THE USE OF F0 RELIABILITY FUNCTION FOR PROSODIC COMMAND ANALYSIS ON F0 CONTOUR GENERATION MODEL

Mitsuru NAKAI and Hiroshi SHIMODAIRA

Japan Advanced Institute of Science and Technology, Hokuriku
1-1 Asahidai, Tatsunokuchi, Nomi, Ishikawa, 923-1292 Japan
E-mail: mit@jaist.ac.jp

ABSTRACT

This paper describes a method of utilizing an "F_0 Reliability Field" (FRF), which we have proposed in our previous work, for estimating prosodic commands on F_0 contour generation model. This FRF is the time-frequency representation of F_0 likelihood, and an advantage of FRF is that it is not necessary to consider F_0 errors that occur during an automatic F_0 determination. Therefore, it is thought that FRF can be a more useful feature for automatic prosody analyses than F_0 contour, and our previous paper has reported the validity of FRF on the analysis of detecting prosodic boundaries in Japanese continuous speech. Moreover, in this paper, we have examined the validity on the prosodic command estimation of superpositional model. Experimental results show that the accuracy of command estimation with FRF is well and it is close to the accuracy of command estimation with ideal F_0 contour that has no F_0 error.

1. INTRODUCTION

Prosody of speech is an important information for speech understanding. It is well known that high quality speech synthesis can be achieved by incorporating accurate prosodic model, and it is also expected that the prosody will be a useful information for high performed speech recognition. In particular, a fundamental frequency (F_0) is widely used for prosody analyses, such as prosodic phrase segmentation, prosodic structure estimation and the superpositional modeling of prosodic command, and the accuracy of these prosody analyses sometimes depends on the accuracy of F_0 extraction. There is a long history of development of F_0 analysis, and various F_0 determination algorithms and their improved method have been proposed, but it may be said that there is no technique that is superior in every aspect to others. Therefore, we have to choose the most suitable F_0 determination algorithm corresponding to each prosody analysis system.

For example, F_0 determination error has a bad influence on the system that employs the technique of pattern matching between the observed F_0 contour and the approximated F_0 contour that the system constructs. This is because the distortion becomes large as the number of F_0 error increases. Therefore, it is necessary to correct F_0 errors and this is one of laborious postprocessing task in any automatic F_0 determination. So we have proposed the "F_0 Reliability Field" (FRF)[1] as a desirable feature for those systems, namely this feature does not need any correction of F_0 errors. This FRF is expressed as a time-frequency function of F_0 likelihood. The frequency that gives maximum likelihood is not always a real F_0 value, but an advantage of FRF is that the frequency that is equivalent to the real F_0 value always gives high F_0 reliability.

In our previous paper, FRF has been applied to an automatic detection system of accent phrase boundaries, which is based on the F_0 contour matching technique, and the validity of FRF has been confirmed. Besides our FRF, some similar features, which are based on a kind of F_0 reliability function, have been proposed. For example, "periodicity diagram"[2] is the time-frequency representation of F_0, and it has been reported that this representation is useful for determining an accurate F_0 value. In addition, "voicograms"[3] method of speech periodicity representation has been used for ensuring the practical correctness of F_0 estimation. Moreover, in this paper, we have applied our FRF to the prosodic command estimation of the F_0 contour generation model[4].

2. *F0* RELIABILITY FIELD

The F_0 reliability field is a temporal sequence of F_0 reliability function, which represents a likelihood of fundamental frequency at each time frame. This F_0 reliability function is based on a short-time autocorrelation function of speech wave, and the process of FRF analysis is quite similar to the F_0 determining process. The outline is shown in Figure.1.

The extraction algorithm is based on the lag-window method[5] that is one of F_0 determination algorithms. In this method, a pitch structure can be separated from the power spectrum. The desirable smoothed function of F_0 reliability can be obtained by incorporating a narrow spectrum band filter on this pitch structure. Here we use Hanning window as a window function on the frequency domain. This F_0 reliability function is analyzed per each time frame, and FRF is represented as its temporal sequence shown in the bottom of Figure.1, in which time is passed from the front to the back and horizontal axis has been converted into the logarithmic frequency domain from the time domain. It can be seen that harmonic contour of F_0 reliability peaks, which means half pitch contour or double pitch contour, lies in a fixed interval of $\ln 2$. Furthermore, as the number of sampling point on frequency domain is finite in this F_0 analysis, the F_0 reliability of

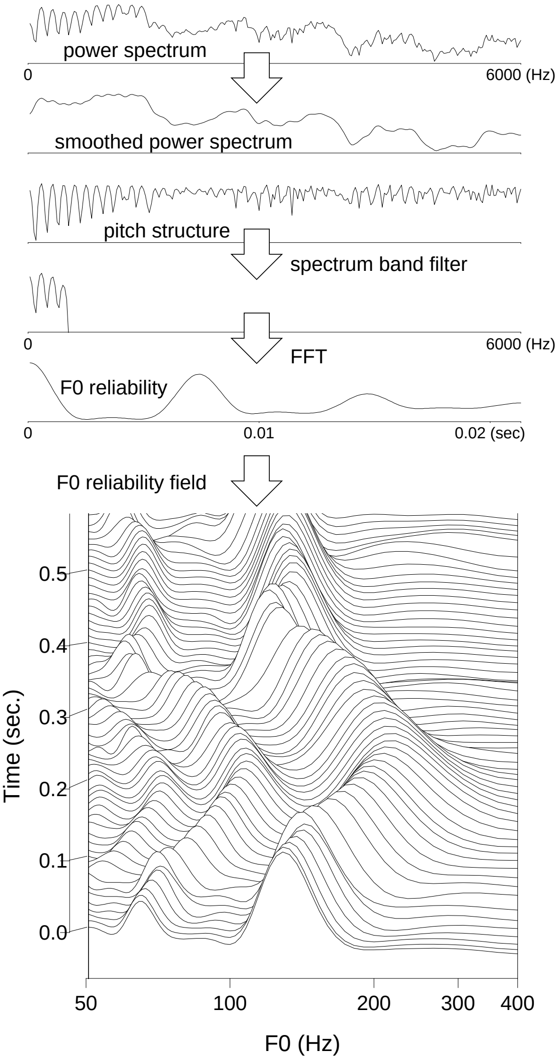


Figure. 1: A process of F_0 reliability analysis.

arbitrary frequency is determined by using Lagrange interpolation, i.e., by interpolating N samples of F_0 reliability near the frequency which we want to obtain. In the following section, we express this FRF as a function $S(t,p)$ with time t and logarithmic frequency p.

3. ESTIMATION OF PROSODIC COMMANDS ON *F0* CONTOUR GENERATION MODEL

F_0 contour generation model which we used in this paper is proposed by Fujisaki[4] and prosodic commands on this model can be estimated by Analysis-by-Synthesis (A-b-S) procedure, i.e., by constructing the best approximation to an observed F_0 feature and by examining the closeness of the approximation. A conventional method employs an F_0 contour as the observed F_0 feature, and the closeness of the approximation is measured by the mean squared error of constructed F_0 contour. On the other hand, using FRF as the observed feature, the optimization carried out by maximizing the mean F_0 reliability.

The Fujisaki's model is given by following equation:

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

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

$$Ga(t) = \begin{cases} \min[1 - (1 + \beta t)e^{-\beta t}, \theta], & (t \geq 0) \\ 0, & \text{(otherwise)} \end{cases} \quad (3)$$

where $Gp(t)$ represents the impulse response function of the phrase control mechanism and $Ga(t)$ represents the step response function of the accent control mechanism. The symbols in these equations indicate

Fb : base value of fundamental frequency,
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 frequency of the phrase control mechanism,
β : natural frequency of the accent control mechanism,
θ : relative ceiling level of accent components.

Here, a set of parameters, which we want to estimate, is defined as

$$\Lambda = (\lambda_1, \lambda_2, \cdots, \lambda_N), \quad (4)$$

and each λ_n is corresponding to some of Ap_i, Aa_j, T_{0i}, T_{1j}, T_{2j}, and sometimes Fb. The parameter α and β are assumed to be constant with in an utterance, while θ is set equal to 0.9. Then, Equation (1) can be replaced with

$$f(\Lambda, t) = \ln F_0(t) \quad (5)$$

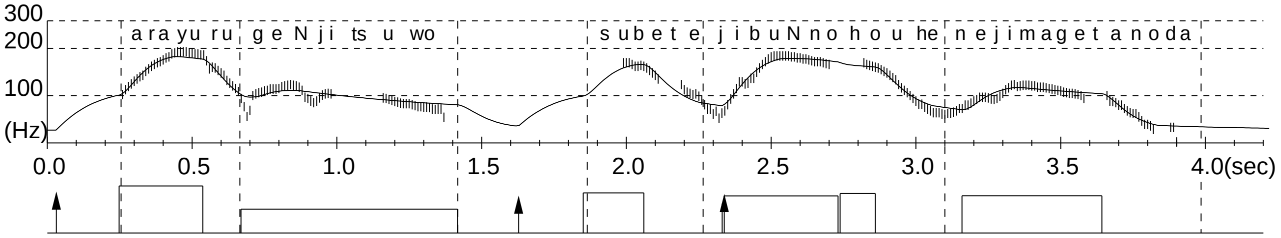
and reliability R_Λ of this F_0 contour becomes

$$R_\Lambda = \sum_t S(t, f(\Lambda, t)) \quad (6)$$

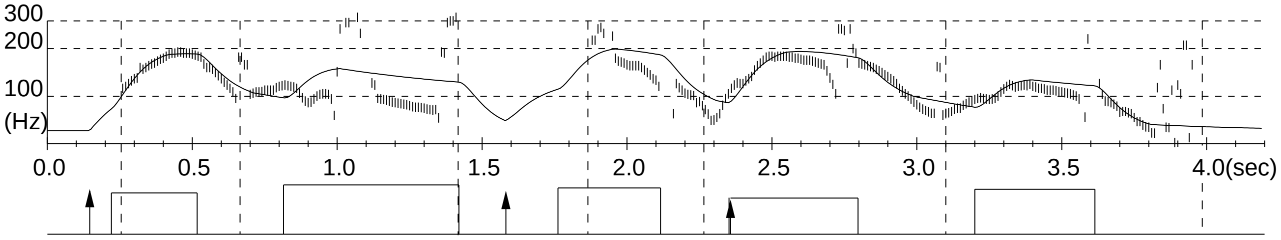
by referring to F_0 reliability field $S(t,p)$. If the reliability R_Λ is not high enough, we have to modify the set of Λ to raise the reliability. The modification value of λ_n is defined as

$$\Delta\lambda_n = g \sum_t \frac{\partial f(\Lambda, t)}{\partial \lambda_n} \Delta S(t, f(\Lambda, t)), \quad (7)$$

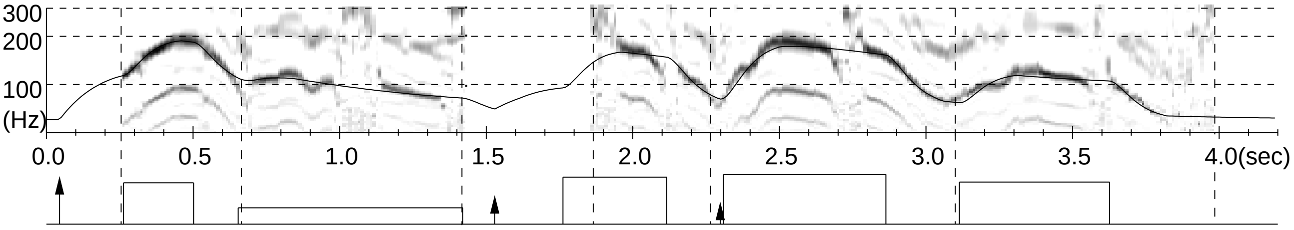
where g is a step gain, and we use a gradient vector of $S(t, f(\Lambda, t))$ as $\Delta S(t, f(\Lambda, t))$. The definition of gradient vector (v_t, v_p) is described in our reference [1] and we use v_p for the



(a) Command estimation result with hand modification F_0 contour.



(b) Command estimation result with automatic extracted F_0 contour ($\kappa = 0.0$).



(c) Command estimation result with F_0 reliability field.

Figure. 2: Examples of estimated prosodic commands. Arrows show the magnitude and the timing of phrase commands. Rectangles show the amplitude and the timing of the accent commands. A solid line in each figure represents an approximation of F_0 contour that is constructed by F_0 contour generation model. In Fig.(a) and Fig.(b), observed F_0 values are plotted with vertical lines, but in Fig.(a), F_0 determination errors are corrected and F_0 values on unvoiced frames are removed by hand operation. In Fig.(c), F_0 reliability is expressed as a density of gray-scaled color. The content of utterance is "*arayuru geNjitsuwo subete jibuNnohouhe nejimagetanoda*" in Japanese.

modification, namely it is defined as

$$\Delta S(t_0, p_0) = \sum_{i=-M}^{M} \sum_{\substack{j=-N \\ j\neq 0}}^{N} w(t_i, p_j) \left(\frac{S(t_i, p_j) - S(t_i, p_0)}{p_j - p_0} \right), \tag{8}$$

where $w(t_i, t_j)$ is a weighting function. Finally, we can obtain the optimized parameter set by iterative computation of the above modification, when an increase of F_0 reliability R_Λ becomes less than some threshold.

4. EVALUATION

4.1. Experimental conditions

Speech database used in this evaluation is the ATR's continuous speech database of phoneme balanced 503 Japanese sentences. Out of them, 50 sentences (A group) uttered by 1 male speaker (MHT) are used for the prosodic command estimation. A set of prosodic command parameters that we would estimate is $\{Ap_i, Aa_j, T_{0i}, T_{1j}, T_{2j}\}$, and initial values of those parameters are given by Hirai's technique[6], in which J_ToBI (Japanese Tone and Break Indices) labels are used. The other parameters are fixed and those values are $\ln Fb = 4.1$, $\alpha = 3.0$, $\beta = 20.0$, and $\theta = 0.9$.

As a comparative experiment, we examine a conventional estimation method, in which the observed F_0 contour and an optimization criterion of least squared error are used. The observed F_0 contour is determined automatically as a temporal sequence of frequency which gives maximum F_0 reliability at each time frame t, namely a sequence of $p_t = \arg\max_p S(t, p)$. If the F_0 reliability of p_t becomes lower than a threshold κ, i.e., $\max_p S(t, p) < \kappa$, it is regarded that there is no F_0 value at that time t. Furthermore, we have prepared ideal F_0 contours to obtain desirable prosodic commands. Here, ideal F_0 contour means the pattern that has no F_0 determination error, and those patterns

	Error†	Score‡
Initial set	0.0401	0.089
F_0 contour		
$(\kappa = 0.0)$	0.0624	0.078
$(\kappa = 0.1)$	0.0618	0.080
$(\kappa = 0.2)$	0.0500	0.124
$(\kappa = 0.3)$	0.0189	0.391
$(\kappa = 0.4)$	0.0164	0.476
$(\kappa = 0.5)$	0.0212	0.423
(ideal)	0.0097	0.437
F_0 reliability field		
	0.0209	0.542

(†) compared with ideal F_0 contour
(‡) F_0 reliability / max F_0 reliability

Table. 1: The approximation error and the F_0 reliability score.

have been created by hand operation of correcting F_0 errors.

4.2. Results

Examples of estimated prosodic commands are shown in Figure.2. In (b), the F_0 contour used for the estimation is automatically extracted with threshold $\kappa = 0.0$, and there are many F_0 determination errors, so the approximation of F_0 contour is extremely bad. Besides, we can see that timings of commands in (b) are greatly different from the estimation result of (a), in which the ideal F_0 contour is used. While, on the estimation (c) with FRF, minute approximation is possible because it is not necessary to consider the correction of F_0 errors and there is no lack of F_0 value on the observed prosodic feature.

Table.1 shows quantitative results of each estimated command set. The "Error" means the mean squared error in comparison with ideal F_0 contour, and the "Score" means the F_0 reliability score. Here, the F_0 reliability score is the ratio of the accumulated F_0 reliability to the accumulated maximum F_0 reliability, so it is defined as

$$\text{Reliability Score} = \frac{\sum_t S(t, f(\Lambda, t))}{\sum_t \max_p S(t, p)}. \tag{9}$$

From the first, a squared error becomes the smallest in the case of the estimation with F_0 contour, because its optimization is based on a criterion of least squared error. Similarly, a reliability score becomes the biggest in the case of the estimation with FRF, because it is optimized by maximizing F_0 reliability. These are expected results. However, we can see that the result of FRF is relatively good with both a squared error and a reliability score, while the estimation accuracy with F_0 contour depends on the F_0 determination accuracy.

But, as a problem of command estimation using FRF, it is pointed out that an establishment of initial parameter value becomes much stricter. This is because the number of local maxima reliability score is increased by harmonic peaks of FRF. Therefore, it may be desirable to use those prosodic features properly in case by case, for example, to estimate roughly by using the F_0 contour at first step, and to optimize by using the FRF at second step.

5. CONCLUSION

We have described that prosodic feature expression like F_0 reliability field is more suitable for prosody analyses than F_0 contour. The validity is shown in both analyses of detecting prosodic boundaries in previous paper and command estimation of F_0 contour generation model in this report. In future works, we would apply FRF for the other prosodic information analyses.

6. ACKNOWLEDGMENT

We would like to thank Mr. Toshio Hirai (OGIS-RI Co.,Ltd.) for providing us with the program, which gives initial command set of F_0 contour generation model.

7. REFERENCES

1. Mitsuru Nakai and Hiroshi Shimodaira. "On representation of fundamental frequency of speech for prosody analysis using reliability function," *EUROSPEECH'97*, pp.243–246, 1997.
2. Edouard Geoffrois. "The multi-lag-window method for robust extended-range F0 determination," *ICSLP-96*, pp.2239–2242, 1996.
3. Serugei Koval, Veronika Bekasova, Michael Khitrov and Andrey Raev. : "Pitch detection reliability assessment for forensic applications," *EUROSPEECH'97*, pp.489–492, 1997.
4. Hiroya Fujisaki and Keikichi Hirose. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Jpn (E)*, vol.5, no.4, pp.233–242, 1984.
5. Shigeki Sagayama and Sadaoki Furui. "A technique for pitch extraction by the lag-window method," *Proc. Conf. IEICE*, 1235, 1978 (in Japanese).
6. Toshio Hirai and Norio Higuchi. "Automatic extraction of the Fujisaki model parameters using the labels of Japanese Tone and Break Indices (J_ToBI) system," *Trans. IEICE*, vol.J81-D-II, no.6, pp.1058–1064, 1998 (in Japanese).