Realization of Prosodic Focuses in Corpus-based Generation of Fundamental Frequency Contours of Japanese Based on the Generation Process Model

Keiko Ochi, Keikichi Hirose, and Nobuaki Minematsu

Department of Information and Communication Engineering,
Graduate School of Information Science and Technology, University of Tokyo, Tokyo
{ochi, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract

A method was developed for generating sentence F_0 contours of Japanese, when a focus is placed in one of the "*bunsetsu*" of an utterance. It controls F_0 based on the F_0 model; not frame-by-frame F_0 prediction as in the case of HMM-based speech synthesis. The method first predicts differences in the F_0 model commands between utterances with and without focus, and then applies them to the F_0 model commands predicted beforehand by the baseline method without focus assignment. The baseline method is trained using a large corpus, while corpus for training command differences can be small and not necessarily be uttered by the same speaker of the large corpus. The validity of the method was proved by the experiment on F_0 contour generation and speech synthesis, including interpolation/extrapolation of the F_0 model commands for focus level control.

Index Terms: Generation process model, F_0 contour, Corpus-based method, Speech synthesis, Prosodic focus

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.

From this consideration, we have developed a corpus-based method of synthesizing F_0 contours in the framework of the generation process model (F_0 model) and realized speech synthesis in reading and dialogue styles with various emotions [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 [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 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].

By handling F_0 contours in the F_0 model framework, a clear relationship is obtainable between generated F_0 contours and their background linguistic (and para-/non-linguistic) information, enabling "flexible" control of prosodic features. 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 [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 proposed 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 our 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 for those of the baseline.

Since, in the proposed method, F_0 contours are controlled by command values, morphing of focus levels can easily be realized by interpolating/extrapolating command magnitudes/amplitudes. This is also an advantage of controlling F_0 contours in the F_0 model framework.

The following sections are organized as follows: In section 2, the baseline method for F_0 contour generation is briefly explained together with our total corpus-based scheme of generating prosodic features from text input. The developed method of realizing prosodic focus is presented in Section 3 together with perceptual experimental results on synthetic speech. Morphing of focus levels is also addressed in Section 3. Section 5 concludes the paper after a brief discussion in Section 4.

2. Baseline method for generating prosodic features

Each sentence of the input text is first parsed into a morpheme sequence using the open source software CHASEN [7]. Parsing using another freeware JUMAN+KNP is also conducted to obtain syntactic structures [8]. The syntactic structure is given as a boundary depth code (BDC) of each *bunsetsu* boundaries, which indicates the distance from the *bunsetsu* immediately before the boundary to the *bunsetsu* directly modified. Here, *bunsetsu* is defined as a basic unit of Japanese syntax and pronunciation consisting of content word(s) followed or not followed by particles. Then the linguistic information thus obtained is used to predict position of pauses and their lengths. Similar processes of predicting phone durations and F_0 model parameters follow.

Since all the timing structures need to be decided before the F_0 contour generation, the prediction of F_0 model parameters is conducted as the last process of prosodic feature generation. Binary decision trees (BDT's) are adopted for the prediction. The CART (Classification And Regression Tree) included in the Edinburgh Speech Tools Library [9] was utilized to construct BDT's. Training corpus (with necessary annotations) is prepared automatically using the above parsers, an HMM-based segmentation scheme, and an F_0 model command extractor [10].

It is known that the information of preceding units has a larger influence on the prosodic features of the current unit than that of following units [2]. Taking these into consideration, information of the directly preceding *bunsetsu* is included in the input parameters for the phrase command predictor as well as that for the current *bunsetsu* in question. Since pauses have a tight relation with phrase commands, information of predicted pauses was included also, while it was not used for the prediction of accent command parameters.

Similar to the case of phrase commands, the parameters on accent commands (position and amplitude) are tightly related to the information of the current and preceding units (prosodic words), such as position in sentence, length, grammatical information of the first and last words of the units, and syntactic boundary between the units. They also change according to the accent types of the units. Taking these into consideration, the input parameters for accent command predictor were selected. All the trainings were conducted using 453 utterances out of 503 utterances of ATR continuous speech corpus by a female narrator.

To investigate the validity of the method of F_0 contour generation when applied in a TTS system, a full speech synthesis system was constructed using the HMM-based speech synthesis as shown in Figure 1. Tri-phone models were trained using the 453 sentence utterances used for the training of the prosodic feature predictors. The segmental features were 75th order vectors consisting of 0th to 24th Mel-cepstrum coefficients and their Δ and Δ^2 values.

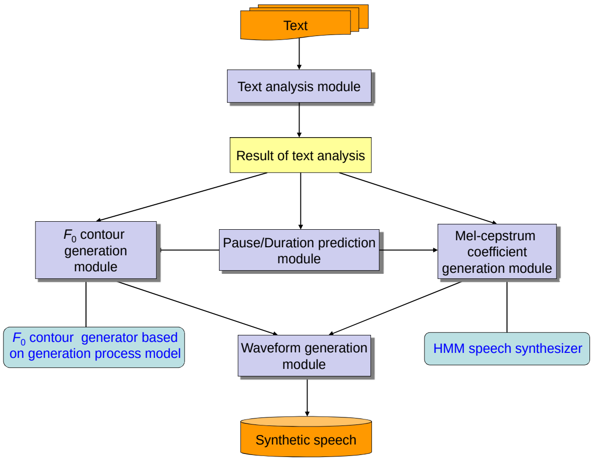
Text
Text analysis module
Result of text analysis
F_0 contour generation module
Pause/Duration prediction module
Mel-cepstrum coefficient generation module
F_0 contour generator based on generation process model
Waveform generation module
HMM speech synthesizer
Synthetic speech

Figure 1: Total configuration of developed speech synthesizer.

3. Focus control

Although emphasis of word(s) is not handled explicitly in most of 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. 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 on one of assigned words (*bunsetsu*'s). For each sentence, 2 to 4 *bunsetsu*'s were assigned depending on the sentence length. Figure 2 shows F_0 contours together with results of 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 commands for different focal conditions. (Although "jibuNnohooe" 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 inspires us to realize focuses only by controlling command magnitudes/amplitudes.

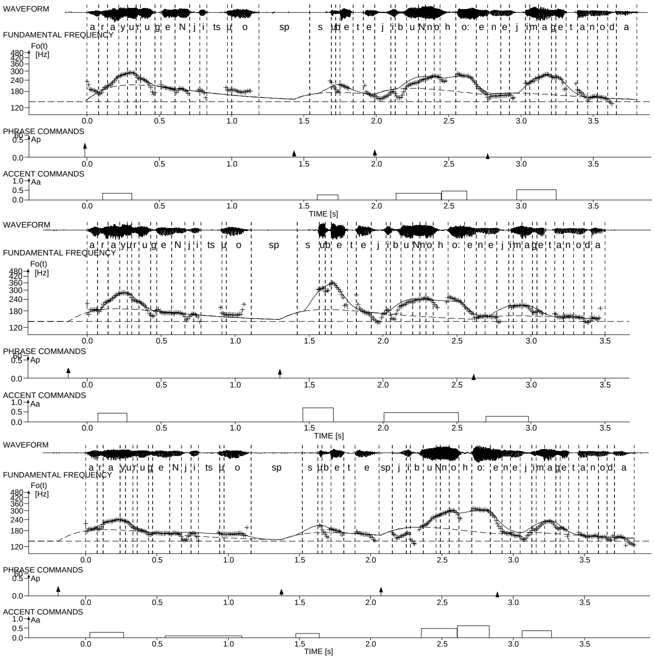

Figure 2: F_0 contours and F_0 model parameters of Japanese sentence "arayuru geNjitsuo subete jibuNnohooe nejimagetanoda ((He) twisted all the reality to his side.)" uttered by a female speaker. From the top to the bottom panels: without specific focus, focus on "subete," and focus on "jibuNnohooe," respectively.

The proposed method for focus control is to modify command magnitudes/amplitudes predicted by the BDT's trained for utterances without specific focuses (baseline method) depending on the differences in command magnitudes/amplitudes between utterances without and with focus. The differences are trained also using BDT's. 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 method, 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.) Tables 1 and 2 show input parameters for the binary decision trees for predicting command magnitude/amplitude differences. Category numbers are reduced from the case of training command magnitudes/amplitudes (of the baseline method), so that training can be done only with a limited speech corpus. For instance, accent type categories are reduced to four by merging several accent types into one group; accent type without accent nuclei, that with accent nuclei at the first *mora*, that with accent nuclei at the second *mora*, and others. This is because accent components show quite different features depending on to which group they belong.

Table 1. 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*."

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

The above utterances for investigation on focus control are used to train the trees for the current experiment. They include 50 utterances without focus and 172 utterances with focus on one of noun phrases (*bunsetsu* including a noun). 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 3 shows examples of generated F_0 contours when the predicted changes are applied to F_0 model parameters predicted by the baseline method. Although the baseline method 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 as can be seen from Figure 2. 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. Speech synthesis was conducted using the system shown in Figure 1. 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 a *bunsetsu* (for each utterance) where they perceived an emphasis. "No emphasis (in the utterance)" 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 %.

Table 2. Input parameters for the prediction of differences in accent command amplitudes. As for accent type inputs, category number of preceding/following prosodic word is larger by one than that of current prosodic word to indicate "no preceding/following prosodic word."

Input parameter	Category
Position of current prosodic word in prosodic phrase	3
Position of current prosodic word in sentence	3
Distance of current prosodic word from *bunsetsu* with focus (in number of *bunsetsu*'s)	5
Accent type (location of accent nucleus) of current prosodic word	4
Accent type (location of accent nucleus) of preceding prosodic word	5
Accent type (location of accent nucleus) of following prosodic word	5
BDC at the *bunsetsu* boundary immediately before current prosodic word	2
Amplitude of current accent command	Continuous
Amplitude of preceding accent command	Continuous
Magnitude of current phrase command	Continuous
Magnitude of preceding phrase command	Continuous

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.

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. 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 3, 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 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; the phrase initial word/*bunsetsu* is usually perceived as focused.

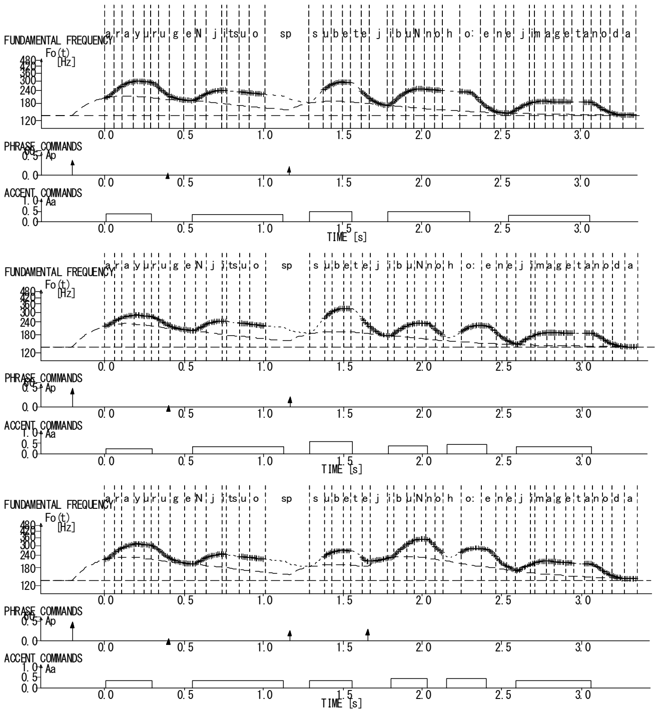

Figure 3: Generated F_0 contours and F_0 model parameters. The sentence and focal conditions are the same with those shown in Figure 2.

Table 3. 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

4. Discussion

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.

5. Conclusion

A method of controlling prosodic focuses is realized as a supplemental process to our corpus-based method of F_0 contour generation. The method is in the framework of the F_0 model and is based on predicting differences in F_0 model command magnitudes/amplitudes with and without focuses. The validity of the method was confirmed through a perceptual experiment on the speech synthesized using the generated F_0 contours. Also focus level control is shown to be possible by interpolating/extrapolating command magnitudes/amplitudes. Controls of duration and amplitude are for future research. We are planning to apply the similar supplemental control of F_0 model commands for realizing various styles including emotional speech.

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] 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).

[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. 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. ICASSP*, pp. 4485-4488 (2009).

[7] Y. Matsumoto, "Morpheme analysis system, Chasen," IPSJ Mag., Vol.41, No.11, pp.12081214 (2000). (in Japanese).

[8] S. Kurohashi, and M. Nagao, 1994. , "A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures," J. Comput. Linguist., Vol.20, No.4, pp.507-534 (1994).

[9] The Edinburgh Speech Tools Library, http://www.cstr.ed.ac.uk/projects/speech_tools/

[10] 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).