Corpus-Based Synthesis of Fundamental Frequency Contours of Japanese Using Automatically-Generated Prosodic Corpus and Generation Process Model

Keikichi Hirose, Takayuki Ono* and Nobuaki Minematsu***

*Graduate School of Frontier Sciences, **Graduate School of Information Science and Technology
University of Tokyo, Japan
{hirose, t-ono, mine}@gavo.t.u-tokyo.ac.jp

Abstract

We have been developing corpus-based synthesis of fundamental frequency (F_0) contours for Japanese. Since, in our method, the synthesis is done under the constraint of F_0 contour generation process model, a rather good quality is still kept even if the prediction process is done poorly. Although it was already shown that the synthesized F_0 contours sounded as highly natural as those using heuristic rules carefully arranged by experts, the F_0 model parameters for the training corpus were extracted with some manual processes. In the current paper, the automatically extracted parameters are used, and a good result is obtained. Also several features are added as the inputs to the statistical method to obtain better results. Some results on the accent phrase boundary prediction in the similar corpus-based framework are also shown.

1. Introduction

Development of new speech technologies such as the selection-based concatenative speech synthesis has largely improved quality of synthetic speech. The improvement, however, was mostly on the segmental features of speech, and rather made clearer the low quality in the prosodic features. Inspired by the success of corpus-based methods in speech processing, several methods were developed for generating prosodic features from linguistic inputs using statistical methods, such as neural net works, hidden Markov models (HMMs), and so on. As for the F_0 contours, some methods try to relate F_0 values of each frame with input parameters. Among them, the most successful method is the one based on HMMs [1]. By introducing delta F_0 features, the method can realize a rather good approximation of F_0 movements, but has possibility of causing unnaturalness especially when the training data are limited.

As a supra-segmental feature, an F_0 movement should be viewed in a wider range, not in a frame. Introduction of the F_0 contour generation process model (henceforth F_0 model [2]) can automatically solve this problem. This model assumes that a sentence F_0 contour can be decomposed into phrase and accent components, and represent them as responses for corresponding commands. With the model, the F_0 movement can be viewed as *mora*-to-*mora* transitions in the case of Japanese. Here, *mora* is a basic unit of Japanese utterance mostly coinciding to a syllable.

Since the model commands has a good correspondence with linguistic information, a set of simple rules can realize F_0 contours close to those of natural speech [3]. However, development of these rules requires experts who have an ample and precise knowledge on F_0 features. Moreover, developing synthesis rules for various styles of utterances is a time-consuming process and even impossible if the expert's knowledge on the style is limited.

From these viewpoints, we have developed a corpus-based synthesis of F_0 contours in the framework of the F_0 model [4-6]. The use of F_0 model is beneficial in that a good constraint will automatically applied on the synthesized F_0 contours; still keeping acceptable speech quality even if the prediction by a statistical method is done incorrectly. Also combination with rule-based synthesis can be realized easily by applying constrains on command position and magnitude/amplitude, though it is not dealt with in the current paper.

The method predicts the F_0 model commands for each accent phrase of the sentence to be synthesized. We already have obtained good results by using linguistic information of the current and the preceding accent phrases and syntactic boundary depth between them [5, 6]. However, the method included one major problem; speech corpus for training should have F_0 model commands and to arrange such a corpus is not an easy task. In the former report, the F_0 model commands were extracted through manual interaction. This process is quite time consuming, and prevents to increase the size of training corpus or to develop a new one for a new speaking style. To cope with this problem, we tried to use the corpus, whose model commands were obtained fully automatic by a method of F_0 model command extraction developed by the author's laboratory [7]. Also, we added several parameters for better prediction. In the current paper, these are explained with experimental results.

In our method, 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 accent phrases. Therefore, given a text, the following 3 processes are necessary before the prediction process: morpheme analysis, accent phrase boundary detection, and accent type prediction of accent phrases. The second process can be performed by a similar corpus-based method. This paper also includes am improvement on the method.

2. F_0 model

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 [3]. 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. The F_0 model is given by the following equation:

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

(1)

In the equation, $G_p(t)$ and $G_a(t)$ 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 α and β) to express functions G_p and G_a. Each components can take different values for α and β, 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. Statistical method

In this paper, prediction of the model parameters were conducted by the binary decision tree constructed using CART (Classification And Regression Tree) software included in the Edinburgh Speech Tools Library [8]. Stop threshold, represented by the minimum number of examples per a leaf node, was set to 40 according to the result of former experiments [5].

4. Prosodic corpus

From the ATR continuous speech corpus of 503 sentences [9], utterances by male speaker MHT were first selected. Then, they were gone through the following processes 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 which *bunsetsu* is 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 [7]. The extraction scheme is first to smooth the observed F_0 contour by a piecewise 3rd order polynomial function and to locate accent command positions by taking the derivative of the function. The bias level F_b was fixed to 63.1 Hz throughout the process.
4. For each accent command extracted by the above process, the *bunsetsu* was identified where the command onset belonged. Such a *bunsetsu* was assumed as the initial morpheme of an accent phrase. If two or more accent commands located at a *bunsetsu*, they were made to one by several rules; if the latter commands had amplitudes less or equal to 0.9 times of the first command amplitude, they are concatenated. For other cases, the latter commands were simply deleted.

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 authors [13], the accent type of each accent phrase could be decided.

Some of the sentences could not be used because of wrong command extraction in the 3rd process. As the result, we got a prosodic corpus with 480 sentences, out of which 428 sentences were used for training and the rest were used for testing in the following experiments.

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 for the baseline method are the information on part-of-speech, conjugation type and form, and length in *mora* of the current and the preceding morphemes [5]. In the new method, KNP codes of the morphemes are introduced to take the syntactic structure into account. It is shown in Table 1 that, by adding KNP codes, the correct prediction rate was increased from 88.7 % to 92.7 % for the test data. It is clear that the deletion errors are reduced a lot.

Table 1: Insertion and deletion error rates (%) and correct prediction rates (%) of accent phrase boundary prediction.

Methods	Closed			Open		
	Ins.	Del.	Cor.	Ins.	Del.	Cor.
Baseline	5.2	3.9	90.9	6.6	4.7	88.7
New	5.1	0.9	94.1	5.9	1.4	92.7

6. Prediction of F_0 model parameters

6.1. Input and output parameters

Table 2 shows the input parameters selected for F_0 model parameter prediction in the baseline method [6]. In the method, prediction of the model parameters is done for each accent phrase, and therefore the first 8 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. Category numbers for these parameters are larger than 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 obtained from KNP codes by an easy calculation [5, 6].

In the current paper, 3 methods, where new parameters are added to the input parameters of the baseline method, are tested. They shall be called as methods A, B and C as shown in Table 3. Possibility of a phrase command at an accent phrase initial increases as the distance from the preceding phrase command comes larger. Parameters for method A were introduced to take this feature into account. Method B is the two-step prediction scheme, where phrase commands are first predicted and their information is added to the input parameters for the prediction of accent commands. This

method is motivated by the compensatory feature between phrase and accent components of the F_0 model; if a phrase command is estimated smaller, accent commands of the phrase are estimated larger. In method C, influence of phoneme identity to the model parameters is taken into account. This method was planned to improve the prediction of timing parameters. When the training samples are limited, adding new input parameters with a large number of categories may have negative effects on the prediction accuracy. To cope with this problem, in the current experiment, the phoneme categories are limited to 3; vowels, voiceless stops and other consonants for phoneme at the initial position of phrase, and without consonant, with voiceless stops and with other consonants for accent nucleus syllable. F_0 contour generation was also conducted for the combination of the methods A, B and C, which shall be called as method D, hereinafter.

Table 2: Input parameters for the baseline method

Accent phrase feature		Category
Current accent phrase	Position in sentence	15
	Number of *mora*e	22
	Accent type (location of accent nucleus)	15
	Number of words	10
	Part-of-speech of the first word	18
	Conjugation type of the first word	20
	Part-of-speech of the last word	18
	Conjugation type of the last word	20
Preceding accent phrase	Number of *mora*e	23
	Accent type (location of accent nucleus)	16
	Number of words	11
	Part-of-speech of the first word	19
	Conjugation type of the first word	21
	Part-of-speech of the last word	19
	Conjugation type of the last word	21
	Boundary depth code	11

Table 3: Input parameters added for the new methods.

Accent phrase feature		Category
Method A	Distance from preceding phrase command (in number of accent phrase)	7
	Distance from preceding phrase command (in number of *mora*e)	26
Method B	Predicted Phrase Command flag (*PF*)	2
	Predicted Phrase Command Magnitude (A_p)	Continuous
Method C	Phoneme at initial position of current phrase	3
	Syllable at accent nucleus	3

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 4. 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* for T_1, and the end of the *mora* containing the accent nucleus for T_2. The first high *mora* of the accent phrase is either the first *mora* for accent phrases of type 1 accent, or the second *mora* for accent phrases of other accent types. There is no change from the baseline method to the new methods.

Table 4: output parameters for the F_0 model parameter prediction.

Accent phrase feature		Category
Phrase command	Flag (*PF*)	2 (1 or 0)
	Magnitude (A_p)	Continuous
	Offset of T_0 (T_{0off})	Continuous
Accent commands	Amplitude (A_a)	Continuous
	Offset of T_1 (T_{1off})	Continuous
	Offset of T_2 (T_{2off})	Continuous

6.2. Experiment

Table 5 shows the root mean square errors for model parameter prediction. As expected, addition of parameters in method A improved phrase command parameter prediction, while that in method B improved accent command amplitude prediction. No significant improvement was observable for method C. The best result was obtained for method D, combination of methods A to C. The effect of distances from preceding phrase command (method A) on phrase command prediction is clearly shown in Table 6, where result of *PF* prediction was summarized as rates of correct detection, insertion error, and deletion error.

Table 5: Root mean square errors for predicted parameters. The target parameter values for prediction are those obtained by the automatic command extraction method.

Output param.	Condition	Base-line	Method A	Method B	Method C	Method D
A_p	Closed	0.176	0.176	-	0.176	0.129
	Open	0.194	0.188	-	0.194	0.143
T_{0off}	Closed	0.164	0.161	-	0.164	0.136
	Open	0.157	0.148	-	0.154	0.134
A_a	Closed	0.162	0.162	0.156	0.161	0.155
	Open	0.205	0.205	0.175	0.207	0.175
T_{1off}	Closed	0.117	0.116	0.114	0.115	0.112
	Open	0.128	0.128	0.126	0.123	0.123
T_{2off}	Closed	0.075	0.075	0.074	0.074	0.074
	Open	0.079	0.079	0.082	0.079	0.082

Table 6: Result of phrase command flag *PF* prediction.

Rate (%)		Method			
		Base-line	A	C	D
Closed	Correct	78.6	85.3	78.8	85.3
	Ins. Err.	11.4	10.9	12.2	10.9
	Del. Err.	10.0	3.5	9.0	3.5
Open	Correct	77.9	83.3	78.8	83.3
	Ins. Err.	7.2	12.6	6.3	12.6
	Del. Err.	14.9	4.1	14.4	4.1

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 generated using the automatically extracted parameter values (target parameter values of the prediction) is defined as:

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

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. In Table 7, the results are shown as averaged values for 52 test sentences. If we compare the result with former one using manually extracted F_0 model parameters [6], we can conclude that the use of automatically extracted F_0 model parameters does not degrade the prediction performance.

Table 7: $F_0 MSE$'s for various methods. Averaged over 52 test sentences.

Baseline	Method A	Method B	Method C	Method D
0.058	0.049	0.056	0.048	0.052

Subjective evaluation was also conducted as listening test for 5 sentences selected from 52 test sentences. Test speech samples were synthesized by converting F_0 contours of the original utterances into those generated by the F_0 model during an analysis-resynthesis process based on log-magnitude approximation (LMA) filter [14]. They were presented to 6 Japanese subjects, who were asked to give a score from 1 to 5 based on the intonation and accent criterion. The scores are associated with subjective criteria as follows:

5: very good, indistinguishable from natural speech,
4: good, although not as much as natural speech,
3: acceptable, although somewhat unnatural,
2: unnatural and not so good,
1: poor.

Table 8: Result of subjective evaluation tests. S.D. depicts standard deviation.

Method	Score (S.D.)
Model parameters by automatic extraction (Target)	4.63 (0,12)
Baseline	2.57 (0.43)
Method D	3.00 (0.14)

Table 8 shows the averaged scores, clearly indicating improvements from the baseline to the method D. Also we should note that the standard deviation largely decreased. Larger standard deviation indicates that the synthetic speech sounds unnatural in larger number of sentences.

7. Conclusions

Corpus-based generation of F_0 contours was successfully conducted using automatically arranged prosodic corpus. Although not mentioned in the paper, we are now trying to generate F_0 contours for various types of speech in the same way, including emotional speech [12].

8. References

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

[2] 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.

[3] Hirose, K. and Fujisaki, H., "A system for the synthesis of high-quality speech from texts on general weather conditions," *IEICE Trans. Fundamentals*, E76-A(11), 1971-1980, 1993.

[4] Sakurai, K. Hirose, and N. Minematsu, "Data-driven generation of F_0 contours using a superpositional model, *Speech Communication*, to be published, 2003.

[5] Hirose, K., Minematsu, N., and Eto, M., "Data-driven synthesis of fundamental frequency contours for TTS systems based on a generation process model," *Proc. Speech Prosody 2002*, 391-394, 2002.

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

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

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

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

[10] Kyoto Univ., Large vocabulary continuous speech recognition decoder Julius, http://winnie.kuis.kyoto-u.ac.jp/pub/julius/.

[11] Nara Institute of Science and Technology, Morphological Analyzer ChaSen, http://chasen.aist-nara.ac.jp/.

[12] Kyoto Univ., 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(1), 550-557, 2003.

[14] Imai, S., "Low bit rate cepstral vocoder using the log magnitude approximation filter," Proc. IEEE ICASSP, 441-444, 1978.