Corpus-based Synthesis of F_0 Contours for Emotional Speech Using the Generation Process Model

Keikichi Hirose*, Toshiya Katsura*, and Nobuaki Minematsu**

*Dept. of Frontier Informatics, School of Frontier Sciences, University of Tokyo, Tokyo

**Dept. of Inf. and Commu. Engg, School of Inf. Science and Tech., University of Tokyo, Tokyo

E-mail: {hirose, katsuya, mine}@gavo.t.u-tokyo.ac.jp

ABSTRACT

A corpus-based generation of fundamental frequency (F_0) contours was realized for emotional speech synthesis. The method, originally developed for read speech, is to predict command values of the F_0 contour generation process model with the input of linguistic information of the sentence to be synthesized. Since the generated F_0 contour is under the model constraint, a certain quality is still kept in synthesized speech even if the prediction is done poorly. The speech corpus used for the F_0 contour generation experiments includes three types of emotional (anger, joy, sad) and calm speech uttered by a female narrator. The command values necessary for the training and evaluation of the method were automatically extracted using a program developed by the authors. We also applied the method to predict segmental durations. The mismatches between the predicted and target contours/durations for emotional speech were similar to those for calm speech.

1. INTRODUCTION

Recent advancement of multimedia interfaces between man and machine largely increased interests on realizing and recognizing emotions conveyed by speech. As for the realization, most of the works tried to build up prosodic control rules based on the analysis of emotional speech. When an expert carefully arranges these rules, the resulting synthetic speech can be in high quality and can convey the designated emotion. However, since prosodic features for emotional speech show large variations due to emotion type and speaker individuality, constructing good rules for prosodic feature control is not an easy task. Therefore, in view of the success of corpus-based methods in speech processing, we newly tried to generate prosodic features from linguistic inputs using a statistical method. (Control of segmental features is also important for the realization of emotion, but is not addressed in the current paper.)

Iida et. al. [1] have already realized a full corpus-based emotional speech synthesis using the ATR selection-based speech synthesis engine CHATR. However, in the framework of CHATR, the precise control of prosodic feature cannot be realized. Using corpus-based methods, we can directly model the fundamental frequency (F_0) of each frame. However, the methods without F_0 constraints theoretically can generate any type of F_0 contours, but have possibility of causing unnaturalness especially when the training data are limited. Also, 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 have developed a corpus-based synthesis of F_0 contours in the framework of the generation process model (henceforth F_0 model) [2]. 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 [3]. By predicting the model commands instead of F_0 values, a good constraint will automatically applied on the synthesized F_0 contours. The method was originally developed for read speech synthesis [4], and, in the current paper, was applied to emotional speech synthesis.

2. MODEL 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 called phrase command, and the accent component is generated by another second-order, critically-damped linear filter in response to a step function called accent command. An F_0 contour is given by the following equation:

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

In the equation, $G_{pi}(t)$ and $G_{aj}(t)$ represent phrase and accent components, respectively. F_{0min} 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 α_i and β_j) to express functions G_{pi} and G_{aj}, but, in the current experiments, they are respectively fixed at 3.0 s^{-1} and 20.0 s^{-1} based on the former F_0 contour analysis results.

ISBN 1-876346-48-5

3. PREDICTION OF F_0 MODEL PARAMETERS

3.1 STATISTICAL METHOD

In this paper, the binary decision tree (BDT) for predicting the model parameters was constructed using CART (Classification And Regression Tree) method included in the Edinburgh Speech Tools Library [5]. Stop threshold, represented by the minimum number of examples per a leaf node, was set to 40 according to the result of former experiments on read speech [6].

3.2. INPUT AND OUTPUT PARAMETERS

The input parameters for BDT were selected as shown in Table 1. In the method, prediction of the model parameters is done for each accent phrase, and therefore the first 9 parameters are those for the phrase in question. Taking into account that the F_0 contour of an accent phrase being influenced by that of preceding phrases, the 7 parameters of the directly preceding accent phrase are added.

Table 1: Input parameters for the F_0 model parameter prediction. The features with "*" are added for two-step prediction.

Accent Phrase Features		Category
Current Phrase	Position in Sentence	27
	Number of Morae	28
	Accent Type	19
	Number of Words	11
	Part-of-Speech of the First Word	14
	Conjugation Type of the First Word	28
	Part-of-Speech of the Last Word	14
	Conjugation Type of the Last Word	28
	Boundary Depth Code	18
	Flag of Phrase Command (PF)*	2 (1 or 0)
	Phrase Command Magnitude (A_p)*	Continuous
	Offset of T_0 (T_{0off})*	Continuous
Preceding Phrase	Number of Morae	29
	Accent Type	20
	Number of Words	12
	Part-of-Speech of the First Word	15
	Conjugation Type of the First Word	29
	Part-of-Speech of the Last Word	15
	Conjugation Type of the Last Word	29

The boundary depth code is to indicate the depth of "bunsetsu" boundary between current and preceding accent phrases [4]. 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 depth of "bunsetsu" boundary was obtained using the Japanese text parser KNP [7]. We also added predicted phrase command parameters for accent command parameter prediction. This two-step prediction scheme was introduced, because compensation between phrase and accent components is often observable; when the phrase command values are estimated smaller than the actual value, the accent command values are estimated larger, and vise versa.

As for the output parameters for each accent phrase, a set of F_0 model parameters (magnitudes/amplitudes and timings) and a binary flag indicating the existence/absence of a phrase command at the head of the accent phrase are selected as shown in Table 2. In the table, T_{0off} is the offset of T_0 with respect to the segmental beginning of the accent phrase. $T_{1\text{off}}$ and T_{2off} are respectively offsets of T_1 and T_2 with respect to segmental anchor points, which are respectively defined as the beginning of the first high mora (basic unit of Japanese pronunciation mostly coincide with a syllable) for T_1, and the end of the mora containing the accent nucleus for T_2.

Table 2: Output parameters for the F_0 model parameter prediction.

Accent Phrase Feature	Category
Flag of Phrase Command (PF)	2 (1 or 0)
Phrase Command Magnitude (A_p)	Continuous
Offset of T_0 (T_{0off})	Continuous
Accent Command Amplitude (A_a)	Continuous
Offset of T_1 (T_{1off})	Continuous
Offset of T_2 (T_{2off})	Continuous

3.3. EXPERIMENT

Speech material for the experiment is utterances by a female narrator recorded at Nara Institute of Science and Technology. It includes 3 types of emotional speech, anger, joy, sad, and calm speech. All the utterances are not spontaneous ones; several hundreds of sentences were prepared for each type as a written text, and the speaker read it. The sentences for calm speech are the 503 sentences used for the ATR continuous speech corpus [8], while those for emotional speech are newly arranged for each emotion type so that the speaker can easily include the intended emotion in the utterances. Some of the ATR 503 sentences are also included in the text of emotional speech. An informal listening test was conducted for all the samples to exclude those without designated emotion from the experiment.

Since no prosodic label was provided for the speech material, after pitch extraction, the F_0 model parameters of the observed F_0 contour were automatically extracted for

ISBN 1-876346-48-5

each utterance sample [9]. After the automatic extraction, all the samples were checked if the extraction was done correctly. If the mean square error between the F_0 counter generated using the extracted parameter values and that observed exceeded a threshold, the automatic extraction was judged incorrect and such samples were excluded from the experiment. Also if more than two accent commands are extracted for one morpheme, such samples were excluded. After the process, we had around 400 sentences for each emotion type, which were divided into two groups to be used for the training and testing of the method as shown in Table 3. When extracting F_0 model parameters and generating F_0 contours, F_{0min} 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 126.43 Hz, 154.10 Hz, 141.17 Hz, and 137.56 Hz, for calm, angry, joy, and sad speech, respectively.

Part-of-speech information (and morpheme boundary information) was extracted from the text using the freeware parser JUMAN [10]. Division into accent phrases, as well as the information related to accent types, were derived from the extracted F_0 model commands.

Table 3: Number of samples used for the experiment.

Type	Category	Number	
		Sentence	Accent Phrase
Calm	Training	332	2128
	Testing	34	162
Anger	Training	395	2409
	Testing	62	330
Joy	Training	240	1021
	Testing	47	305
Sad	Training	291	1243
	Testing	74	303

Table 4: Correct rate, insertion error rate, and deletion error rate of flag PF prediction (in %).

Type	Condition	Correct Rate	Insertion Error Rate	Deletion Error Rate
Calm	Closed	81.5	14.3	4.2
	Open	76.4	18.2	5.4
Anger	Closed	71.9	17.7	10.4
	Open	65.7	23.7	10.6
Joy	Closed	78.6	17.4	4.0
	Open	85.0	13.0	2.0
Sad	Closed	76.6	18.5	4.9
	Open	75.7	19.9	4.3

Table 4 summarizes the results of *PF* prediction, and Table 5 shows root mean square errors for model parameter prediction. From these results, it can be said that the prediction is done for emotional speech with the similar accuracy as the calm speech, except for *PF* and T_{0off} of angry speech. Table 6 shows the result after two-step prediction. The effect of two-step prediction is clear if we compare it with Table 5.

Table 5: Root mean square errors for predicted parameters before two-step prediction. Unit of timing parameters is "second."

Parameter	Condition	Calm	Angry	Joy	Sad
A_p	Closed	0.204	0.213	0.223	0.195
	Open	0.213	0.221	0.223	0.195
T_{0off}	Closed	0.110	0.178	0.129	0.127
	Open	0.105	0.203	0.128	0.134
A_a	Closed	0.167	0.164	0.174	0.120
	Open	0.156	0.163	0.160	0.117
T_{1off}	Closed	0.088	0.086	0.096	0.093
	Open	0.089	0.093	0.091	0.100
T_{2off}	Closed	0.060	0.056	0.062	0.070
	Open	0.072	0.056	0.062	0.067

Table 6: Root mean square errors for predicted parameters after two-step prediction. Unit of timing parameters is "second."

Parameter	Condition	Calm	Angry	Joy	Sad
A_a	Closed	0.141	0.153	0.160	0.108
	Open	0.143	0.157	0.147	0.107
T_{1off}	Closed	0.078	0.075	0.084	0.083
	Open	0.074	0.081	0.079	0.092
T_{2off}	Closed	0.060	0.056	0.062	0.069
	Open	0.071	0.054	0.054	0.066

Importance of phrase command magnitude A_p for the prediction of accent command amplitude A_a is clear from the constructed decision tree shown in Fig. 1. Also phrase command timing T_{0off} is found to be important for accent command onset timing T_{1off} estimation. 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:

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

where $\ln F_0(t)$ 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 7, where F_0MSE values are averaged over all the training/testing sentences. Again the results for emotional speech are as good as those for calm speech. According to the former experiment on the relationship between F_0MSE and subjective evaluation score for the synthetic speech [4], synthetic speech with F_0MSE around 0.06 will be ranked "acceptable, although somewhat unnatural" or higher, which was supported by an informal listening test for the current paper.

ISBN 1-876346-48-5

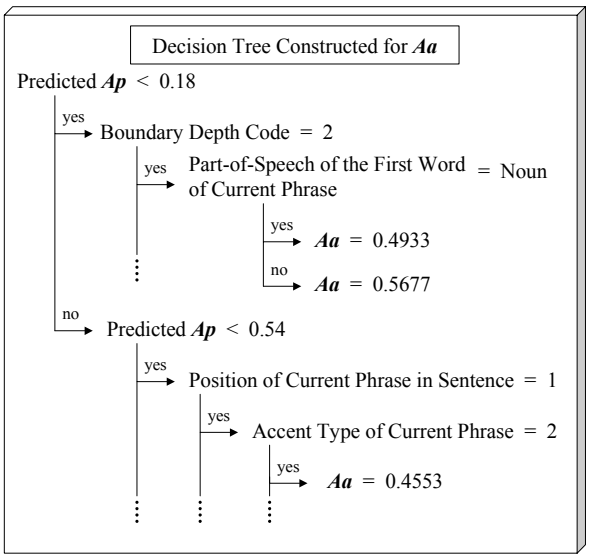

Figure 1: Part of decision tree constructed for accent command amplitude A_a prediction.

Table 7: AverageF_0*MSE*'s of F_0 contours generated using the predicted model parameters.

Condition	Calm	Angry	Joy	Sad
Closed	0.041	0.054	0.068	0.053
Open	0.049	0.049	0.050	0.048

4. PREDICTION OF SEGMENTAL DURATIONS

Control of segmental duration is also important to include the intended emotion in synthetic speech. Prediction of phoneme duration was also conducted in a similar framework as the prediction of the F_0 model commands. The input parameters are the identity of the phoneme in question, preceding and following phoneme identities, position of the mora (to which the phoneme in question being included) in the accent phrase, together with linguistic information of the accent phrase (as shown in Table 1). We also added part-of-speech and conjugation type information of the morpheme and location of the current phoneme in the accent phrase. The methods before and after the addition of parameters shall be called Method 1 and Method 2, respectively.

Table 8: Root mean square errors for predicted phoneme durations (in second).

Condition		Calm	Angry	Joy	Sad
Closed	Method 1	0.035	0.029	0.039	0.053
	Method 2	0.034	0.029	0.039	0.053
Open	Method 1	0.033	0.029	0.033	0.052
	Method 2	0.032	0.029	0.032	0.052

The root mean square errors of the phoneme length prediction are summarized in Table 8. The result was worst for sadness, indicating a rather large durational variation. Although the effect of additional parameters is not clear from the table, a higher correlation was obtained.

5. CONCLUSIONS

A scheme of corpus-based F_0 contour synthesis under the F_0 model constraints was applied successfully for emotional speech. Further experiments, such as subjective evaluation of synthetic speech, are necessary. Although the current experiments are speaker dependent, we are planning to apply the deviation between emotional speech features and calm speech features to other speaker's calm speech to generate his/her emotional speech.

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

REFERENCES

[1] A. Iida, F. Higuchi, N. Campbell, and M. Yasumura, "A corpus-based speech synthesis system with emotion," *Speech Communication*, vo.40, no.1-2, pp. 161-187, 2002.

[2] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Japan*, vol. 5, no. 4, pp. 233-242, 1984.

[3] K. Hirose, N. Minematsu, and H. Kawanami, "Analytical and perceptual study on the role of acoustic features in realizing emotional speech," *Proc. ICASSP*, Beijing, vol.2, pp.369-372, 2000.

[4] K. Hirose, M. Eto, and N. Minematsu, "Improved corpus-based synthesis of fundamental frequency contours using generation process model," *Proc. ICSLP*, pp. 2085-2088, 2002.

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

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

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

[8] Speech Corpus Set B. http://www.red.atr.co.jp/database_page/digdb.html

[9] 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*, pp. 509-512, 2002.

[10] Kyoto University, Japanese Morpheme Analysis System JUMAN http://www-nagao.kuee.kyoto-u.ac.jp/projects/nl-resource/.