ACCENT PHRASE SEGMENTATION USING TRANSITION PROBABILITIES BETWEEN PITCH PATTERN TEMPLATES *

Hiroshi SHIMODAIRA and *Mitsuru NAKAI* †

School of Information Science, Japan Advanced Institute of Science and Technology,
Tatsunokuchi, Ishikawa, 923-12 Japan
E-mail: sim@jaist-east.ac.jp
†Dept. of Information Eng., Faculty of Eng., Tohoku University,
Sendai-shi, 980 Japan

Abstract

This paper proposes a novel method for segmenting continuous speech into accent phrases by using a prosodic feature 'pitch pattern'. The pitch pattern extracted from input speech signals is divided into the accent segments automatically by using the One-Stage DP algorithm, in which reference templates representing various types of accent patterns and connectivity between them are used to find out the optimum sequence of accent segments. In case of making the reference templates from a large number of training data, the LBG clustering algorithm is used to represent typical accent patterns by a small number of templates.

Evaluation tests were carried out using the ATR continuous speech database of a male speaker. Experimental results showed more than 91 % of phrase boundaries were correctly detected.

Keywords: prosodic phrase, accent phrase segmentation, pitch pattern clustering

1 Introduction

It is well known that prosodic features, such as accent, intonation and pauses, which are noticed as super segmental features of spoken language, provide important information for comprehending speech, especially spontaneous speech [1][2][3]. But these prosodic features have not been used successfully in speech recognition and speech understanding, because of their unreliable and unstable properties, in contrast to steady progress of phonetic based recognition.

As speech recognition tasks become difficult and complicated, conventional speech recognition systems based on phonemic segment classification without using prosodic features could not achieve reliable performance any more.

Under these situation, some statistical approaches to find out prosodic segments by using the prosodic features have been proposed [4][5][6][7]. Among the approaches, our approach is based on clustering of accent patterns and optimal boundary determination in the meaning of the least squared error (LSE) by using the One-Stage DP search algorithm [6]. The segmentation operation based on LSE criterion is not done locally in speech signals but done globally over long duration of speech signals, such as a sentence. Therefore our approach is expected to be robust against the variety of utterance speed and the fluctuation of F_0 contours.

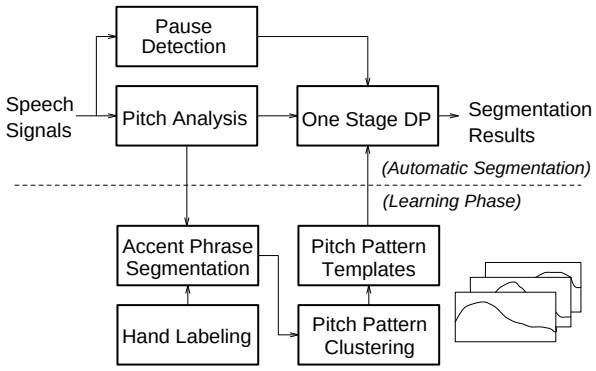

Figure 1: Block diagram of the system

Fig. 1 shows a block diagram of the whole system. In learning phase, F_0 contours extracted from input speech signals are divided into accent phrases according to hand labeled boundaries, and then a large number of pitch patterns of accent phrases are classified into a small number of patterns called 'pitch-pattern templates' by using the LBG clustering algorithm. In testing phase (automatic segmentation phase), the One-Stage DP matching between F_0 contours of input speech signals and the pitch pattern templates is carried out to find out the optimum sequence of segments, which are expected to be corresponding to real accent phrases. In the operation of the One-Stage DP matching, we further take account of transition probabilities between the consecutive templates not to permit improbable template connections.

*This work was begun with Dr. Shigeki SAGAYAMA when one of the authors visited ATR Interpreting Telephony Research Laboratories.

2 Classification of Accent Pitch Patterns

2.1 Clustering algorithm

It is noticed that the syntactic structure of spoken language have good relationship with accent phrases. The accent phrase, which is sometimes called a 'prosodic phrase', is the unit which has a single accented syllable ('accent cue').

From the observations on F_0 contours, most of the boundaries of accent phrases can be found around the '*fall-rise patterns*' of F_0 contours. However, all of the fall-rise patterns of F_0 contours are not caused by the accent phrases but also caused by the the unvoiced sounds and detection errors of F_0 and so on. Therefore finding the boundaries of accent phrases from F_0 contours is not so easier than expected.

In phonetics, the accent patterns of accent phrases can be classified into a small number of accent types according to the location of the accented syllable in the phrase. If we have some templates of pitch patterns corresponding to the accent types. we could find the boundaries of accent phrases much easier and more reliable than searching the fall-rise pattern. From the view point of pattern recognition, it is much interesting and perhaps effective to use pitch-pattern templates designed by using the clustering technique rather than by using the a priori knowledge of phonetics. This is why we make clustering of the pitch patterns of accent phrases for prosodic segmentation.

To keep consistency with the algorithm of accent pattern segmentation in section 3, the "LBG" vector quantization algorithm is implemented to perform the accent pattern clustering without using a priori knowledge of accent types. Only boundaries of accent phrases are required in the algorithm.

To apply the VQ algorithm to the accent pattern clustering, we must at first define the distance measure between the two accent patterns P_j, P_k, which have the different frame length each other. To avoid difficulty in comparing with the two patterns of different frame length, we divided the operation into the comparison of the shapes of pattern, and the comparison of the frame length.

For the comparison of the shapes, we transform P_j and P_k linearly into $\hat{P}_j$ and $\hat{P}_k$ respectively, and both of which have the same frame length of L. Let $\hat{p}_{ji}$ be a F_0 of the i-th frame of $\hat{P}_j$ in a logarithmic scale. Then the distance between the two patterns $\hat{P}_j$ and $\hat{P}_k$ will be defined by

$$D_S(\hat{P}_j, \hat{P}_k) = \sum_{i=1}^{L} |\hat{p}_{ji} - \hat{p}_{ki}|^2. \quad (1)$$

Considering the case that $\hat{P}_k$ is similar to $\hat{P}_j$ except that $\hat{p}_{ji} = \hat{p}_{ki} + const.$ for all $k = 1, \cdots, L$, then the two patterns might fail to be classified into the same cluster. This characteristic is not desirable for good distance measure to classify accent patterns based on the shapes of pitch patterns. To overcome this obstacle, the approach we took is to shift the height of the all patterns by adjusting the values of the beginning frame to the same value. Denoting the new pattern by $\tilde{P}_j = (\tilde{p}_{ji})$, the shifting operation is given by

$$\tilde{p}_{ji} = \hat{p}_{ji} - (\hat{p}_{j1} - b), \quad (2)$$

where b is any constant value. Then the distance between the shifted pattern $\tilde{P}_j$ and $\tilde{P}_k$ is now

$$D_S(\tilde{P}_j, \tilde{P}_k) = \sum_{i=1}^{L} |\tilde{p}_{ji} - \tilde{p}_{ki}|^2. \quad (3)$$

For the comparison of the frame lengths, we just defined the distance between the two patterns by

$$D_L(\tilde{P}_j, \tilde{P}_k) = (L_j - L_k)^2 \quad (4)$$

where L_j, L_k denote the original frame length of P_j, P_k.

Using the two types of distance measures above, we define the distance between the two patterns of different frame length by

$$D_\lambda(\tilde{P}_j, \tilde{P}_k) = (1-\lambda) D_S(\tilde{P}_j, \tilde{P}_k) + \lambda C_L D_L(\tilde{P}_j, \tilde{P}_k) \quad (5)$$

where C_L is a normalizing factor defined by

$$C_L = \frac{\sum_{\tilde{P}_n \in \tilde{P}} D_S(\tilde{P}_n, \bar{P})}{\sum_{\tilde{P}_n \in \tilde{P}} D_L(\tilde{P}_n, \bar{P})}, \quad (6)$$

λ is a weighting factor for D_L, $sPitch$ is a set of $\tilde{P}_j$ and $\bar{P}$ is the average pattern of $\tilde{P}$.

Once defining the distance above, we can perform clustering of accent pitch patterns, and finally we have a set of reference templates as the centroid vectors of the VQ, $R = \{R_1, R_2, \cdots, R_M\}$. At this time, the frame length of each reference pattern R_m is set to the average frame length in the m-th cluster.

2.2 Pitch determination algorithm

Because both of the algorithm of accent pattern clustering described above and the algorithm of prosodic segmentation explained in the next section are using the LSE criterion, accuracy and continuity of F_0 contours is an very important factor for the performance. Especially, discontinuity of F_0 may lead the system undesirable results. So the pitch determination algorithm (PDA) used in our system is the one characterized by estimating F_0 contours as continuous as possible. Fig. 2 shows the block diagram of our PDA [8]. The basic idea of the algorithm is the multiple-band F_0 analysis and integration of the F_0 candidates. For every frame, and for each frequency window $k = 1 \sim K$, $F_0(k)$ is extracted from the k-th narrow-band. Then the optimal F_0 contours which satisfies continuity best are chosen by using the DP algorithm.

3 Segmentation Algorithm

Let $R = \{R_1, R_2, \cdots, R_M\}$ be the set of reference patterns created by the clustering operation, then segmentation of the pitch pattern into accent phrases is a question of solving the optimum time warping function between the input pitch pattern and the super-reference pattern:

$$\mathcal{R} = R_{q(1)} \oplus R_{q(2)} \oplus \ldots \oplus R_{q(N)}$$

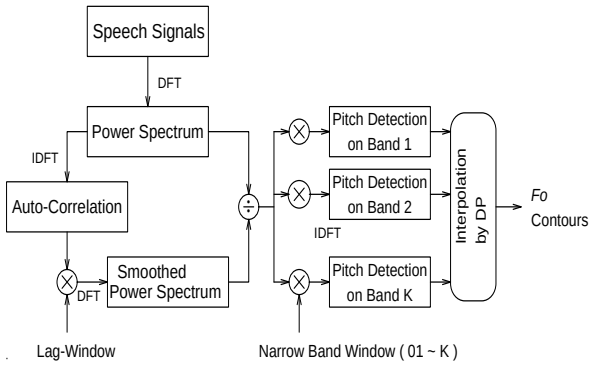

Figure 2: Block diagram of pitch pattern estimation algorithm

where $q(1), \cdots, q(i), \cdots, q(N)$ $(1 \leq q(i) \leq M)$ is a sequence of indices of the templates, and $\oplus$ means a binary operator connecting the two reference templates. The One Stage DP algorithm [9] can be applied to solve the optimum time warping paradigm.

Using the One Stage DP algorithm enables us to find the optimum segmentation in the meaning of the least squared error (LSE), but the result sometimes show inappropriate segmentation. To reduce the segmentation errors, transition probabilities between accent patterns can be incorporated into the One Stage DP algorithm by the following way.

Let $\mathcal{S} = S_1 \oplus S_2 \oplus \cdots \oplus S_N$ be a candidate sequence of pitch pattern segments of input pitch pattern, then the reference template which matches the S_n best is defined by

$$R_{q(n)} = \arg \min_{R_k \in R} [D(S_n, R_k)] . \tag{7}$$

where $D(S_n, R_k)$ is the DTW distance between S_n and R_k. Denoting $D_n = D(S_n, R_{q(n)})$, the overall distance between the input pitch pattern and the super reference pattern $\mathcal{R}$ is given by

$$D(\mathcal{S}, \mathcal{R}) = \sum_{n=1}^{N} D_n. \tag{8}$$

The optimum sequence of segments is now given by

$$\mathcal{S}^* = \arg \min_{\mathcal{S}} [D(\mathcal{S}, \mathcal{R})] . \tag{9}$$

To take account of the transition probabilities into the above definition, the Eq. (8) can be modified as follows

$$D(\mathcal{S}, \mathcal{R}) = \sum_{n=1}^{N} W_{q(n-1),q(n)} D_n \tag{10}$$

where $W_{i,j}$ is an weighting factor giving a penalty for the transition from template R_i to R_j, and which is defined by

$$W_{i,j} = \left(\frac{1}{(1-\gamma) Pr(j|i) + \gamma} \right)^{\beta}, \qquad (\gamma = 10^{-6}) \tag{11}$$

where $Pr(j|i)$ is the conditional transition probability from template R_i to R_j, and β is a kind of weighting factor of the transition probability to the distance of Eq. (10).

4 Experiments

4.1 Experimental conditions

The speech database used in the experiments is the ATR continuous speech database of phoneme balanced 503 Japanese sentences uttered by a single male speaker 'MYI'. The speed of utterance in the database is about 8.7 mora/s, and the frame length of accent phrases is 543 ms in average. This database provides labeling information of phonemes and prosodic phrase structures.

F_0 is calculated every 10ms from input speech signals sampled at 12 kHz sampling rate.

Pitch patterns from the sentences no.101 $\sim$ 503 were used to make the pitch pattern templates, and the remains no.1 $\sim$ 100 were used to test the performance of our system. In case of making the reference patterns, the parameter λ in Eq. (5) was set to 0.5.

In the One-Stage DP operation, slope for searching the best path was restricted in the range $1/2 \sim 2$. In case of defining accuracies of detecting phrase boundaries, detected boundaries located within 100 ms from the hand labeled boundaries were treated as correct.

4.2 Results

Fig. 3 shows an example of the pitch pattern templates obtained by the clustering algorithm when the number of clusters is eight. The frame length of each templates represents the average frame length of the accent patterns belonging to the cluster.

An example of detecting phrase boundaries from speech signals is shown in Fig. 4, in which (a) shows the wave form of the input signals of a Japanese sentence. The vertical bars represents the boundaries of accent phrases marked by hand in the database. (b) shows the F_0 contours extracted, and (c) is the segmentation result where segmentation boundaries are marked by the vertical bars and patterns between the two vertical bars show corresponding pitch pattern templates determined by the One-Stage DP algorithm. No transition probabilities were used in this case. We can see some boundaries are marked correct, but some are not. (d) shows the segmentation result when the transition probabilities of pitch-pattern templates were used. Most of the miss segmentations in (c) are now corrected and correspondence of pitch pattern templates to F_0 seems reasonable.

Performance of detecting accent phrase boundaries is shown in Fig. 5. The horizontal axis of the figure displays the value of β in Eq(11). $\beta = 0.0$ means no transitions probabilities are used. It can be seen from the figure that the improvement of the performance by the use of the transition probabilities is not so high than expected. But if we could chose the optimal parameters such as β and the number of templates, the performance could become better.

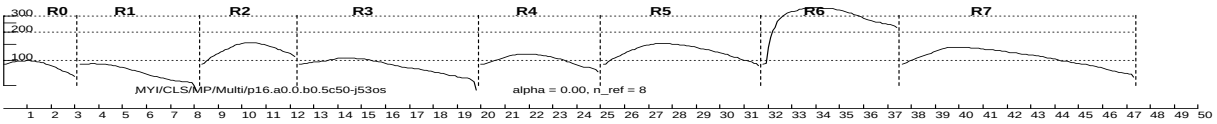

Figure 3: An example of the pitch pattern templates (n=8)

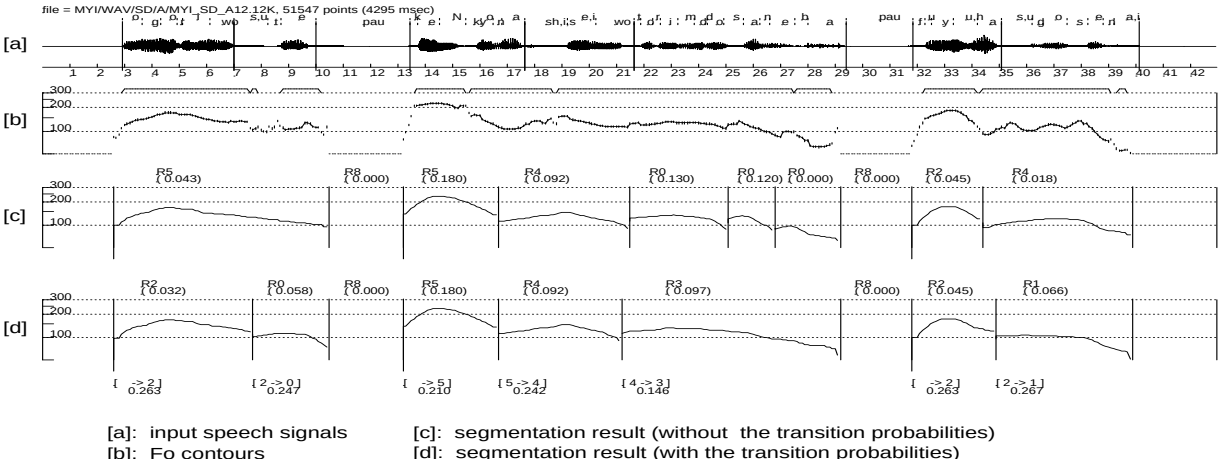

Figure 4: An example of pitch pattern segmentation

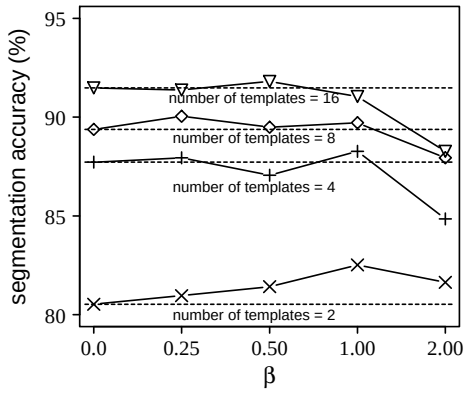

Figure 5: Segmentation Accuracy

5 Conclusions

We have developed an accent phrase segmentation system using accent phrase clustering for training and the One Stage DP algorithm for automatic segmentation. Experimental results using the ATR continuous speech database of one male speaker showed that more than 91 % of phrase boundaries were correctly detected. Incorporating connectivity between the accent pattern templates into the DP operation showed a slight effect of improving the segmentation accuracy. But we got some results encouraging us that the restriction by the connectivity on DP operation could improved the performance if we could had enough data to train the system. Although the results reported here is only for one speaker, we are now testing the system for multiple-speaker environment and have got good results, which will be reported in the next time.

REFERENCES

[1] W. A. Lea, M. F. Medress and T. E. Skinner: "A Prosodically Guided Speech Understanding Strategy", IEEE ASSP-23,1, pp.30–37 (1975-02)

[2] Y. Kitahara, S. Takeda, A. Ichikawa and Y. Tohkura, "Role of Prosody in Cognitive Process of Spoken Language", Trans. IEICE, J70-D, 11, pp.2095–2101 (1987-11) (in Japanese)

[3] A. Komatsu, E. Oohira and A. Ichikawa: "Conversational Speech Understanding Based on Sentence Structure Inference Using Prosodics, and Word Spotting", Trans. IEICE, J71-D,7, pp.1218–1228 (1988-07) (in Japanese)

[4] Y. Suzuki, Y. Sekiguchi and M. Shigenaga: "Detection of Phrase Boundaries Using Prosodics for Continuous Speech Recognition", Trans. IEICE, J72-D,10, pp.1609–1617 (1989-10) (in Japanese)

[5] C. W. Wightman and MÖstendorf: "Automatic Recognition of Prosodic Phrases", ICAASP-91, pp.321–324 (1991)

[6] H. Shimodaira and M. Kimura: "Accent Phrase Segmentation Using Pitch Pattern Clustering", ICASSP-92, I-217 (1992)

[7] K. Shirai, S. Okawa and T. Kobayashi: "Phoneme Recognition in Continuous Speech Based on Mutual Information Considering Phonemic Duration and Connectivity", ICSLP-92, pp.1479–1482 (1992-10)

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

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