



Adaptation of Prosody in Speech Synthesis by Changing Command Values of the Generation Process Model of Fundamental Frequency

Keikichi Hirose, Keiko Ochi, Ryusuke Mihara, Hiroya Hashimoto, Daisuke Saito, and Nobuaki Minematsu

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

Abstract

A method was developed to adapt prosody to a new speaker/style in speech synthesis. It is based on predicting differences between target and original speakers/styles and applying them to the original one. Differences in fundamental frequency (F_0) contours are represented in the framework of the generation process model; differences in the command magnitudes/amplitudes. While the original one requires a certain amount of training corpus, while corpus for training command differences can be small. Furthermore, in the case of style adaptation, it is not necessarily the corpus being uttered by the same speaker of the original style. Speech synthesis was conducted using HMM-based speech synthesis system, where prosody was controlled by the method. Listening experiments on synthetic speech with style adaptation and voice conversion both showed the validity of the method.

Index Terms: prosody adaptation, generation process model, speech synthesis

1. Introduction

Recently, HMM-based speech synthesis attains a special concern among speech synthesis researchers, since a flexible control in voice quality and/or speech style is possible by adapting phone HMMs to the target speakers/styles. 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 includes an inherent 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 causes over-smoothed F_0 contours, and occasionally 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]. 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. A corpus-based method of synthesizing F_0 contours in the framework of F_0 model was developed, and speech synthesis in reading and dialogue styles with various emotions was realized already [3]. By predicting the model commands (timings and magnitudes/amplitudes) instead of frame-by-frame F_0 values, a good constraint is automatically applied on

the generated F_0 contours; still keeping acceptable speech quality even if the prediction is done somewhat incorrectly. In order to generate F_0 contours from given texts, the method first predicts pauses and phone durations in other corpus-based ways, and then uses obtained information on phone/syllable boundaries for the F_0 model command prediction [4].

The F_0 model 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 and “flexible” control can be realized for F_0 contour generation than the frame-by-frame control. It is rather easy to analyze the prosodic controls obtained by statistical methods and to modify generated F_0 contours manually or in a corpus-based way, which is trained using a small speech corpus. Although several sophisticated methods have already been developed for converting voice-quality/speaking-style to a new one in HMM-based speech synthesis, they are mostly on segmental features of speech. For prosody, only a simple linear conversion is conducted; conversion based on the average F_0 values and deviations of original and target speakers/styles. This conversion has a serious problem in that it does not take prosodic structures into account.

As an example of such flexible controls, we have developed a method of focus control in speech synthesis [5, 6]. 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 developed method realizes prosodic focus as a supplemental process to our corpus-based method of F_0 contour generation; to train binary decision trees 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 following to the differences. By concentrating on 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 with those for the baseline.

In the current paper, the method is checked if it is applicable also to style adaptation and voice conversion, where F_0 contours may largely change over the whole utterance. (In this paper, we use “adaptation” when linguistic information is used during the adaptation process, and “conversion” when not used.)

The rest of the paper is organized as follows: in section 2, following to explanation on the F_0 model, explanation on the method of prosody adaptation is given, sections 3 and 4 respectively show style adaptation and voice conversion with experimental results, and section 5 concludes the paper with a brief discussion.

2. F_0 model and adaptation/conversion of prosody

The F_0 model is a command-response model that describes F_0 contours in logarithmic scale as the superposition of phrase and accent components. The phrase component is generated by a second-order, critically-damped linear filter in response to an impulse-like phrase command, while the accent component is generated by another second-order, critically-damped linear filter in response to a stepwise accent command. An F_0 contour is given by the following equation:

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

In the equation, $G_{pi}(t)$ and $G_{aj}(t)$ represent phrase and accent components in response to their corresponding commands, respectively. F_b is the bias level, i is the number of phrase commands, j is the number of accent commands, A_{pi} is the magnitude of the 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. Although $G_{pi}(t)$ and $G_{aj}(t)$ include natural angular frequencies for phrase and accent control mechanisms, they are set constant throughout the experiment. This is because they are known to show minor changes depending on speakers/styles.

The proposed method for prosody adaptation/conversion consists of the following two processes [6]:

1. Extract F_0 model commands for original and target utterances of the same sentence, and calculate differences in magnitudes/amplitudes of corresponding commands. Train binary decision trees (BDTs) to predict these differences. The CART (Classification And Regression Tree) included in the Edinburgh Speech Tools Library is utilized. No adaptation is conducted for the timings of commands, since they are known to be less affected by the speaker/style differences [2]. Although there are cases where one-to-one correspondences between commands of original and target utterances are not available, such samples are excluded from the training for the current experiments. Such cases are left for the future work.
2. Apply differences to phrase command magnitudes of the original utterances, and then apply differences to accent command amplitudes taking modified phrase commands into account.

3. Style adaptation

Although the method of prosody adaptation was shown valid in emphasizing specific words in synthetic speech, changes in command magnitudes/amplitudes are mostly around the words to be emphasized. In order to check if the proposed method is applicable when the changes cover for entire utterance, an experiment is conducted to adapt the neutral reading style (original style) to “polite” and “impolite” ones (target styles). Realizing emotional speech from calm speech is also our concern, but it was left for the future work. This is because a rather large change in prosodic structures is often observable between calm and emotional speech, and it is not included in the method currently.

The input parameters for predicting command magnitude/amplitude differences are shown in Table 1, which are similar to the case of focus control.

Table 1. Input parameters for predicting differences in phrase/accnt command magnitudes/amplitudes. “*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. Parameters listed in the first 8 lines are used both for phrase and accent commands. Those in the 9th to 11th lines and those in 12th to 16th lines are only for phrase commands and accent commands, respectively. All parameters are for the original style other than specified.

Position of current <i>bunsetsu</i> in sentence
Number of <i>mora</i> e of current/preceding <i>bunsetsu</i>
Accent type (accent nucleus position) of current/preceding <i>bunsetsu</i>
Number of <i>mora</i> e between preceding and adjacent phrase commands
BDC at the boundary immediately before current <i>bunsetsu</i>
With or without pause at the boundary immediately before current <i>bunsetsu</i>
Pause length if applicable
With or without phrase command for the preceding <i>bunsetsu</i>
Phrase command magnitude of current <i>bunsetsu</i>
Phrase command magnitude of preceding <i>bunsetsu</i>
Position of current <i>bunsetsu</i> in prosodic clause
Phrase command magnitude of current <i>bunsetsu</i> (Target utterance for training and predicted magnitude for prediction)
Accent command amplitude of current <i>bunsetsu</i>
Number of words of the current/preceding <i>bunsetsu</i>
Part-of-speech of the first/last word of the current/preceding <i>bunsetsu</i>
Conjugation form of the first/last word of the current/preceding <i>bunsetsu</i>

The method also includes adaptation of pause and phone durations in a similar way based on the differences between styles. As for the phone durations, phones are categorized into 7 groups depending on the manner of articulation to cope with the limited size of the training corpus. Adaptation of pause and phone durations come first, followed by that for F_0 contours (F_0 model commands).

Utterances used for the experiment are those by female narrator FTY, who uttered 503 sentences of ATR continuous speech corpus in three styles; neutral reading (original), polite and impolite styles. First, baseline speech synthesis is conducted using the whole utterances of original style following to the method developed formerly [3]. The method combines segmental features generated by HMM-base speech synthesis and prosodic features generated from the F_0 model. The BDTs predicting differences are trained using 40 sentences, where one-to-one correspondence of commands is obtained between two styles; original and polite styles, or original and impolite styles. Style adaptation is conducted for ten sentences not included in training sentences. Bias level F_b for speaker FTY is set constant to 140 Hz throughout the experiment. No adaptation is conducted for the segmental features. Six native Japanese speakers are asked to score synthetic speech of two versions; one with original style and another with target (polite/impolite) style. First, 10 natural (sentence) utterances for each of original and target styles are offered to informants so that they get an idea on “polite” and

“impolite” styles. These 10 sentences are different from 10 sentences of style adaptation. Then, they are asked to select one of two versions, which they feel closer to the target style, and assign their confidence in two levels. So the resulting scores are as follows; 5-speech with style adaptation, 4-likely speech with style adaptation, 3-no difference, 2-likely original speech, 1: original speech). Figure 1 show the scores, indicating the adaptation is effective for “polite” but not for “impolite.” Observation of the results suggests that changes in prosodic structure need to be taken into account to realize impolite style.

Modification of F_0 contours may cause degradation in synthetic speech quality. In order to check this point, the same 6 speakers are 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 during original-to-polite style adaptation (Fig. 2). The degradation is larger for impolite style adaptation.

As already mentioned, our method does not require the speaker of the utterances to train differences being the same person for the training corpus of baseline. In order to check this, experiments on polite style conversion are further conducted when differences are trained from 10 paired utterances by male narrator MMI and applied to FTY’s (synthetic) speech. The score for the polite style realization is 4.2, indicating that the method works as expected.

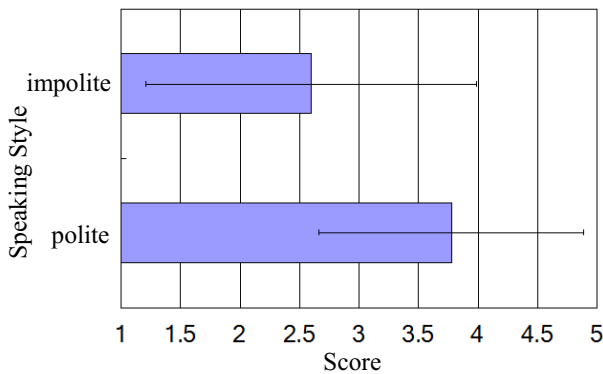


Figure 1: Average scores and standard deviations for style realization.

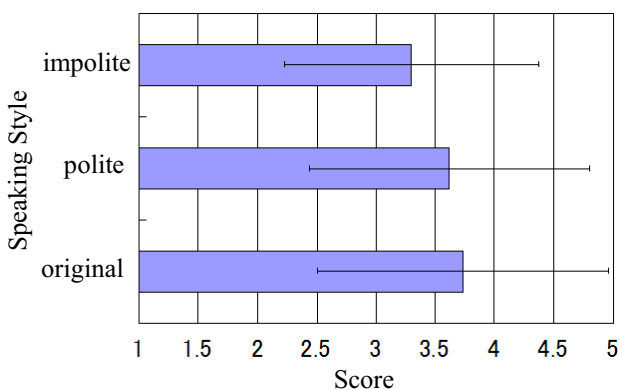


Figure 2: Average scores and standard deviations for naturalness of synthetic speech.

4. Voice conversion

Voice conversion is a technique to convert someone’s voice to another’s keeping the linguistic (and para-/non-linguistic) contents of the utterances without knowing these contents. Among various methods for voice conversion, those based on

Gaussian Mixture Modeling (GMM) are widely used. In this paper, we take the method by Kain et. al. [7] (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 kept by DP matching. In the method, F_0 ’s are linearly converted using the following equation:

$$\hat{y}_t = \frac{\sigma^{(y)}}{\sigma^{(x)}} (x_t - \mu^{(x)}) + \mu^{(y)}, \quad (2)$$

where x_t and \hat{y}_t are log F_0 value for the original speaker and expected log F_0 value for the target speaker. $\mu^{(x)}$ and $\sigma^{(x)}$ are mean and standard deviation of training data of original speaker, and $\mu^{(y)}$ and $\sigma^{(y)}$ are mean and standard deviation of training data of target speaker.

We replace the method of linear conversion to our method based on the differences in the F_0 model commands (henceforth, proposed method). Pause and phone durations are kept unchanged. Tables 2 and 3 show input parameters of BDTs to predict differences of phrase and accent commands, respectively. Although better prediction is possible taking linguistic information of the utterances, such as part of speech, syntactic structure, and so on, into account, it is not included here to check how the proposed method works only with parameters obtainable from acoustic features of utterances.

Table 2. Input parameters for predicting phrase command magnitude differences. All parameters are for the original speaker’s utterances.

Magnitude of the command in question
Preceding pause length
Time from the preceding phrase command
Number of accent commands included in the phrase
Count of phrase commands from the sentence initial
Count of phrase commands from the preceding respiratory pause

Table 3. Input parameters for predicting accent command amplitude differences. All parameters are for the original speaker’s utterances other than the third one.

Amplitude of the command in question
Magnitude of the phrase command where the accent component belongs
Magnitude of the phrase command where the accent component belongs (Target utterance for training and predicted magnitude for prediction)
Count of accent commands in the phrase

Speech synthesis experiments are 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 503 sentences, 200 sentences and 53 sentences are selected, and used for training and testing (evaluation), respectively. As for GMM for conversion, one with feature vectors of 1st to 24th Mel-cepstrum coefficients in 128 mixtures is trained. F_0 model parameters are automatically extracted by the method previously developed by the authors [8], and corrected manually. Bias level F_b is set constant for each speaker; 60 Hz for MHT and 120 Hz for FKS.

Figure 3 shows the result of F_0 conversion for sentence “chiisana unagiyan nekkino yoonamonoga minagiru (A small

eel shop is filled with a kind of hot air.)” Although the results by the original and proposed methods are similar, an improvement is observable in F_0 contour around the sentence initial by the proposed method.

As an objective measure to evaluate the F_0 contour generated by the baseline and proposed methods, the root mean square error between the generated contour and the target contour is defined as:

$$F_0RMSE = \left(\frac{\sum_t (\Delta \ln F_0(t))^2}{T} \right)^{\frac{1}{2}}, \quad (3)$$

where $\Delta \ln F_0(t)$ is the F_0 distance in logarithmic scale at frame t between the two F_0 contours. The summation is done only for voiced frames and T denotes their total number in the sentence. The averaged value for the 53 test sentences is 20.73 by the proposed method, while it is 23.05 by the baseline method.

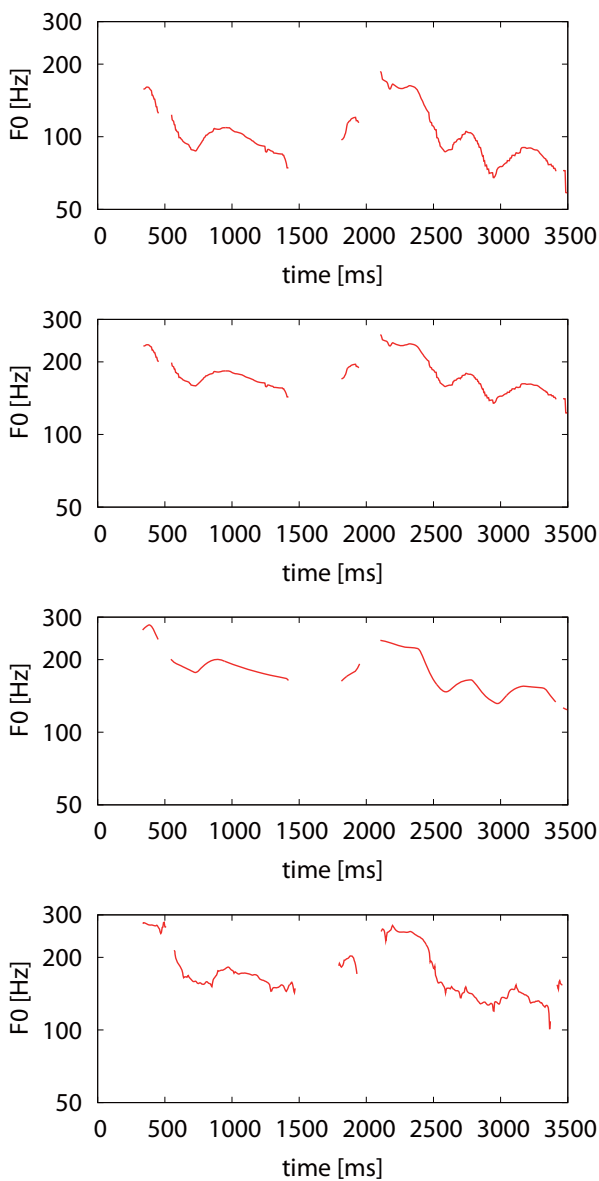


Figure 3: Comparison of F_0 contours. From top to bottom, F_0 contour of original speech, that generated by the baseline method, that generated by the proposed method, and that of target speech.

Since minor undulations in F_0 contour, such as those due to micro-prosody, are not represented by the F_0 model, F_0RMSE by the proposed method may appear larger than the actual speech quality. A listening experiment is conducted from this viewpoint. Ten native speakers of Japanese are asked to select one (A or B) which is closer to X in AB-X test. A and B are synthetic speech by the baseline and proposed methods, while X is the target speech. In order to avoid order effect, both cases of “A: original and B: proposed” and vice versa are included in the stimuli. Score “1” or “-1” is assigned when speech by the proposed method is judged closer or farther to the target speech. When an informant cannot judge, score 0 is allowed. The average score over the 53 test sentences is 0.419 with ± 0.09 confidence interval in significance level of 5%. The result clearly shows the advantage of the proposed method.

5. Conclusions

The method of focusing on differences of F_0 model commands for prosody adaptation is successfully applied for style adaptation and voice conversion. Two issues need to be studied for the future work. First one is on how to handle the cases where a large difference in prosody structure is observable between original and target utterances, which is the case when emotional speech is targeted. Another issue is on how to handle minor undulations, which is not modeled in the F_0 model. If we assume such undulations are speaker/style independent, they can be kept as residuals of F_0 model approximation, and be added to the F_0 contours obtained by the prosody adaptation. We have developed a method to reshape the F_0 contours generated by the HMM-based speech synthesis in the framework of the F_0 model [9]. The proposed method can be easily combined with this method giving flexible prosody control to the HMM-based speech synthesis.

6. 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 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).
- [3] 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).
- [4] K. Hirose, K. Ochi, and N. Minematsu, "Corpus-based generation of prosodic features from text based on generation process model," *Proc. Interspeech*, pp.1274-1277 (2007).
- [5] K. Ochi, K. Hirose, and N. Minematsu, "Control of prosodic focus in corpus-based generation of fundamental frequency based on the generation process model," *Proc. Interspeech*, p.1216 (2008).
- [6] 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).
- [7] A. Kain and M.W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. IEEE ICASSP*, pp.285-288 (2002).
- [8] 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).
- [9] T. Matsuda, K. Hirose, and N. Minematsu, "HMM-based synthesis of fundamental frequency contours using the generation process model," *Journal of Signal Processing*, vol.14, no.4 pp.277-280 (2010).