Fundamental Frequency Contour Reshaping in HMM-based Speech Synthesis and Realization of Prosodic Focus Using Generation Process Model

Keikichi Hirose, Hiroya Hashimoto, Jun Ikeshima, and Nobuaki Minematsu

Department of Information and Communication Engineering, the University of Tokyo, Tokyo
{hirose, hiroya, ikeshima, mine}gavo.t.u-tokyo.ac.jp

Abstract

Frame-by-frame representation is not appropriate for prosodic features, which are tightly related to speech units spreading a wide time span, such as words, phrases and so on. This causes an inherit problem in fundamental frequency (F_0) contour generation by HMM-based speech synthesis. Our formerly-developed method, which modify generated F_0 contours in the framework of the generation process model, is improved to allow plural phrase components in a breath group. Since the model can clearly relate its commands with linguistic (and para-/non- linguistic) information, the method further enables flexible controls of prosody through manipulating model commands. Prosodic focus is realized in HMM-based speech synthesis as a supplemental process; viewing the differences of command magnitudes/amplitudes between utterances without and with focus. Validity of the method was confirmed by listening experiments of synthetic speech.

Index Terms: fundamental frequency contour, generation process model, HMM-based speech synthesis, prosodic focus

1. Introduction

Recently, in speech synthesis community, HMM-based speech synthesis attains researchers' special attentions, since it enables a flexible control in speech styles by adapting phone HMMs to a new style. The method processes segmental and prosodic features of speech 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 generally produces over-smoothed F_0 contours with occasional F_0 undulations not observable in human speech especially when the training data are limited. Moreover, relation of the generated F_0 contours with linguistic (and para-/non- linguistic) information conveyed by them is unclear, preventing further processing, such as to add additional information to speech, to change speaking styles, etc. 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 (henceforth, 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/commands 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 speech, manipulation of generated F_0 contours is possible, leading to a flexible control of prosodic features.

We already have developed a corpus-based method of synthesizing F_0 contours in the framework of F_0 model and have combined it with HMM-based speech synthesis to realize speech synthesis in reading and dialogue styles with various emotions [4]. As an example for the flexible control, we also have developed a method of focus control [5]. 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 system from the beginning. The method realizes prosodic focus as a supplemental process to our corpus-based method of F_0 contour generation; train binary decision trees (BDT's) for differences in phrase command magnitudes and accent command amplitudes between utterances with and without focuses. The command values predicted by the baseline method (for utterances without specific focuses) are modified applying the differences. By concentrating on the predicted 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 speech need not be the same for those of the baseline.

However, in the method, 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 are controlled independently, which may cause speech quality degradation due to mismatches between the two types of features. Furthermore, F_0 model commands need to be extracted for utterances included in the training corpus. Automatic command extraction facilitates the process, but is erroneous without (time-consuming) manual correction. Performance of automatic extraction was improved by applying constraints on the command locations to meet the linguistic information of the utterances [6, 7], but it is not enough for making the process fully automatic.

In order to solve this situation, a method has been developed; to reshape the F_0 contours generated by the HMM-based speech synthesis under the F_0 model framework; modifying the F_0 contours to those generated by the F_0 model [8]. This may reduce mismatches between segmental and prosodic features as compared to separately generating both features, though the maximum likelihood condition is not strictly satisfied. The most significant advantage of the method is that F_0 contours are represented by F_0 model commands, and are further controlled easily to realize "flexible" controls in speech synthesis, viz. adding linguistic and/or para-/non- linguistic information, changing speaking styles, and so on. The above-mentioned method of focus control was applied successfully to the modified F_0 contours.

The rest of the paper is constructed as follows; after brief explanations on the F_0 model in section 2 and HMM-based speech synthesis in section 3, the method of F_0 contour reshaping is given in section 4. In section 5, focus control is conducted with listening experiments of synthetic speech. Section 6 concludes the paper.

2. Generation process model

Movements of F_0 along time axis are well represented by the F_0 model, which is a command-response model that describes F_0 contours in logarithmic scale as the superposition of phrase and accent components [3]. 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:

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

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

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, which 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:

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

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 i^{th} phrase command, A_{aj} is the amplitude of the j^{th} accent command, T_{0i} is the time of the i^{th} phrase command, T_{1j} is the onset time of the j^{th} accent command, and T_{2j} is the reset time of the j^{th} accent command.

3. HMM-based speech synthesis

HMM-based speech synthesis was conducted for utterances by a male Japanese narrator (MHT) included in ATR continuous speech corpus. Out of 503 sentence utterances, 450 utterances were selected for the HMM training. The utterances were sampled at 16 [kHz] with 16 [bit] accuracy, and windowed with a 25[ms] Hamming window with a 5 [ms] shift. The feature vector consisted of spectral and F_0 parameter ones. The spectral parameter vector consisted of 25 mel-cepstral coefficients (including the 0th one) and their *delta* and *delta*^{2} coefficients, while the F_0 parameter vector consisted of log F_0 and its *delta* and *delta*^{2} values. Five-state left-to-right HMMs with three emitting states were used. Each state has single diagonal Gaussian output distribution. Decision-tree-based clustering was conducted.

4. F_0 contour reshaping under F_0 model framework

The method of F_0 contour reshaping first decides initial positions of the F_0 model commands from the linguistic information (boundaries and accent types of accent phrases) of the sentence for speech synthesis, and then estimates their magnitudes/amplitudes from the F_0 contours generated by the HMM-based speech synthesis. The method is similar to the one to find out the F_0 model parameters for an observed F_0 contour [9], but a better extraction is possible. This is because initial positions for the F_0 model commands are decided from the linguistic information, and HMM-based speech synthesis generates F_0 contours free from pitch extraction errors.

The original reshaping method [10] first applies cubic spline interpolation to F_0 sequence $\hat{\boldsymbol{p}}_c$ generated by the HMM-based speech synthesis to obtain a smoothed continuous F_0 contour for each breath group (period delimited by long pauses (>300 ms)). Then, accent command amplitudes are calculated by taking derivative of the smoothed F_0 contour. Phrase command magnitudes are calculated from the residual of F_0 contour; smoothed F_0 contour minus accent components. Timing parameters are decided from corresponding syllable initial positions by a simple calculation. These parameters are then optimized by the steepest descent method under the following criterion;

$$\hat{\boldsymbol{p}}_{\boldsymbol{p}} = \arg\min_{\boldsymbol{p}} (\boldsymbol{p} - \hat{\boldsymbol{p}}_c)^{\mathrm{T}} U^{-1} (\boldsymbol{p} - \hat{\boldsymbol{p}}_c) \tag{4}$$

where $\boldsymbol{p}$ is the F_0 sequence generated by the F_0 model. U is the diagonal matrix of variances, which are obtained through the process of F_0 sequence generation in the HMM-based speech synthesis. To avoid over-adjustment, certain limitations are set to parameters during search (Timing commands T_{0i}, T_{1j} and T_{2j} are searched in ±200 [ms] range from the initial values, for instance). The numbers of commands I and J are fixed to their initial values. Other parameters α_i, β_j and γ are fixed to constant values as explained in section 2. All the processes are conducted separately for each breath group.

Finally, F_0's are calculated from the F_0 model using the optimized parameters, and are used for the speech synthesis.

Since estimation of phrase commands from the F_0 contour residual is erroneous, phrase commands are assumed only at breath group boundaries with no additional phrase commands in breath groups. This assumption may over-estimate amplitudes of accent commands locating other than breath group initial, but the F_0 contour approximation by the F_0 model is still good and the method improves synthetic speech quality. However, when realizing focus control as a supplemental process, the above situation need to be avoided for a better performance. This is because BDT's for estimating command magnitude/amplitude differences is trained using natural utterances, where phrase commands may occur frequently at breath group medial positions; causing mis-matching between training and application.

In order to solve this situation, search process of F_0 model commands is modified as follows to allow breath-group medial phase commands:

- For each vowel segment, when its F_0 contour has large discontinuities, delete F_0's far from F_0 median so that the F_0 variance of the segment comes below a preset value.
- Label vowel segments either L (low) or H (high) according to the accent types of accent phrases. Decide phrase command magnitude for each accent phrase using F_0 values of L segments. The method requires L segments before and after H segment(s). Since no L segments exist before H segment for type 1 accent and after H segment(s) for type 0 accent, they are estimated by subtracting a fixed value from the F_0 value of an H segment (first H segment for type 1

accent, and last H segment for type 0 accent).
- Accent command amplitudes are calculated from the F_0 (residual) value of H segment with the highest F_0.

The major changes are; no interpolation of F_0 sequences and estimating phrase commands first. The process of parameter optimization is also by the steepest descent method, but counts only vowel segments. This is because sharp undulations out of F_0 model are mostly observed at voiced consonants. The F_0 variances of equation (4) are not counted, since their effect is minor at vowel segments. When phrase/accent command magnitudes/amplitudes come to zero, such commands are deleted. Figure 1 shows the result of F_0 model reshaping by the original and the new methods. The new method generates an F_0 contour with clear declination and with two additional phrase commands at the second breath-group. Informal listening by authors indicates an advantage of the new method.

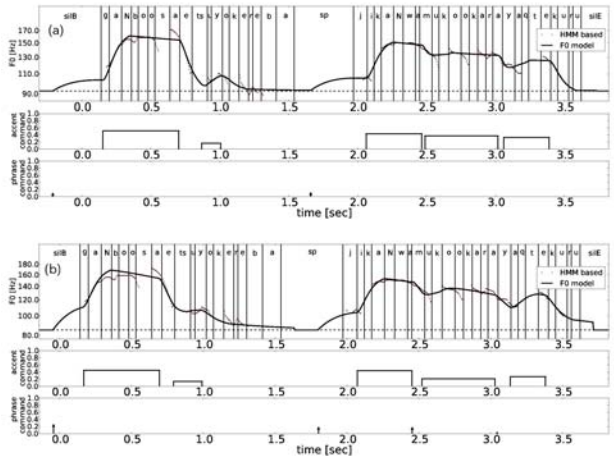

Figure 1: *F_0 contour reshaping by the F_0 model approximation for Japanese sentence "gaNboosae tsuyokereba jikaNwa mukookara yattekuru (Time will automatically come if (you) have a strong wish.)." F_0 contours and their F_0 model commands for (a) original method, and (b) new method.*

5. Focus control

As mentioned already, one of the most significant advantages of the method of F_0 reshaping is that the modified F_0 contours are represented by the F_0 model, allowing further modifications by changing F_0 model commands. Here, we try to add an emphasis on selected accent phrase to synthetic speech generated by the HMM-based speech synthesis.

Emphasis associated with narrow focus in speech can be achieved by contrasting the F_0's of the accent phrase to be focused from those of neighboring accent phrases. 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 accent phrase, by increasing the accent command amplitudes of the accent phrase, and by decreasing the accent command amplitudes of the neighboring accent phrases.

A corpus-based method was developed to predict differences in F_0 model commands between two versions of utterances of the same linguistic content [5, 11]. Applying the predicted differences to the baseline version of speech, the new version of speech can be realized. A large scale speech corpus is not necessary to train F_0 model command differences. This method is applied to realize prosodic focus. BDT's are first trained to predict command magnitude/amplitude differences between utterances without and with focuses. Since a new phrase command usually appears when a focus is placed where no phrase command initially exists (for an utterance without focus), a phrase command with zero magnitude is assumed there. Tables 1 and 2 show input parameters for BDT's. Fifty sentences were selected from ATR continuous speech corpus, and a female speaker (different from speakers of ATR corpus) was asked to utter them without focus and with focus on one of accent phrases (mostly those including a noun). BDT's are trained using these utterances; 50 utterances without focus and 183 utterances with focus.

Table 1. *Input parameters for the prediction of differences in phrase command magnitudes.*

Input parameter	Category
Distance of current accent phrase from accent phrase with focus (in number of accent phrases)	5
Number of *morae* of current accent phrase	3
Number of *morae* of preceding accent phrase	4
Number of *morae* of next accent phrase	4
Accent type (location of accent nucleus) of current accent phrase	4
Accent type (location of accent nucleus) of preceding accent phrase	5
Accent type (location of accent nucleus) of next accent phrase	5
Pause immediately before current accent phrase	2 (yes or no)
Magnitude of current phrase command	Continuous
Magnitude of preceding phrase command	Continuous

Table 2. *Input parameters for the prediction of differences in accent command amplitudes.*

Input parameter	Category
Distance of current accent phrase from accent phrase with focus (in number of accent phrases)	5
Number of *morae* of current accent phrase	3
Number of *morae* of preceding accent phrase	4
Number of *morae* of next accent phrase	4
Accent type (location of accent nucleus) of current accent phrase	4
Accent type (location of accent nucleus) of preceding accent phrase	5
Accent type (location of accent nucleus) of next accent phrase	5
Pause immediately before current accent phrase	2 (yes or no)
Amplitude of current accent command	Continuous
Amplitude of preceding accent command	Continuous

Figure 2 shows examples of F_0 contours before and after applying the predicted differences to F_0 model command magnitudes/amplitudes. Although prosodic focus involves changes also in pauses and phone durations, they are not counted in the current experiment to see the effect of F_0 contours. Three controls, viz., addition of new phrase

command, increase of accent command amplitude for the focused accent phrase, and decrease of accent command amplitudes of neighboring accent phrases, can be seen in the figure.

In order to check the effect of the focus control for realizing emphasis, a perceptual experiment was conducted for the synthetic speech. Eighteen 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 two versions of synthetic speech are played sequentially for each sentence. The first one is always utterances without focus (before F_0 change), and the second one is with focus for 16 sentences and without focus for the rest 2 sentences. (These numbers are not known by listeners.) Eleven native speakers of Japanese were asked to listen to these utterances and check an accent phrase where they perceived an emphasis. "No emphasis" answer was allowed. On average, in 77.3 % cases, the accent phrases focused by the proposed method were perceived as "with emphasis."

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; 3.07 (standard deviation 0.96) for utterances with focus and 3.25 (standard deviation 0.93) for those without.

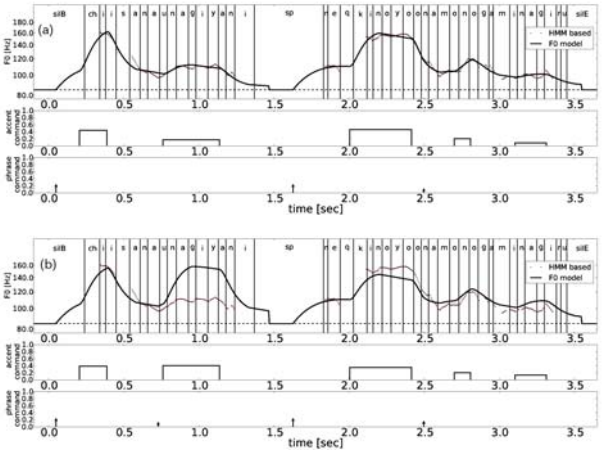

Figure 2: *F_0 contours and F_0 model parameters for Japanese sentence "chiisana unagiyani nekkino yoonamonoga minagiru (A small eel shop is filled with a kind of hot air.)." (a) without specific focus and (b) focus on "unagiyani." F_0 contour by HMM-based speech synthesis (without specific focus) is shown for comparison.*

6. Conclusions

F_0 model is ideal to view F_0 contours in wider time spans with good and clear relations to linguistic and para-/non- linguistic information of speech. A method was developed to increase the naturalness of prosody in synthetic speech generated by HMM-based speech synthesis. It is to modify F_0 contours through representing them with the F_0 model. One of major merits of representing F_0 contours by the F_0 model is that manipulation of prosody is possible as a supplemental process. The method of representing F_0 changes between two styles of utterances as command magnitude/amplitude differences was successfully applied to add prosodic focuses to HMM-based synthetic speech. Better results will be obtainable by inserting pauses (increasing pause lengths when pauses already existing) and changing phone durations. We can also concentrate on differences when handling these [12].

F_0 model stylization may include some problems. One is how to extract F_0 model command parameters automatically for given F_0 contours. The other is that F_0 features are not fully modeled by the F_0 model; F_0 movements due to micro-prosody are not modeled for instance. This situation is partly the reason of the first problem. These problems can be tackled by two ways; (1) to delete/discard F_0 movements not modeled by the F_0 model when searching F_0 model command parameters, and (2) to keep such F_0 movements as residuals of the F_0 model estimation and treat separately. Also, we plan to apply our method of F_0 contour modification to realize style/voice conversions in HMM-based synthesis.

7. 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] 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 Commu.*, Vol.46, Nos.3-4, pp.385-404 (2005).

[5] 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).

[6] 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).

[7] K. Hirose, Y. Furuyama, and N. Minematsu. "Corpus-based extraction of F_0 contour generation process model parameters," *Proc. INTERSPEECH*, pp. 3257-3260 (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, No.4, pp.277-280 (2010).

[9] 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).

[10] T. Matsuda, K. Hirose, and N. Minematsu, "HMM-based F_0 contour synthesis using the generation process model," Acoustical Science and Technology, Acoustical Society of Japan, to be published (2012).

[11] 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. Int. Conf. on Speech Prosody*, CD-ROM (2010).

[12] K. Hirose, K. Ochi, R. Mihara, H. Hashimoto, D. Saito, and N. Minematsu, "Adaptation of prosody in speech synthesis by changing command values of the generation process model of fundamental frequency," *Proc. INTERSPEECH*, pp.2793-2796 (2011).