USING F_0 CONTOUR GENERATION PROCESS MODEL FOR IMPROVED AND FLEXIBLE CONTROL OF PROSODIC FEATURES IN HMM-BASED SPEECH SYNTHESIS

Keikichi Hirose, Keiko Ochi, Miaomiao Wang, Tatsuya Matsuda, Miaomiao Wen, and Nobuaki Minematsu

University of Tokyo
hirose@gavo.t.u-tokyo.ac.jp

Abstract: Generation process model of fundamental frequency contours known as Fujisaki's model is ideal to represent global features of prosody. It is a command response model, where the commands have clear relations with linguistic and para/non linguistic information included in the utterance. Therefore, by controlling fundamental frequency contours in the framework of the generation process model, a more flexible control of prosodic features comes possible in speech synthesis. Also, the model can be used to solve the problems of HMM-based speech synthesis, which arise from frame-by-frame treatment of fundamental frequencies. In this paper, two methods for improved control of prosodic features in HMM-based speech synthesis, and one method for flexible fundamental frequency control to realize prosodic focuses in synthetic speech, are presented. All these methods are based on the generation process model.

1 Introduction

Recently, in the speech synthesis community, a special attention has been placed on HMM-based speech synthesis, where a flexible control in speech styles is possible by adapting phone HMMs to a new style. In the method, both segmental and prosodic features of speech are processed together in a frame-by-frame manner, and, therefore, it has an advantage that synchronization of both features is kept automatically [1]. Although utterances conveying various attitudes and emotions are possible with rather high quality by the method, frame-by-frame processing of prosodic features, however, includes an inherit problem. It 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 (not observable in human speech) especially when the training data are limited. Prosodic features cover a wider time span than segmental features, and should be treated differently.

One possible solution to this issue is to use the generation process model (F_0 model) developed by Fujisaki and his co-workers [2, 3]. The model represents a sentence F_0 contour as a superposition of accent components on phrase ones; each type of components assumed to be responses to step-wise accent commands and impulse-like phrase commands, respectively. These components are known to have clear correspondences with linguistic and para/non linguistic information, which is conveyed by prosody. Thus, using this model, a better control can be realized for F_0 contour generation than the frame-by-frame control. Because of clear relationship between generated F_0 contours and linguistic and para/non linguistic information of input texts, manipulation of generated F_0 contours is possible, leading a flexible control of prosodic features.

However, in order to fully extract the benefit of the F_0 model in speech synthesis, two major problems should be solved. One is to analyze and to extract the F_0 model commands from observed F_0 contours of utterances in the training corpus. This process needs to be done at least semi-automatically to avoid a time consuming process of manual extraction [4]. Since

the content of utterances is known for the training corpus, it can be used to facilitate the command extraction. We have developed several methods for automatic extraction of F_0 model commands, which apply constraints on phrase and accent commands [5, 6]. The current paper, however, focuses on the other problem: how to incorporate the F_0 model in HMM-based speech synthesis. We have developed a corpus-based method of synthesizing F_0 contours in the framework of F_0 model and have combined with HMM-based speech synthesis to realize speech synthesis in reading and dialogue styles with various emotions [7]. However, the F_0 contours generated by HMM-based speech synthesis are simply substituted by those generated by the method before the speech synthesis. Although, a better quality is obtained for synthetic speech by the method, the segmental and prosodic features do not satisfy the maximum likelihood condition in the HMM framework anymore; losing a benefit of simultaneous control of the segmental and prosodic features of HMM-based speech synthesis.

In order to solve this situation, two methods have been developed; to reshape the F_0 contour generated by the HMM-based speech synthesis using the F_0 model, and to avoid degradation of synthetic speech due to erroneous voiced and unvoiced decision of the training corpus. The former reshape the F_0 contours taking the probabilistic factor of HMM-based speech synthesis into account [8]. This may reduce the mismatch between segmental and prosodic features as compared to separately generating both features, though the maximum likelihood condition is not satisfied. The latter satisfies the condition, because it only modifies the training corpus [9].

By handling F_0 contours in the F_0 model framework, a "flexible" control of prosodic features comes possible. It is rather easy to analyze the prosodic controls obtained by statistical methods and to modify generated F_0 contours in another corpus-based way, which is trained using a small speech corpus. As an example for the flexible control, we have developed a method of focus control [10]. Given a speech synthesis system without specific focus control, it is not efficient to prepare a large speech corpus with focus control and train the speech synthesis system from the beginning. The method predicts the differences in F_0 model commands between utterances with and without focuses, and modifies the F_0 model commands of the synthetic speech without focus.

2 Modeling F_0 Contours

The movement of F_0 is well represented by the generation process model (henceforth F_0 model) [3]. As shown in Fig. 1, it is a command-response model that describes F_0 contours in logarithmic scale as the superposition of phrase and accent components. The *i*th phrase component $G_{pi}(t)$ is generated by a second-order, critically-damped linear filter in response to an impulse-like phrase command, while the *j*th accent component $G_{aj}(t)$ is generated by another second-order, critically-damped linear filter in response to a stepwise accent command:

$$G_{pi}(t) = \begin{cases} \alpha_i^{\,2} t\, e^{-\alpha t} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{1}$$

$$G_{aj}(t) = \begin{cases} \min\left[1 - (1 + \beta_j\, t)\, e^{-\beta t}, \gamma\right] & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{2}$$

Based on the analysis of Japanese utterances, time constants α_i and β_j are known to be almost fixed to 3.0 s^{-1} and 20.0 s^{-1}, respectively. The parameter γ thresholds accent components can also be set to a fixed value around 0.9. An F_0 contour is then given by the following equation:

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} A_{aj} \{G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j})\} \quad (3)$$

where, F_b is the bias level, I is the number of phrase components, J is number of accent components, 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.

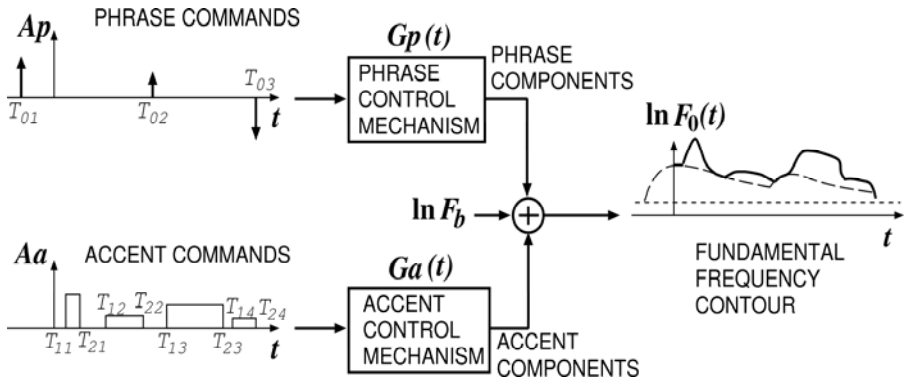

Figure 1 – F_0 model

The model has been shown to be applicable for many languages. In the case of Mandarin, each syllable can have up to four tone types with different F_0 contours, and F_0's often take values below the phrase components. To represent this F_0 movement, tone components are introduced instead of accent components: tone components have negative values corresponding to negative commands [11].

3 Improving Prosodic Control in HMM-based Speech Synthesis

In HMM-based speech synthesis, an F_0 sequence is modeled by multi-space probability distribution (MSD) HMMs [1]. It combines discrete HMMs (for voiced/unvoiced signs) and continuous HMMs (for F_0's and their Δ and Δ^2 values). F_0 contours are generated from these HMMs under the maximum likelihood criterion. In order to avoid over-smoothed F_0 contours, global variances (GVs) of the F_0 sequences are modeled by a single Gaussian distribution and taken into account.

3.1 Reshaping F_0 contours

In spite of the inclusion of deliberative values of F_0 in HMM parameters and consideration of global variances, there are still cases of over-smoothed F_0 contours and F_0 undulations not corresponding to the linguistic information of input texts. In order to solve this situation, a method was developed to generate F_0 contours using the F_0 model and to use them for speech synthesis. The method first decides initial positions of F_0 model commands from the linguistic information and estimates their magnitudes/amplitudes from the F_0 contours generated by the HMM-based speech synthesis. The estimation is done by taking derivatives of F_0 sequences smoothed by piece-wise third order polynomials [4]. Then the F_0 model parameter values are optimized recursively so as to minimize the following value;

$$(\boldsymbol{p} - \hat{\boldsymbol{p}}_c)^T U^{-1} (\boldsymbol{p} - \hat{\boldsymbol{p}}_c) \quad (4)$$

where $\boldsymbol{p}$ is the F_0 sequence generated by the F_0 model and $\hat{\boldsymbol{p}}_c$ is that of generated by HMM-based speech synthesis. U is the diagonal matrix of variances, which are determined by the HMMs. The method is similar to the method to find out optimum F_0 model parameters for an

observed F_0 contour [4], but different in that it takes variances of F_0 contours generated by the HMM-based speech synthesis.

To evaluate the method, speech synthesis was conducted for two Japanese native speakers' utterances (one male and one female) included in ATR continuous speech corpus. Out of 503 sentence utterances for each speaker, 450 utterances were used for the HMM training. Two versions of speech were generated for the rest of 53 sentences; one by the original HMM-based speech synthesis and the other by the proposed method. Their qualities were compared through RAB test by 12 native speakers of Japanese. For R sounds, natural utterances were used. A 5-point scoring was employed; 2 (proposed method is much better) and -2 (original HMM-based speech synthesis is much better). As shown in Fig. 2, significant improvements are observed for the proposed method in 8 sentences (male speaker) and 9 sentences (female speaker), though the difference between two methods are not clear for the rest of sentences. The total mean scores are 0.252 with a 95 % confidence interval [0.168, 0.335] and 0.230 with a 95 % confidential interval [0.148, 0.311] for male and female speakers, respectively. Clear improvements by the proposed method are observable especially when the original HMM-based speech synthesis generates erroneous F_0 contours; F_0 contours with local undulations not corresponding to the linguistic information of input texts.

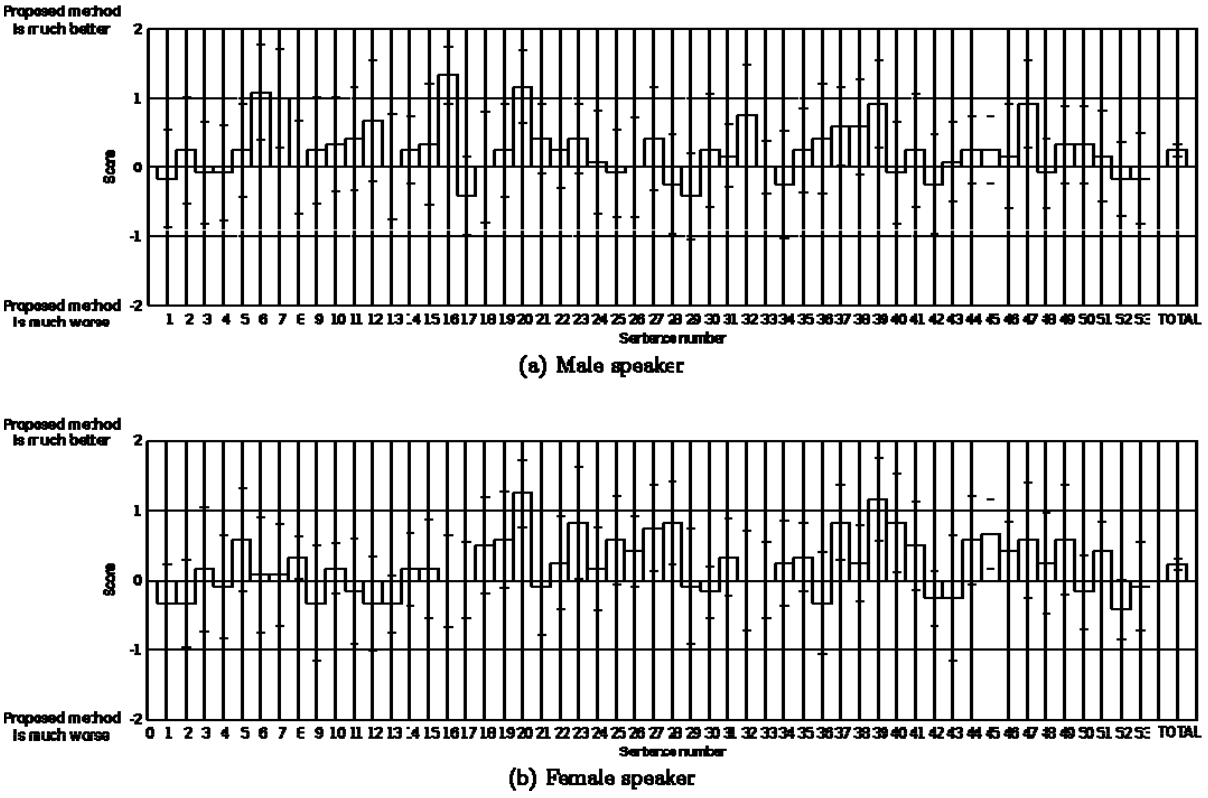

Figure 2 – Result of subjective comparison of synthetic speech quality (RAB test).

3.2 Coping with voiced and unvoiced decision errors

In HMM-based speech synthesis system, the Voiced/Unvoiced (*VU*) decision of each state is independently made based on the multi-space distribution of F_0 parameters of that state. The multi-space distribution of F_0 parameters of one state is estimated by traversing the decision tree by the contextual features till a leaf node. Due to pitch tracking errors or badly pronounced vowels, a leaf for a state belonging to a vowel may contain more unvoiced occurrences than voiced occurrences. Thus, if that leaf is chosen, the corresponding state is decided as unvoiced. Then the voice quality is degraded not only by the pitch tracking errors, but also by the *VU* decision errors in HMM training.

Due to larger dynamic F_0 ranges, the above problem becomes a serious issue for tonal languages such as Chinese. When automatic F_0 extraction is conducted (using ESPS RAPT algorithm) for the Mandarin speech corpus with 300 sentences by a female speaker, almost 22.37% syllables of the total include the *VU* decision errors; among these errors, 33% failures are occurred in T4, 39% in T3, 11% in T0, 12% in T2 and 5% in T1. After training process of MSD-HMM, the errors will increase due to hard *VU* decision of states.

This consideration lead us to an idea of generating continuous F_0 contours (for utterances in training corpus) assuming F_0's in unvoiced regions. F_0 model interpolation is used for the purpose. (F_0 model commands are extracted by *FujiPara* Editor [12].) When generating F_0 contours, *VU* decision is done according to the phoneme segmental information; we defined Mandarin phonemes with either voiced or unvoiced as shown in Table 1.

Table 1 – Mandarin initial and tonal final units (in pinyin) with voiced or unvoiced decision.

Unvoiced Initials	b, c, ch, d, f, g, h, j, k, p, q, s, sh, t, x, z, zh
Voiced Initials	l, m, n, r, u, y
Voiced Tonal Finals	a, ai, an, ang, ao, e, ei, en, eng, er, i, ia, ian, iang, iao, ie, ii, iii, in, ing, iong, o, ong, ou, u, ua, uai, uan, uang, uei, uen, uo, v, van, ve, vn

To evaluate the performance of our method as compared to the MSD-HMM, the above 300 sentences of Mandarin speech corpus are divided into 270 sentences for HMM training and 30 sentences for testing. The labels of unvoiced initials attached to the corpus are used as the boundaries of *VU* switch. The input text to the system includes symbols on pronunciation and prosodic boundaries, which can be obtained from orthogonal text using a natural language processing system, developed at University of Science and Technology of China.

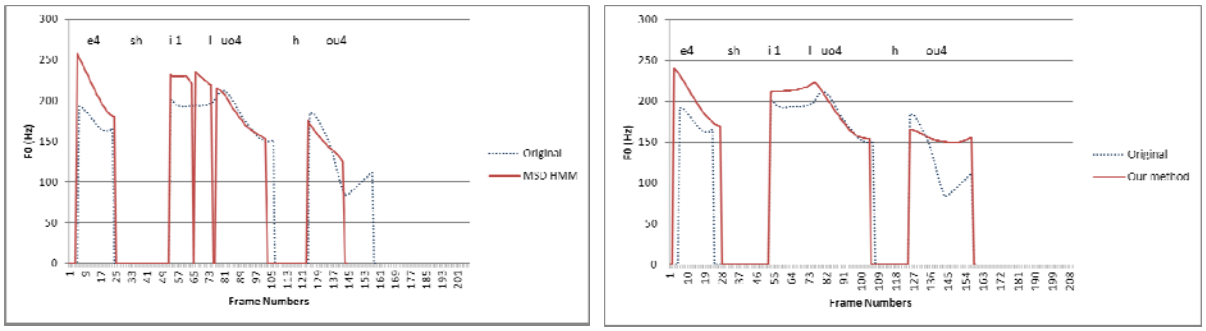

Figure 3 - F_0 contours predicted by MSD and our method, along with corresponding original F_0 contour of natural utterance.

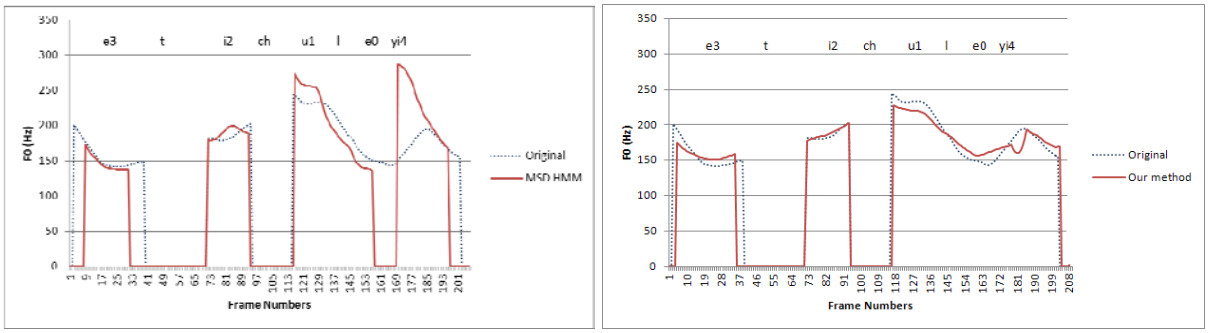

Figure 4 - F_0 contours predicted by MSD and our method, along with corresponding original F_0 contour of natural utterance.

Figures 3 and 4 show examples of F_0 contours generated by MSD-HMM and by our approach, overlaid onto those of the corresponding original (natural) utterances. The sentences shown in Figs. 3 and 4 consist of 4 and 5 Mandarin syllables, respectively: “she4+shi1+luo4+hou4.”

and “zhe3+ti2+chu1+le0+yi4.” Here, the syllables “zhe3” in Tone 3 and “hou4” in Tone 4 are difficult to be synthesized correctly because of large dynamic ranges of their F_0 contours and occasional creaky phonation. The syllable “le0” in neutral tone is also hard to be synthesized correctly; reduced power and highly contextually-dependent F_0 contour make accurate pitch tracking difficult. As shown in the figures, the MSD-HMM-based synthesizer has *VU* decision errors in “shi1,” “luo4,” “zhe3,” and “le0” syllables with inaccurate F_0 contours. On the contrary, our method can generate F_0 contours closer to original utterances with less *VU* decision errors.

Table 2 compares the root mean squared errors (RMSEs) of F_0 and phone duration predictions for MSD-HMM and our method. The RMSE of F_0 is calculated in voiced regions using the forced-aligned state durations. The RMSE of phone durations is calculated between the forced-aligned phone durations and the synthesized phone durations in both voiced and unvoiced regions except the silences. The RMSEs of F_0's and phone durations shown in the table are those averaged over all the test samples. When continuous F_0 contours are used in the HMM training (our method), the *VU* decision errors are significantly reduced. This situation contributes to the better prediction of phone durations by our method. Advantage of our method over MSD-HMM is clear from the reduced RMSE in F_0 prediction.

Table 2 – RMSEs of F_0's and phone durations predicted by MSD-HMM and by HMM trained using continuous F_0 contours (our method).

	RMSE of F_0	RMSE of phone duration
MSD-HMM	52.8 Hz	28 ms
Continuous F_0	29.7 Hz	24 ms

4 Realizing focuses in speech synthesis

Although emphasis of word(s) is not handled explicitly in most current speech synthesis systems, its control comes important in many situations, such as when the systems are used for generating reply speech in spoken dialogue systems: words conveying key information to the user's question need to be emphasized. Emphasis associated with narrow focus in speech can be achieved by contrasting the F_0's of the word(s) to be focused from those of neighboring words.

This contrast can be achieved by placing a phrase command (or increasing phrase command magnitude, when a command already exists) at the beginning of the word(s), by increasing the accent command amplitudes of the word(s), and by decreasing the accent command amplitudes of the neighboring words. The way of using these three controls may be different from language to language.

Although it is possible to realize prosodic focuses in speech synthesis by preparing a speech corpus with focus control and training the speech system from the beginning, it is time consuming. Given an HMM-based speech synthesis system without focus control, an efficient way to realize prosodic focus is to adapt HMMs to speech samples with focuses. However, an efficient and better adaptation is possible in the F_0 model frame-work.

As mentioned briefly in section 1, a corpus-based method of predicting F_0 model commands from input text was already developed [7]. In the method, a binary decision tree (BDT) is trained for each model parameter and used for the prediction. Training was done for ATR continuous speech corpus and speech synthesis system was constructed by combining with HMM-based speech synthesis. Since the corpus did not include apparent focus control, the resulting system was without focus control.

To realize prosodic focus, a method was proposed to modify the command magnitudes/amplitudes predicted by the above baseline system. In the method, first, training

of BDT's is conducted for differences in phrase command magnitudes and accent command amplitudes between utterances with and without focuses. Then, the command values (magnitudes/amplitudes) predicted by the baseline system (for utterances without specific focuses) are modified using the differences. The modification is first applied to the phrase command magnitudes and then to the accent command amplitudes taking the (modified) phrase command information into account. (In the current experiment, training is conducted using phrase command information observable in the corpus for the baseline system, while prediction is done using modified phrase commands. It is possible to use the modified phrase component also for the training, which is more consistent.) Table 3 shows input parameters for the binary decision trees for predicting phrase command magnitude differences. By concentrating to the differences, a better training for F_0 change due to focal position comes possible only with a limited speech corpus. Moreover, speakers for the training need not be the same for those of the baseline. (HMM-based speech synthesis requires the target speech for adaptation.)

Table 3 - Input parameters for the prediction of differences in phrase command magnitudes. Category numbers of "number of *morae*" and "accent type" for preceding *bunsetsu* are larger by one than those of current *bunsetsu* to indicate "no preceding *bunsetsu*." Here, *bunsetsu* is defined as a basic unit of Japanese syntax and pronunciation consisting of content word(s) followed or not followed by particles. Boundary depth code (BDC) indicates the distance from the *bunsetsu* immediately before the boundary to the *bunsetsu* directly modified.

Input parameter	Category
Position of current *bunsetsu* in prosodic phrase	3
Position of current *bunsetsu* in prosodic clause	4
Position of current *bunsetsu* in sentence	5
Distance of current *bunsetsu* from *bunsetsu* with focus (in number of *bunsetsu*'s)	5
Number of *morae* of current *bunsetsu*	4
Number of *morae* of preceding *bunsetsu*	5
Accent type (location of accent nucleus) of current *bunsetsu*	4
Accent type (location of accent nucleus) of preceding *bunsetsu*	5
BDC at the boundary immediately before current *bunsetsu*	9
Pause immediately before current *bunsetsu*	2 (yes or no)
Length of pause immediately before current *bunsetsu*	Continuous
Phrase command for the preceding *bunsetsu*	2 (yes or no)
Number of *morae* between preceding phrase command and head of current *bunsetsu*	4
Magnitude of current phrase command	Continuous
Magnitude of preceding phrase command	Continuous

We selected 50 sentences from the 503 sentences of the ATR continuous speech corpus, and asked a female speaker (different from the speaker for the baseline system) to utter each sentence without (specific) focus and with focus on one of assigned words (*bunsetsu*'s). For each sentence, 2 to 4 *bunsetsu*'s were assigned depending on the sentence length. As the result, 50 utterances without focus and 172 utterances with focus on one of noun phrases (*bunsetsu* including a noun) are obtained. These utterances are used to train BDT's for command magnitude/amplitude prediction. There are cases where phrase command magnitudes take minus values after modification. Since minus magnitudes are not allowed in the F_0 model, they are set to zero for the current experiment.

Figure 5 shows examples of generated F_0 contours when the predicted changes are applied to F_0 model parameters predicted by the baseline system. Although the baseline system includes prediction of pauses and phone durations, no modification is applied to those values. This is because changes in pauses and phone durations due to focuses are not significant in the case of Japanese. The three controls, viz., increasing phrase command magnitudes, increasing accent command amplitudes for focused words, and decreasing accent command amplitudes

of neighboring words, can be seen in the figure. Here we should note that the speaker to train the command differences is different from one (the narrator) for training the baseline method.

In order to check the effect of the focus control for realizing emphasis, a perceptual experiment was conducted for the synthetic speech. Twenty six sentences not included in the 50 sentences for training command magnitude/amplitude differences are selected from the 503 sentences of the ATR continuous speech corpus, and one synthetic utterance is selected for each sentence; 19 utterances with focus and 7 utterances without focus. Eleven native speakers of Japanese were asked to listen to these utterances and check *bunsetsu* where they perceived an emphasis. "No emphasis" answer was allowed. On average, in 76.1 % cases, the *bunsetsu*'s focused by the proposed method were perceived as "with emphasis." If "no emphasis" answers are excluded from the statistics, the rate increases to 83.7 %.

Modification of F_0 contours may cause degradation in synthetic speech quality. In order to check this point, the same 11 speakers were also asked to evaluate the synthetic speech from naturalness in prosody in 5-point scoring (5: very natural, 1: very unnatural). No apparent degradation is observed from the result; 3.03 (standard deviation 1.00) for utterances with focus and 3.12 (standard deviation 0.93) for those without.

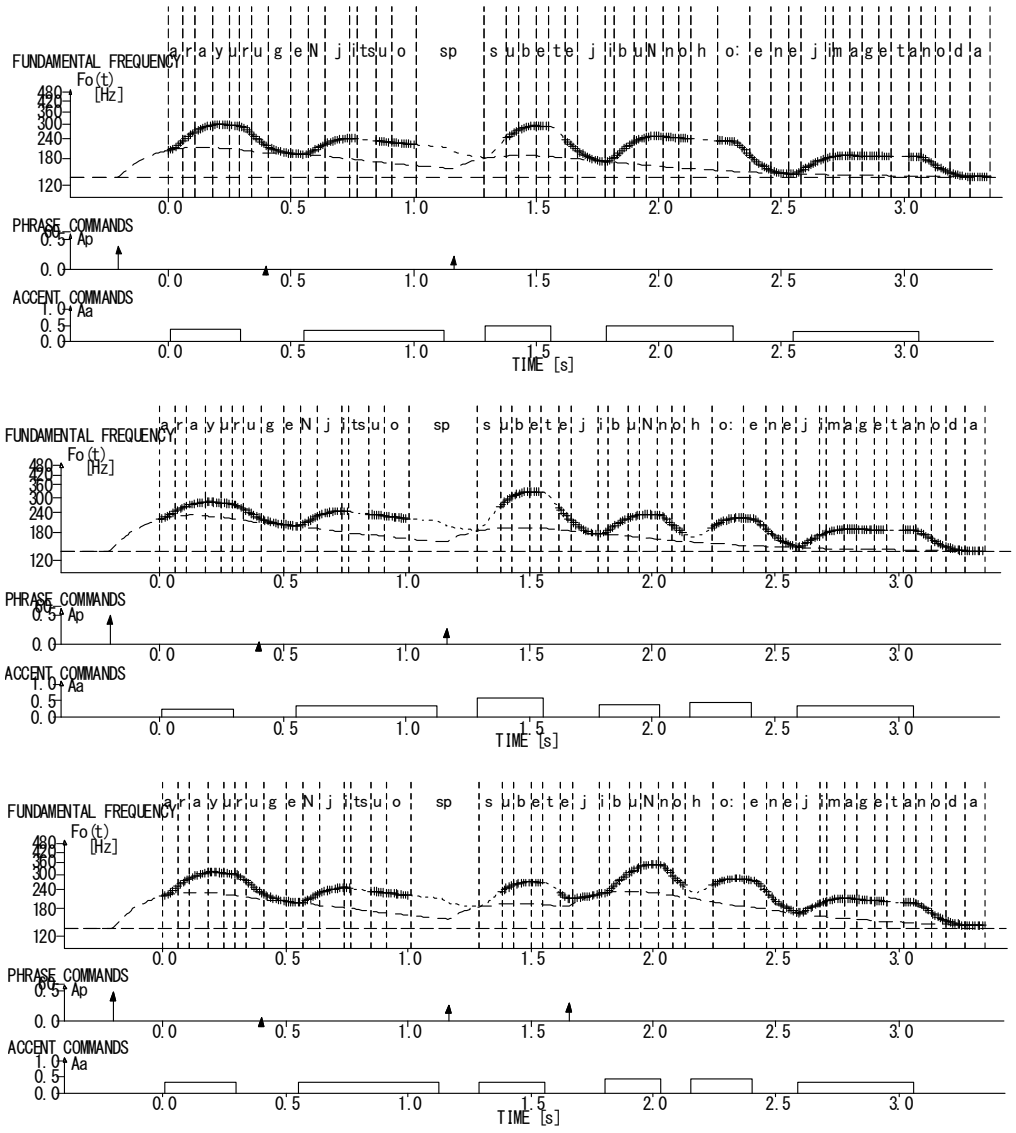

Figure 5 - Generated F_0 contours and F_0 model parameters for Japanese sentence "arayuru geNjitsuo subete jibuNnohooe nejimagetanoda ((He) twisted all the reality to his side.)." From the top to the bottom panels: without specific focus, focus on "subete," and focus on "jibuNnohooe," respectively.

Since focus is represented as changes in the F_0 model command magnitudes/amplitudes, emphasis levels can be controlled easily by interpolating/extrapolating the changes [13]. Experiments were conducted by selecting 64 sentences (from the 503 sentences of the ATR continuous speech corpus) not included in the 50 sentences for training command magnitude/amplitude differences. Their predicted differences in command magnitudes/amplitudes were multiplied by the scale factor r before applied to the command magnitudes/amplitudes predicted by the baseline method. For each sentence, one scale factor r was selected from 8 levels ranging from 0 (baseline) to 1.7 as shown in Table 4, so that the same sentence did not appear in a series of perceptual experiment. Speech synthesis was conducted for each generated F_0 contours, and totally 64 speech samples were prepared. (Eight speech samples for each scale factor.) Four native speakers of Japanese were asked to evaluate the naturalness and to judge emphasis levels for the synthetic speech. The evaluation/judgment was done again in 5-point scoring. As for the emphasis levels, score 5 is for strong emphasis and score 1 is for no emphasis. Scores for naturalness is the same with the former experiment. As shown in Table 4, emphasis levels can be changed by the interpolation/extrapolation without serious degradation in naturalness. The emphasis level is perceived as 2.68 in the case $r = 0$ (no focus). This may be due to the default focus; the phrase initial word/*bunsetsu* is usually perceived as focused.

The proposed method assumes no change in the prosodic structures for utterances with and without focuses; prosodic words are the same for the both cases. Although, in Japanese, it is true for most cases, focuses can be realized also by raising F_0 only for particles of the *bunsetsu*'s to be focused, for instance. The situation will be more complicated when we try to realize attitudes and emotions as the differences in the F_0 model command level; changes in prosodic structures should be taken into account. The situation will be different for languages. Since, in Japanese, each word/*bunsetsu* has its own "accent type," F_0 rise/fall timings respect to the syllable boundaries should not change depending on the focuses. However, this may not be true for other languages, where each word needs not necessarily have a specific F_0 rise/fall pattern.

Table 4 - Result of perceptual experiment for synthetic speech with various interpolation/extrapolation levels on the command magnitudes/amplitudes.

r	Naturalness	Emphasis
1.70	2.91	4.13
1.50	3.22	3.97
1.30	3.50	3.89
1.00	3.71	4.06
0.75	3.19	3.75
0.50	3.50	3.50
0.25	3.44	3.47
0 (without focus)	3.18	2.68

5 Conclusion

Two methods for improving prosody control in HMM-based speech synthesis and one method of adding flexibility in speech synthesis are developed. All the methods are based on the F_0 model, which provides us clear relations between F_0 contours and linguistic and para/non linguistic information conveyed by spoken language. The experimental results (listening tests of synthetic speech) show advantages of the developed methods over the baseline (original) HMM-based speech synthesis. Further researches are necessary for; to incorporate F_0 model constraints directly in HMM-based speech synthesis, and to realize flexible F_0 control other than for prosodic focuses.

Our sincere thanks are due to Prof. Renhua Wang and his colleagues in the University of Science and Technology of China for their providing us the Mandarin speech corpus.

References

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

[2] H. Fujisaki and H. Sudo, "A model for the generation of fundamental frequency contours of Japanese word accent," *J. Acoust. Soc. Japan*, Vol.27, pp.445-453 (1971).

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

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

[5] K. Hirose, Y. Furuyama, S. Narusawa, N. Minematsu, and H. Fujisaki, "Use of linguistic information for automatic extraction of F_0 contour generation process model parameters," *Proc. Oriental COCOSDA*, pp. 38-45 (2003).

[6] K. Hirose, Y. Furuyama, and N. Minematsu. "Corpus-based extraction of F_0 contour generation process model parameters," *Proc. INTERSPEECH*, pp. 3257-3260.

[7] K. Hirose, K. Sato, Y. Asano, and N. Minematsu, "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).

[8] T. Matsuda, K. Hirose, and N. Minematsu, "HMM-based synthesis of fundamental frequency contours using the generation process model," *J. Signal Processing*, Vol.14, to be published (2010).

[9] M. Wang, M. Wen, K. Hirose, and N. Minematsu, "Improved generation of fundamental frequency in HMM-based speech synthesis using generation process model," Proc. INTERSPEECH, to be published (2010).

[10] K. Ochi, K. Hirose, and N. Minematsu, "Control of prosodic focus in corpus-based generation of fundamental frequency contours of Japanese based on the generation process model," *Proc. IEEE ICASSP*, pp.4485-4488 (2009).

[11] K. Hirose, H. Lei, and H. Fujisaki, "Analysis and formulation of prosodic features of speech in standard Chinese based on a model of generating fundamental frequency contours," *J. Acoust. Soc. Japan*, Vol.50, No.3, pp.177-187 (1994).

[12] H. Mixdorff, Y. Hu, and G. Chen, "Towards the automatic extraction of Fujisaki model parameters for Mandarin," *Proc. INTERSPEECH*, pp.873-876 (2003).

[13] K. Ochi, K. Hirose, and N. Minematsu, "Realization of prosodic focuses in corpus-based generation of fundamental frequency contours of Japanese based on the generation process model," *Proc. International Conf. on Speech Prosody*, CD-ROM (2010).