Chapter 10
Use of Generation Process Model for Improved Control of Fundamental Frequency Contours in HMM-Based Speech Synthesis

Keikichi Hirose

Abstract The generation process model of fundamental frequency contours is ideal to represent the global features of prosody. It is a command response model, where the commands have clear relations with linguistic and para/nonlinguistic information conveyed by the utterance. By handling fundamental frequency contours in the framework of the generation process model, flexible prosody control becomes possible for speech synthesis. The model can be used to solve problems resulting from hidden Markov model (HMM)-based speech synthesis, which arise from the frame-by-frame treatment of fundamental frequencies. Methods are developed to add constraints based on the model before HMM training and after the speech synthesis processes. As for controls with increased flexibility, a method is developed to focus on the model differences in command magnitudes between the original and target styles. Prosodic focus is realized in synthetic speech with a small number of parallel speech samples, uttered by a speaker not among the speakers forming the training corpus for the baseline HMM-based speech synthesis. The method is also applied to voice and style conversions.

10.1 Introduction

Synthetic speech close to the quality of human utterances is now available through concatenation-based speech synthesis. However, the method requires a large speech corpus of the speaker and style to be synthesized. Thus, it is necessary to prepare a large speech corpus to realize each new voice quality with a new speaking style. Therefore, hidden Markov model (HMM)-based speech synthesis has garnered special attention from researchers, since it can generate synthetic speech with rather high quality from a smaller sized speech corpus, and can realize flexible control in voice qualities and speech styles. In this method, both segmental and prosodic features of speech are processed together in a frame-by-frame manner, and, therefore, it has

K. Hirose (✉)
Graduate School of Information Science and Technology,
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
e-mail: hirose@gavo.t.u-tokyo.ac.jp

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_10

the advantage that synchronization of both features is kept automatic (Tokuda et al. 2000). However, because of this, the frame-by-frame processing also includes an inherit problem when viewed from the aspect of prosodic features. Although it has the merit that fundamental frequency (F_0) of each frame can be used directly as the training datum, it generally produces oversmoothed F_0 contours with occasional F_0 undulations not observable in human speech, especially when the training data are limited. Moreover, the relationship of the generated F_0 contours with the linguistic (and para/nonlinguistic) information conveyed by them is unclear, preventing further processing, such as adding additional information, changing of speaking styles, etc. Prosodic features cover a wider time span than segmental features, and should be processed differently.

One possible solution to this problem is to apply the generation process model (F_0 model) developed by Fujisaki and his coworkers (Fujisaki and Hirose 1984). Details of the model are given in Chap. 3 of this book along with a method for the automatic extraction of model parameters from observed F_0 contours. The model represents a sentence F_0 contour in a logarithmic scale as the superposition of accent components on phrase components. These components are known to have clear correspondences with linguistic and para/nonlinguistic information, which are conveyed by prosody. Thus, using this model better control of prosody can be realized for F_0 contour generation than with the frame-by-frame control. Because of the clear relationship between generated F_0 contours and linguistic and para/nonlinguistic information of utterances, manipulation of generated F_0 contours is possible, leading to a flexible control of prosody.

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 (Hirose et al. 2005). In that method, F_0 contours generated by HMM-based speech synthesis were simply substituted with those generated by that method. Although, improved quality is obtained for the synthetic speech generated by the method, the controlling of segmental and prosodic features independently may cause speech quality degradation due to mismatches between the two types of features.

Introducing the F_0 model into HMM-based speech synthesis is not an easy task, since F_0 model commands cannot be well represented in a frame-by-frame manner. An effort has been reported that represents F_0 model in a statistical framework to cope with the above problem, but its implementation into HMM-based speech synthesis requires some additional works (Kameoka et al. 2013). Here, two simple procedures are adopted; one to approximate the F_0 contours of training speech data with the F_0 model, and to use these F_0s for HMM training (Hashimoto et al. 2012), and the other to reshape the generated F_0 contour under the F_0 model framework (Matsuda et al. 2012).

In order to fully reap the benefits of the F_0 model in speech synthesis, a critical problem must be solved, namely how to extract the F_0 model (command) parameters from observed F_0 contours. This process needs to be done at least semiautomatically to avoid the overly time-consuming process of manual extraction. Although several methods have been developed already, their performance is less than satisfactory,

suffering from two major problems: over-extraction of accent components resulting in minor accent components not corresponding to the linguistic content and under-extraction of phrase components resulting in missing phrase components. A new method has been developed for the automatic extraction of F_0 model parameters. It uses the linguistic information of text as constraints during the F_0 model parameter extraction (Hashimoto et al. 2012).

By handling F_0 contours in the framework of F_0 model, "flexible" control of prosodic features becomes possible. A corpus-based method has been developed to predict the differences in F_0 model commands between two versions of utterances containing the same linguistic content (Ochi et al. 2009). By applying the predicted differences to the baseline version of the speech, a new version of the speech can be realized. A large speech corpus is not necessary to train F_0 model command differences. This method is applied to realize prosodic focus (Ochi et al. 2009; Hirose et al. 2012), and speaking style and voice conversions (Hirose et al. 2011).

10.2 Automatic Extraction of F_0 Model Commands

Several methods have already been developed for the automatic extraction of F_0 model commands from given F_0 contours (Narusawa et al. 2002; Mixdorff et al. 2003). The basic idea behind them is as follows: smoothing to avoid microprosodic and erroneous F_0 movements, interpolating to obtain continuous F_0 contours, and taking derivatives of F_0 contours to extract accent command locations and amplitudes. Phrase commands are extracted from the residual F_0 contours (F_0 contour minus extracted accent components) or low-pass filtered F_0 contours. Extracted phrase and accent commands are optimized by an iterative process. These methods, however, are not robust against pitch extraction errors, and produce commands not corresponding to the linguistic information of the utterances to be analyzed. Although attempts have been carried out to reduce extraction errors by introducing constraints (on command locations) induced from the linguistic information, their performance was still not satisfactory.

Interpolation of F_0 contours has a drawback since it relies on F_0s around voiced/unvoiced boundaries, where F_0 extraction is not always precise. This situation leads to the extraction of false commands. Microprosodic F_0 movements during voiced consonants may also degrade the command extraction performance, since they are not expressed in the F_0 model. To avoid false extractions, a new method is developed where F_0 contours only of vowel segments are considered. The downside being that since no F_0 is available between vowels, it becomes difficult to extract accent commands from F_0 contour derivatives. Therefore, the method is designed to take the features of Japanese prosody into account (Hashimoto et al. 2012). In Japanese, F_0s of a syllable take either high or low values corresponding to accent types. The method extracts phrase commands first viewing "Low" parts and then estimates the accent command amplitudes from the "High" parts. The method can extract minor phrase commands that are difficult to be found from the residual

F_0 contours. We can say that the method is motivated from the human process of command extraction. Since it is developed taking Japanese prosody into account, further investigations are necessary to make it applicable to other languages.

10.3 Prosodic Control in HMM-Based Speech Synthesis

10.3.1 Using F_0 Contours Approximated by F_0 Model for HMM Training

F_0 contours usually appear as quasicontinuous curves, since F_0 values are unobservable in the unvoiced parts of speech. To cope with this situation, the multispace probability distribution HMM (MSD-HMM) scheme (Tokuda et al. 1999) is widely used, where discrete HMMs (for voiced/unvoiced signs) and continuous HMMs (for F_0 values and their Δ and Δ^2 values in voiced segments) are combined. When synthesizing, F_0 contours are generated from these HMMs under the maximum likelihood criterion with voiced/unvoiced signs. Using this scheme, efficient processing both in training and synthesis is realized. It is pointed out, however, that the method has a rather limited ability to do pitch tracking and is not robust against F_0 extraction errors (including voiced/unvoiced errors) of the training corpus. Due to F_0 tracking errors or poorly pronounced vowels, a leaf for a state belonging to a vowel may contain more unvoiced occurrences than voiced ones. Thus, if that leaf is chosen, the corresponding state is judged as unvoiced. This leads to the voice quality being degraded not only by the F_0 tracking errors, but also by the *VU* decision errors in HMM training. Due to their larger dynamic F_0 ranges, the problem becomes a serious issue for tonal languages such as Chinese.

To rectify this situation, continuous F_0 HMMs have been proposed (Yu and Young 2011, see Chap. 9). In order to obtain continuous F_0 contours for the training corpus, the method selects the "most probable" F_0 values for unvoiced regions during F_0 extraction processes. Interpolation of F_0 contours can also be used for that purpose. However, the resulting continuous F_0 contours still contain unstable F_0 movements due to microprosody and F_0 extraction errors. By using F_0 contours generated by the F_0 model for HMM training, the F_0 contours generated by the HMM-based speech synthesis can be stabilized.

This idea is first applied to Mandarin speech synthesis, where F_0 extraction often includes serious voiced/unvoiced errors especially for tones with low F_0s (Wang et al. 2010). To evaluate the performance of our method as compared to the MSD-HMM, a Mandarin speech corpus of 300 sentences is 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 the *VU* switch. The input text to the speech synthesis system includes symbols on pronunciation and prosodic boundaries, which can be obtained from orthographic text using a natural language processing system developed at the University of Science and Technology of China.

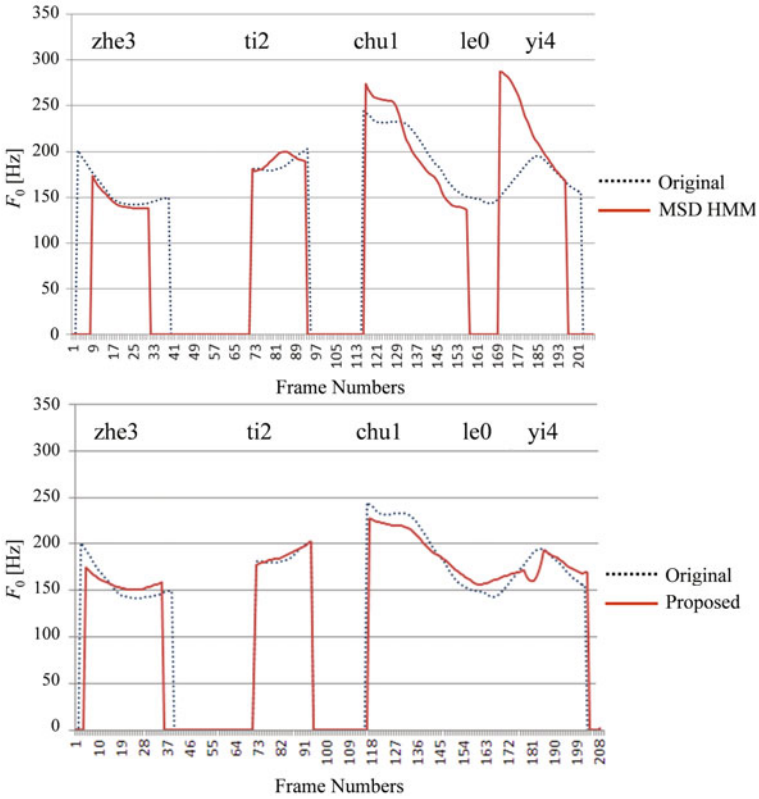

Fig. 10.1 F_0 contours generated by MSD-HMM and by the proposed method, along with corresponding original F_0 contour of natural utterance

Figure 10.1 shows examples of F_0 contours generated by MSD-HMM and by our approach, overlaid onto that of the corresponding original target utterance. The sentence is: "zhe3 + ti2 + chu1 + le0 + yi4." Here, the syllable "zhe3" (Tone 3) is difficult to synthesize correctly because of the large F_0 dynamic range in their contours and occasional creaky phonation. The syllable "le0" (neutral tone) is also hard to be synthesized correctly; reduced power and highly context-dependent F_0 contours make accurate F_0 tracking difficult. On the contrary, our method can generate F_0 contours closer to those of the original utterances with less *VU* decision errors. The validity of the method was demonstrated also through a listening experiment.

The method is then applied to Japanese. Since *VU* decision errors are few in number compared to Chinese, the MSD-HMM may work well in the case of Japanese. So a speech synthesis experiment was done using MSD-HMM. Two cases are compared; using F_0 contours extracted from speech waveforms (the original HMM) and using F_0 contours approximated by the F_0 model (proposed method). The F_0 model approximation is conducted through the automatic model command extraction explained in Sect. 10.2. It should be pointed out that accent phrase boundaries

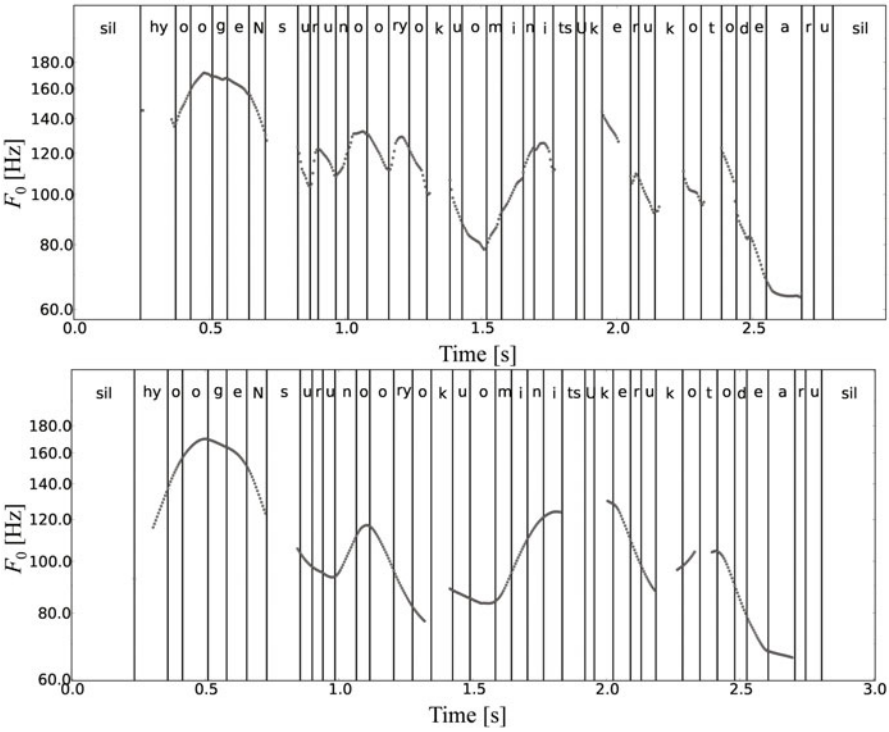

Fig. 10.2 Comparison of F_0 contours for Japanese sentence: "hyoogeNsurunooryokuo minitsukerukotodearu (It is to obtain an ability of expressing)." From *top to bottom*: F_0 contour generated by the original HMM and that generated by the proposed method

and accent types, which are given in the speech corpus, are both used in command extraction and HMM-based speech synthesis processes.

Speech synthesis experiments are conducted using the ATR continuous speech corpus of 503 sentences by speaker MHT. Out of the 503 sentences, 450 sentences are used for HMM training and rest 53 sentences are used for testing. HMM training is conducted using open software HTS-2.11[1]. Two versions of F_0 contours are prepared for the training: F_0 contours extracted from speech waveforms (original HMM), those generated by the F_0 model (proposed method). Speech signals are sampled at 16 kHz sampling rate, and STRAIGHT analysis is used to extract the spectral envelope, F_0, and aperiodicity with a 5 ms frame shift. The spectral envelope is converted to mel-cepstral coefficients using a recursive formula. The feature vector is 138 dimensional and consists of 40 mel-cepstral coefficients including the 0th coefficient, the logarithm of fundamental frequency, five band-aperiodicity (0–1, 1–2, 2–4, 4–6, 6–8 kHz) and their delta and delta–delta coefficients. A five-state left-to-right model topology is used for the HMM. Figure 10.2 compares F_0 contours

[1] http://hts.sp.nitech.ac.jp/.

generated by the original HMM and proposed method. It is clear from the figure that the F_0 contour by the proposed method is smooth.

The quality of synthetic speech from the original HMM and the proposed method is evaluated by means of a listening test with eight native speakers of Japanese. They are asked to listen to pairs of synthetic speech (one by the original HMM and the other by the proposed method) and select one from 5 scores (2: proposed method is clearly better, 1: proposed method is better, 0: no difference, -1: original HMM is better, -2: original HMM is clearly better). The average score over the 53 test sentences is 0.143 with $\pm$ 0.124 confidence interval, significant at a level of 5 %.

10.3.2 Reshaping F_0 Contours

A method was also developed to add an F_0 model constraint on the HMM-based speech synthesis before the speech waveform generation is carried out (Matsuda et al. 2012). The approach is to approximate F_0 contours generated by HMM-based speech synthesis using the F_0 model. The method first makes a decision on the initial positions of the 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. During the optimization process of F_0 model commands, F_0 variance obtained through the HMM-based speech synthesis process is used to normalize the F_0 mismatch between observed F_0s and the F_0s of F_0 model; F_0 mismatch is weighted less for frames with larger variances.

To evaluate the method, speech synthesis was conducted on two Japanese native speakers' utterances (one male and one female) included in the ATR continuous speech corpus. Out of the 503 sentence utterances for each speaker, 450 utterances were used for the HMM training. Two versions of speech were synthesized for the rest of the 53 sentences; one by the original HMM-based speech synthesis and the other by the proposed method. The difference in quality between them was calculated through a listening test with 12 native speakers of Japanese. A 5-point scoring method was employed; 2 (proposed method is much better) and -2 (original HMM-based speech synthesis is much better). 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 in the proposed method are observable especially in the cases when the original HMM-based speech synthesis generates erroneous F_0 contours. Figure 10.3 shows the reshaped F_0 contour compared with one generated by HMM-based synthesis.

10.4 Conversion of Prosody by F_0 Model Command Manipulation

The most significant advantage to adding an F_0 model constraint during speech synthesis is that the resulting F_0 contours are represented by F_0 model commands and can further be adjusted easily to realize flexible controls in speech synthesis. The

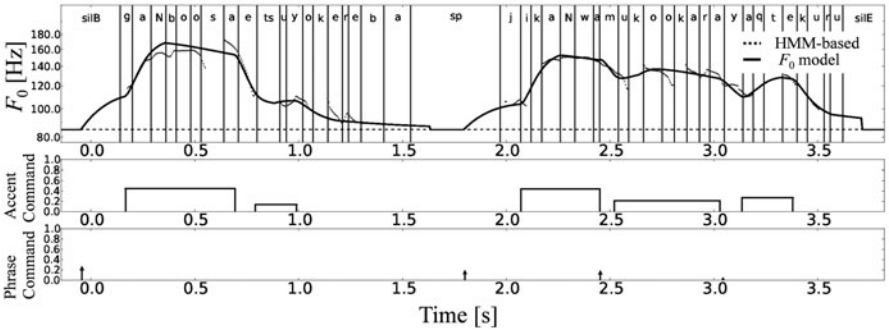

Fig. 10.3 F_0 contour reshaping by the F_0 model approximation for Japanese sentence "gaNboosae tsuyokereba jikaNwa mukookara yattekuru" (time will naturally come if (you) have a strong wish.)

method developed for prosody conversion consists of the following two processes (Hirose et al. 2011):

1. Extract F_0 model commands for the original and target utterances of the same sentence and calculate the differences in magnitudes/amplitudes of corresponding commands. Train binary decision trees (BDTs) to predict these differences.
2. Apply the differences to the phrase command magnitudes of the original utterances and then apply the differences to accent command amplitudes taking the modified phrase commands into account.

10.4.1 Prosodic Focus (Ochi et al. 2009)

Although emphasis of a word(s) is not handled explicitly in most current speech synthesis systems, its control becomes important in many situations, such as when the systems are used for generating response speech in spoken dialogue systems: words conveying information key to the user's question need to be emphasized. Emphasis associated with narrow focus in speech can be achieved by contrasting the F_0s of the word(s) to be focused on 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. These three controls may manifest differently in each language.

In order to investigate the situation for Japanese, we selected 50 sentences from the 503 sentences of the ATR continuous speech corpus and asked a female speaker to utter each sentence without (specific) focus and with focus assigned on one of the *bunsetsus*[2]. For each sentence, 2–4 *bunsetsus* were assigned depending on the

[2] "*bunsetsu*" is defined as a basic unit of Japanese syntax and pronunciation consisting of content word(s) followed or not followed by particles.

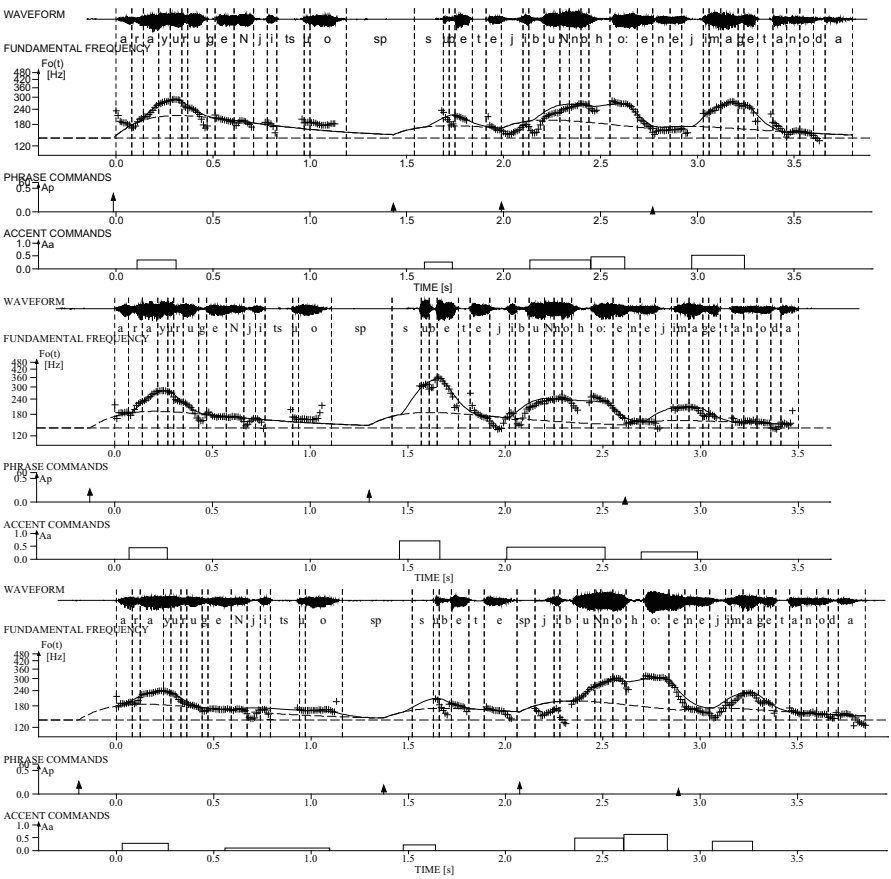

Fig. 10.4 F_0 contours and F_0 model parameters of Japanese sentence "arayuru geNjitsuo subete jibuNno hooe nejimagetanoda" ((He) twisted all the reality to his side.) uttered by a female speaker. The panels from *top* to *bottom*: without specific focus, focus on "subete," and focus on "jibuNno," respectively

sentence length. Figure 10.4 shows F_0 contours together with the results of the F_0 model approximations for utterances of the same sentence in different focal conditions. From the figure it is clear that the above three controls occur in the case of Japanese. It is also clear that there are one-to-one correspondences in phrase and accent command for different focal conditions. (Although "jibuNno hooe" has one accent command when focus is placed on "subete," it can be processed to have two commands with the same amplitude.) This one-to-one correspondence has inspired us to realize focus by controlling command magnitudes/amplitudes.

Tables 10.1 and 10.2 show input parameters for BDTs for predicting command magnitude/amplitude differences between utterances with and without focus. "BDC" in the tables denotes Boundary Depth Code, which represents the depth of syntactic

Table 10.1 Input parameters for the prediction of differences in phrase command magnitudes. The category numbers in parentheses are those for the directly preceding *bunsetsu*

Input parameter	Category
Position in prosodic phrase of current *bunsetsu*	3
Position in prosodic clause of current *bunsetsu*	4
Position in sentence of current *bunsetsu*	5
Distance from focal position (in *bunsetsu* number)	6
Length of *bunsetsu* (in number of *morae*)	4 (5)
Accent type of *bunetsu* (location of accent nucleus)	4 (5)
BDC at the boundary immediately before current *bunsetsu*	9
Existence of pause immediately before current *bunsetsu*	2
Length of pause immediately before current *bunsetsu*	Continuous
Existence of phrase command for the preceding *bunsetsu*	2
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

Table 10.2 Input parameters for the prediction of differences in accent command amplitudes. The category number in parentheses is those for the directly preceding and following *bunsetsu's*

Input parameter	Category
Position in sentence of current prosodic word	3
Position in prosodic phrase of current prosodic word	3
Position of prosodic phrase to which the current prosodic word belongs	2
Distance from focal position (in number of *bunsetsu*)	5
Accent type of *bunsetsu* (location of accent nucleus)	4 (5)
BDC at the boundary immediately before current *bunsetsu*	2
Amplitude of directly preceding accent command	Continuous
Amplitude of current accent command	Continuous
Magnitude of current phrase command	Continuous
Magnitude of preceding phrase command	Continuous

boundary and is obtainable by analyzing input sentences using the natural language parser KNP[3] (Hirose et al. 2005). The above utterances for investigation on focus control are used to train BDTs. They include 50 utterances without focus and 172 utterances with focus on one of the noun phrases (*bunsetsu* including a noun).

[3] http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html.

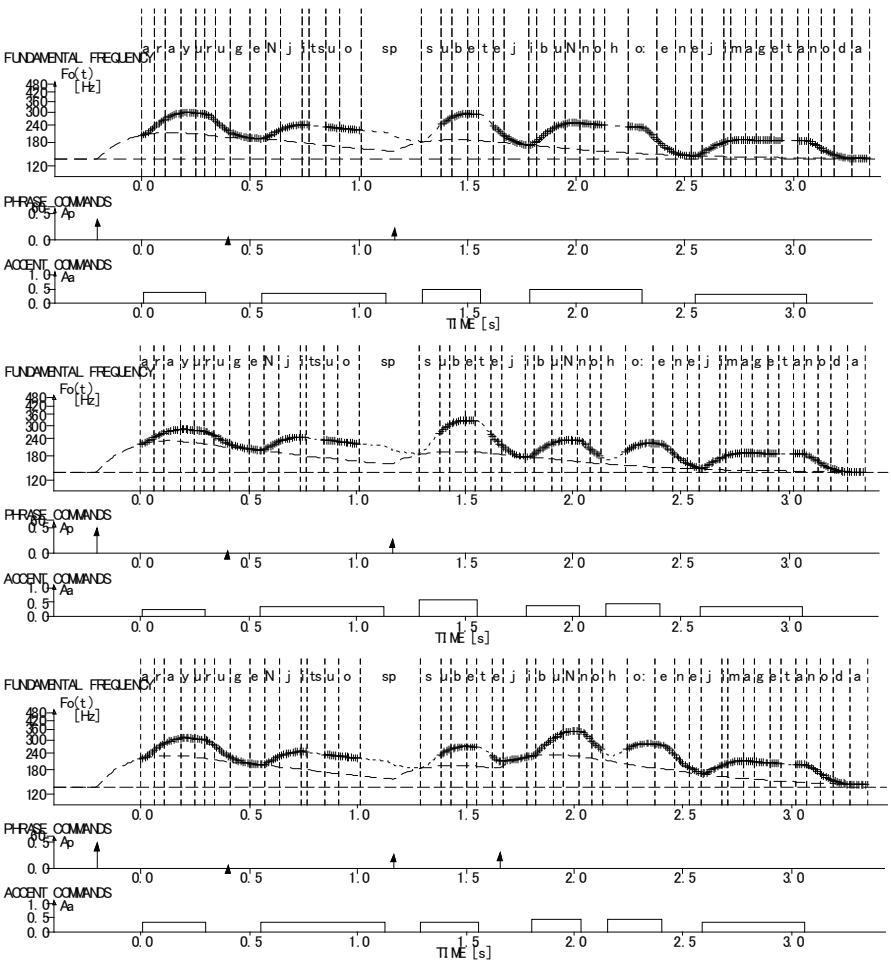

Fig. 10.5 Generated F_0 contours and F_0 model parameters. The sentence and focal conditions are the same with those shown in Fig. 10.2

As for the baseline speech synthesis on which focus control is applied, a combined method is adopted; F_0 model-based generation for F_0s with other acoustic features generated by HMM-based speech synthesis (Hirose et al. 2005). Figure 10.5 shows examples of generated F_0 contours when the predicted changes are applied to F_0 model parameters predicted by the baseline synthesis. Although prosodic focus also involves changes in pause and phone durations, they are not factored into the current experiment to focus on the effect of F_0 contours. The three controls listed above for focus control can be seen in the figure. Here we should note that the speaker used to train the command differences can be different from the one (the narrator) used for training baseline method.

In order to check the effect of the focus control for realizing emphasis, a perceptual experiment was conducted on the synthetic speech. Twenty-six sentences not

Table 10.3 Results of perceptual experiment for synthetic speech with various interpolation/extrapolation levels on the command magnitudes/amplitudes

	r	Naturalness	Emphasis
Extrapolation	1.70	2.91	4.13
	1.50	3.22	3.97
	1.30	3.50	3.89
Interpolation	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

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 the *bunsetsu* on which they perceived an emphasis. An answer specifying "No emphasis" was also made available. On average, in 76.1 % of the cases, the *bunsetsus* with focus placed 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 investigate this point, the same 11 speakers were also asked to evaluate the synthetic speech from the viewpoint of naturalness in prosody with 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.

Since focus is represented with the changes in the F_0 model command magnitudes/amplitudes, emphasis levels can be controlled easily by interpolating/extrapolating the changes (Ochi et al. 2010). Experiments were conducted by selecting 64 sentences (from the 503 sentences of the ATR continuous speech corpus) not included in the set of 50 sentences for training command magnitude/amplitude differences. The predicted differences in command magnitudes/amplitudes were multiplied by the scale factor r before applying it to the command magnitudes/amplitudes predicted by the baseline method. For each sentence, a scale factor r was selected from eight levels ranging from 0 (baseline) to 1.7 as shown in Table 10.2, so that the same sentence would not appear in a series of perceptual experiment. Speech synthesis was conducted for each generated F_0 contour and in total 64 speech samples were prepared (Eight speech samples for each scale factor). Four native speakers of Japanese were asked to evaluate the naturalness andjudge the emphasis levels for the synthetic speech. The evaluation/judgment was done in this case as well with 5-point scoring. As for the emphasis levels, a score of five means strong emphasis and score of one means no emphasis. Scoring for naturalness

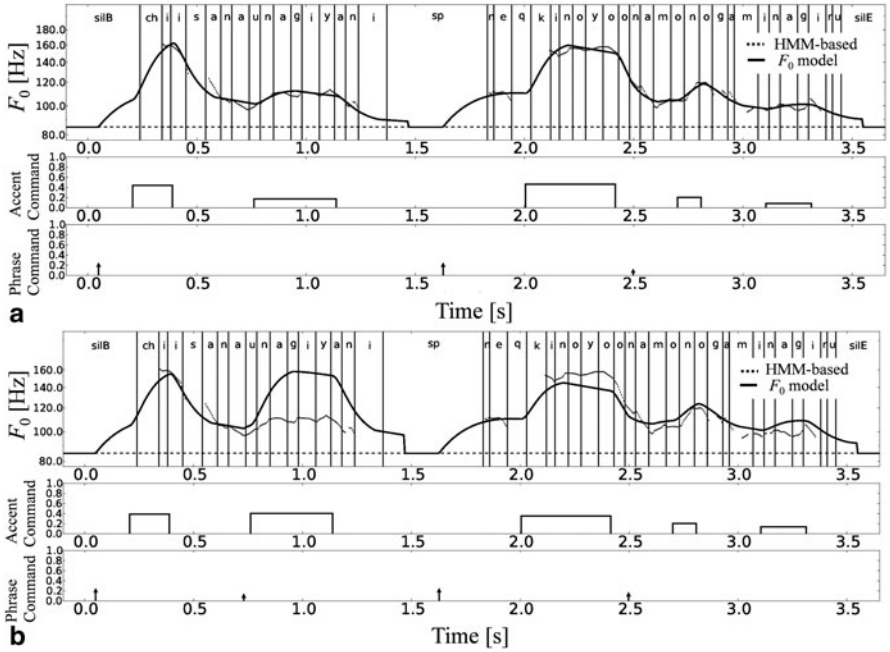

Fig. 10.6 F_0 contours and F_0 model parameters for Japanese sentence "chiisana unagiyani nekkino yoonamonoga minagiru (A small eel shop is filled with a kind of hot air)." **a** without specific focus and **b** focus on "unagiyani." F_0 contour by HMM-based speech synthesis (without specific focus) is shown for comparison

was done the same as in the former experiment. As shown in Table 10.3, 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, for which the phrase-initial word/*bunsetsu* is usually perceived as focused.

Prosodic focuses can be added in a similar way to F_0 contours reshaped by the F_0 model in Sect. 10.3.2. Figure 10.6 shows examples of (a) reshaped F_0 contour and (b) F_0 contour with prosodic focus on "*unagiyani*." It is assumed that prosodic focus can also be added to F_0 contours generated by HMM-based speech synthesis trained using F_0 model-based F_0 contours (Sect. 10.3.1). Although the F_0 model command extraction process is necessary for F_0 contours generated by the HMM-based speech synthesis before the command manipulation, from Fig. 10.2, it is expected to be achieved easily.

10.4.2 Voice Conversion (Hirose et al. 2011)

Voice conversion is a technique used to convert one voice to another without altering the linguistic (and para/nonlinguistic) contents of utterances, despite no knowledge

of these contents. Among various methods for voice conversion, those based on Gaussian mixture modeling (GMM) are widely used. In this Chapter, we take the method by Kain et al. (Kain et al. 2002) as the baseline method, where the cepstral features of original and target speakers' utterances of the same contents are tied to form joint feature vectors. Time synchrony between feature vectors is maintained through DP matching. In the method, F_0s are linearly converted using the means and standard deviations of the original and target speakers.

We replace this method with ours, which makes use of the differences in the F_0 model commands. Pause and phone durations are left unchanged. Although better prediction is possible by taking into account the linguistic information of the utterances, such as part of speech, syntactic structure, and so on, it is not included here to determine how the proposed method works with only parameters obtainable from the acoustic features of utterances.

Speech synthesis experiments were conducted using ATR continuous speech corpus of 503 sentences. Utterances by male narrator MHT are used as original utterances and those by female narrator FKS are used as target utterances. Out of the 503 sentences, 200 sentences and 53 sentences are selected, and used for training and testing (evaluation), respectively.

Ten native speakers of Japanese are asked to select the one (A or B) which is closer to X in AB-X test. A and B are synthetic speech produced by the baseline and proposed methods respectively, while X is the target speech. In order to avoid order effect, both cases with "A: original and B: proposed" and "A: proposed and B: original" are included in the stimuli. A score "1" or "$-$ 1" is assigned when speech by the proposed method is judged as being closer or farther to the target speech, respectively. When a subject cannot judge, a score of 0 is allowed. The average score over the 53 test sentences is 0.419 with $\pm$ 0.09 confidence interval at a significance level of 5 %.

10.5 Conclusions

Two methods are developed to improve the naturalness of prosody in HMM-based synthetic speech. Both are based on the F_0 model: one is to use F_0 contours approximated by the F_0 model for HMM training and the other is to reshape F_0 contours generated by the HMMs using the F_0 model. Prosodic focus is realized by manipulating the F_0 model command magnitudes/amplitudes, indicating that the F_0 model can add flexibility in prosody control. Voice conversion is also realized in the same framework.

Although the model constraint provides an improved control of F_0 contours in HMM-based speech synthesis, it has a drawback that F_0 movements not represented by the model are missing in synthetic speech. Effects of these fractional movements, such as microprosody, etc., on synthetic speech quality are assumed to be minor, but a scheme still needs to be developed to handle them properly. One possible solution is to predict these movements also in HMM-based speech synthesis.

References

Fujisaki, H., and K. Hirose. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustic Society of Japan. (E)* 5 (4): 233–242.

Hashimoto, H., K. Hirose, and N. Minematsu. 2012. Improved automatic extraction of generation process model commands and its use for generating fundamental frequency contours for training HMM-based speech synthesis. *Proceedings of the INTERSPEECH*, 4.

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

Hirose, K., K. Ochi, R. Mihara, H. Hashimoto, D. Saito, and N. Minematsu. 2011. Adaptation of prosody in speech synthesis by changing command values of the generation process model of fundamental frequency. *Proceedings of the INTERSPEECH*, 2793–2796.

Hirose, K., H. Hashimoto, J. Ikeshima, and N. Minematsu. 2012. Fundamental frequency contour reshaping in HMM-based speech synthesis and realization of prosodic focus using generation process model. *Proceedings of the International Conference on Speech Prosody*, 171–174.

Kain, A., and M. W. Macon. 2002. Spectral voice conversion for text-to-speech synthesis. *Proceedings of the IEEE ICASSP*, 285–288.

Kameoka, H., K. Yoshizato, T. Ishihara, Y. Ohishi, K. Kashino, and S. Sagayama. 2013. Generative modeling of speech F_0 contours. *Proceedings of the INTERSPEECH*, 1826–1830.

Kawahara, H., M. Morise, T. Takahashi, R. Nishimura, T. Irino, and H. Banno. 2008. Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F_0 and aperiodicity estimation. *Proceedings of the IEEE ICASSP*, 3933–3936.

Kurematsu, A., K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano. 1990. ATR Japanese speech database as a tool of speech recognition and synthesis. *Speech Communication* 9 (4): 357–363.

Matsuda, T., K. Hirose, and N. Minematsu. 2012. Applying generation process model constraint to fundamental frequency contours generated by hidden-Markov-model-based speech synthesis. *Acoustical Science and Technology, Acoustics Society of Japan* 33 (4): 221–228.

Mixdorff, H., Y. Hu, and G. Chen. 2003. Towards the automatic extraction of Fujisaki model parameters for Mandarin. *Proceedings of the INTERSPEECH*, 873–876.

Narusawa, S., N. Minematsu, K. Hirose, and H. Fujiaski. 2002. A method for automatic extraction of model parameters from fundamental frequency contours of speech. *Proceedings of the IEEE ICASSP*, 509–512.

Ochi, K., K. Hirose, and N. Minematsu. 2009. Control of prosodic focus in corpus-based generation of fundamental frequency contours of Japanese based on the generation process model. *Proceedings of the IEEE ICASSP*, 4485–4488.

Ochi, K., K. Hirose, and N. Minematsu. 2010. Realization of prosodic focuses in corpus-based generation of fundamental frequency contours of Japanese based on the generation process model. *Proceedings of the International Conference on Speech Prosody*, 4.

Tokuda, K., T. Masuko, N. Miyazaki, and T. Kobayashi. 1999. Hidden Markov models based on multispace probability distribution for pitch pattern modeling. *Proceedings of the IEEE ICASSP*, 229–232.

Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. *Proceedings of the IEEE ICASSP*, 1315–1318.

Tokuda, K., T. Mausko, N. Miyazaki, and T. Kobayashi. 2002 Multispace probability distribution HMM. *IEICE Transactions on Information and Systems* E85-D (3): 455–464.

Wang, M., M. Wen, K. Hirose, and N. Minematsu. 2010. Improved generation of fundamental frequency in HMM-based speech synthesis using generation process model. *Proceedings of the INTERSPEECH*, 2166–2169.

Yu, K., and Steve Young. 2011. Continuous F_0 modeling for HMM based statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing* 19 (5): 1071–1079.