ON REPRESENTATION OF FUNDAMENTAL FREQUENCY OF SPEECH FOR PROSODY ANALYSIS USING RELIABILITY FUNCTION

Mitsuru NAKAI and Hiroshi SHIMODAIRA

Japan Advanced Institute of Science and Technology, Hokuriku
1-1 Asahidai, Tatsunokuchi, Nomi, Ishikawa, 923-12 Japan

ABSTRACT

This paper highlights on a method that provides a new prosodic feature called 'F_0 reliability field' based on a reliability function of the fundamental frequency (F_0). The proposed method does not employ any correction process for F_0 estimation errors that occur during automatic F_0 extraction. By applying this feature as a score function for prosodic analyses like prosodic structure estimation or superpositional modeling of prosodic commands, these prosodic information could be acquired with higher accuracy. The feature has been applied to 'F_0 template matching method', which detects accent phrase boundaries in Japanese continuous speech. The experimental results show that compared to the conventional F_0 contour, the proposed feature overcomes the harmful influence caused by F_0 errors.

1. INTRODUCTION

Prosody is a very important information for speech understanding. Several researches have been performed for obtaining prosodic information and they are reported in both of speech synthesis and speech recognition field. We have also proposed an automatic scheme called 'F_0 template matching method[1]' for estimating prosodic phrase boundaries of Japanese continuous speech. It is largely hoped that these kind of prosodic information will be useful for constructing a high-performanced speech recognition system with low-costed CPU power.

Among several prosodic features, F_0 contour (pitch pattern) has been widely used for prosodic analysis and many Pitch Determination Algorithms (PDA) have been proposed. However, PDAs tend to yield some gross errors, such as harmonic errors of double pitch or half pitch, and error correction is one of the laborious postprocessing task in any automatic PDA.

In the present approach, our segmentation system is based on a pattern matching technique that employs some distance measure between reference templates of typical accent F_0 pattern and an observed F_0 contour for which the prosodic boundaries are unknown. Therefore, the previously mentioned errors can largely effect the boundary segmentation results as the number of boundary insertion increase. However, this problem is inevitable as long as the F_0 value is fixed to some unique value per temporal analyzing frame. In the present approach, this problem has been thoroughly investigated and an 'F_0 *reliability field'* has been designed in which there is no process of deciding F_0 value distinctly.

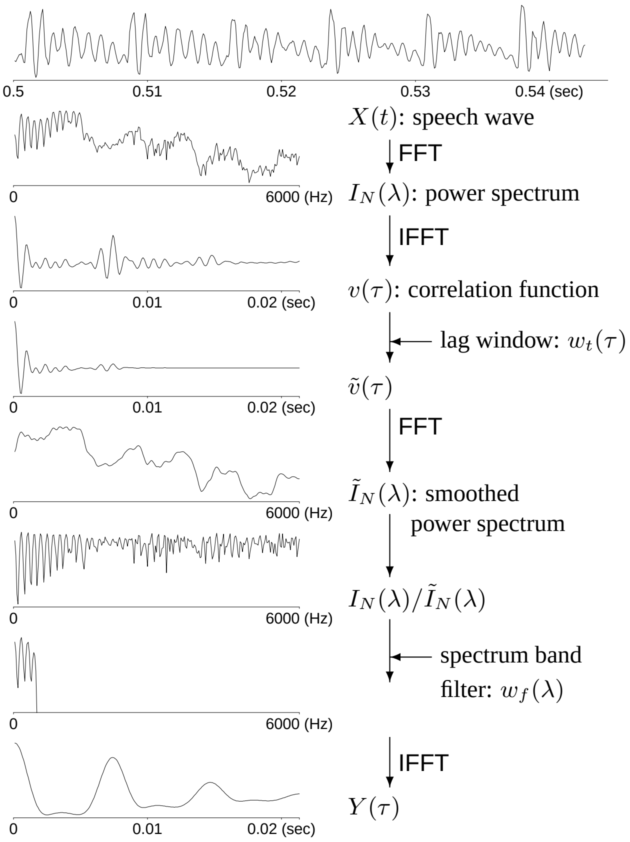


Fig. 1: A process of F_0 reliability analysis based on lag-window method of F_0 determination.

2. ANALYSIS OF *F0* AND *∆F0* RELIABILITY FIELD

Euclidean distance is one of the most widely used measure for pattern matching technique. The proposed prosodic segmentation system is based on Dynamic Programming (DP) method that employs Euclidean distance. The distance between two F_0 values at time t is defined by

$$d(p_t, \bar{p}_t) = (p_t - \bar{p}_t)^2, \tag{1}$$

where $\bar{p}_t$ is a reference $\ln F_0$ value and p_t is an observed $\ln F_0$ value. This distortion function has a continuous distribution on the frequency axis. However, as the detected F_0 value p_t sometimes has an error of half pitch or double pitch, the distortion does not vary monotonously on the temporal sequence. In the present approach, the F_0 reliability field does not have any process for F_0 determination and therefore, it is not necessary to consider such problems.

Fig.1 shows a process of F_0 reliability analysis based on lag-window method[2] which is one of PDAs. The desirable smoothed function $Y(\tau)$ can be obtained by incorporating a narrow spectrum band filter[3]. We have applied rectangle band filter as a window function in previous works, but here we use the Hanning window $w_f(\lambda)$ on the frequency domain.

By using a temporal sequence of the function $Y(\tau)$, the F_0 reliability field can be represented as shown in Fig.2. Here, the horizontal axis represents the logarithmic frequency $\ln(1/\tau)$ and it can be seen that the contours of the harmonic peak lie in a fixed interval. This field is defined as a distribution of the form $S(t,p)$, where t and p stand for time and the logarithmic value of F_0 respectively. Here, similar to the characteristic of lag-window method, the maximum value of the field $S(t,p)$ is normalized to 1.0. If we determine the frequency value by the following equation,

$$f_t = \arg\max_p S(t,p), \tag{2}$$

then f_t would be the exact F_0 value at time t and the algorithm would become equivalent to PDA.

We have also defined ΔF_0 as a regression coefficient of $\ln F_0$. Similarly, ΔF_0 vector can be represented as a direction along the ridge of an F_0 reliability field and it is shown in Fig.3(a). If we assume that the vector (v_t, v_p) lies in the direction of the maximum slope on the point $\ln F_0$ with value p_0 at time t_0, then the component of the vector along the temporal direction can be represented by the following equation,

$$v_t = \sum_{\substack{i=-M \\ i\neq 0}}^{M} \sum_{j=-N}^{N} w_t(t_i,p_j)\left(\frac{S(t_i,p_j)-S(t_0,p_j)}{t_i-t_0}\right), \tag{3}$$

and the component of the vector along the frequency direction can be represented by the following equation.

$$v_p = \sum_{i=-M}^{M} \sum_{\substack{j=-N \\ j\neq 0}}^{N} w_p(t_i,p_j)\left(\frac{S(t_i,p_j)-S(t_i,p_0)}{p_j-p_0}\right). \tag{4}$$

Here, M and N are the window width, w_t and w_p are weighting function for each direction. In Fig.3(b) the ΔF_0 vector is represented by a vector that is orthogonal

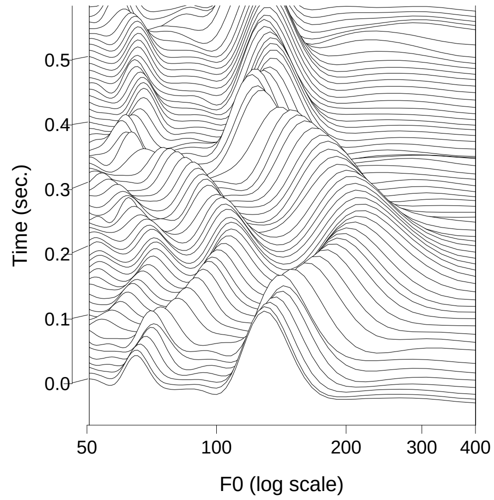
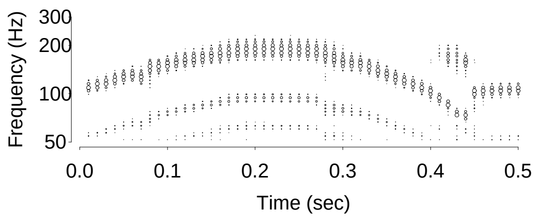
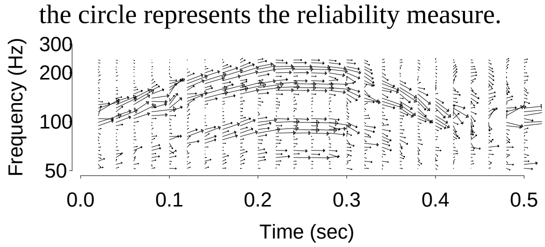


Fig. 2: An example of F_0 reliability filed.



(a) Observation of F_0 reliability pattern. The radius of the circle represents the reliability measure.



(b) A vector field of F_0 reliability pattern.

Fig. 3: F_0 reliability pattern and its ΔF_0 vector field.

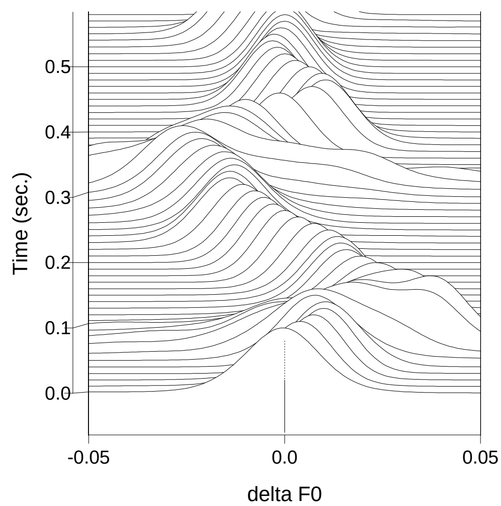


Fig. 4: An example of ΔF_0 reliability filed. The maximum value is normalized to 1.0.

to (v_t, v_p) and it has an angle of $a_{(t,p)} = -v_t/v_p$ with the temporal axis. Fig.4 shows an example of ΔF_0 reliability field which is calculated by the summation of Gaussian distribution per each temporal frame:

$$\sum_p \sqrt{v_t^2 + v_p^2}\, N(a_{(t,p)}, \sigma^2). \qquad (5)$$

3. PROSODIC SEGMENTATION USING RELIABILITY FIELD

The proposed prosodic segmentation system[1] has two phases; training phase and segmentation phase, and the reliability field is only used during the segmentation phase. During training, a set of K templates $\{R_0, \cdots, R_{K-1}\}$ are created from a large number of accent F_0 patterns using k-means clustering algorithm, and they are approximated semi-automatically by the commands of the superpositional model[4]. On the other hand, the segmentation phase is performed automatically by One-Stage DP matching between the reference F_0 templates and the target F_0 reliability field. The location of acquired reference template connection boundary is considered to be the phrase boundary. The N-best sequences of F_0 templates can also be searched by using the criterion of the N most scores. The matching score between the reference template $R_i = (\bar{p}_1, \cdots, \bar{p}_n)$ and an input speech started at time t_s is defined as

$$D_i = \sum_{t=1}^{n} S(t_s + t, \bar{p}_t), \qquad (6)$$

where $S(t, p)$ is a reliability field of F_0 (or ΔF_0). If the matching template forms the sequence $R_{c_1} \oplus R_{c_2} \oplus \cdots \oplus R_{c_i} \oplus \cdots \oplus R_{c_M}$ $(0 \leq c_i \leq K-1)$, then the score is summed up to

$$D = \sum_{i=1}^{M} D_{c_i} + \gamma \sum_{i=1}^{M-1} \ln P(c_{i+1}|c_i) \qquad (7)$$

where $P(c_{i+1}|c_i)$ is the transition probability from the template R_{c_i} to the template $R_{c_{i+1}}$ and γ is the strength factor of their bigram constraints. Furthermore, if $S(t, p)$ is replaced to $-d(p_t, \bar{p}_t)$ in Eq.(1), then the algorithm becomes equivalent to the conventional segmentation method which is based on Euclidean measure.

4. EVALUATION OF PROSODIC FEATURES

For the evaluation of prosodic features, prosodic segmentation accuracy has been used under the environment of 1-best search for F_0 template sequence. The speech database used in the evaluation test is the ATR's continuous speech database of phoneme balanced 503 Japanese sentences uttered by each of 3 male speakers and 1 female speaker. Out of them, a total of 575 utterances from 3 male speakers (MHT, MSH, MTK) is used for constructing 8 F_0 templates and the bigram probabilities between the templates were estimated. Automatic prosodic segmentation was performed for each 50 sentences from the speaker MHT (male) and the speaker FKN (female), which are different in contents from the training sentences.

The result is shown in Fig.5 and Fig.6. The y-axis is the correctly detected rate which is defined by

$$\text{correct rate} = \frac{\text{\# correctly detected boundaries}}{\text{\# accent boundaries}}$$

and the x-axis is the boundary insertion error rate per accent phrase:

$$\text{insertion rate} = \frac{\text{\# incorrectly detected boundaries}}{\text{\# accent phrases}}.$$

Here, we treat the detected boundaries located within 100 ms from the hand labeled boundaries as the correct one. Varying the bigram strength γ, we can reduce many insertion errors while the correct detected rate decrease a little.

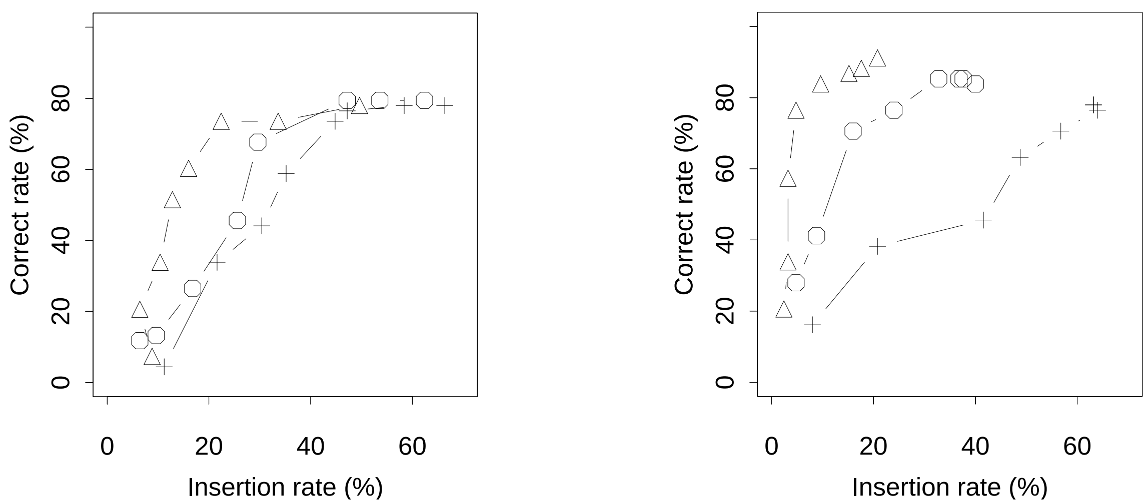
Apart from the proposed reliability field ('$\bigcirc$' in figure), the following two features have been used as the target pattern for DP matching. One is an automatically detected F_0 contour using PDA and the postprocessing of F_0 error correction is not applied ('$+$' in figure). The other is an ideal F_0 contour produced by F_0 generating function[4], and it is regarded as the best case with no F_0 errors ('$\triangle$' in figure).

In reality, it is not possible to detect all of the prosodic boundaries, because the perception of the same depend on the listener and the hand labeled boundaries used during the experiments are not always correct. Therefore, 70% correct rate with less than 25 % insertion rate is considered to be sufficiently efficient for prosodic segmentation accuracy of Japanese speech. It is also reported in the experimental results of Takahashi and Matsunaga[5], that these numerical values are similar to the human hearing ability for distinguishing accent segments.

It can be seen that the segmentation accuracy using the reliability field ($\triangle$) is better than the conventional F_0 contour ($+$) and the result does not depend on F_0 or ΔF_0, and male or female speakers. In the experiments of speaker MHT, the segmentation accuracy came up to 70 % of correct rate with low insertion error rate, and it is close to the accuracy of ideal F_0 contour which has no F_0 estimation error. On the other hand, due to several F_0 extraction errors, the segmentation accuracy of speaker FKN has been always lower than the other speakers. However, in case of her utterance, we can see the improvement of segmentation accuracy in Fig.6. It is concluded that the reliability field is enable to overcome the harmful influences caused by F_0 errors.

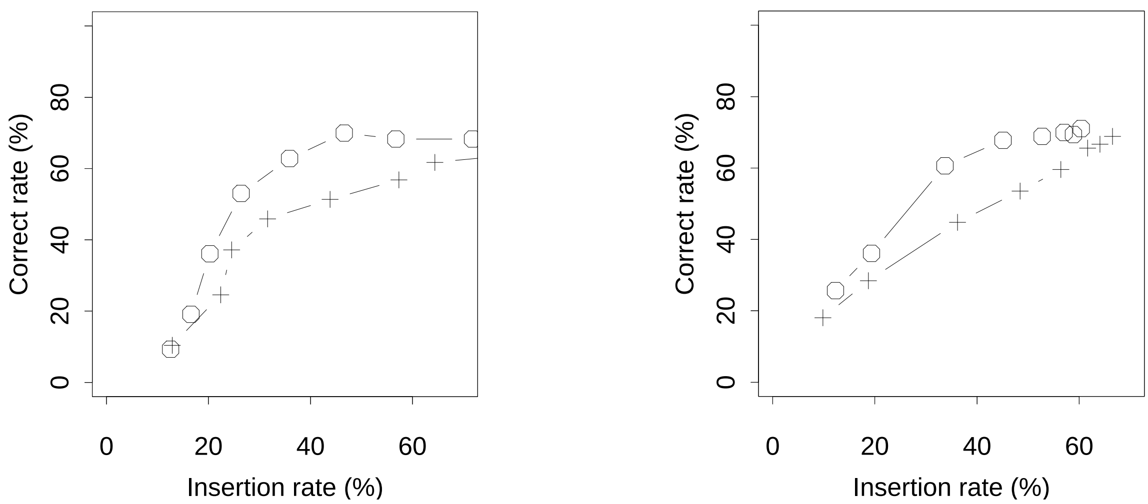
5. CONCLUSION

A new prosodic feature 'F_0 *reliability field*' has been proposed. The analysis of this feature is very simple and it is basically the same as the conventional F_0 determination



(a) Segmentation accuracy with F_0 reliability field
(b) Segmentation accuracy with ΔF_0 reliability field

Fig. 5: A relationship of correctly detected rate and boundary insertion rate for speaker MHT. '$+$' is using an automatically detected F_0 contour, '$\triangle$' is using an ideal F_0 contour without F_0 errors, and '$\bigcirc$' is using a reliability field.



(a) Segmentation accuracy with F_0 reliability field
(b) Segmentation accuracy with ΔF_0 reliability field

Fig. 6: A relationship of correctly detected rate and boundary insertion rate for speaker FKN. '$+$' is using an automatically detected F_0 contour, and '$\bigcirc$' is using a reliability field.

algorithm, like lag-window method. In case of the proposed reliability field feature, it is not necessary to correct the F_0 extraction errors due to the cause of focusing on a specific F_0 decision value. It has been also shown through experiments using prosodic segmentation, that the proposed feature is effective for prosodic analysis. The next step would be to apply this feature for other prosodic information analysis, like prosodic structure estimation or superpositional modeling of prosodic commands.

REFERENCES

[1] M. Nakai and H. Singer and Y. Sagisaka and H. Shimodaira: "Automatic Prosodic Segmentation by F_0 Clustering Using Superpositional Modeling", *ICASSP-95*, pp.624–627, (1995).

[2] S. Sagayama and S. Furui: "A technique for pitch extraction by the lag-window method", Proc. Conf. IEICE, 1235 (1978) (in Japanese).

[3] H. Shimodaira and M. Nakai: "Robust Pitch Detection by Narrow Band Spectrum Analysis", *ICSLP-92*, pp.1597–1600, (1992).

[4] H. Fujisaki and H. Kawai: "Realization of Linguistic Information in the Voice Fundamental Frequency Contour of the Spoken Japanese", *ICASSP-88*, pp.663–666, (1988).

[5] S. Takahashi and S. Matsunaga: "Stochastic Prosody Modeling for Accent Phrase Boundary Detection in Continuous Speech", In Papers of Technical Group of Speech, IEICE, SP90-71 (1990) (in Japanese).