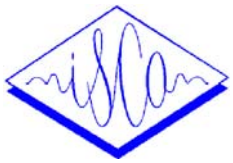

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

PROSODIC WORD BOUNDARY DETECTION USING MORA TRANSITION MODELING OF FUNDAMENTAL FREQUENCY CONTOURS –SPEAKER INDEPENDENT EXPERIMENTS–

Koji Iwano

Department of Information and Communication Engineering
School of Engineering, University of Tokyo
Bunkyo-ku, Tokyo, 113-8656, Japan
iwano@gavo.t.u-tokyo.ac.jp

ABSTRACT

We have been developing a reliable method for prosodic word boundary detection for Japanese continuous speech based on the discrete hidden Markov modeling of fundamental frequency (F_0) contours in mora unit. Although a favorable result was obtained for ATR continuous speech corpus as reported already, experiments were done only on closed conditions. This paper reports the results on open and speaker-independent cases using database by two speakers. On average, detection rate reached around 76.0% with insertion error rate of 18.4%. Degradation from the closed condition experiment was only a little, showing the validity of the method for open conditions.

10.21437/Eurospeech.1999-61

1. INTRODUCTION

Although a rather large number of methods have been developed for the use of prosodic features in automatic speech recognition with some favorable results, the usage is limited to small parts of recognition process. A more positive use is necessary for the further advancement of speech recognition.

In view of rather low performances in conventional methods for detecting prosodic events using prosodic features, we have been questioning: "what is the best unit to statistically model prosodic features?" and "how we can utilize segmental information obtainable through recognition process?" Statistical modeling of fundamental frequency (F_0) contours of mora (a basic unit of Japanese pronunciation mostly coinciding with a syllable) transition was our answer for this question. It is based on segmenting F_0 contours in mora units using mora boundary information obtained from speech recognition process, and then assigning discrete codes to the mora F_0 contours. Inputs to the statistical models are sequences of these codes. This method has several advantages over conventional frame based methods: 1. Since prosodic features are supra-segmental and should be treated in longer units, better performance is obtainable. 2. The method decreases the necessary size of database for model training and may simplify its introduction to the total recognition process.

This modeling was already applied for syntactic boundary detection of Japanese sentences [1] and accent type recognition of Japanese 4-mora words [2]. More recently, we have developed a method to detect prosodic word boundaries and to recognize lexical accent types simultaneously and effectively for continuous speech of Japanese by modeling F_0 contours of prosodic words separately according to their accent types and presence/absence of succeeding pauses [3]. However, in [3], we did not conduct the experiment on speaker-independent conditions. This is because, in ATR speech corpus we used, database with the prosodic labels such as prosodic word boundaries and lexical accent types is available only for one speaker (MYI, 503 sentences).

Therefore, since the database with J-ToBI labels is available for another speaker (MHT, 500 sentences), we used these data after converting their J-ToBI labels into prosodic word boundaries and lexical accent types. In this paper, after an explanation of the prosodic word boundary detection method in [3], the experimental results of speaker-dependent and independent cases on open conditions will be reported.

2. STATISTICAL MODELING OF MORAIC TRANSITION

2.1. Outline

Figure 1 shows the method of syntactic boundary detection based on the mora transition modeling of F_0 contours. For an input speech, the extracted F_0 contour on logarithmic frequency scale is first segmented into moraic units to produce moraic F_0 contours. Information on segmental boundaries is supposed to be given by the preceding process of phoneme recogni-

tion. Each moraic F_0 contour is represented by a combination of two kinds of codes: one for representing contour shape (shape code) and the other for representing whether the average moraic F_0 value is higher or lower as compared to that of the preceding mora (ΔF_0 code). Both codes are selected from 11 candidates, and double code-book technique is used.

This method models prosodic words, which is defined as a word or word chunk corresponding to one accent component of F_0 contour, differently according to their accent types and presence/absence of succeeding pauses. As for the statistical modeling, discrete HMM of HTK software [4] was utilized. An utterance was regarded as a prosodic word sequence and a simple heuristic grammar or bigram was used. Finally, the obtained code sequence is matched against the prosodic word models with Viterbi algorithm. The results is obtained as accent types of constituting prosodic words and prosodic word boundaries.

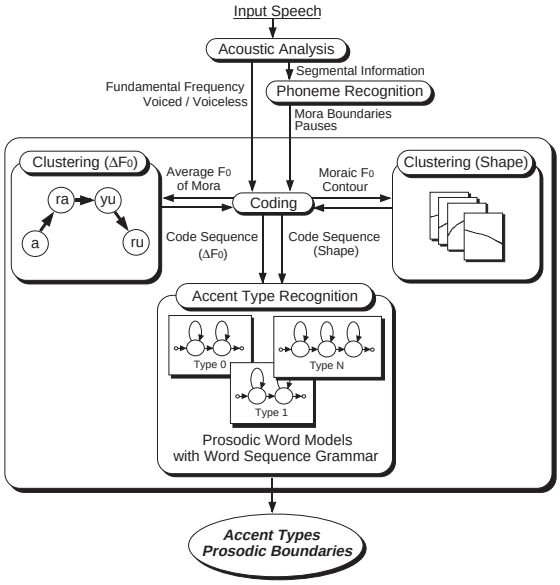

Figure 1. Method of prosodic word boundary detection based on the statistical modeling of F_0 contours of mora transition

2.2. Normalization of Moraic F_0 contours

Each segmented F_0 contour may differ in length and frequency range and should be normalized. Currently, normalization was conducted simply by shifting the average value of a moraic F_0 contour to zero and by linearly warping the contour to a fixed length. Since the derivative of F_0 contour is an important information in characterizing F_0 contour, it was preserved during the warping process by conducting the same warping also along the log-frequency axis.

2.3. Shape Coding

Shape codes were decided by clustering moraic F_0 contours without voiceless part. The clustering scheme was that based on the single linkage method and the leader method. As the result, 9 clusters were obtained and named as codes 3 to 11. Two additional codes 1 and 2 were also prepared respectively for pauses and voiceless mora.

These 11 codes were assigned to moraic F_0 contours of input speech as follows:

(1) A pause period is divided into 100 ms segments (pause mora) from the top of the period and code 1 (pause code) is assigned to each segment. Code 1 is also assigned to the last segment which may be shorter than 100 ms.
(2) Code 2 is assigned to a mora whose voiced portion does not exceed 10% of the whole length of the mora (voiceless mora).
(3) For other mora (voiced mora), one of the codes 3 to 11 is assigned based on the minimum distances between its moraic F_0 contour and the averaged F_0 contours of the clusters. Different from the case of clustering, a moraic F_0 contour may include voiceless regions. Such regions are excluded from the distance calculation.

2.4. ΔF_0 Coding

Clustering was conducted by selecting pairs of voiced mora adjacent to each other. After calculating average F_0 for voiced portion of each voiced mora, differences between the averages of the first to the second mora were calculated for all the pairs. Then, the standard deviation SD of the distance was used for the index of clustering; simply dividing 3SD region centered 0 distance into 9 parts of equal ranges and assigning one of codes 2 to 10 to each part. Codes 1 and 11 were used to represent the distances exceeding the 3SD region.

In order to assign one of these codes to each moraic F_0 contour, we defined average F_0 of a voiceless mora as follows:

(1) For a pause mora, its average F_0 is assumed 0.
(2) For a voiceless mora, its average F_0 is calculated as the interpolation between the average F_0 of its preceding voiced (or pause) mora and that of its succeeding voiced (or pause) mora.

2.5. Prosodic Word Modeling

In Tokyo dialect Japanese, an n-mora word is uttered with one of $n+1$ accent patterns. These accent pat-

terns are denoted as type i ($i = 0 \sim n$) accents and are distinguishable to each other from their high-low combinations of F_0 contours of the consisting mora. Letter "i" indicates the location of dominant downfall in F_0 contour. For instance, type 1 denotes the accent type with an F_0 downfall at the end of the first mora. Type 0 accent shows no apparent downfall in its F_0 contour.

The following 7 HMMs were arranged as prosodic word models.

T0, T0_P : type 0 (or type n) prosodic words
T1, T1_P : type 1 prosodic words
TN, TN_P : types 2 to $n-1$ prosodic words
P pauses

T0, T1, TN are for prosodic words not followed by a pause, while T0_P, T1_P, TN_P are for prosodic words followed by a pause. Model "P" was prepared to absorb pause periods in an utterance, though a pause is actually not a prosodic word. The number of states was 3 for TN and TN_P, 2 for T0, T0_P, T1 and T1_P, and 1 for P. A double code-book scheme was adopted to assign the shape and the ΔF_0 codes to each moraic F_0 contours. In this scheme, for state j the probability $b_j(\boldsymbol{o}_t)$ of generating observation $\boldsymbol{o}_t$ is given by:

$$b_j(\boldsymbol{o}_t) = [P_{js}(o_{st})]^{\gamma_s}[P_{jr}(o_{rt})]^{\gamma_r} \quad (1)$$

where $P_{js}(o_{st})$ is probability of state j generating the shape code o_{st}, and $P_{jr}(o_{rt})$ is probability of state j generating the ΔF_0 code o_{rt}. γ_s and γ_r are stream weights for shape codes and ΔF_0 codes.

2.6. Grammar

As for the grammar of word sequences, a simple heuristic grammar or bigram was used. The heuristic grammar describes the constraint on linking prosodic word to a pause, that is, "X_P must precede P, and the final prosodic word of a sentence must be X_P (X = T0, T1, TN)."

3. EXPERIMENTS

3.1. Training and Testing Data

In order to conduct boundary detection experiments in speaker independent and open data conditions, utterances of two speakers were selected from ATR continuous speech corpus and were divided into training and testing data sets as follows:

T(MYI) : training data of 450 utterances by speaker MYI, including 3,023 prosodic words and 586 pauses.

R(MYI) : testing data of 50 utterances by speaker MYI, including 326 prosodic words and 70 pauses.

T(MHT) : training data of 450 utterances by speaker MHT, including 3,167 prosodic words and 915 pauses.

R(MHT) : testing data of 50 utterances by speaker MHT, including 325 prosodic words and 99 pauses.

Lexical contents of **T** and **R** for speaker MYI are identical with those for speaker MHT. Training data (**T**) were used not only to train prosodic word models, but also to cluster shape and ΔF_0 codes, and to build up prosodic word bigram. Since prosodic labels necessary for the experiments, such as lexical accent types and prosodic word boundaries (accent phrase boundaries), are not included in the data by speaker MHT, they are converted from tone and break indices of J-ToBI labels [5] attached to the data. Strictly speaking, this means the prosodic labels used for the experiments are not assigned based on the same criterion for two speakers, leading to a degradation of the detection performances.

3.2. Features and Distribution of Codes

Table 1 shows features and distribution of shape codes in the training data. "Convex #1" means the peak of convex locates at the former half of the contour, while "Convex #2" means the peak of convex locates at the latter half. Table 2 shows these similarly for ΔF_0 codes.

3.3. Prosodic Word Boundary Detection

Mora boundaries were obtained from phone labels of the corpus and used to segment utterances into mora units. Code weightings γ_s and γ_r were both set to 1.0.

Boundary detection rates R_d and insertion error rates R_i are defined as:

$$R_d = N_{cor}/N_{bou} \quad (2)$$

$$R_i = N_{ins}/N_{bou} \quad (3)$$

where N_{bou}, N_{cor}, N_{ins} respectively indicate the numbers of total prosodic word boundaries, boundaries detected inside the $\pm$100ms region from the correct position and insertion errors.

Table 3 shows these prosodic boundary detection rates for the two types of grammars. Four experiments are conducted on open conditions :

(1a) : T(MYI) for training data and **R(MYI)** for testing data. (speaker-dependent case)

Table 1. Features and distribution of shape codes in the training data.

Code	Feature of the Shape	Number of Mora	
		T(MYI)	**T(MHT)**
1	Pause	2,070	3,974
2	Voiceless	1,814	1,430
3	Flat	2,238	3,694
4	Slightly Rising	1,057	1,338
5	Rising	348	454
6	Sharply Rising	301	404
7	Slightly Falling	4,368	3,804
8	Falling	2,121	2,044
9	Sharply Falling	1,525	787
10	Convex #1	292	255
11	Convex #2	300	154
	Total	16,434	18,338

Table 2. Features and distribution of ΔF_0 codes in the training data.

Code	ΔF_0	Number of Mora	
		T(MYI)	**T(MHT)**
1	Negative(falling)	1,372	1,461
2	!&	80	47
3	!&	390	233
4	!&	1,357	1,360
5	!&	3,817	4,111
6	Zero(no change)	5,578	7,203
7	!&	1,583	1,596
8	!&	764	581
9	!&	235	154
10	!&	90	57
11	Positive(rising)	1,168	1,535
	Total	16,434	18,338

Table 3. Results of prosodic word boundary detection for the open experiments of speaker dependent and independent cases.

Experiment	Constraint		Bigram	
	R_d (%)	R_i (%)	R_d (%)	R_i (%)
(1a)	82.52	26.99	75.15	16.26
(1b)	84.92	25.23	80.31	14.46
(2a)	83.69	31.38	77.85	21.23
(2b)	79.75	26.69	74.23	15.64

(1b) : T(MHT) for training data and **R(MHT)** for testing data. (speaker-dependent case)

(2a) : T(MYI) for training data and **R(MHT)** for testing data. (speaker-independent case)

(2b) : T(MHT) for training data and **R(MYI)** for testing data. (speaker-independent case)

4. DISCUSSION AND CONCLUSION

We conducted open and speaker-independent experiments on prosodic word boundary detection. In the former experiment on closed condition for speaker MYI, boundary detection rate R_d of 77.0% was obtained with insertion error rate R_i of 14.7%. The open condition experiment for speaker MYI this time, showed a performance degradation (sum of detection rate decrease and insertion error rate increase) of around 3.4%. As for the speaker independent cases (experiments **(2a)** and **(2b)**), detection rate reached 76.0% with insertion error rate of 18.4%, on the average. Comparing the results with those of speaker dependent cases (experiments **(1a)** and **(1b)**), average performance degradation was about 4.8%. Taking the fact that the prosodic labeling was done in slightly different criterion for speaker MYI and speaker MHT into consideration, the obtained results are rather favorable.

We also applied the boundary detection method to speech recognition system with unlimited vocabulary. A few percentage improvement was observed in mora recognition rate [6]. In order to improve the results on speaker-independent condition, training data with prosodic labels is necessary for increased number of speakers. From this point of view, we are now planning to construct speech database with prosodic labeling using an unified rule.

Acknowledgement: Special thanks are due to Keikichi Hirose (University of Tokyo) for his valuable advice.

REFERENCES

[1] K. Hirose, K. Iwano, "A Method of Representing Fundamental Frequency Contours of Japanese Using Statistical Models of Moraic Transition," *Proc. EUROSPEECH'97*, Rhodes, Vol.1, pp.311-314 (1997-9).

[2] K. Hirose, K. Iwano, "Accent Type Recognition and Syntactic Boundary Detection of Japanese Using Statistical Modeling of Moraic Transitions of Fundamental Frequency Contours," *Proc. IEEE ICASSP*, Seattle, Vol.1, pp.25-28 (1998-5).

[3] K. Iwano, K. Hirose, "Representing Prosodic Words Using Statistical Models of Moraic Transition of Fundamental Frequency Contours of Japanese," *Proc. ICSLP'98*, Sydney, Vol.3, pp.599-602 (1998-12).

[4] S. Young, J. Jansen, J. Odell, D. Ollason, P. Woodland, The HTK Book, v2.1, Cambridge University (1997).

[5] J. J. Venditti, "Japanese ToBI Labelling Guidelines," Technical report, Ohio-State University, Columbus, U.S.A. (1995).

[6] K. Iwano, K. Hirose, "Prosodic Word Boundary Detection Using Statistical Modeling of Moraic Fundamental Frequency Contours and Its Use for Continuous Speech Recognition," *Proc. IEEE ICASSP*, Arizona, Vol.1, pp.133-136 (1999-3).