Emotional Speech Synthesis with Corpus-Based Generation of F_0 Contours Using Generation Process Model

Keikichi Hirose, Kentaro Sato* & 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 method was developed for the corpus-based synthesis of emotional speech. Fundamental frequency (F_0) contours were synthesized by predicting command values of the generation process model 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 poorly. Prediction of the accent phrase boundaries for the input text, a necessary process for the synthesis, was also realized in a similar statistical framework. The HMM synthesis scheme was used to generate segmental features. The speech corpus used for the synthesis includes three types of emotional speech (anger, joy, sadness) and calm speech uttered by a female narrator. The command values of the model necessary for the training and testing of the method were automatically extracted using a program developed by the authors. For the better prediction, accent phrases where the automatic extraction was done poorly were excluded from the training corpus. The mismatches between the predicted and target contours for angry speech were similar to those for calm speech. Larger mismatches were observed for sad speech and joyful speech. Perceptual experiment was conducted using synthesized speech, and the result indicated that the anger could be well conveyed by the developed method.

1. Introduction

Recent advancement of speech recognition and synthesis highly improved the performance of man-machine interface through spoken language. The synthesized speech from interface systems, however, is mostly in reading style, which is not appropriate in a certain situations. A technology enabling speech synthesis with various styles is required. As an example of various styles, emotional speech is targeted in the current paper.

Up to now, a rather large number of research works have been reported on the realization of emotions in speech synthesis. Most of them 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 types 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 are trying to generate prosodic features from linguistic inputs using a statistical method.

A full corpus-based emotional speech synthesis has already realized using the ATR selection-based speech synthesis engine, CHATR [1]. However, in the framework of CHATR, the precise control of prosodic features cannot be realized. Tokuda *et al.* [2] developed a method of corpus-based speech synthesis, where all the acoustic features, including prosodic features, were handled in the HMM framework, and succeeded to generate synthetic (calm read) speech with highly natural F_0 movements by counting F_0 delta features. The method was applied to emotional speech synthesis with a certain success [3, 4]. The method controls F_0 value of each frame and, therefore, can model (and generate) any types of F_0 movements. However, in tern, it has possibility of causing un-naturalness 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) [5]. 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 [6]. 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 somewhat incorrectly. The method was originally developed for read speech synthesis [5, 7, 8], and was applied to emotional speech synthesis [9].

In the current paper, experiments on F_0 contour synthesis are further conducted by developing a scheme of automatic preparation of the speech corpus. In order to realize emotional speech synthesis with text input, prediction of accent phrase boundaries for the input text was conducted in a similar statistical way. The segmental features for synthetic speech were generated by an HMM-based method.

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-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 β_j) to express functions G_{pi} and G_{aj}, but, in the current experiments, they are fixed to 3.0 s^{-1} and 20.0 s^{-1} respectively 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. All the 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 [10], 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 [11].
2. From the content (text) of each utterance, its morphemes and part-of-speech information were obtained using the Japanese parser Chasen [12]. Another parser KNP [13] 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 [14]. 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 126.43 Hz, 154.10 Hz, 141.17 Hz, and 137.56 Hz, for calm, angry, joyful, and sad speech, respectively.
4. For each accent command extracted by the above process, the accent phrase was decided using a set of simple rules. The rules include such as; a morpheme is the initial morpheme of an accent phrase if an accent command onset locates at the former half of the morpheme. Also a *bunsetsu* boundary was assumed to be an accent phrase boundary if no accent command was extracted nearby.
5. For each accent phrase thus obtained, an accent type was assigned by referring to the accent type dictionary. The dictionary had accent type and attribute information, and, using a system developed by the authors [15], the accent type of each accent phrase could be decided.

Some of the samples thus obtained cannot be used for the experiment, because of wrong command extraction in the 3rd process. Therefore, all the samples were checked if the extraction was done correctly from the viewpoint of the mean square error between the F_0 counter generated using the extracted parameter values and that observed. If it 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. The exclusion can be sentence basis (method 1) or accent phrase basis (method 2). As the result, around 400 sentences 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. The numbers in the parentheses are those for method 2.*

Type	Category	Number	
		Sentence	Accent Phrase
Calm	Training	353 (434)	2497 (2801)
	Testing	50	342
Anger	Training	494 (587)	3524 (3895)
	Testing	50	341
Joy	Training	387 (477)	2650 (3008)
	Testing	50	283
Sadness	Training	283 (385)	2049 (2467)
	Testing	50	363

4. Prediction of F_0 model parameters

4.1. Input and output parameters

A binary decision tree (BDT) was trained to predict each of F_0 model parameters. The CART (Classification And Regression Tree) method included in the Edinburgh Speech Tools Library [16] was utilized. 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 [7].

The input parameters for BDT were selected as shown in Table 2. In the method, prediction of the model parameters is done for each accent phrase. Besides the features of the accent phrase in question, those of directly preceding accent phrase are added, taking into account that the F_0 contour of an accent phrase being influenced by that of preceding phrase. Their category numbers, shown in the parentheses, are larger than those of the corresponding parameters of the current phrase by one to represent "no preceding phrase." The boundary depth code is to indicate the depth of *bunsetsu* boundary between current and preceding accent phrases and can be calculated easily from the KNP code. We also added predicted phrase command parameters for accent command parameter prediction. This two-step prediction scheme was introduced, because there is compensation between phrase and accent components; 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 3. In the table, T_{0off} is the offset of T_0 with respect to the segmental beginning of the accent phrase. T_{1off}

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: *Input parameters for F_0 model parameter prediction. The last three features are for the two-step prediction and used to predict accent command parameters only.*

Accent phrase features	Category
Position in sentence	27
Number of *morae*	28 (29)
Accent type (location of accent nucleus)	19 (20)
Number of words	11 (12)
Part-of-speech of the first word	14 (15)
Conjugation form of the first word	21 (22)
Part-of-speech of the last word	14 (15)
Conjugation form of the last word	21 (22)
Boundary depth code	18
Predicted Phrase Command flag (*PF*)	2
Predicted Phrase Command Magnitude (A_p)	Continuous
Predicted Phrase Command offset of T_0 (T_{0off})	Continuous

Table 3: *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

4.2. Experiment

Experiment of F_0 contour generation was conducted using the corpus arranged in Section 3. 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:

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

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 4, 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 calm speech, and the result came worse in the order of anger, sadness and joy. If we compare methods 1 and 2, slightly better results are obtained by method 2 for all the emotions.

Table 4: *AverageF_0MSE's of F_0 contours generated using the predicted model parameters.*

	Closed		Open	
	Method 1	Method 2	Method 1	Method 2
Calm	0.082	0.084	0.081	0.086
Anger	0.084	0.084	0.114	0.111
Joy	0.207	0.204	0.208	0.203
Sadness	0.138	0.133	0.133	0.133

5. Prediction of accent phrase boundaries

The predictor examines each morpheme boundary of input text, and outputs a binary flag indicating whether the current morpheme boundary is an accent phrase boundary or not. The input parameters are those indicating part-of-speech, conjugation type and form, and length in *mora* of the current and the preceding morphemes. Also information on the boundary in question being a *bunsetsu* boundary or not and the KNP codes of the morphemes are included to take the syntactic structure into account. The results are slightly worse for joy and sadness as shown in Table 5.

Table 5: *Insertion and deletion error rates (%) and correct prediction rates (%) of accent phrase boundary prediction.*

	Closed			Open		
	Ins.	Del.	Cor.	Ins.	Del.	Cor.
Calm	9.4	6.4	84.3	13.2	6.9	79.9
Anger	10.9	6.5	82.6	12.6	6.9	80.5
Joy	12.4	8.7	79.0	16.5	8.9	74.6
Sadness	14.6	7.7	77.8	15.8	8.9	75.3

6. Prediction of segmental durations

Prediction of phoneme duration was 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 2). We also added part-of-speech and conjugation type information of the morpheme and location of the current phoneme in the accent phrase. The detail was given in [9].

7. Speech synthesis and evaluation

Using the developed methods for F_0 model command prediction (method 1) and phoneme duration prediction, speech synthesis from text was conducted for the 3 types of emotional speech and the calm speech. Segmental features were generated using the HMM-based speech synthesis toolkit [17]. Tri-phone models were trained for each type of emotion using the training sentences (for method 1) shown in Table 1. The segmental features were 75^{th} order vectors consisting of 0^{th} to 24^{th} cepstrum coefficients and their *delta* and *delta*^{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 without errors in accent phrase prediction 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 "correct (target)" F_0 contours and phoneme durations. The synthesized speech was presented to 15 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; how well they can perceive the emotion designated for each sample (5: quite well, 1: poorly, 0: other types). Table 7 shows the result. A good result was obtained for anger, but the results were unsatisfactory for joy and sadness. These results coincide with the tendency in F_0*MSE* results shown in Table 4.

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. "Cor." indicates the results when the "correct" prosodic features are used, while "Pre." indicates those when the predicted prosodic features are used. The results were averaged over all 10 sentences and 15 speakers for each emotion style.*

	Anger		Joy		Sadness	
	Cor.	Pre.	Cor.	Pre.	Cor.	Pre.
Calm	2.2	5.2	0.0	47.3	8.9	30.9
Anger	*93.3*	*87.4*	2.2	9.3	2.2	6.0
Joy	2.2	2.2	*97.8*	*38.7*	2.2	6.7
Sadness	2.2	5.2	0.0	4.7	*86.7*	*56.4*

Table 7: *Scores for the realization of the designated emotion.*

	Anger		Joy		Sadness	
	Cor.	Pre.	Cor.	Pre.	Cor.	Pre.
Score	3.09	3.08	3.38	1.03	3.64	1.34

8. Conclusions

A method of corpus-based F_0 contour synthesis under the F_0 model constraints, originally developed for read speech, was applied successfully for emotional speech. Perceptual experiment was conducted for the speech synthesized by the HMM-based synthesis using the F_0 contours and phoneme durations generated by the method. The results were satisfactory for angry speech, but they indicated the necessity of further study for joy and sadness. Although the current experiments are speaker dependent, we are planning to apply the deviation of emotional speech features from calm speech features to other speaker's calm speech to generate his/her emotional speech.

In the current experiments, level of each emotion is not accounted. Since utterances with different levels show quite different acoustic features [18], such information needs to be included in the input parameters of the method.

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

9. References

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

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

[3] 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.

[4] Tsuduki, R.; Zen, H.; Tokuda, K; Kitamura, T.; Bulut, M.; Narayanan, S., 2003. A study on emotional speech synthesis based on HMM, *Record of Autumn Meeting, Acoustical Society of Japan*, 241-242. (in Japanese)

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

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

[7] 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.

[8] 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.

[9] Hirose, K.; Katsura, T.; Minematsu, N., 2003, Corpus-based synthesis of F_0 contours for emotional speech using the generation process model, *Proc. ICPhS*, Barcelona, 2945-2948.

[10] Set B, http://www.red.atr.co.jp/database_page/digdb.html

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

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

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

[14] 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.

[15] 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.

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

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

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