Control of Prosodic Focus in Corpus-based Generation of Fundamental Frequency based on the Generation Process Model

Keiko Ochi[1], *Keikichi Hirose* [1], *and Nobuaki Minematsu* [2]

[1] Department of Information and Communication Engineering, the University of Tokyo, Tokyo
[2] Department of Electrical Engineering and Information Systems, the University of Tokyo, Tokyo

{Ochi, hirose, minematsu}@gavo.t.u-tokyo.ac.jp

Abstract

A method was developed for generating sentence F_0 contours, when a focus is placed in one of *bunsetsu* of an utterance. The method is to predict differences in F_0 model commands between with and without focus utterances, and applies them to the F0 model commands predicted beforehand by the baseline method. The validity of the method was proved by the experiment on F_0 contour generation and speech synthesis.

Index Terms: speech synthesis, F_0 contour, Focus

1. Introduction

In spontaneous speech, speakers may frequently place focuses on selected word(s) in their utterances. Therefore, focus control is becoming one of important issues in spontaneous speech synthesis. However, 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. While we have developed a corpus-based method of synthesizing F_0 contours in the framework of the generation process model (F_0 model) [1] and realized speech synthesis in reading and dialogue styles with various emotions [2, 3]. 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.

In this paper, we propose a method of realizing prosodic focus as a supplemental process to our corpus-based method of F_0 contour generation; 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 using the differences. By concentrating to the differences, a better training for F_0 change due to focal position comes possible only with a limited speech corpus. Moreover, speakers for the training need not be the same for those of the baseline.

2. Method and Result

Observations of F_0 contours imply that we can represent F_0 change due to focal positions as changes in F_0 model command magnitudes/amplitudes. The differences in phrase command magnitudes are predicted using CART with inputs of distance from focused *bunsetsu*, position of current *bunsetsu* in phrase/sentence, *mora* numbers and accent types, distance in *mora* from the preceding phrase command, depth of syntactic boundary, and pause between preceding and current *bunsetsu*s. The differences in accent command amplitudes are also predicted in a similar way. 172 utterances by a female speaker were used for the prediction. Figure 1 shows an example of generated F_0 contour when the predicted changes are applied to F_0 model parameters predicted by the baseline method. Result of preliminary perceptual experiment confirmed that the focus was perceived on the right place by the method.

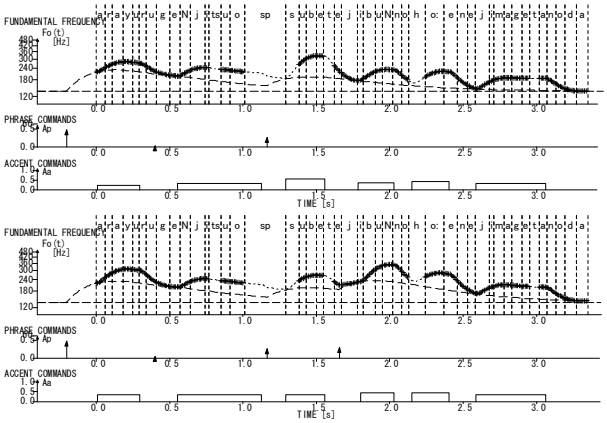

Figure 1: *Generated F_0 contours and F_0 model parameters of Japanese sentence "arayuru genjitsuo subete jibunno hooe nejimagetanoda (They bended all the realities to their side)". The first and second panels show when "subete" and "jibunno" are focused, respectively.*

3. Conclusion

A method was developed to generate F_0 contours with prosodic focuses. The method works only with an additional small speech corpus with various focal positions. Detailed results including variations of applying predicted differences to F_0 model commands will be shown in the presentation.

4. References

[1] Hirose, K., Sato, K., Asano, Y., and Minematsu, N., "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 (2005-7).

[2] Hirose, K., Ochi, K., and Minematsu, N., "Corpus-based generation of prosodic features from text based on generation process model," *Proc. Interspeech*, pp.1274-1277 (2007-8).

[3] Fujisaki, H. and Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Japan (E)*, Vol.5, No.4, pp.233-242 (1984-10).