IMPROVED CORPUS-BASED SYNTHESIS OF FUNDAMENTAL FREQUENCY CONTOURS USING GENERATION PROCESS MODEL

Keikichi Hirose, Masaya Eto* 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, eto, mine}@gavo.t.u-tokyo.ac.jp

ABSTRACT

We have been developing corpus-based synthesis of fundamental frequency (F_0) contours for Japanese text-to-speech (TTS) conversion systems. 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 incorrectly. Although it was already shown that the synthesized F_0 contours sounded as highly natural as those using heuristic rules arranged by experts, there were occasional cases with low quality depending on sentences to be synthesized. Several features, including a code representing syntactic boundary depth obtainable through an automatic parsing process, were added to input parameters of the statistical methods, and a better prediction was realized. The boundary depth code was shown to be very effective for improving especially phrase component parameter prediction.

1. INTRODUCTION

When experts carefully arrange synthesis rules, a high quality, hard to be surpassed by statistical 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 a new style of utterance is a time-consuming process and even impossible if the expert's knowledge on the style is limited. Therefore, in view of the success of corpus-based methods in speech processing, 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, but have possibility of causing un-naturalness especially when the training data are limited. Several methods are reported under the ToBI labeling strategy. 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 have developed a corpus-based synthesis of F_0 contours in the framework of the generation process model (henceforth F_0 model) [2, 3]. 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 [4]. 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 current constraints are limited to the model's command response features, further constrains are possible based on various knowledge on model commands, such as on command timing as compared to the segmental boundary locations.

The use of F_0 model in a statistical approach was already tried in [5], where multiple split regression trees are used to derive rules to generate F_0 model parameters. However, the timing parameters are excluded from the mapping and have to be externally assigned. Moreover, it uses high-level syntactic information as the statistical model input, which is difficult to be automatically obtained in a TTS system. Our method estimates magnitudes/amplitudes and timings of F_0 model commands from linguistic information automatically obtainable through parsing of input texts. In the current paper, several features, including information on which word directly modifying which word, are added to input parameters of the statistical methods, and their effects on the F_0 model parameter prediction are examined.

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 constituting 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. We are developing the second and the third processes, which will not be addressed in the current paper. The second process can be performed by similar statistical ways [3]. As for the first process, freeware parsers can be utilized.

2. F_0 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 [4]. 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:

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

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 15.0 s^{-1} based on the former F_0 contour analysis results.

3. PREDICTION OF F_0 MODEL PARAMETERS

3.1. Statistical methods

In the current paper, statistical methods based on binary decision tree (BDT) and multiple linear regression analysis (MLRA) were used. Neural networks were not checked, since their performance was almost the same with BDT and MLRA in the former experiments [2]. As for the BDT, the freeware Wagon [6] from the Edinburgh Speech Tools Library was used. Stop threshold, represented by the minimum number of examples per a leaf node, was set to 40 according to the result of former experiments [3].

3.2. Input and output parameters

In the original method for F_0 model parameter prediction, taking the fact that the prediction of F_0 model parameters is done in accent phrase basis, input parameters were selected from those related only to the accent phrase in question as shown in Table 1. In the current paper, 3 new methods (methods A, B, and C) are developed and checked by adding several input parameters as summarized in Table 2. Method A added directly-preceding accent phrase features, method B added a code to indicate the depth of "bunsetsu" boundary between current and preceding accent phrases, and method C added predicted phrase command information for accent command parameter prediction. 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] with no manual correction. KNP tells us which "bunsetsu" directly modifies which "bunsetsu." 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 the boundary depth codes are obtained by simply shifting the distances rightward. In method A, one category is added to all features to represent the accent phrase in question locating at sentence initial (and no preceding accent phrase existing). Method C 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. If *PF* is predicted as "1," "distance from preceding predicted phrase command" takes 1, and if *PF* is "0," the distance (in accent phrase number) to the preceding accent phrase with *PF*=1 is added. 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. F_0 contour generation was also conducted for the combination of the methods A and B (henceforth, method D), and that of methods A, B and C (henceforth, method E). 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 1. 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. 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 original method to the new methods.

Table 1: Input and output parameters for the original method of F_0 model parameter prediction.

Accent Phrase Feature		Category
Input Para-Meter	Position of Accent Phrase in Sentence	18
	Number of Morae	15
	Accent Type	10
	Number of Words	7
	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
Output Para-Meter	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

Table 2: Input parameters added for the new methods.

Accent Phrase Feature		Category
Method A	Number of Morae of Preceding Phrase	16
	Accent Type of Preceding Phrase	11
	Number of Words of Preceding Phrase	8
	Part-of-Speech of the First Word of Preceding Phrase	15
	Conjugation Type of the First Word of Preceding Phrase	29
	Part-of-Speech of the Last Word of Preceding Phrase	15
	Conjugation Type of the Last Word of Preceding Phrase	29
Method B	Boundary Depth Code	11
Method C	Predicted Phrase Command Magnitude (A_p)	Continu-ous
	Distance (in Accent Phrase Number) from Preceding Predicted Phrase Command	7

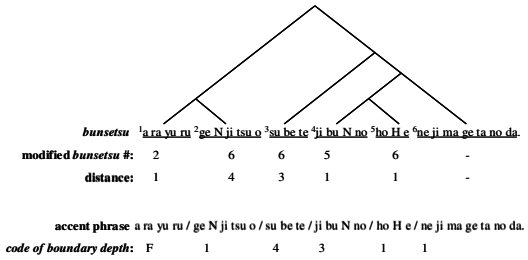

Figure 1: Result of syntactic analysis by KNP and codes to represent *bunsetsu* boundary depth.

3.3. Experiments

From the ATR continuous speech corpus of 503 sentences [8], utterances by male speaker MHT with prosodic labels were selected for the experiments. Among them, 388 sentences (2803 accent phrases) and 48 sentences (262 accent phrases) were used as the training data and test data, respectively. The F_0 model parameters for the training and test data were derived from J-ToBI labels attached to the corpus. The F_0 model parameters for the test data are necessary to evaluate the results of prediction. First, timing parameters were estimated using J-ToBI labels as suggested in [9], and, then, the analysis-by-synthesis (AbS) process was carried out for F_0 contours extracted form the speech waveform. F_{0min} was fixed to 51.0 Hz.

Division into accent phrases, as well as the information related to accent types, were also derived from J-ToBI labels. Mora boundaries were obtained from the phoneme boundaries of the corpus using simple rules. Part-of-speech information was extracted from the text of ATR corpus using the freeware parser JUMAN [10]. No manual correction was added.

Table 3 shows the multiple correlation coefficients for MLRA. Although improvements are observable for all the output parameters from the original to the new methods, they are significant for those of phrase components by method B. This result is quite natural, since the boundary depth, as an index representing the syntactic structure of a sentence, mostly related to the phrase features. Importance of the syntactic structure for phrase command prediction is also clear from Fig. 2, which shows the decision tree (BDT-40) for the phrase command flag PF prediction. Table 4 summarizes the results of *PF* prediction for BDT-40 and MLRA.

Table 3: Multiple correlation coefficients of the training data for multiple linear regression analysis.

Output Para-meter	Method					
	Original	A	B	C	D	E
PF	0.60	0.66	0.70	-	0.72	-
A_p	0.65	0.70	0.74	-	0.76	-
T_{0off}	0.58	0.64	0.65	-	0.68	-
A_a	0.44	0.49	0.46	0.52	0.50	0.57
T_{1off}	0.44	0.49	0.44	0.47	0.49	0.52
T_{2off}	0.42	0.44	0.43	0.44	0.42	0.43

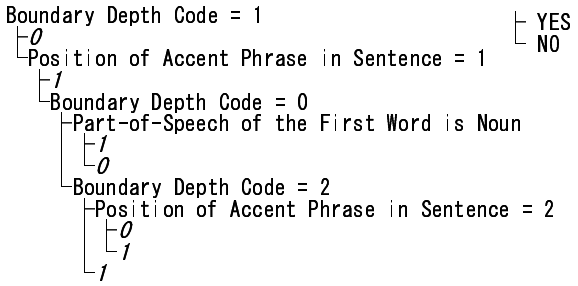

Figure 2: Part of decision tree (BDT-40) for phrase command flag *PF* prediction.

Table 4: Result of phrase command flag *PF* prediction.

Rate (%)		Method			
		Ori-ginal	A	B	D
BDT-40	Correct	74.4	76.7	84.7	84.7
	Ins. Err.	14.5	10.7	6.9	6.9
	Del. Err.	11.1	12.6	8.4	8.4
MLRA	Correct	74.4	79.0	83.6	84.7
	Ins. Err.	9.9	7.6	7.3	5.3
	Del. Err.	15.7	13.4	9.2	9.9

As an objective measure to totally evaluate the predicted F_0 model parameters, mean square error between a sentence F_0 contour generated using the predicted parameters and that of the "best" approximation by the model is defined as:

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

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. Here, "best" approximation means the parameter set obtained after the AbS process using the J-ToBI labels as explained already. The results are summarized in Table 5, where F_0MSE values are averaged over all the test sentences. Clearly better results were obtained by the new methods B and E.

Subjective evaluation was also conducted as listening test for 10 sentences selected from 48 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 [11]. They were presented to 10 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 6 shows the averaged scores, clearly indicating improvements from the original to the new methods. When the subjective score and F_{0min} were compared for each synthetic speech (after averaging over subjects), the correlation coefficient between them was -0.84. The rather high negative correlation indicates that F_0MSE can be utilized as a good measure for the evaluation of the methods. Formerly, we used F_0 contour of target speech as the reference to calculate F_0MSE. The correlation coefficient was -0.73, though it was not based on this listening test.

Table 5: Mean square errors (F_0MSE's) between F_0 contours generated using F_0 model command values predicted by the methods and those generated using the "best" estimated parameter values. Averaged over 48 test sentences.

F_0MSE	Method				
	Ori-ginal	A	B	C	E
BDT-40	0.074	0.079	0.059	0.078	0.061
MLRA	0.081	0.078	0.058	0.090	0.057

Table 6: Result of subjective evaluation tests. Averaged over 10 sentences and 10 subjects.

Method		Score
Best Estimation		4.1
BDT-40	Original	2.4
	E (A+B+C)	3.2
MLRA	Original	2.6
	E (A+B+C)	3.1

4. CONCLUSIONS

Several parameters are newly added to improve our corpus-based F_0 contour synthesis scheme under the F_0 model constraints. Through experiments, it was clarified that the boundary depth code between current and preceding accent phrases served a lot especially to improve the phrase command parameter prediction. Although not addressed in the paper, we are now trying to generate training corpus with F_0 model command values automatically [12]. A preliminary experiment for speaker MHT's utterances resulted in F_0MSE=0.0796, which roughly corresponded to subjective score 2.

The work is partly supported by Grant in Aid for Scientific Research of Priority Areas (#746).

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

[3] 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.

[4] 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.

[5] Hirai, T., Iwahashi, N., Higuchi, N., and Sagisaka, Y., "Automatic extraction of F_0 control rules using statistical analysis," in *Advances in Speech Synthesis*, Springer, 333-346, 1996.

[6] Edinburgh University, Edinburgh Speech Tools Library - Wagon, http://www.cstr.ed.ac.uk/projects/speeech_tools/manual.

[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] Hirai, T. and Higuchi, N., "Automatic extraction of the Fujisaki model parameters using the labels of Japanese tone and break indices (J-ToBI) system," Trans. IEICE, J81-D-II, 1058-1064, 1998.

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

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

[12] Narusawa, N., 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.