PROSODIC PHRASE SEGMENTATION BY PITCH PATTERN CLUSTERING *

Hiroshi SHIMODAIRA *Mitsuru NAKAI* †

Japan Advanced Institute of Science and Technology (JAIST),
Tatsunokuchi, Nomi, 923-12 Japan
E-mail: sim@jaist.ac.jp
† Tohoku University, Sendai, 980 Japan

ABSTRACT

This paper proposes a novel method for detecting the optimal sequence of prosodic phrases from continuous speech based on data-driven approach. The pitch pattern of input speech is divided into prosodic segments which minimized the overall distortion with pitch pattern templates of accent phrases by using the One Pass search algorithm. The pitch pattern templates are designed by clustering a large number of training samples of accent phrases. On the ATR continuous speech database uttered by 10 speakers, the rate of correct segmentation was 91.7 % maximum for the same sex data of training and testing, 88.6 % for the opposite sex.

1 INTRODUCTION

It is well-known that humans recognize speech by integrating many information not only short-time spectral information of speech and language knowledge but also prosodic information such as accent, intonation, stress and pause. A number of studies have proposed that prosody has a great impact on intelligibility and naturalness of human speech communication, while quite a few speech understanding system have incorporated prosodic information into sentence processing and language understanding [1, 2]. The reason for it is that prosodic features are strongly influenced by speakers, sentence types etc. and it is difficult to use them for speech understanding.

The goal of our research is to develop a practical way of incorporating prosodic information into speech understanding. As the first step, our current interest is to detect prosodic segments by means of F_0 contours. It is noticed that the syntactic structure of spoken language has good relationship with accent phrases, each of which has a single accent cue, and most of the boundaries of the accent phrases can be found around the '*fail-rise patterns*' of F_0 contours (pitch patterns). For this reason, accent phrase is used as the unit of prosodic segment in this study.

Some approaches for prosodic segmentation have been proposed. One of them is based on finding out dips among fall-rise patterns of Fo contours by using piece-wise line fitting [1, 3], and the other is to use phoneme durations given by a speech recognizer [4]. Another interesting approach is a model-based one in which prosodic events are estimated by using the 'Fujisaki-model' [5]. On the other hand, our approach is based on the ideas that

1. Behavior of pitch patterns of accent phrases is not chaos, but it forms some classes. Each class can be represented by some typical pitch patterns called 'pitch pattern templates'.

2. A pitch pattern of breath group is expressed by the concatenation of the templates.

We realize these ideas by classifying a large number of training pitch patterns of accent phrases using clustering method, and also segmenting input pitch pattern by finding out the optimal sequence of pitch pattern templates using dynamic programming [6].

It should be noted that our approach is a kind of bottom-up approach or what is called 'data-driven' approach. Perfect prosodic segmentation, of course, is not accomplished by this approach, but it is important to show how one can go without any model such as 'Fujisaki model'. In respect to the model-based approach, the capacity of it highly depends on the capacity of the prosodic model. Unfortunately, prosodic models available today do not seem to have enough power to handle variability of real speech data.

2 PROSODIC PHRASE SEGMENTATION

As is shown in Fig. 1, segmentation of input speech into prosodic segments is carried out by finding out

*This work is cooperated with Shigeki Sagayama of NTT Human Interface Research Labs.

0-7803-1775-0/94 $3.00

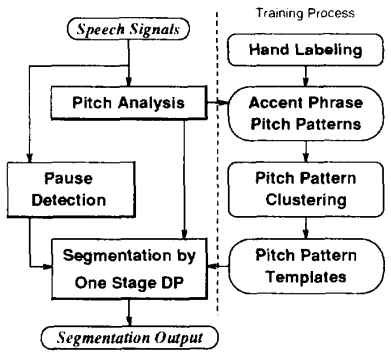

Figure 1: Prosodic phrase segmentation system

the optimum segmentation between the pitch pattern of input speech and the pitch pattern templates.

2.1 Pitch Contour Estimation

Pitch determination algorithm used in our system is the multiple-band lag-window method [7]. Difference between the original lag-window method and ours is that the former extracts a single F_0 for each frame from the power spectrum of entire short-time frequency domain, while the latter divides the frequency domain into multiple sub-bands and extracts multiple F_0 candidates from them and unify the F_0 candidates into a single one. The unification operation of the F_0 candidates for each frame and for each sub-band is done by a kind of DP-based pitch contour smoothing operation. Dividing frequency domain into multiple sub-bands enables accurate and robust pitch analysis.

2.2 Clustering

From the linguistic point of view, accent patterns of accent phrases are classified into a small number of accent types according to the location of the accented syllable ('accent cue') in the phrase. But the pitch pattern observed from the speech is not modeled by a simple concatenation of accent patterns but by an unknown complex dynamical system in which phrase patterns and context effects as well as accent patterns are to be taken into account at the same time. This is another reason why we use the clustering approach instead of using accent types as a priori knowledge. Therefore we do not expect to obtain some meaningful relationships between the accent types and the pitch pattern templates given by the clustering.

For the algorithm of clustering pitch patterns of accent phrases, the LBG vector quantization algorithm which approximately minimizes quantizing distortion is used to keep consistency of pattern comparison criterion with the algorithm of the prosodic segmentation described in the next section.

Before applying the VQ algorithm to the accent pattern clustering, distance measure between any two accent patterns must be defined. 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 length.

For shape comparison the distance between any two pitch patterns P_j and P_k can be defined by the expression:

$$D_{SA}(P_j, P_k) = \| \hat{P}_j - \hat{P}_k \|^2 = \sum_{i=1}^{L} |\hat{p}_{ji} - \hat{p}_{ki}|^2 \qquad (1)$$

where $\hat{P}_j$ and $\hat{P}_k$ are the pitch patterns on log scale normalized in frame lengths to L.

Considering the case that $\hat{P}_k$ is similar to $\hat{P}_j$ except that $\hat{p}_{ji} = \hat{p}_{ki} + const$ for all $k = 1, \cdots, L$, in that case those two patterns might fail to be classified into the same cluster. One approach to overcome the obstacle is to use 'delta pitch' [6], which is the first-order differential regression of pitch patterns, in conjunction with the distance between pitch patterns. But the approach taken this time is to shift the height of $\hat{P}_j$ and $\hat{P}_k$ to make the both patterns have the same value at their beginning frame. Denoting the shifted patterns by $\tilde{P}_j = (\tilde{p}_{ji})$ and $\tilde{P}_k = (\tilde{p}_{ki})$, the distance between the shifted pattern $\tilde{P}_j$ and $\tilde{P}_k$ becomes

$$D_{SR}(P_j, P_k) = \| \tilde{P}_j - \tilde{P}_k \|^2 = \sum_{i=1}^{L} |\tilde{p}_{ji} - \tilde{p}_{ki}|^2. \qquad (2)$$

For the comparison of the frame lengths, distance can be simply defined by

$$D_L(P_j, P_k) = (L_j - L_k)^2 \qquad (3)$$

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

Using the two types of distance measure above, the distance between the two patterns is now defined as a linear combination of the two types of distance

$$D_\beta(P_j, P_k) = (1 - \beta) D_S(P_j, P_k) + \beta\gamma D_L(P_j, P_k) \quad (4)$$

where D_S represents either of D_{SA} or D_{SR}, β is a weighting factor for D_L and γ is a normalizing factor.

Once defining the above distance, 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 stage, the frame length of each reference pattern R_m is set to the average frame length in the m-th cluster.

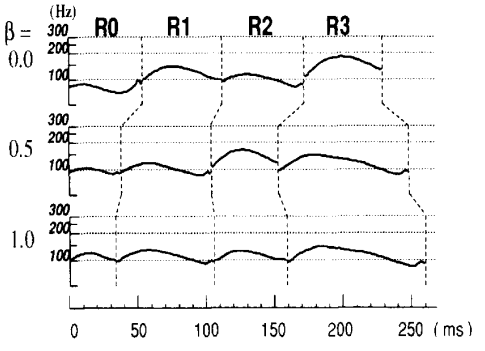

Figure 2: An example of pitch pattern templates

An example of clustering accent pitch patterns is shown in Fig. 2. We can see the templates of $\beta = 0.0$ have almost the same frame length but the shape is quite different from one another. On the other hand, the templates of $\beta = 1.0$ have different frame lengths while the shape is very similar.

2.3 Prosodic Segmentation

As is described in the introduction, we assume that the pitch pattern of the breath group is the concatenation of the pitch pattern templates, prosodic segmentation is defined as a problem of finding out the optimal sequence of pitch pattern templates ($\mathcal{R}^*$), which minimizes the accumulative distance with the input pitch pattern ($\mathcal{S}$):

$$\mathcal{R}^* = \arg\min_{R} D(\mathcal{S}, \mathcal{R}) \tag{5}$$

where $\mathcal{R} = R_{q(1)} \oplus R_{q(2)} \oplus \ldots \oplus R_{q(N)}$ and $q(1), \cdots, q(i), \cdots, q(N)$ $(1 \leq q(i) \leq M)$ is a sequence of indices of the templates, and $\oplus$ means a binary operator connecting the two adjacent reference templates.

The One Pass DP algorithm can be applied to solve the optimum time warping paradigm and as a result of that the optimum sequence of segmentation is obtained.

3 EXPERIMENTS

3.1 Experimental conditions

The speech database used is the ATR continuous speech database of phoneme balanced 503 Japanese sentences uttered by 10 speakers. The speakers are divided into three distinct speaker-groups:

Group	# speakers (speaker ID)	used for (# sentences)
G1 (male)	3 (mho, mht, mmy)	training (453)
G2 (male)	3 (msh, mtk, myi)	testing (50)
G3 (female)	4 (fkn, fks, fym, ftk)	testing (50)

As is shown in the table, data from the group G1 are used for training the system, or making the pitch pattern templates, and the data from G2 and G3 are used to evaluate the performance of the system. It should be noted that in order to make the evaluation test open for speakers and texts we have made the training sentences distinctly different from the test sentences.

Each sentence in the database is hand-labeled with the phonetic transcription and prosodic phrase structures by the labellers of professional skill. F_0 is calculated every 10 ms from input speech signals sampled at 12 kHz sampling rate. In the One Pass DP operation, slope for searching the best path was restricted to the range $1/2 \sim 2$. In case of defining the rate of correct segmentation, which means the rate of correct detection of phrase boundaries, detected boundaries located within 100 ms from the hand labeled boundaries were treated as correct.

3.2 Results

An example of the segmentation is shown in Fig. 3. The vertical bars in the speech waveform in the figure show the correct boundaries of accent phrase labeled in the database.

Segmentation performance is shown in Fig. 4 $\sim$ Fig. 6. In these figures, 'type-A' denotes the use of distance measure D_{SA} of (1), 'type-R' denotes the use of distance measure D_{SR} of (2). It can be seen from the figures that the rate of correct segmentation increases as the number of template increases except for the case of 'type-A' with $\beta = 0.0$. Comparing 'type-A' and 'type-R' distance measure, we may see that type-R shows better rates of correct segmentation than that of type-A especially for the test environment of opposite sex (Fig. 5), while insertion rate of type-R is much worse than that of type-A (Fig. 6). For the effect of the weighting factor β of (4), $\beta \approx 0.5$ gives reasonable performance.

4 CONCLUSIONS

We have proposed the data-driven approach for prosodic phrase segmentation, in which the optimum segmentation is given by using the One Pass DP search between input pitch pattern and the pitch pattern templates obtained by clustering a large number of training accent pitch patterns. We have used speech database uttered by 10 speakers for the purpose of training and testing and in case of same sex the rate of correct segmentation was 85 $\sim$ 92 % while for the opposite sex it was 80 $\sim$ 89 %.

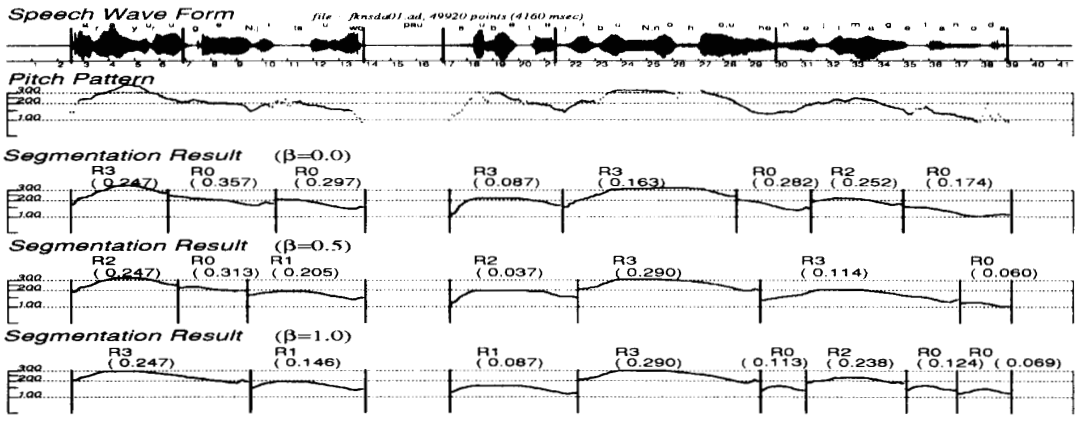

Figure 3: An example of prosodic segmentation

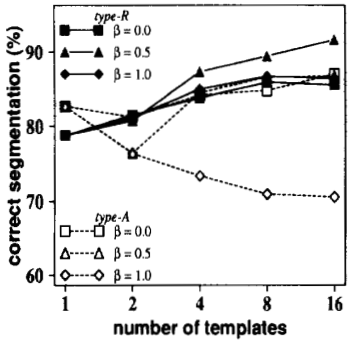

Figure 4: Rate of correct segmentation for G2 (male)

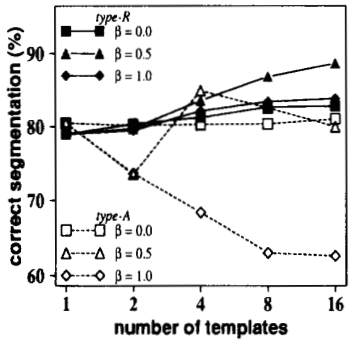

Figure 5: Rate of correct segmentation for G3 (female)

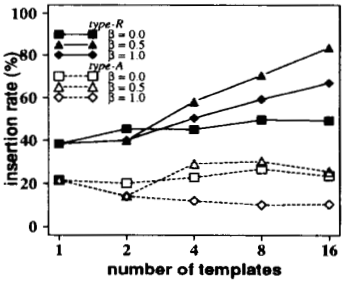

Figure 6: Insertion rate for G3 (female)

REFERENCES

[1] 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)

[2] R. Kompe, A. Kießling, T. Kuhn, M. Mast, H. Niemann, E. Nöth, K. Ott, A. Batliner : "Prosody Takes Over: A Prosodically Guided Dialog System", Proc. of Eurospeech'93, pp.2003–2006 (1993-09)

[3] 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)

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

[5] E. Geoffrois, "A Pitch Contour Analysis Guided by Prosodic Event Detection," Proc. of Eurospeech'93, pp.793–796 (1993-09)

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

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