AUTOMATIC PROSODIC SEGMENTATION BY F_0 CLUSTERING USING SUPERPOSITIONAL MODELING

Mitsuru Nakai, Harald Singer †, Yoshinori Sagisaka † and Hiroshi Shimodaira ‡

Tohoku University, Sendai, 980 Japan
† ATR Interpreting Telecommunications Research Labs. Seika, Soraku, Kyoto 619-02 Japan
‡ Japan Advanced Institute of Science and Technology, Tatsunokuchi, Ishikawa, 923–12 Japan

ABSTRACT

In this paper, we propose an automatic method for detecting accent phrase boundaries in Japanese continuous speech by using F_0 information. In the training phase, hand labeled accent patterns are parameterized according to a superpositional model proposed by Fujisaki, and assigned to some clusters by a clustering method, in which accent templates are calculated as centroid of each cluster. In the segmentation phase, automatic N-best extraction of boundaries is performed by One-Stage DP matching between the reference templates and the target F_0 contour. About 90% of accent phrase boundaries were correctly detected in speaker independent experiments with the ATR Japanese continuous speech database.

1. INTRODUCTION

Continuous speech recognition and understanding is a difficult task and it is indispensable to use phrase boundary information for raising the recognition accuracy. But extraction of phrase boundaries in continuous speech by a preprocessor has not yet been developed, and thus speech recognition is very costly in terms of CPU time and memory. Therefore, the estimation of boundary positions either directly from the input speech or from extracted prosodic parameters is a very important problem.

In this paper we propose a new segmentation scheme using structured expressions of F_0 patterns based on superpositional modeling. This structured expression enables stochastic modeling of the correlation between adjacent prosodic phrases, and has achieved significantly higher performance than a previous extraction scheme using plain F_0 clustering[1].

Both our previous method and this new method are based on the assumption, that all F_0 patterns can be expressed by a limited number of typical pattern templates and that a whole F_0 contour can be approximated by a connection of these patterns. We can thus reformulate the problem of phrase boundary extraction as a recognition of accent phrase patterns. We have implemented this approach as a One-Stage DP matching of the F_0 contour against a sequence of templates.

In our previous research, templates were constructed by clustering the accent phrases of an F_0 contour without a feature model for accent phrase patterns. By contrast, our new method is based on a widely-used F_0 control model in which the accent phrase patterns can be expressed by very few parameters, so that templates can be constructed using comparatively little training data.

Furthermore, a major benefit of using an accent model is that by using constraints on the generation of the F_0 pattern, both constraints on the path in the One-Stage DP and structural restrictions on the transitions between templates can be applied, and the calculation cost can be considerably reduced.

In the following, we will describe the algorithm and give some experimental results.

2. ALGORITHM

2.1. Superpositional model

We use the accent phrase model proposed by Fujisaki[2]. In this model, the fundamental frequency F_0 as a function of time t is given by

$$\begin{aligned}\ln F_0(t) &= \ln F_{\min} + \sum_{i=1}^{I} A_{p_i} G_{p_i}(t - t_{pi}) \\ &+ \sum_{j=1}^{J} A_{a_j} \left\{ G_{a_j}(t - t_{aj}) - G_{a_j}(t - (t_{aj} + \tau_{aj})) \right\},\end{aligned}$$

where

$$\begin{aligned}G_{p_i}(t) &= \alpha_i t e^{-\alpha_i t}, \\ G_{a_j}(t) &= \min[1 - (1 + \beta_j t) e^{-\beta_j t}, \theta_j],\end{aligned}$$

indicate the impulse response function of the phrase control mechanism and the step response function of the accent control mechanism. The symbols in the above equations indicate

$F_{\min}$: bias level,
I, J: number of phrase and accent commands,

current accent command (a_1)	
τ_b:	duration of accent pattern
A_{a1}:	amplitude of accent command
t_{a1}:	onset of accent command
τ_{a1}:	duration of accent command
previous phrase command (p_0)	
A_{p0}:	magnitude
t_{p0}:	instant of occurrence
current phrase command (p_1)	
A_{p1}:	magnitude
t_{p1}:	instant of occurrence

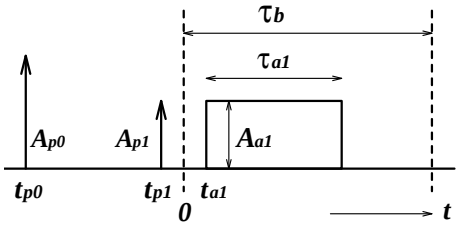

Figure 1: Parameter set for accent phrase model

A_{p_i}: magnitude of the ith phrase command,
A_{a_j}: amplitude of the jth accent command,
t_{pi}: instant of occurrence of the ith phrase command,
t_{aj}, τ_{aj}: onset and duration of the jth accent command,
α_i, β_j: natural angular frequency of the phrase and accent control mechanisms,
θ_j: ceiling level of the accent component.

Among these parameters we decided to keep $\alpha_i, \beta_j, \theta_j$ fixed because there is no large variation of these parameters between different speakers or speaking styles. An accent phrase pattern in this research is thus represented by the 8 parameters shown in Figure 1.

2.2. Training phase

Hand labeled F_0 accent phrase patterns are parameterized according to the mentioned above model and from these parameterized patterns a new set of F_0 patterns P_j is regenerated, where $P_j = (p_{j1}, \cdots, p_{ji}, \cdots, p_{jL})$ with p_{ji} as the logarithmic F_0 value of frame i for the j-th accent phrase and L as a fixed length in common for all patterns. Then, the distance between a pair of patterns, P_j and P_k can be defined by

$$D(P_j, P_k) = \sum_{i=1}^{L} (p_{ji} - p_{ki})^2.$$

After the LBG clustering operation, the average model parameters for each cluster are calculated and a set of templates $R = \{R_0, R_1, \cdots, R_{K-1}\}$ is constructed. Figure 2 shows the reference templates in the case of $K = 8$.

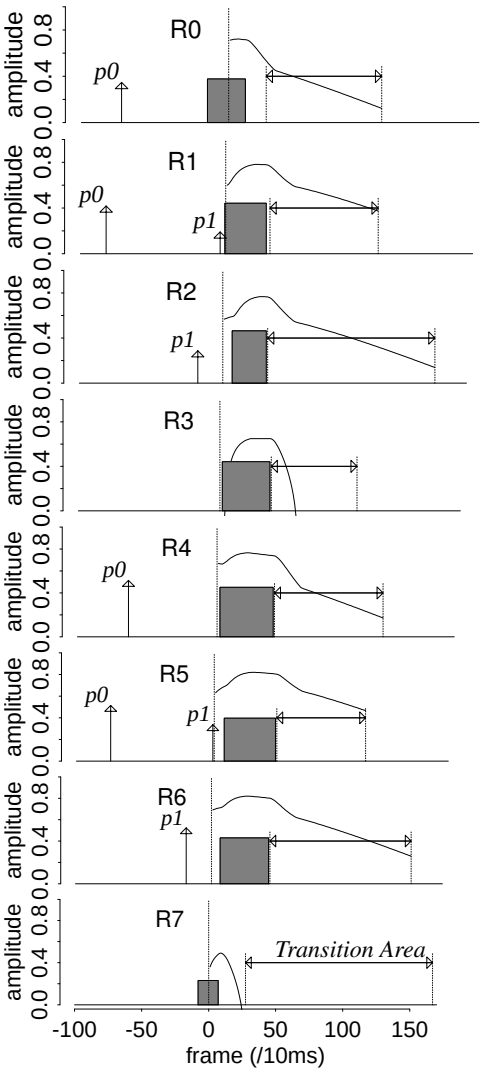

Figure 2: F_0 pattern and corresponding parameters (see Figure 1) for each accent phrase cluster. Values for A_{p0} and A_{p1} may be 0. For example, a sentence with one phrase command and three accent commands might be represented by the template sequence R2-R0-R3.

2.3. Segmentation phase

Automatic extraction is performed by One-Stage DP[3] between the reference templates and the target F_0 contour. The DP path can be constrained to 45 degrees shown in Figure 3 as in the superpositional model with fixed angular frequency (α, β) any F_0 value in an accent phrase pattern is completely defined by amplitude of commands and time from onset of commands. A certain template is accepted as a valid match only if the length of the corresponding F_0

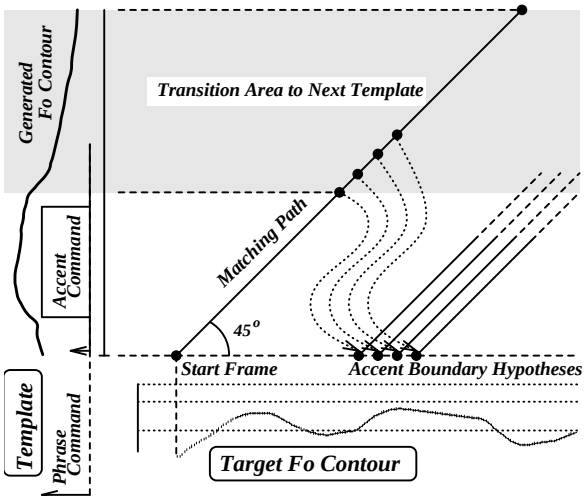

Figure 3: Matching Path

contour segment is shorter than the maximal length of all accent phrase patterns in the cluster for this template, and at the same time longer than the following three lengths: (1) the minimal length of all accent phrase patterns in the cluster for this template, (2) half of the average accent phrase pattern length, (3) end of accent command. For an adaptation of the minimal fundamental frequency ($F_{\min}$), we add a variable offset value to the reference templates.

As there is a strong correlation between adjacent templates, we use this additional information by introducing bigram probabilities of accent phrases as a template connection cost defined by

$$C(k^*, k) = -\gamma * \ln\left(P(k \mid k^*)\right),$$

where $P(k \mid k^*)$ is a transition probability from k^*-th template to k-th template, and γ is a factor of strength of bigram constraints.

Figure 4 is an example of the segmentation result in which the 8 templates are matched against the F_0 contour in [c] and the 10 best results are given in [d]. [a] displays the input speech wave and the vertical lines show the hand-labeled accent phrase boundaries. [b] shows the reliability of the pitch values, which are used as a weighting coefficient for the squared error between reference template and F_0 contour. The labels on top of each accent phrase candidate in [d] refer to the templates given in Figure 2.

3. EVALUATION

3.1. Experimental Condition

The speech database used in this evaluation test consists of a continuous speech database of phoneme balanced 503 Japanese sentences uttered by 4 male speakers [5].

Table 1: Experimental condition

Pitch extraction	
Window Length	512 point (42.7ms)
Analysis Interval	120 point (10.0ms)
F_0 Search	(male) 50 ~ 300 Hz
Extraction Method	lag-window method
Automatic segmentation	
# template	8
# candidate	10-best

For a total of 565 sentences from 3 speakers (MHT, MSH, MTK), model parameters were semi-automatically extracted[4] and then 8 accent phrase templates were constructed and bigram probabilities between phrase templates were estimated. Automatic phrase segmentation was then performed with 50 sentences from a different speaker (MYI), which are also different in contents from the training sentences, and the 10 best candidates were retained.

Detected boundaries located within 100 ms from the hand labeled boundaries are treated as correct. Correct rate (R_c) and insertion rate (R_i) for each candidate are defined by

$$R_c = \frac{\text{\# correct detected boundaries}}{\text{\# hand labeled boundaries}},$$

$$R_i = \frac{\text{\# incorrect detected boundaries}}{\text{\# hand labeled boundaries}}.$$

In the example of Figure 4, the number of hand labeled boundaries in the first part of the sentence before the pause is two, and the correct rate R_c of the first candidate is 100% (2/2). On the average over 10 candidates, the correct rate $\bar{R}_c$ is 90% (16/20), and the insertion rate $\bar{R}_i$ is 10% (2/20). Also, when we merge all boundaries on the 10-best candidates into 1 sequence together, the correct rate of the sequence, which we call"10-best" correct rate R_c^{10}, becomes 100% (2/2).

3.2. Results

Figure 5 shows segmentation accuracy when varying the strength of bigram constraints γ. As γ increases from 0.0 to 1.0, both the averaged correct rate $\bar{R}_c$ and the averaged insertion rate $\bar{R}_i$ decrease, but the "10-best" correct rate R_c^{10} does not decrease so rapidly because undetected boundaries for higher ranking candidates can be detected in lower ranking candidates. Varying γ between 0.0 and 0.1, we notice a reduction of the insertion errors $\bar{R}_i$ from 75.3% to 39.1% while R_c^{10} remains above 90%. Thus the template bigram is a useful constraint for insertion error control.

Furthermore, we carried out additional experiments for comparison of our previous method[1] for $\gamma = 0.1$, and obtained $R_c^{10} = 87.5\%$ and $\bar{R}_i = 67.9\%$. Half of the undetected boundaries in the previous segmentation system can thus be

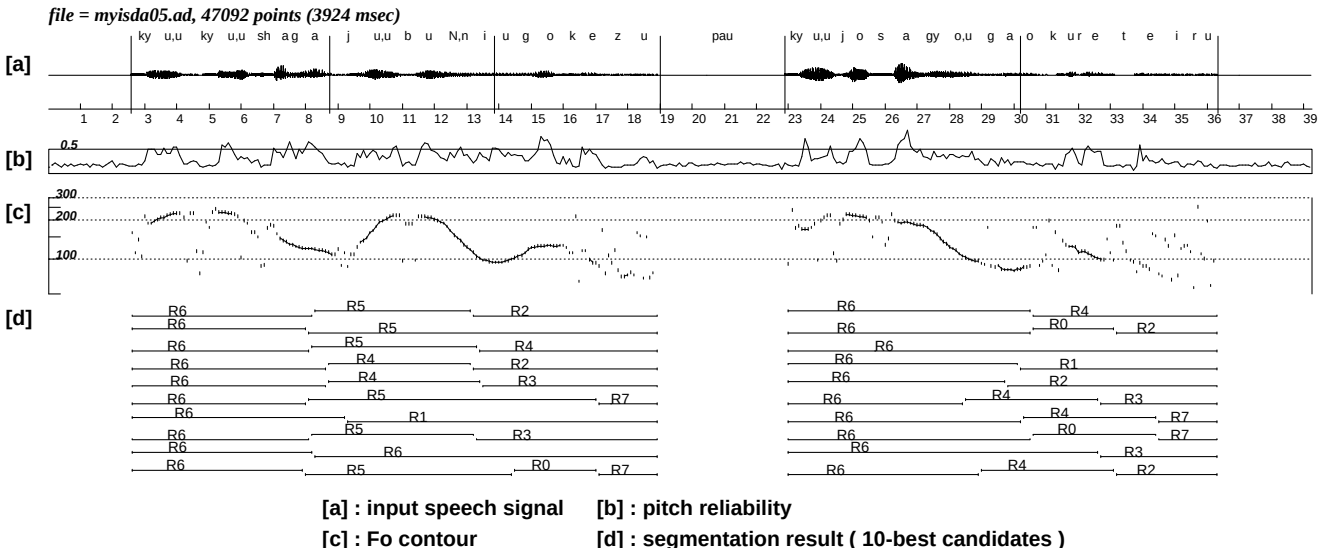

Figure 4: Example of segmentation result

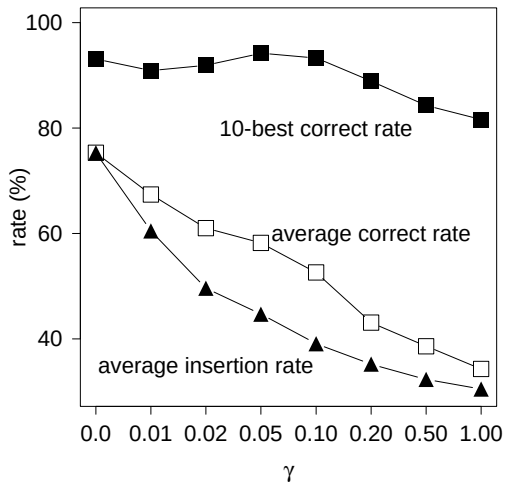

Figure 5: Segmentation accuracy (MYI)

detected by our new segmentation scheme (R_c^{10} = 93.3%) with the additional benefit of reducing the insertion error rate $\bar{R}_i$ from 67.9% to 38.6%.

4. CONCLUSION

We have proposed a new segmentation scheme using structured expressions of F_0 patterns based on superpositional modeling. These structured expressions enable stochastic modeling of the correlation between adjacent prosodic phrases and permit significantly higher performance than a previous extraction scheme using plain F_0 clustering.

Another interesting aspect of our method is that we do not rely on automatic extraction of parameters for the superpositional model during "recognition". These parameters are used only during training and can thus be hand-corrected.

However, we notice that there is still room for improvement of the template training process. For example, there is a large difference between the centroid F_0 contours calculated by LBG clustering and the F_0 contours generated from average accent model parameters for each cluster. As a second step, we started to develop a continuous speech recognition search algorithm to use this phrase boundary information effectively.

REFERENCE

[1] M. Nakai and H. Shimodaira: "Accent Phrase Segmentation by Finding N-best Sequences of Pitch Pattern Templates", *ICSLP-94*, pp.347–350, (1994).

[2] 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).

[3] Hermann Ney: "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition", *IEEE ASSP-32*, 2, pp.263–271 (1984-04)

[4] T. Hirai, N. Iwahashi, H. Valbert, N. Higuchi and Y. Sagisaka: "Fundamental Frequency Contour Modeling Using Statistical Analysis", *Proc. Acoust. Soc. Jpn. Autumn 93*, pp.225–226, (1993-10)

[5] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, H. Kuwabara : "A Large-Scale Japanese Speech Database", *ICSLP-90*, pp.1089–1092, (1990).