. Perceptual experiments for synthetic speech showed that the designated emotions could be conveyed with the F_0 contours generated by the newly developed method better than with those generated by our original method.

The authors' sincere thanks are due to Hiromichi Kawanami, Nara Institute of Science and Technology for providing emotional speech database.

REFERENCES

Fujisaki, H. & Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese, *J. Acoust. Soc. Japan 5 (4)*, 233-242.

Hirose, K., Eto, M., Minematsu, N. & Sakurai, A. (2001). Corpus-based synthesis of fundamental frequency contours based on a generation process model, *Proc. EUROSPEECH, Aalborg*, 2255-2258.

Hirose, K., Ono, T. & Minematsu, N. (2003). Corpus-based synthesis of fundamental frequency contours of Japanese using automatically-generated prosodic corpus and generation process model, *Proc. EUROSPEECH, Geneva*, 333-336.

Hirose, K., Sato, K. & Minematsu, N. (2004). Emotional speech synthesis with corpus-based generation of F_0 contours using generation process model, *Proc. International Conference on Speech Prosody, Nara*, 417-420, 2004.

Iida, F. Higuchi, N. Campbell & Yasumura, A. (2002). Corpus-based speech synthesis system with emotion, *Speech Communication 40 (1-2)*, 161-187.

Minematsu, N., Kita, R. & Hirose, K. (2003). Automatic estimation of accentual attribute values of words for accent sandhi rules of Japanese text-to-speech conversion, *IEICE Trans. Information and Systems E86-D (3)*, 550-557.

Matsumoto, Y. (2000). Morpheme analysis system "Chasen," *IPSJ Magazine 41 (11)*, 1208-1214. (in apanese)

Narusawa, N., Minematsu, N., Hirose, K. & Fujiaski, H. (2002). A method for automatic extraction of model parameters from fundamental frequency contours of speech, *Proc. ICASSP, Orlando*, 509-512.

Tsuduki, R., Zen, H., Tokuda, K, Kitamura, T., Bulut, M. & Narayanan, S. (2003). A study on emotional speech synthesis based on HMM, *Record of Fall Meeting, Acoust. Soc. Japan*, 241-242. (in Japanese)

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