Corpus-based Generation of Fundamental Frequency Contours Using Generation Process Model and Considering Emotional Focuses

Keikichi Hirose[1], Yasufumi Asano[2], & Nobuaki Minematsu[2]

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

Abstract

We formerly conducted emotional speech synthesis using our corpus-based method of generating fundamental frequency (F_0) contours from text. The method predicts command values of F_0 contour generation process model instead of directly predicting F_0 value of each time frame. A better control of F_0 contours was realized by taking the emotional level of each *bunsetsu* into account: adding information on which *bunsetsu*(s) the emotion is especially placed to the command predictor inputs. In the case of anger, F_0 contours closer to the target contours are obtained by adding emotional levels. Speech synthesis was conducted by generating F_0 contours in two ways: using commands predicted by taking emotional levels into account and those not. The result of perceptual experiment indicated that emotion was conveyed well by adding emotional levels.
Index Terms: speech synthesis, emotion, F_0 contour

1. Introduction

Quality of synthetic speech was improved to a "close to human" level by the introduction of corpus-based scheme. However, aside several exceptions [1], the available speech is mostly limited to reading style and rather monotonous. This situation still prevents the use of synthetic speech in various situations, and limits the usability of spoken dialogue systems in "real-world applications." To cope with this situation, a technology enabling speech synthesis with various styles is required.

Recently, a certain number of research works have been conducted for realizing various styles in synthetic speech. Among those, an attention is given to works on HMM-based speech synthesis. In the method, both of segmental and prosodic features of speech are processed together in frame-by-frame manner [2]. Various styles can be realized from a limited data by adapting phone HMMs to a new style [3].

However, although various attitudes and emotions were realized with rather high quality by the HMM-based speech synthesis, frame-by-frame processing of prosodic features includes some problems. Frame-by-frame processing has a merit that fundamental frequency (F_0) of each frame can be used directly as the training data, but, in turn, it sometimes causes sudden F_0 undulations (which are not characteristics of human speech) especially when the training data are limited. Prosodic features cover a wider time span than segmental features, and should be treated differently.

From these considerations, we have developed a corpus-based method of synthesizing F_0 contours in the framework of the generation process model (F_0 model) and realized speech synthesis in reading and dialogue styles and various emotional styles [4-6]. The model represents a sentence F_0 contour as a superposition of accent components on phrase components, each type of components assumed to be responses to accent commands and phrase commands, respectively [7]. By predicting the model commands instead of frame-by-frame F_0 values, a good constraint is automatically applied on the generated F_0 contours; still keeping acceptable speech quality even if the prediction is done somewhat incorrectly. Also, it is rather easy to analyze the prosodic controls obtained by statistical methods and to modify the generated F_0 contours according to our knowledge obtained through observations of natural utterances.

Although a rather good quality was realized in the synthetic speech with generated F_0 contours by the method, the realization of emotions is still not satisfactory. When predicting F_0 model command for emotional speech, we assumed that the emotion was placed evenly over a whole sentence. This will not be the case in human utterances: humans may place a focus of emotion on a part (or parts) of a sentence. In order to reflect this human nature to synthetic speech, we newly took the level of emotion into account. By labeling the degree of emotion for each *bunsetsu* (basic unit of Japanese syntax consisting of content word(s) followed or not followed by particles), and by adding it as inputs to the F_0 model command predictors, a better emotional control was realized.

The following sections are organized as follows: Section 2 describes the F_0 model. Section 3 explains speech corpus used in the experiment. The methods of F_0 model parameter prediction are shown in section 4. Methods with and without emotion levels of *bunsetsu* unit are compared with some experimental results. Section 5 gives results of perceptual experiments on emotion conveyed by speech synthesized using F_0 contours generated by the methods. Section 6 concludes the paper.

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.

3. Speech corpus

A voice actress uttered the 503 sentences used for the ATR continuous speech corpus in 3 types of emotion (anger, joy, sadness), and calmly. As a professional voice actress, she is well trained to express various types of emotion even when reading written texts. After recording, she was offered a written text of the 503 sentences with spaces for all *bunsetsu* boundaries, and was asked to mark the *bunsetsu*'s, where she placed emotion especially. During the marking, she was allowed to listen her recordings. The markings were checked through a listening test by a Japanese female, who is a speech therapist doing work on emotional speech. She was asked to judge if the marked parts included the parts of speech, where she felt as emotionally focused. The answer was mostly "yes" for "anger," but a certain degree of mismatch was found for other emotion types. Based on this result, the current experiment was done only on "anger."

After pitch extraction, F_0 model commands were extracted from observed F_0 contour for all the 503 sentence utterances. The extraction was done automatically using a method developed by the authors, where linguistic information of sentences was utilized as the constraints of command locations [6]. The 503 utterances with labels on F_0 model commands thus obtained served as speech (prosodic) corpus of the experiment of F_0 model parameter prediction. The corpus was divided into two groups: 453 sentences for training and 50 sentences for testing.

4. Prediction of F_0 model parameters

The parameters of F_0 model are predicted through the following processes:

1. Prediction of phrase command: each *bunsetsu* boundary is checked whether it is accompanied by a phrase command or not. If it is, the magnitude A_p and the timing T_0 of the command are predicted.
2. Prediction of prosodic word boundary location: each morpheme boundary is checked for whether it is also a prosodic word boundary or not.
3. Decision of accent types: for each prosodic word, an accent type is assigned.
4. Prediction of accent command: for each prosodic word, the amplitude A_a and the timings T_1 and T_2 of an accent command are predicted.

Processes 1, 2 and 4 are conducted using binary decision trees (BDT's). The CART (Classification And Regression Tree) included in the Edinburgh Speech Tools Library [8] was utilized to construct BDT's. Stop threshold, represented by the minimum number of examples per leaf node, was set to 40. One BDT was constructed for each model parameter for the processes 1 and 4. So predictors for these processes consisted of plural BDT's. In the current work, the same input parameters were selected for each BDT in a predictor, though it is possible to select differently. In the reminder of this session, phrase and accent command prediction processes (processes 1 and 4) are addressed, since the emotional level information is assumed to be effective for these processes.

It is known that the information of preceding units has a larger influence on the prosodic features of the current unit than that of following units. Taking these into consideration, the information of the current *bunsetsu* in question and that of directly preceding *bunsetsu* were included in the input parameters for the phrase command predictor as shown in Table 1. The category numbers 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*." Since the depth of syntactic boundary has a tight relation with the phrase command, boundary depth code (BDC) between the preceding and current *bunsetsu*'s was added to the input parameters. BDC denotes the depth of the boundary between the current and preceding *bunsetsu*'s, and was obtained by a simple calculation from the corresponding KNP code [9]. Punctuation marks of the text were not included, because of large variation according to writing styles. Instead, information of pause location was utilized: pauses have a tight relation with phrase commands. Although, in the whole process of text-to-speech synthesis, pauses should also be predicted in a similar way from input text, in the current experiment, pauses in the target speech were utilized. This is to place the research focus only on the F_0 control issue. The last three parameters in the table were added to count for the influence of the preceding phrase command on the current phrase command.

Table 1. *Input parameters for the 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
Pause immediately before the current *bunsetsu*	2
Phrase command for the 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

Similar to the case of phrase command, the parameters on accent commands (position and amplitude) are tightly related to the information of the current and preceding units (prosodic words), such as position in sentence, length, grammatical information of the first and last words of the units, and syntactic boundary between the units. They also change according to the accent types of the units. Taking these into consideration, the input parameters for accent command predictor were selected as shown in Table 2. Through the process of F_0 model parameter prediction, BDC's are assigned only for *bunsetsu*'s, and no BDC is given for a prosodic word when it belongs to a *bunsetsu* with two or more prosodic words and is not the first one in the *bunsetsu*. BDC is set to 1 (which means left branching boundary) for such a case.

Table 2. *Input parameters for the 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 (BDC)	22
Number of *morae* between the preceding phrase command and the head of the current prosodic word	23
Magnitude of the preceding phrase command	Continuous
Number of *morae* between the preceding accent command reset and the head of the current prosodic word	17
Amplitude of the preceding accent command	Continuous
Duration of the preceding accent command in *morae*	13

Table 3. *Use of emotional levels in F_0 model parameter prediction. Symbols* "o" *and* "x" *indicate when the levels are used and not used, respectively.*

Predic-tion	Original method (Condition 0)	New methods		
		Condition 1	Condition 2	Condition 3
Phrase command	x	o	x	o
Accent command	x	x	o	o

We added emotion levels of the current and preceding *bunsetsu*'s/prosodic-words into the input parameters of the phrase and accent command predictors: 1 when *bunsetsu*/prosodic-word coincides with a marked part by the speaker (see section 3) even in part, and 0 otherwise.

To check the validity of the emotional level for F_0 model parameter prediction, experiments are conducted in the four conditions as shown in Table 3. Henceforth, the methods with and without emotional levels are depicted as new methods and original method (condition 0), respectively.

Figures 1 and 2 show the F_0 contours and F_0 model commands for "watashitachiwa shizukani ayumiyori atamao sageta (We gently walked up and bowed.)," predicted by the original and new methods (condition 2), respectively. The solid lines indicate the F_0 contours (top panels) and F_0 model commands (second and third panels) predicted, while the dashed lines indicate those obtained from the observed F_0 contour by the automatic F_0 model command extraction. The gray thick lines at the top panels are observed F_0 contours (after smoothing using piecewise polynomial curves and interpolating portions corresponding to voiceless consonants). A better prediction is observable at the sentence-end *bunsetsu* "sageta" for the new method with emotional levels in accent command prediction. The smaller accent command amplitude for the new method as compared to the original method indicates that the focus in emotion is somewhat different from discourse focus, which usually increases the accent command amplitude.

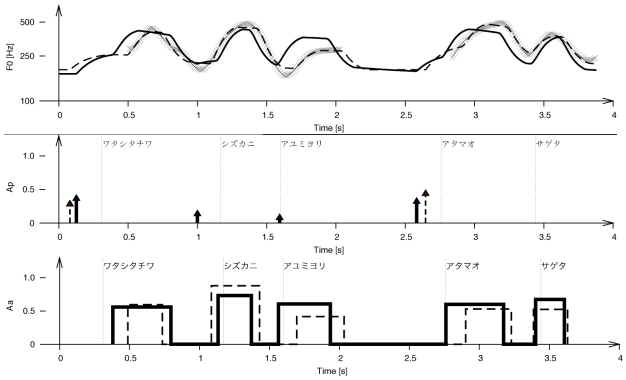

Figure 1. *The F_0 contours and F_0 model commands predicted by the original method not taking emotional levels into account. From top to bottom panels, F_0 contour, phrase commands, and accent commands are shown with their targets.*

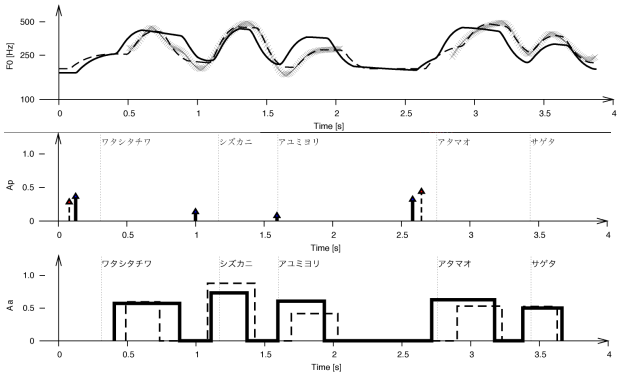

Figure 2. *The F_0 contours and F_0 model commands predicted by the new method (condition 2) taking emotional levels into account. From top to bottom panels, F_0 contour, phrase commands, and accent commands are shown with their targets.*

As an objective measure to evaluate the F_0 contour generated using the predicted F_0 model parameters, the mean square error between the generated contour and the target contour 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. Table 4 shows average F_0MSE values for the 4 conditions. The results are shown separately for data used for training (close) and for testing (open). For the open case, a better prediction was realized by taking the emotional levels into account. Better results for conditions 2 and 3 indicate that the effect is larger for accent components as compared to phrase components.

Table 4. *AverageF_0MSE's of F_0 contours generated using the model parameters predicted in four different conditions.*

Predic-tion	Original method	New methods		
		Condition 1	Condition 2	Condition 3
Close	0.0696	0.0713	0.0692	0.0714
Open	0.0755	0.0750	0.0745	0.0742

5. Speech synthesis and evaluation

Two versions of synthetic speech were compared: one with F_0 contours generated by the original method (without emotional levels) and the other by the new method (with emotional levels). Segmental features were generated using the HMM-based speech synthesis toolkit [10]. Tri-phone models were trained using the 453 sentence utterances used for the training of the F_0 model command predictors. 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.

Synthetic speech samples for 20 sentences by the new method (condition 2) and those for 10 sentences by the original method were randomized and presented to 12 Japanese, who were asked to check *bunsetsu*'s where they felt the designated emotion (anger) well as compared to other parts. The option of no such *bunsetsu* was allowed. When the checked parts coincided with those checked by the speaker of speech corpus (see section 3), even if they are partly, the sentences including the parts were counted as the emotional levels being correctly realized in synthetic speech. This judgment was conducted separately for each of 12 informants. When F_0 contours were generated by the new method, emotional levels were correctly realized in 92.7 % of sentences. The rate decreased to 78.2 % when F_0 contours were generated by the original method.

In order to evaluate how well the designated emotion was conveyed by the new method, another listening test was conducted. Each of 30 sentences was synthesized by both new (condition 2) and original methods and the two versions of synthesized speech were presented to 9 Japanese speakers. They were asked to select the version, to which they felt the designated emotion (anger, for the current experiment) clearer. The version by the new method was selected in 79.3 % probability. These results on the listening tests indicate the validity of adding *bunsetsu*-based emotional levels in realizing designated emotion in synthetic speech.

6. Conclusion

An improvement was realized in the ability of expressing designated emotions in our corpus-based method of generating F_0 contours of emotional speech. By labeling the levels of emotion for each *bunsetsu*, and by adding the labels in input parameters of the F_0 model command predictors, a better expression of emotion was realized in synthetic speech.

Currently, the method is only trained for a speech corpus, and used for realizing the same emotion in the same voice quality. Further research is planned to realize emotional speech for a speaker without speech corpus of that emotion: applying deviations in acoustic features between emotional speech and calm speech of an actor/actress to other speaker's calm speech to generate his/her emotional speech. By doing so, any person can perform like an actor/actress.

7. References

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

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

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

[4] Hirose, K., Sakata, M., Kawanami, H., "Synthesizing dialogue speech of Japanese based on the quantitative analysis of prosodic features," *Proc. ICSLP*, Philadelphia, Vol.1, pp.378-381 (1996).

[5] Sakurai, A., Hirose, K., and Minematsu, N., "Data-driven generation of F_0 contours using a superpositional model," *Speech Communication*, Vol.40, No.4, pp.535-549 (2003).

[6] Hirose, K., Sato, K., Asano, Y., and Minematsu, N., "Synthesis of F_0 contours using generation process model parameters predicted from unlabeled corpora: Application to emotional speech synthesis," *Speech Communication*, Vol.46, Nos.3-4, pp.385-404 (2005-7).

[7] Fujisaki, H. and Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Japan (E)*, Vol.5, No.4, pp.233-242 (1984-10).

[8] The Edinburgh Speech Tools Library, http://www.cstr.ed.ac.uk/projects/speech_tools/

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

[10] Galatea Project, http://hil.t.u-tokyo.ac.jp/~galatea/regist-jp.html (in Japanese).