PROSODIC PARAMETERIZATION OF SPOKEN JAPANESE BASED ON A MODEL OF THE GENERATION PROCESS OF F_0 CONTOURS

Hiroya Fujisaki and Sumio Ohno

Department of Applied Electronics, Science University of Tokyo
2641 Yamazaki, Noda, 278 Japan

ABSTRACT

The process of generating an F_0 contour from a small number of linguistically meaningful parameters, has been modeled quite accurately, and the model has been used extensively in speech synthesis. The present study deals with the inverse problem, i.e., that of extracting the model parameters from a given contour, which can only be solved by successive approximation. This paper presents a method for deriving a first-order approximation to a given F_0 contour from the linguistic information of the utterance, and refining the approximation by Analysis-by-Synthesis. The validity of the method has been confirmed experimentally.

1. INTRODUCTION

Prosody of the spoken language plays important roles in conveying various types of information, and is therefore quite important both for the intelligibility and naturalness of speech. In many languages including Japanese and Chinese, the most important manifestation of prosody is known to be the contour of the voice fundamental frequency (henceforth the F_0 contour).

The relationship between the F_0 contour and the underlying information can be quantitatively analyzed if we have a model that describes the process of generation of F_0 contours in mathematical terms. Such a model has been presented and extensively used by Fujisaki and his coworkers in the analysis and synthesis of F_0 contours of Japanese, as well as of several other languages [1]. The model can generate very close approximations to the actual F_0 contours from two kinds of commands. Since these commands directly correspond to phrasing and accentuation, the model provides a powerful tool not only for synthesis of natural speech, but also for finding the location of pitch accents and phrase boundaries in automatic speech recognition.

The extraction of these underlying commands from an observed F_0 contour is an inverse problem for the generative model. Since it cannot be solved analytically, the solution has to be obtained by Analysis-by-Synthesis, i.e., by successive approximation. For the purpose of computational efficiency, it is quite important to find an appropriate first-order approximation. Finding such an approximation requires both the knowledge of the linguistic content of the utterance and the knowledge on the general characteristics of an F_0 contour. In the present paper, we propose a method for automatic extraction of the F_0 contour parameters using the linguistic information of the utterance to derive a good first-order approximation, which is then refined by Analysis-by-Synthesis.

2. A FUNCTIONAL MODEL FOR THE PROCESS OF GENERATION OF THE F_0 CONTOUR OF AN UTTERANCE OF JAPANESE

The F_0 contour of an utterance of Japanese can be regarded as the response of the mechanism of vocal cord vibration to a set of commands which carry information concerning lexical word accent, syntactic and discourse structures of the utterance. Two different kinds of command have been found to be necessary to account for the formation of an F_0 contour of an utterance of the common Japanese; one is an impulse-like command for the onset of relatively slow rise-fall pattern of the fundamental frequency over a time span roughly corresponding to a syntactic phrase, while the other is a stepwise command for the onset and end of a relatively rapid rise-fall pattern of the fundamental frequency over a time span corresponding to the accented mora or morae of a word or a string of words. Consequences of these two types of command have been shown to appear as the phrase component and the accent component, each being approximated by the response of a second-order linear system to the respective commands. If we represent an F_0 contour as a pattern of the logarithm of the fundamental frequency along the time axis, it can be approximated by the sum of these components. The entire process of generating an F_0 contour of a sentence can thus be modeled by the block diagram of Fig. 1.

In this model, the F_0 contour can be expressed by

$$\begin{aligned}\ln F_0(t) = \ln Fb + \sum_{i=1}^{I} Ap_i Gp(t - T_{0\,i}) \\ + \sum_{j=1}^{J} Aa_j \{Ga(t - T_{1j}) - Ga(t - T_{2j})\},\end{aligned} \tag{1}$$

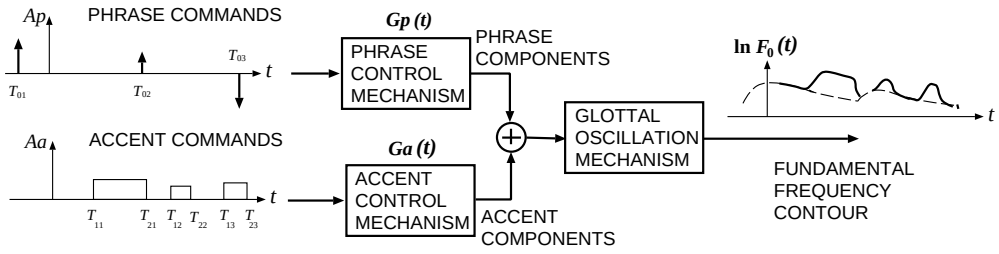

Figure 1: Block diagram of a functional model for the process of generating the F_0 contour of an utterance.

where

$$Gp(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & t \geq 0, \\ 0, & t < 0, \end{cases} \quad (2)$$

and

$$Ga(t) = \begin{cases} \min[1 - (1+\beta t)\exp(-\beta t), \gamma], & t \geq 0, \\ 0, & t < 0, \end{cases} \quad (3)$$

respectively indicate the impulse response function of the phrase control mechanism and the step response function of the accent control mechanism. The symbols in Eqs. (1), (2), and (3) indicate

Fb: asymptotic value of fundamental frequency in the absence of accent components, I: number of phrase commands, J: number of accent commands, Ap_i: magnitude of the ith phrase command, Aa_j: amplitude of the jth accent command, T_{0i}: timing of the ith phrase command, T_{1j}: onset of the jth accent command, T_{2j}: end of the jth accent command, α: natural angular frequency of the phrase control mechanism to the phrase commands, β: natural angular frequency of the accent control mechanism to the accent commands, γ: a parameter to indicate the ceiling level of the accent component (generally set equal to 0.9).

By the technique of Analysis-by-Synthesis, it is possible to decompose a given F_0 contour into its constituents, i.e., the phrase components and the accent components, and estimate the magnitude and timing of their underlying commands by deconvolution. Analysis of F_0 contours of a large number of Japanese utterances has indicated that the model is always capable of generating, from a small number of linguistically meaningful parameters, very close approximations to observed F_0 contours [2]. It has also been shown that the model can be modified to apply to F_0 contours of a number of other languages including Chinese [3], English [4], German [5], Spanish [6], and Swedish [7]. In fact, a specific modification of the model to apply for F_0 contours of a particular language reveals the language-specific characteristics and constraints of the prosody of that language, while the unchanged properties of the model correspond to characteristics common to speakers of various languages. It has further been shown that the model actually represents the mathematical formulation of the dynamic characteristics of the physiological and physical mechanisms of the larynx [8].

3. AUTOMATIC EXTRACTION OF F_0 CONTOUR PARAMETERS

3.1. The problem and previous approaches

Because of the physiological and physical foundations of the model just mentioned, and of its capability of generating extremely close approximations to observed F_0 contours, one can say that the parameters of the model giving the closest approximation to a given F_0 contour is actually the parameters of the observed F_0 contour itself. Hence the model can be used for the automatic extraction of F_0 contour parameters.

It does not mean, however, that the extraction is simple and straightforward, nor that the uniqueness of results is guaranteed. In fact, the model is a formulation of the process of generating an F_0 contour from a set of parameters. Hence the process of obtaining a set of parameters from a given F_0 contour is an *inverse* problem. Inverse problems involving complex relations such as Eq. (1) can only be solved by successive approximation, i.e., by starting from a first-order approximation and successively refining the approximation until it converges. Due to the multi-dimensionality of the parameter space, an exhaustive search for an optimum solution is impossible, so that only a small portion of the space should actually be examined. On the other hand, an excessive limitation of the search space may always increase the danger of being trapped at a local optimum. The problem is further complicated by the presence of measurement noise as well as factors that are not accounted for in the model's formulation. In order to assure the accuracy and efficiency of automatic extraction, therefore, we need a reasonable first-order approximation that is close to the true optimum, and an efficient algorithm for successive approximation.

There have been a few attempts toward automatic extraction of F_0 contour parameters using the above-mentioned model. Fujisaki and his coworkers have proposed a method using the derivative of the F_0 contour to obtain first-order approximations for the timing and magnitude of the accent commands, subtracting the corresponding accent components from the F_0 contour, and then estimating the timing and magnitude of the phrase commands. These first-order approximations are then refined by recursive optimization [9].

On the other hand, Geoffrois has presented a method based on recursive optimization followed by prosodic event detection by incrementing the analysis interval from the onset toward the end of an utterance [10].

These two methods rely solely on the information contained in the F_0 contour. Both the accuracy and the efficiency will be much higher, however, if the linguistic content of the utterance is known and can be used to generate the first-order approximation.

3.2. The proposed method

Here we propose a new method for parameter extraction of F_0 contours using information on the linguistic content of the utterance, which may either be given by a text, or by automatic speech recognition. Although the method can in principle be applied to any language in which systematic correspondence exists between the linguistic content and the F_0 contour, we shall describe the method in the case of Japanese.

Figure 2 illustrates the principle of the proposed method. The speech signal is analyzed to produce both segmental boundaries and the F_0 contour, while the corresponding text is used as the input to a text-to-speech system which generates segmental boundaries and F_0 contour commands for the message. After time alignment of the synthetic parameters with the observed segmental boundaries of the speech signal, the phrase and accent commands generated by the text-to-speech system are used as the first-order approximation to those of the observed F_0 contour. Recursive optimization of the parameters of these commands and other parameters such as α, β and Fb are conducted by Analysis-by-Synthesis of the F_0 contour, i.e., by minimizing the mean squared error between the observed F_0 contour and the model-generated contour on the logarithmic scale of F_0.

The text-to-speech system consists of four processing stages: 1) linguistic processing, 2) phonological processing, 3) control parameter generation, and 4) speech waveform generation. Since the system is described in detail elsewhere [11], we will not give the details. The present method utilizes the first three stages to obtain information both on time alignment and on commands for F_0 contour generation.

3.3. Experiment and results

The speech material is a set of utterance of declarative sentences of the common Japanese uttered by two native male speakers. Each speech sample was digitized at 10 kHz with 12-bit precision. The F_0 contour was extracted by a modified autocorrelation analysis, while segmental boundaries were marked automatically using the short-time average intensity, zero-crossing counts and the spectrum.

Figure 3 shows an example of the analysis results for an utterance of /aoiaoinoewa jamanouenoienniaru/ (The picture of a blue hollyhock is in a house on the top of the mountain.). In both panels (a) and (b), the + symbols indicate measured F_0 values, the curve in a solid line indicates a model-generated F_0 contour, and the curve in a broken line indicates the phrase components. The panels also show the phrase and accent commands for the model-generated contour. Panel (a) shows the comparison of the measured F_0 contour and its first-order approximation given by the proposed method, and indicates that the method can produce a first-order approximation which is already quite close to the given contour in its essential characteristics. Panel (b) shows the comparison of the measured F_0 contour with the final result of recursive optimization of model parameters by Analysis-by-Synthesis, and indicates that the approximation is almost perfect. Thus the model's parameters in panel (b) can be considered as those of the actual F_0 contour itself. The validity of the proposed method has been confirmed by analysis of a number of utterances.

4. SUMMARY AND CONCLUSION

In this paper we have described a new method for automatic extraction of the parameters of F_0 contours which utilizes the linguistic information of the utterance to be analyzed. Experimental results of analysis of a number of an F_0 contour have shown that the method can always produce a good first-order approximation which, after recursive optimization, gives quite accurate values for the parameters of a given F_0 contour. Thus the method is especially useful for automatic prosodic parameterization and labeling of a speech corpora for which linguistic transcription is given. Even in cases where the linguistic content is not available, the method

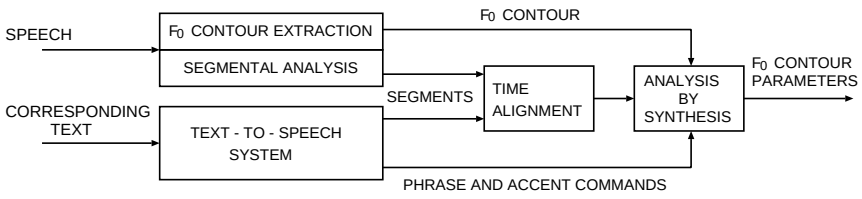

Figure 2: Block diagram of a method for automatic extraction of F_0 contour parameters of speech using the corresponding text.

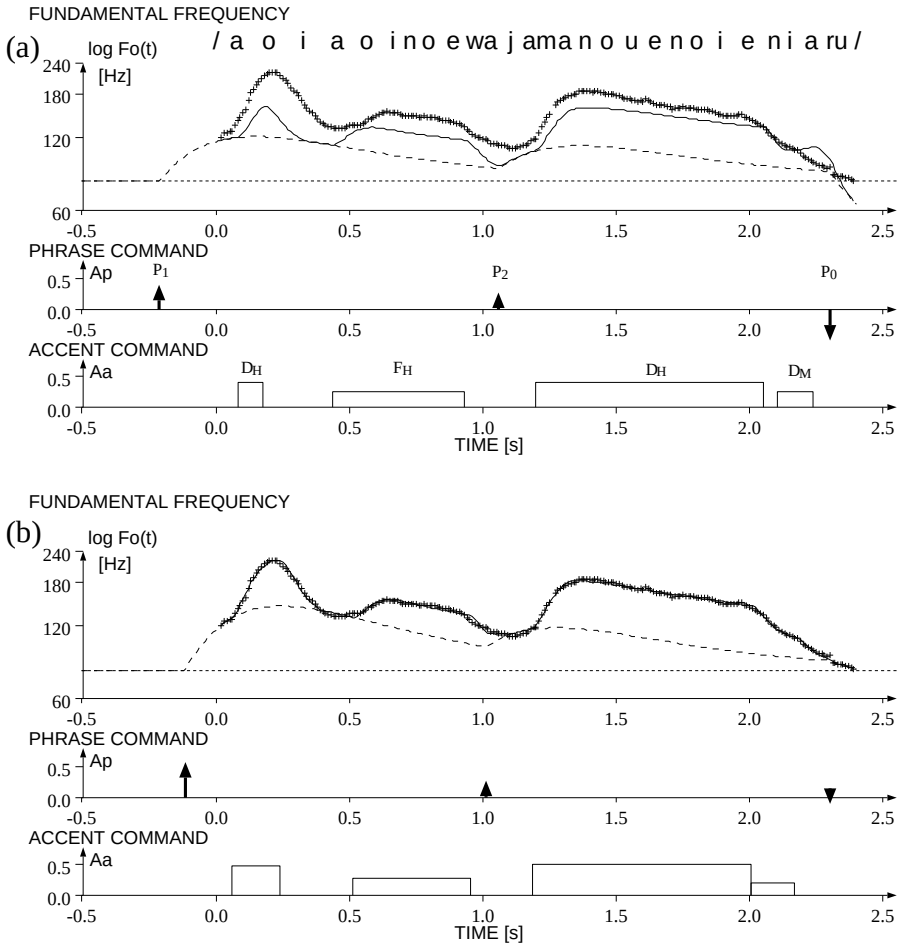

Figure 3: An example of automatic extraction of F_0 contour parameters from an utterance of the Japanese sentence /aoiaoinoewa jamanouenoieniaru/. (a) First-order approximation derived from the linguistic information of the utterance. (b) Final result of Analysis-by-Synthesis, yielding an almost perfect approximation to the given F_0 contour. Details in the text.

will be useful when we combine it with a reliable method for automatic speech recognition. Work is under way to extend the method to these cases, as well as to larger speech material with greater variety.

5. REFERENCES

1. Fujisaki, H., "From information to intonation," *Proc. ISSD-93*, pp. 7–18, 1993.
2. Fujisaki, H. and Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Jpn. (E)*, **5**, pp. 233–242, 1984.
3. Fujisaki, H., Hirose, K. and Lei, H., "Prosody and syntax in spoken sentences of Standard Chinese," *Proc. ICSLP-92*, **1**, pp. 433–436, 1992.
4. Fujisaki, H. and Ohno, S., "Analysis and modeling of fundamental frequency contours of English utterances," *Proc. EUROSPEECH'95*, **2**, pp. 985–988, 1995.
5. Mixdorff, H. and Fujisaki, H., "Analysis of voice fundamental frequency contours of German utterances using a quantitative model," *Proc. ICSLP94*, **4**, pp. 2231–2234, 1994.
6. Fujisaki, H., Ohno, S., Nakamura, K., Guirao, M. and Gurlekian, J., "Analysis of accent and intonation in Spanish based on a quantitative model," *Proc. ICSLP94*, **1**, pp. 355–358, 1994.
7. Fujisaki, H., Ljungqvist, M. and Murata, H., "Analysis and modeling of word accent and sentence intonation in Swedish," *Proc. ICASSP-93*, **2**, pp. 211–214, 1993.
8. Fujisaki, H., "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," In O. Fujimura, ed., *Vocal Physiology: Voice Production, Mechanisms and Functions*, Chapter 30, Raven Press, New York, 1988.
9. Hirose, K., Fujisaki, H. and Yamaguchi, M., "Automatic estimation of characteristic parameters of fundamental frequency contours," *Reports of the 1983 Spring Meeting, Acoust. Soc. Jpn.*, **1**, pp. 93–94, 1983.
10. Geoffrois, E., "A pitch contour analysis guided by prosodic event detection," *Proc. EUROSPEECH'93*, **2**, pp. 793–796, 1993.
11. Hirose, K. and Fujisaki, H., "A system for the synthesis of high-quality speech from texts on general weather conditions," *Trans. IEICE on Fundamentals of Electronics, Communications and Computer Sciences*, **E76-A**, pp. 1971–1980, 1993.