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

Abstract

In our fully automatic corpus-based method of generating fundamental frequency (F_0) contours for emotional speech synthesis, an improvement was realized related to the process of corpus preparation. The method assumes the generation process model and predicts its command parameters using binary regression trees with inputs 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 newly applied on the position of phrase commands during the extraction. Also, since performance of phrase command prediction dominates the overall accuracy of generated F_0 contours, the method was modified to predict phrase commands first. 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.

1. Introduction

Introduction of corpus-based methods to speech synthesis largely improved the quality of synthetic speech. However, most speech synthesizers only offer speech in monotonous reading style, and this poorness in prosody 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 [1]. However, emotional speech often shows a rather large dynamic range in the movement of fundamental frequency (henceforth, F_0) and its synthesis requires a precise control of prosodic features, which is rather difficult in the framework of CHATR. 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 [2, 3]. 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 [4, 5] in the framework of the generation process model [6]. 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.

The command values are predicted 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 [7, 8], 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 [9]. 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.

2. Prosodic Corpus

Utterances by a female narrator recorded at Nara Institute of Science and Technology include 3 types of emotional speech, anger, joy, sadness, and calm speech. These utterances are not spontaneous ones; the speaker read several hundreds of sentences, which were prepared for each type 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 recognizer *Julius* [10].
2. From the content (text) of each utterance, its morphemes and part-of-speech information were searched using the Japanese parser *Chasen* [11]. Another parser KNP [12] 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, indicating the *bunsetsu*, which are directly modified by the current *bunsetsu*.
3. For the F_0 contour extracted from the speech waveform, F_0 model parameters were estimated using the model parameter extraction method [8]. To increase the accuracy of extraction, a constraint was added to the location of the phrase command; a phrase command should locate before a *bunsetsu* boundary. If the *bunsetsu* boundary is accompanied by a pause, the phrase command should be located in the period 300 ms to 100 ms before the *bunsetsu* boundary, and if not, in the period 100 ms to 0 ms before the 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 (for each word), and, using a system developed by the authors [13], the accent type of each prosodic word can be decided automatically.

Table 1: Number of samples used for the experiment.

Type	Category	Number	
		Sentence	Prosodic word
Calm	Training	333	2340
	Testing	50	338
Anger	Training	472	3247
	Testing	50	346
Joy	Training	358	2391
	Testing	50	271
Sadness	Training	305	2185
	Testing	50	389

The constraint on the phrase command location in the 3rd process may cause some errors if sentences include long compound words, where phase commands can locate in *bunsetsu*. However, this is not the case in the speech corpuses used in the experiment.

After the above processes, around 400 sentence samples with prosodic labels (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 as shown in Table 1.

3. F_0 Contour Generation

In our original method [7, 8], 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. Therefore, given a text, accent phrase boundary detection was conducted before the F_0 model command prediction. In the current paper, method was modified to fit to the processes in section 3 (henceforth, the new method). From the text, the following four processes were conducted:

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 section 3.
4. Prediction of accent command: for each prosodic word, an accent phrase was predicted.

The processes 1, 2 and 4 are done using a scheme based on binary decision trees (BDT's). The CART (Classification And Regression Tree) method included in the Edinburgh Speech Tools Library [14] was utilized to construct BDT's. In the following sub-sections, the processes 1 and 4 are explained in detail.

3.1. Phrase Command Prediction

Table 2: 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

Table 3: Result of phrase command flag *PF* prediction. (in %)

	Closed	Open
Calm	67.4	64.9
Anger	69.0	66.1
Joy	66.1	63.2
Sadness	74.5	74.8

The input parameters for BDT of phrase command prediction were selected as shown in Table 2. Besides the features of the current *bunsetsu* in question and those of directly preceding *bunsetsu*, boundary depth code (BDC) between the two *bunsetsu*'s, which can be derived from KNP codes with a simple calculation, 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*."

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

Table 3 shows the correct prediction rate of *PF*. The rates are slightly higher for sadness than for other cases.

3.2. Accent Command Prediction

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. Table 4 shows root mean square errors of accent command amplitude prediction. Rather low values for sadness are mostly because of the smaller command amplitudes than other cases. Since, in our former analyses of sentence F_0 contours, accent command amplitudes and phrase command magnitudes showed negative correlation and the preceding accent command position and amplitude influenced the current accent command amplitude, parameters indicating the phrase command and preceding accent command were added to the input parameters. However, contrary to our expectation, these parameters had shown no effect on the prediction accuracy. The merit of new parameters will be cancelled by their prediction errors. Further research should be conducted to clarify this point.

Table 4: Root mean square errors of accent command amplitude prediction.

	Closed	Open
Calm	0.158	0.170
Anger	0.162	0.181
Joy	0.153	0.130
Sadness	0.112	0.127

3.3. Result

Figure 1 shows F_0 contours generated using model commands predicted by the original and new methods for "zeikiNno mudazukaida yametekure (Stop to waste taxes.)" Clearly the F_0 contour closer to the target counter is generated when the commands estimated by the new method are used.

As an objective measure to totally evaluate the predicted F_0 model parameters, mean square error between the F_0 contour generated using the predicted parameters and that of the target by the model 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. The results are summarized in Table 5, where F_0MSE values are averaged over all the training and testing sentences for closed and open cases, respectively. The best result was obtained for sad speech. Again this is due to the rather low command values in sadness.

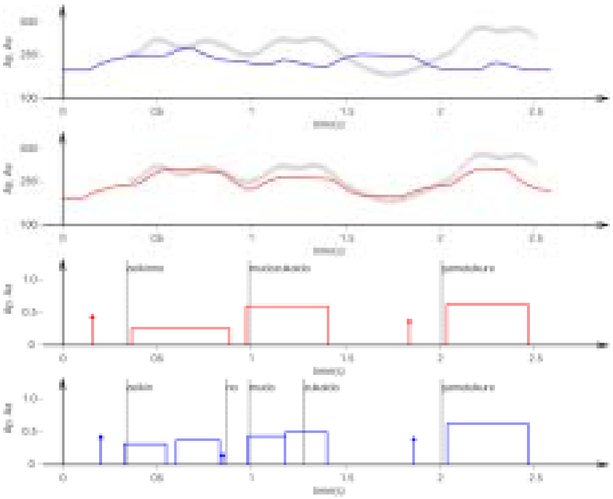

Figure 1: F_0 contours generated by the original method (1st panel) and the new method (2nd panel). 3rd and the 4th panels show the model commands predicted by the new method and those extracted for the target F_0 contour, respectively.

Table 5: Average F_0MSE's of F_0 contours generated using the predicted model parameters.

	Closed	Open
Calm	0.049	0.048
Anger	0.051	0.065
Joy	0.052	0.078
Sadness	0.035	0.043

4. 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 for the synthesis was predicted in a similar way as the command prediction. Segmental features were generated using the HMM-based speech synthesis toolkit [15]. Tri-phone models were trained for each type of emotion using the training sentences shown in Table 1. 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.

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 for F_0 contours generated using 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. The result is shown in Table 6. 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 7 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.

Table 6: Percentages showing how correctly the designated emotion (anger, joy, sadness) in synthetic speech is perceived. The italic numbers indicate the percentages when the designated emotion is perceived correctly. "Ori." indicates the results when the commands predicted by the original method are used, while "New" indicates those predicted by the new method are used. The results are averaged over all 10 sentences and 9 speakers for each emotion.

	Anger		Joy		Sadness	
	Ori.	New	Ori.	New	Ori.	New
Calm	10.0	7.8	30.0	23.3	32.2	26.7
Anger	*78.3*	*83.3*	11.1	10.0	11.7	10.0
Joy	6.1	5.6	*56.7*	*57.8*	11.7	7.8
Sadness	5.6	3.3	2.2	8.9	*44.4*	*55.6*

Table 7: Scores for the realization of the designated emotion and naturalness of prosody.

	Anger		Joy		Sadness	
	Ori.	New	Ori.	New	Ori.	New
Degree	4.01	4.21	3.26	3.36	3.07	3.12
Quality	2.06	2.48	1.76	1.90	1.61	2.32

5. Conclusion

A corpus-based method of generating F_0 contours of emotional speech from text was developed. With a text input, the method generates F_0 contours through prediction of phrase command, prediction of prosodic word boundary location, decision of accent types, and prediction of accent command. 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.

Although the developed method is a corpus-based one, systematical modification is possible to the generated F_0 contours by manipulating F_0 model commands. This flexibility will be one of the major merits of using the F_0 model. By generating F_0 contours for emotional speech and calm speech, and by comparing them for the same sentences, we can build up rules such as to include emotions in a speaker's normal readings. For most people, it is not so easy to realize various styles in their speech. The above procedure will make it possible for ordinary persons to control their speaking styles as professionals do.

6. Acknowledgements

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

7. References

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

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

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

[4] Sakurai, A., Hirose, K. and Minematsu, N., "Data-driven generation of F0 contours using a superpositional model," *Speech Communication* 40 (4), pp. 535-549, 2003.

[5] Hirose, K., Eto, M., Minematsu, N. and Sakurai, A., "Corpus-based synthesis of fundamental frequency contours based on a generation process model, " *Proc. EUROSPEECH, Aalborg*, 2255-2258, 2001.

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

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

[8] Hirose, K., Sato, K. and Minematsu, N., "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.

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

[10] Julius, Open Source real-time large vocabulary speech recognition engine. http://julius.sourceforge.jp/

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

[12] Kyoto University, Japanese Syntactic Analysis System KNP http://www-nagao.kuee.kyoto-u.ac.jp/projects/nl-resource/.

[13] Minematsu, N., Kita, R., and Hirose, K., "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), pp. 550-557, 2003.

[14] Edinburgh University, The Edinburgh Speech Tools Library, http://www.cstr.ed.ac.uk/projects/speeech_tools/.

[15] Galatea Project, http://hil.t.u-tokyo.ac.jp/~galatea/regist-jp.html