CORPUS-BASED SYNTHESIS OF FUNDAMENTAL FREQUENCY CONTOURS WITH VAROUS SPEAKING STYLES FROM TEXT USING F_0 CONTOUR GENERATION PROCESS MODEL

Keikichi Hirose, Kentaro Sato* and Nobuaki Minematsu***

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

ABSTRACT

A corpus-based method of generating fundamental frequency (F_0) contours of various speaking styles from text was developed. Instead of directly predicting F_0 values, the method predicts command values of the F_0 contour generation process model. Because of the model constraint, the resulting F_0 contour keeps certain naturalness even when the prediction is done incorrectly. The method includes a scheme of automatic extraction of the model commands, which is necessary to prepare the training corpuses for various speaking styles. By introducing constraints on phrase command locations, a better extraction was realized, led to a better performance of the method. Speech synthesis was conducted using HMM speech synthesizer for calm speech and three types of emotional speech. The perceptual experiment showed the designated emotions could be well conveyed with the F_0 contours generated by the developed method.

1. INTRODUCTION

Due to recent advancement of speech recognition and synthesis technology, a number of dialogue systems have been developed. However, the speech outputs from these systems are mostly in reading style and monotonous, which sometimes discourages people to use the systems. A technology enabling speech synthesis with various speaking styles is required. As an example of various speaking styles, emotional speech is targeted in the current paper.

Current speech synthesis methods are mostly those of corpus-based. Surely, when experts carefully arrange synthesis rules, a high quality, hard to be surpassed by corpus-based methods, can be realized in synthetic speech. This is especially true for fundamental frequency (F_0) contours, for which several models capable of closely approximating natural F_0 movements have been developed already. However, developing synthesis rules for various speaking styles is a time-consuming process and even impossible if the expert's knowledge on the styles is limited. Therefore, a rather large number of researchers try to generate prosodic features from linguistic inputs using statistical methods, such as neural networks, binary decision trees and so on.

In corpus-based methods for F_0 contour generation, F_0 movements can be directly related to linguistic information of the input texts. An HMM-based method successfully generated synthetic speech with highly natural prosodic features by counting F_0 delta features [1]. These methods without F_0 model constraints theoretically can generate any type of F_0 contours [2], but have possibility of causing un-naturalness especially when the training data are limited. Several methods are reported under the ToBI labeling scheme. Constraints by the ToBI system are beneficial in avoiding unlikely F_0 contours being generated. The major problem of ToBI system is that it is not a full quantitative description of F_0 contours, which causes some limitations to the quality of synthesized F_0 contours.

From these considerations, we already have developed a corpus-based synthesis of F_0 contours in the framework of the generation process model (henceforth F_0 model) [3, 4]. The model assumes two types of commands, phrase and accent commands, as model inputs, and these commands are proved to have a good correspondence with linguistic (and para-/non-linguistic) information of speech [5]. 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, since we already have knowledge on model commands, such as on command timing as compared to the segmental boundary locations.

Prediction of the command values was 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 [6, 7], 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 [8]. 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, in the current paper, 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 [9].

2. MODELING AND PARAMETRIC REPRESENTATION OF F_0 CONTOURS

The F_0 model is a command-response model that describes F_0 contours in logarithmic scale as the superposition of phrase and accent components. The phrase component is generated by a second-order, critically-damped linear filter in response to an impulse-like phrase command, while the accent component is generated by another second-order, critically-damped linear filter in response to a stepwise accent command. An F_0 contour is given by the following equation:

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} A_{aj} \{ G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j}) \} \quad (1)$$

In the equation, $G_{pi}(t)$ and $G_{aj}(t)$ represent phrase and accent components, respectively. F_b is the bias level, i is the number of phrase commands, j is the number of accent commands, A_{pi} is the magnitude of the ith phrase command, A_{aj} is the amplitude of the jth accent command, T_{0i} is the time of the ith phrase command, T_{1j} is the onset time of the jth accent command, and T_{2j} is the reset time of the jth accent command. The F_0 model also makes use of other parameters (time constants a_i and $ß_j$) to express functions G_{pi} and G_{aj}, but, in the current experiments, they are respectively fixed at 3.0 s^{-1} and 15.0 s^{-1} based on the former F_0 contour analysis results.

3. 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 recognition software Julius [10].
2. From the content (text) of each utterance, its morphemes and part-of-speech information were obtained 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, 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 [8]. 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. 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. Throughout the process, the bias level F_b was fixed to a value for each emotion type, which was calculated as the F_0 average of all the samples of the emotion type minus 3 standard deviations. The values were 147.67 Hz, 182.49 Hz, 210.30 Hz, and 182.49 Hz, for calm, angry, joyful, and sad speech, respectively.
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 that 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 [13], the accent type of each prosodic word can be decided automatically.

The constraint on the phrase command location in the 3rd process may cause some errors if sentences include long compound words, where phase commands are possible to be located 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.

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

4. PREDICTION OF F_0 CONTOURS

In our original method [6. 7], 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. 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 and the timing of the command are 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 is assigned using the same process as process 5 in section 3.
4. Prediction of accent command: for each prosodic word, the amplitude and the timings of an accent command are 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.

4.1. Prediction of phrase command

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 was 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. Figure 1 shows an example of parsing for the sentence "*arayuru geNjitsuo subete jibuNno hoHe nejimagetanoda* ([He] twisted all the reality to his side.)." In the example, one *bunsetsu* corresponds one accent phrase, and BDC's are obtained by simply shifting the distances rightward.

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

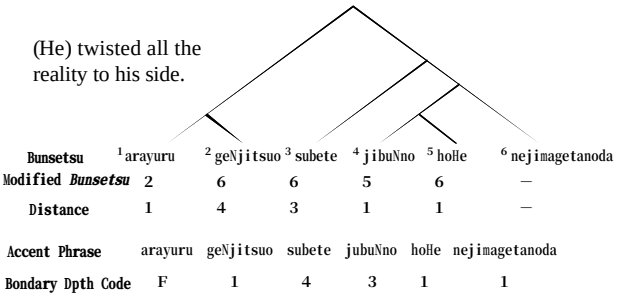

Fig. 1. Result of syntactic analysis by KNP and *bunsetsu* boundary depth codes.

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.

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

4.2. Prediction of prosodic word boundary

As shown in Table 4, linguistic information of current and directly preceding morphemes and phrase command information predicted in the 1st process are selected as input parameters for prosodic word boundary prediction.

Table 4: Input parameters for prosodic word boundary prediction. The category numbers in the parentheses are those for the directly preceding morpheme.

Input parameter	Category
Part-of-speech	15 (16)
Conjugation form	24 (25)
Conjugation type	35 (36)
Number of *morae*	9 (10)
Position in sentence	63
BDC of *bunsetsu* where the current morpheme belongs	22
Bunsetsu boundary at the head of the current morpheme	14 (15)
PF for current morpheme	2
Number of morae between the preceding phrase command and the head of the current morpheme	31
Magnitude of the preceding phrase command	Continu-ous

The output parameter is the binary flag indicating whether the boundary between current and preceding morphemes is a prosodic word boundary or not.

Table 5 summarizes the result. More than 80 percent of correct prediction was obtained for all the cases.

Table 5. Result of prosodic word boundary prediction. (in %)

	Closed	Open
Calm	88.5	87.7
Anger	87.0	83.7
Joy	86.7	85.3
Sadness	85.3	85.0

4.3. Prediction of accent command

Similar parameters as the phrase command prediction were selected as input parameters for accent command prediction as shown in Table 6. The output parameters are amplitudes and timings of the accent commands. Table 7 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 6. Input parameters for accent command prediction. The category numbers in the parentheses are those for the directly preceding prosodic word.

Input parameter	Category
Position in sentence	27
Number of *morae*	17 (18)
Accent type (location of accent nucleus)	16 (17)
Number of words	9 (10)
Part-of-speech of the first word	14 (15)
Conjugation form of the first word	23 (24)
Part-of-speech of the last word	14 (15)
Conjugation form of the last word	23 (24)
Boundary depth code	22

Table 7. 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

4.4. F_0MSE

Figure 2 shows F_0 contours generated using model commands predicted by the original method and new method for "*oyajiwa gaNkodakeredomo soNna ekohiikiwa senu otokoda* (Although my dad is a tough person, he never shows such a prejudice to any person.)" Clearly the F_0 contour closer to the target counter is generated when the commands estimated by the new method are used.

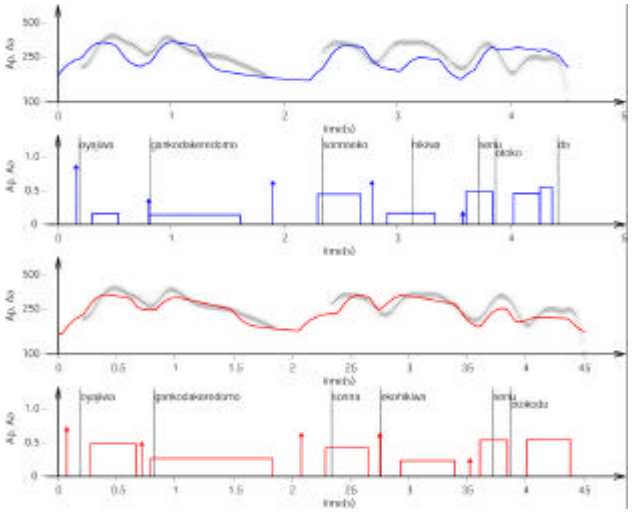

Fig. 2. F_0 contours generated by the original method (first panel) and the new method (third panel). The 2nd and the 4th panels show the model commands.

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:

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

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 8, 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.

Table 8. 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

5. 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 [7]. Segmental features were generated using the HMM-based speech synthesis toolkit [9]. 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 delta and delta2 values. The sampling frequency, the frame period, and the frame length were set to 16 kHz, 5 ms, and 25 ms, respectively.

Table 9. 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 10. 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

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. The result is shown in Table 9. 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 10 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.

6. CONCLUSIONS

A new corpus-based method of generating F_0 contours from text was developed for enabling various speaking styles in speech synthesis. 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. Experiments are planned for other styles of speech, including various levels of emotion. We are also trying to apply constraints to model parameters for better performances.

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

7. REFERENCES

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

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

[3] A. Sakurai, K. Hirose, and N. Minematsu, "Data-driven generation of F_0 contours using a superpositional model," *Speech Communication* 40 (4), pp. 535-549, 2003.

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

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

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

[7] K. Hirose, K. Sato, and N. Minematsu, "Emotional speech synthesis with corpus-based generation of F_0 contours using generation process model," *Proc. International Conference on Speech Prosody*, Nara, pp. 417-420, 2004.

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

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

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

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

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

[13] N. Minematsu, R. Kita, and K. Hirose, "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/.