ACCENT PHRASE SEGMENTATION BY FINDING N-BEST SEQUENCES OF PITCH PATTERN TEMPLATES

Mitsuru Nakai[†] and Hiroshi Shimodaira[‡]

† Dept. of Information Eng. Faculty of Eng., Tohoku University,
Sendai-shi, 980 Japan
‡ School of Information Science, Japan Advanced Institute of Science and Technology,
Tatsunokuchi, Ishikawa, 923–12 Japan

ABSTRACT

This paper describes a prosodic method for segmenting continuous speech into accent phrases. Optimum sequences are obtained on the basis of least squared error criterion by using dynamic time warping between F_0 contours of input speech and reference accent patterns called 'pitch pattern templates'. But the optimum sequence does not always give good agreement with phrase boundaries labeled by hand, while the second or the third optimum candidate sequence does well. Therefore, we expand our system to be able to find out multiple candidates by using N-best algorithm.

Evaluation tests were carried out using the ATR continuous speech database of 10 speakers. The results showed about 97% of phrase boundaries were correctly detected when we took 30-best candidates, and this accuracy is 7.5% higher than the conventional method without using N-best search algorithm.

Figure 1: Block diagram of the system

1 INTRODUCTION

Recently, the use of prosodic features of spoken language, such as accents, intonations and pauses, take on more importance, as recognition tasks become difficult. On the study of prosody, various segmentation methods for prosodic phrase boundaries have been proposed, and it has been already reported that phrase boundaries provide useful information for speech understandings, for example, estimating structure of sentence or spotting important words[1].

Among these studies, an accent phrase which has a single accented syllable is usually used as a segmentation unit, because accent phrases on prosodic structure have similar aspects to syntactic phrases and we can reduce recognition tasks by replacing continuous speech processings with discrete word-like speech processings. It is well known that these boundaries are usually observed as a ravine on the F_0 contour. There are many studies which attach importance to characterize the local feature of F_0 contour[2].
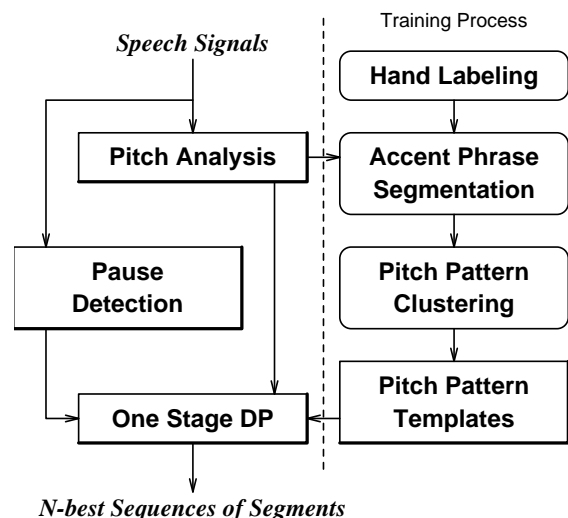
On the other hand, we regard it as important to find out an optimum sequence of accent phrases over a whole input sentence under the criterion of the least squared error[3][4], and it is not necessary to extract the local feature of F_0. Fig.1 shows a block diagram of our system. In a training process, 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 optimum sequence of segments is found out by the One-Stage DP matching between F_0 contours of input speech signals and the pitch pattern templates, and they are expected to be corresponding to real accent phases. In the One-Stage DP processing, we can obtain multiple sequences of segments by using the N-best algorithm[5], and among these candidates we can expect to find a sequence which gives best agreement with real accent phrases.

2 TRAINING PHASE

2.1 Clustering pitch templates

An accent phrase, which is sometimes called a 'prosodic phrase', can be usually observed as a *fall-rise pattern* of F_0, and the F_0 contour of a sentence uttered continuously is considered as a connected series of these prosodic patterns. Based on this idea, we extract phrase boundaries as a result of recognizing *fall-rize patterns* of the F_0 contour. Because there are so many types of accent patterns that we cannot represent their contours precisely by our knowledge of accent phrase, we use a statistical clustering algorithm to learn some typical accent patterns.

The LBG VQ algorithm has been employed for this clustering and distance measure between any two accent patterns is defined as follows. In order to avoid difficulty of comparing two patterns with different frame length, the operation is divided into two types; one is comparison of the shapes of pattern, and the other is comparison of the frame lengths.

Here, we have a set of training accent patterns,

$$P_j = (p_{j0}, \cdots, p_{ji}, \cdots, p_{jL_j})$$

where p_{ji} is a logarithm F_0 value of frame i at the j-th accent phrase. For comparison of the shape of patterns, to define the distance based on a squared error criterion, each training patterns are transformed linearly into fixed frame length pattern:

$$\hat{P}_j = (\hat{p}_{j0}, \cdots, \hat{p}_{ji}, \cdots, \hat{p}_{jL}).$$

Then the distance between a pair of patterns, $\hat{P}_j$ and $\hat{P}_k$ can be defined by

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

where a is a offset value about $\hat{p}_{j1} - \hat{p}_{k1}$. Therefore, if $\hat{p}_{ji} = \hat{p}_{ki} + a$ for all i ($1 \leq i \leq L$), we consider $\hat{P}_j$ and $\hat{P}_k$ have a same shape. For comparison of the frame lengths, we define the distance by

$$D_L(\hat{P}_j, \hat{P}_k) = (L_j - L_k)^2.$$

Using these two types of distance measures, we define the distance between two patterns by

$$D_\lambda(\hat{P}_j, \hat{P}_k) = (1 - \lambda)D_S(\hat{P}_j, \hat{P}_k) + \lambda\gamma D_L(\hat{P}_j, \hat{P}_k)$$

where λ is a weighting factor for D_L, and γ is a normalized coefficient given by

$$\gamma = \frac{\sum_{\hat{P}_n \in \hat{P}} D_S(\hat{P}_n, \bar{P})}{\sum_{\hat{P}_n \in \hat{P}} D_L(\hat{P}_n, \bar{P})},$$

and $\bar{P}$ is an average pattern of $\hat{P}$.

After the clustering operation, we obtain a set of pitch templates $R = \{R_1, R_2, \cdots, R_M\}$ by transforming each centroid of the VQ linearly to the average frame length of patterns in each cluster.
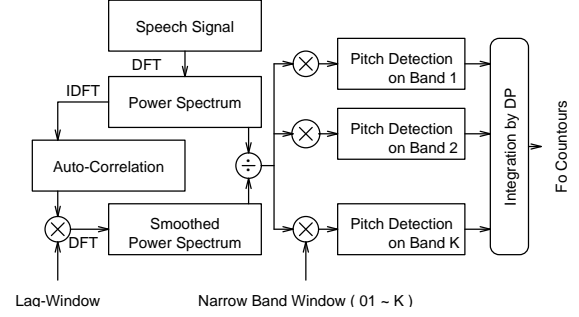

Figure 2: multiple-band F_0 analysis

2.2 Pitch detection

In our system, continuity of F_0 contours is a very important factor for the segmentation accuracy, because we have defined the optimum sequence of pitch templates based on the least squared error (LSE) criterion, and discontinuity of F_0 contour increase an error between a desired sequence of pitch template and the pitch pattern. Therefore, we use the multiple-band F_0 analysis[6] to estimate F_0 contours as continuous as possible. Fig.2 shows the block diagram of multiple-band analysis. The basic idea of this algorithm is to extract multiple pitch candidates from each frequency band and to choose a reliable pitch among the candidates. On the assumption that F_0 value changes slowly and continuously, we integrate pitch candidates into F_0 contour by using a DP algorithm.

3 SEGMENTATION PHASE

From the training operation, we obtain a set of reference patterns $R = \{R_1, R_2, \cdots, R_M\}$. Then segmentation of the pitch pattern into accent phrases is regarded as a problem of finding out the optimum sequence of pitch pattern templates ($\mathcal{S}^*$), which minimizes the accumulative distance with the input pitch pattern ($\mathcal{P}$):

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} D(\mathcal{P}, \mathcal{S})$$

where $\mathcal{S} = R_{q(1)} \cdot R_{q(2)} \cdot \ldots \cdot R_{q(N)}$ and $q(1), \cdots, q(i), \cdots, q(N)$ ($1 \leq q(i) \leq M$) is a sequence of indices of the templates. The One Stage DP algorithm[7] can be applied to solve the optimum time warping paradigm.

Using the One Stage DP algorithm enable us to find the optimum sequence in the meaning of the LSE, but it does not always give desirable phrase boundaries. To obtain a reasonable sequence which gives correct boundaries, we search multiple candidates for the sequence of templates by using N-best algorithm.

The basic idea of the N-best algorithm is to keep N candidates at each frame of input speech signals. Fig.3 shows
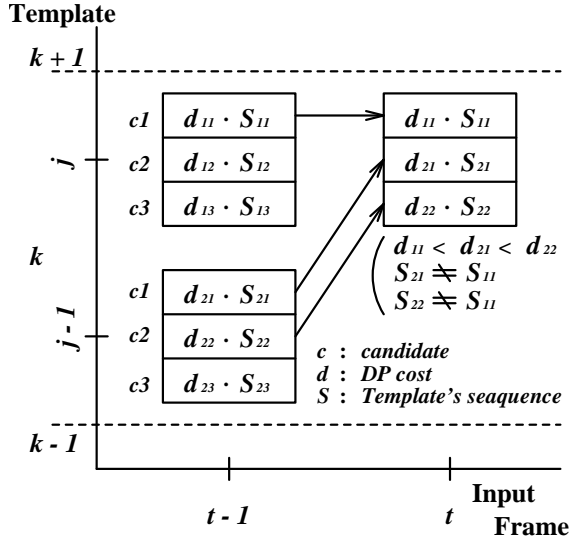
Figure 3: N-best operation

Figure 4: Segmentation accuracy

how to obtain the best three paths arriving at frame t of input speech and frame j of k-th template. In the case that there are two directions to arrive at frame t, we have six paths in total, and choose three paths in order of DP cost d. If we save three cost d_{11}, d_{21}, d_{22} ($d_{11}, \le d_{21}, \le d_{22}$), three sequences of S_{11}, S_{21}, S_{22} must be different each other. In this way, the N candidates of templates' sequence are time-synchronously found from the pitch pattern of input speech signals.

4 EXPERIMENTS

4.1 Condition

The speech database used in this evaluation test is the ATR continuous speech database of phoneme balanced 503 Japanese sentences uttered by 10 speakers. We divided the 10 speakers into three groups, G1, G2 and G3. Speech data from three male speakers (G1) are used for training, and the other three male speakers (G2) and four female speakers (G3) are used for segmentation tests. The detail is shown as Table 1.

In making the reference patterns, the LBG clustering

Table 1: ATR continuous speech database

Group	G1(male):	'mho', 'mht', 'mmy'
	G2(male):	'msh', 'mtk', 'myi'
	G3(female):	'fkn', 'fks', 'ftk', 'fym'
Text	Training:	No.051 ~ 503
	Test:	No.001 ~ 050

was driven at $M = 8$ and the parameter λ 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 phrase segmentations, detected boundaries located within 100 ms from the hand labeled boundaries are treated as correct. So, correct rate and insertion rate of the n-th candidate is defined by

$$R_c^N(n) = \frac{\text{\# correct detected boundaries in the } n\text{-th candidate}}{\text{\# hand labeled boundaries}}$$

$$R_i^N(n) = \frac{\text{\# incorrect detected boundaries in the } n\text{-th candidate}}{\text{\# hand labeled boundaries}}$$

where N is a number of candidates to keep in N-best operation. We search an n^*-th candidate which gives maximum $R_c^N(n^*)$ among N candidates, and segmentation accuracy in the case of size N is defined by

$$\bar{R}_c^N = R_c^N(n^*), \qquad \bar{R}_i^N = R_i^N(n^*).$$

4.2 Results

Fig.5 shows an example of detecting phrase boundaries from a speech signal in case of the candidate size $N = 10$, the reference template size $M = 4$ (actually, we use $M = 8$ in this experiments). [a] shows a wave form of input Japanese speech, and the vertical bars are boundaries of accent phrase labeled by hand in this database. [b] shows the F_0 contours extracted by our multiple-band F_0 analysis. [d] shows four pitch pattern templates. [c] shows 10-best sequences of pitch templates in order of DP distance.
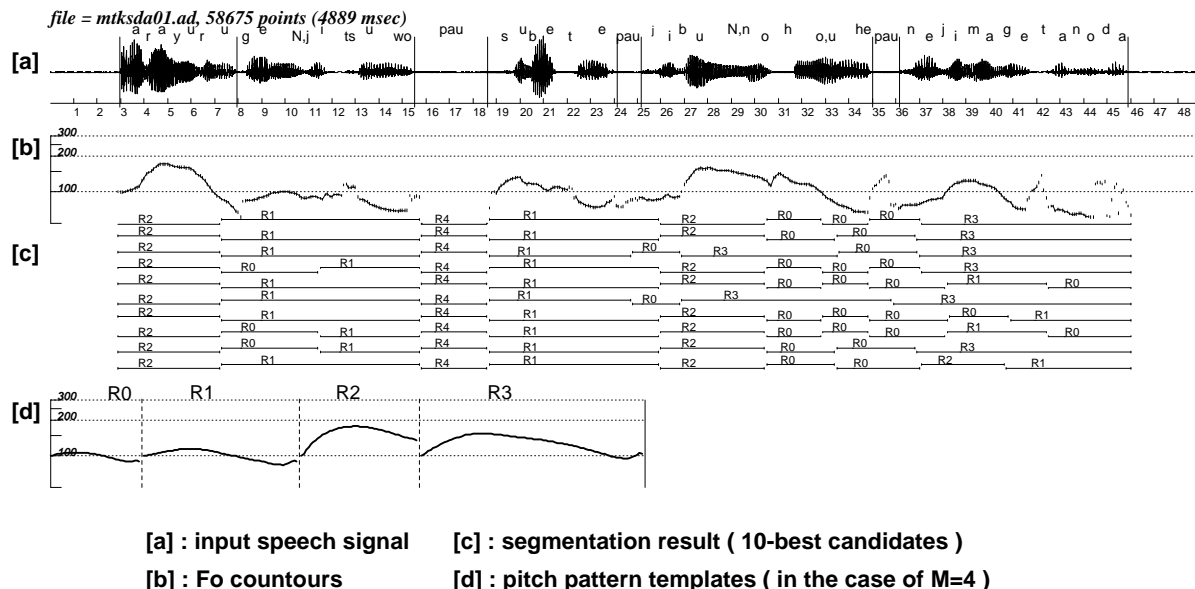
[a] : input speech signal   **[c]** : segmentation result (10-best candidates)
[b] : Fo countours   **[d]** : pitch pattern templates (in the case of M=4)

Figure 5: An example of pitch pattern segmentation

The words which begin with 'R' are identical with pitch templates shown in [d] and horizontal bars indicate accent phrase regions, but only 'R4' denotes a pause region detected directly from input speech signals. Among the 10-best sequences, we can see that the third candidate gives better segmentation accuracy than the first candidate.

Fig.4 shows the segmentation accuracy in the case of $M = 8$. Although our pitch templates were made from the three male speakers of G1, results for the open test groups, G2(male) and G3(female), have a similar tendency in segmentation performance. As a number of candidates N increases from 1 to 30, the correct rate $\bar{R}_c^N$ increases by about 7.5%, and the insertion rate $\bar{R}_i^N$ decreases by about 15%. In the case of $N = 30$, maximum correct rate is about 97%, which seemed to be a superior limit of our segmentation system, because we think that the remained boundaries labeled in the ATR database seemed to be hard to detect.

5   CONCLUSION

In this paper, we have presented a multiple-candidates search method for detecting prosodic phrase boundaries by using the N-best algorithm. Experimental results using the ATR continuous speech database have shown that this system performed well and about 97% of prosodic boundaries have been correctly detected. For further studies, we should exploit the way to incorporate these results into speech understandings, such as estimating a structure of sentences or spotting important words.

References

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

[2] 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).

[3] H.Shimodaira and M.Kimura: "Accent Phrase Segmentation Using Pitch Pattern Clustering", *ICASSP-92*, pp.I-217–220, (1992-03).

[4] M.Nakai, H.Shimodaira and S.Sagayama: "Prosodic Phrase Segmentation Based on Pitch-Pattern Clustering", *Trans. IEICE*, J77-A, 2, pp.206–214 (1994-02) (in Japanese).

[5] R.Schwartz, and Y.-L. Chow: "The N-best algorithms: an efficient and exact procedure for finding the N most likely sentence hypotheses", *ICASSP-90*, pp.81–84, (1990).

[6] H.Shimodaira and M.Nakai: "Robust Pitch Detection by Narrow Band Spectrum Analysis", *ICSLP-92*, pp.1597–1600, (1992-10)

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