PROSODIC WORD BOUNDARY DETECTION USING STATISTICAL MODELING OF MORAIC FUNDAMENTAL FREQUENCY CONTOURS AND ITS USE FOR CONTINUOUS SPEECH RECOGNITION

Koji Iwano and Keikichi Hirose

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 hirose@gavo.t.u-tokyo.ac.jp

ABSTRACT

A new method for prosodic word boundary detection in continuous speech was developed based on the statistical modeling of moraic transitions of fundamental frequency (F_0) contours, formerly proposed by the authors. In the developed method, F_0 contours of prosodic words were modeled separately according to the accent types. An input utterance was matched against the models and was divided into constituent prosodic words. By doing so, prosodic word boundaries can be obtained. The method was first applied to the boundary detection experiments of ATR continuous speech corpus. With mora boundary locations given in the corpus, total detection rate reached 91.5 %. Then the method was integrated into a continuous speech recognition scheme with unlimited vocabulary. A few percentage improvement was observed in mora recognition for the above corpus. Although all the experiments done in closed conditions due to the corpus availability, the results indicated the usefulness of the proposed method.

1. INTRODUCTION

In view of the importance of prosodic features in the human process of speech perception, several works have been conducted aiming at using prosodic features in machine speech recognition process. In the recognition system developed under the Verbmobil project, for instance, prosodic features are used to determine whether the input utterance is declarative or interrogative [1]. Although the reported results were favorable, we should say that the usage is limited only to a small part of the recognition process. More positive use of prosodic features is necessary for future advancements in speech recognition.

Several methods have already been developed to find out syntactic structures of input speech using prosodic structures. However, their performance is rather limited and, accordingly, they are not utilized in current speech recognition systems. This is because most of them attempted to detect prosodic events relating to syntactic boundaries only prior to the main recognition process. From this point of view, we have been developing methods to utilize segmental information also, which is assumed to be obtained through the phoneme recognition process. One such method is the statistical modeling of F_0 contour transitions in mora units [2]. Different form the case of segmental features, modeling in frame units will not give a good result. This is because prosodic features are those of supra-segmentals and should be treated in longer periods. Taking into account that "mora" is the basic unit of Japanese pronunciation (mostly coinciding with a syllable) and that its relative F_0 value is important for accent-type perception, we have developed the moraic transition modeling [2]. Since the modeling is time-aligned to segmental boundaries, it can be rather easily incorporated into phoneme-based speech recognition process.

We already have applied this modeling for syntactic boundary detection and accent type recognition of Japanese [2], [3], [4]. Especially, in [4], we modeled F_0 contours of prosodic words and succeeded to simultaneously detect and recognize prosodic word boundaries and accent types with rather high rates. A prosodic word is defined as a word or a word chunk corresponding to an accent component and can be expressed clearly by the F_0 contour generation model [5]. No results, however, were given there when we applied the developed method to speech recognition. We newly conducted continuous speech recognition experiments by incorporating the method into the recognition process.

In the current paper, after a brief explanation on the prosodic word models, results of boundary detection experiments will be given. Also, we will show that the speech recognition rate can be improved by the method.

2. STATISTICAL MODELING OF F_0 CONTOURS

2.1. Outlines

The developed method models prosodic words, differently according to their accent types and presence/absence of succeeding pauses. The prosodic word models are then matched against input utterances to obtain prosodic word sequences with their accent types. Since an input utterance can be regarded as a sequence of prosodic words, prosodic word boundaries can be detected simultaneously. Each moraic F_0 contour is represented by two codes: one for representing the contour shape (shape code) and the other representing the average F_0 shift from the preceding mora (ΔF_0 code).

Figure 1 shows the process for the prosodic word boundary detection and accent type recognition. For an input speech, its F_0 contour in logarithmic scale is first extracted and then segmented into mora units using the mora boundary information obtained by the phoneme recognition process. A set of shape and ΔF_0 codes is assigned to each moraic F_0 contours to obtain a double code sequence. Finally, this sequence is matched against the prosodic word models and the results is obtained as accent types of constituting prosodic words and prosodic word boundaries.

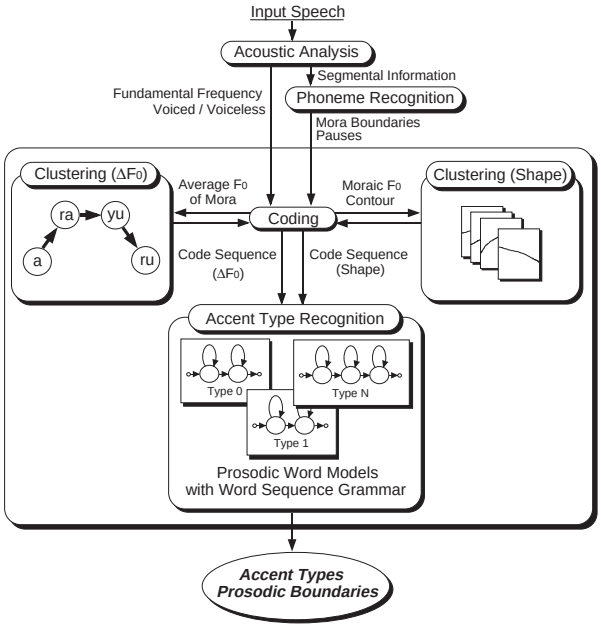

Figure 1: Method of prosodic word boundary detection and accent type recognition.

2.2. Shape Codes

Each segmented moraic F_0 contour may differ in length and frequency range and should be normalized before shape coding. Currently, normalization is 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 an F_0 contour is an important feature characterizing prosodic events, it was preserved during the warping process by conducting the same warping also along the log-frequency axis.

Shape codes were decided by clustering 983 moraic F_0 contours without voiceless part, selected from 85 sentence utterances by a male announcer (speaker MYI) on task SD (a pile of sentences without definite context to each other) included in the ATR continuous speech corpus. As the result, 9 clusters were obtained and named as codes 3 to 11 [4]. 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 was divided into 100 ms segments from the beginning. They were named as pause morae and code 1 (pause code) was assigned. Code 1 was assigned to the last segment, which would be shorter than 100 ms.
2. A mora whose voiced portion does not exceed 10 % of the whole length was called as voiceless mora and code 2 was assigned.
3. For other mora (voiced mora), one of the codes 3 to 11 was assigned based on the minimum distances between its moraic F_0 contour and the averaged F_0 contour of each cluster. Voiceless regions included in moraic F_0 contours were excluded from the distance calculation.

2.3. ΔF_0 Codes

Clustering was conducted by selecting two consecutive moraic F_0 contours from the same corpus as used in the shape code clustering. Only pairs of voiced morae were selected, and, consequently, 11,779 pairs were used for the clustering. After calculating average F_0 for voiced portion of each voiced mora, differences between the averages of the first and the second morae were calculated for all the pairs. Then, the standard deviation σ of the distance was used for the index of clustering; simply dividing 3σ region centered 0 distance into 9 parts of equal ranges and assigning one of codes 2 to 10 to each part [4]. Codes 1 and 11 were used to represent the distances exceeding 3σ 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 as 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.

2.4. Prosodic Word Models

In Tokyo dialect of Japanese, an n-mora word is uttered with one of $n+1$ accent types, which are usually 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 morae. Letter "i" indicates the dominant downfall in F_0 contour occurring at the end of ith mora. Type 0 accent shows no apparent downfall. Discrete HMMs with left to right configuration in HTK software (version 2.0) were adopted to model the prosodic words. The training and the recognition were done by EM algorithm and Viterbi algorithms respectively.

The following 7 models were trained using 503 sentence utterances selected from the ATR continuous speech corpus also by speaker MYI and on task SD. Total number of prosodic words is 3,365 with 15,966 (non-pause) morae and 658 pause periods.

T0 and T0_P : for type 0 (or type n) prosodic words,

T1 and T1_P : for type 1 prosodic words,

TN and TN_P : for types 2 to $n-1$ prosodic words,

P : for pauses.

T0, T1 and TN models are for prosodic words not followed by a pause, while T0_P, T1_P and TN_P are for prosodic words followed by a pause. "P model" 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 models, 2 for T0, T0_P, T1 and T1_P models, and 1 for P models. A double code-book scheme was adopted to assign a pair of shape and ΔF_0 codes to each moraic F_0 contours. The probability $b_j(o_t)$ of observation o_t being generated at state j at time t is given by:

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

where $P_{js}(o_{st})$ and $P_{jr}(o_{rt})$ are probabilities of state j generating the shape code o_{st} and the ΔF_0 code o_{rt} respectively. Symbols γ_s and γ_r are stream weights for shape codes and ΔF_0 codes, both of which were set to 1.0 for the current experiments.

2.5. Grammar for prosodic words

As for the grammar of prosodic word sequences, a simple heuristic grammar or bi-gram was used. The heuristic grammar describes the constraint on linking prosodic word to a pause, that is, "An X_P model must precede a P model, and the final prosodic word of a sentence must be modeled by X_P (X=T0, T1 or TN)." Bi-gram was constructed using the same training data for the prosodic word models.

3. DETECTION OF PROSODIC WORD BOUNDARIES

Fifty utterances were selected out of 503 utterances explained in section 2.4 and were used as the testing data for the boundary detection experiment. The total number of prosodic words in the testing data was 326. Although accent type recognition is included in the method, only the results of prosodic word boundary detection will be given here. Refer [4] for the results on accent type recognition. Different from the next section, experiments in this section were conducted using the mora boundary information given in the database. Table 1 shows prosodic word boundary detection rate C_b, non-boundary detection rate C_n and total detection rate C. These rates are defined as:

$$C = \frac{H_b + H_n}{N_b + N_n} \quad (2)$$

$$C_b = \frac{H_b}{N_b} \quad (3)$$

$$C_n = \frac{H_n}{N_n} \quad (4)$$

where N_b, N_n, H_b and H_n respectively denote the numbers of total prosodic word boundaries in the testing data, mora boundaries not prosodic word boundaries, mora boundaries correctly judged as prosodic word boundaries and mora boundaries correctly judged as not prosodic word boundaries. The table also shows the results obtained by the former method [3]. Since the results should be evaluated as a compromise of C_b and C_n, it is not clear which prosodic word grammar will give the better result. However, in both cases, the proposed method gave better results as compared to the former method. In the recognition experiments in the next section, only the bi-gram was used.

Table 1: Result of prosodic word boundary detection

		Detection Rate (%)		
		C	C_b	C_n
Proposed Method	Constraint	89.85	76.99	92.75
	Bigram	91.49	72.70	95.72
Method Formerly Proposed		87.66	72.39	91.09

4. CONTINUOUS SPEECH RECOGNITION

4.1. Outlines

The developed method was integrated with a continuous speech recognition scheme as shown in Figure 2. In order to clarify the effects using prosodically obtainable word boundary information in speech recognition, word dictionary was not used (unlimited-vocabulary). In the system shown in Figure 2 recognition is conducted in two stages. The first stage operates without prosodic information and the resulting information on mora boundary locations is fed to the process of prosodic word boundary detection. In the second step, input speech is first segmented into prosodic words using the prosodic word boundary information thus obtained, and then mora recognition is re-conducted to get the final results. All the recognition process is programmed utilizing HTK (version 2.0) software. Conditions of acoustic analysis are summarized in Table 2.

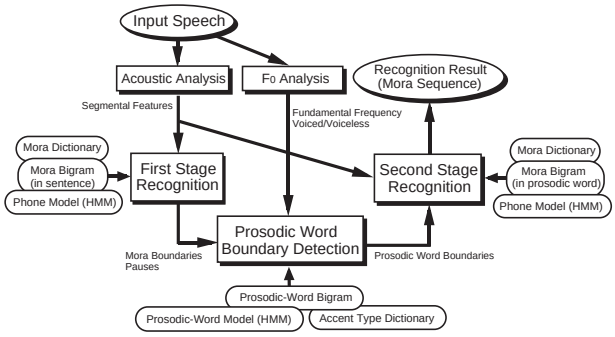

Figure 2: Integrated speech recognition system

The following items were arranged for the both stages:

1. Mora dictionary defining all the possible morae of Japanese, including the pause mora.
2. Phoneme HMMs selected from Japanese tri-phone models trained elsewhere [6].
3. Two types of mora bi-gram: one obtained without taking prosodic word boundaries into account and the other obtained with taking into account. The

Table 2: Conditions of acoustic analysis

Sampling frequency	20 kHz
Analysis window	Hamming window
Window size	25 ms
Frame shift	10 ms
Pre-emphasis coefficient	0.97
Feature vector	12MFCC + 12ΔMFCC + ΔPower
Number of filterbank channels	24

former one was used in the first stage and the latter in the second stage. The bi-gram was constructed by the back-off smoothing technique using the same database used for the prosodic word model training.

4.2. Experimental results

Mora recognition experiments were conducted for the same 50 sentences used in the boundary detection experiments in section 3. These includes a total of 1,541 morae.

Results are shown in Figure 3, where mora recognition rates before and after the second stage C_{bm} and C_{am} are defined as:

$$C_{am}, C_{bm} = \frac{N_{mora} - N_{del} - N_{subst} - N_{ins}}{N_{mora}} \quad (5)$$

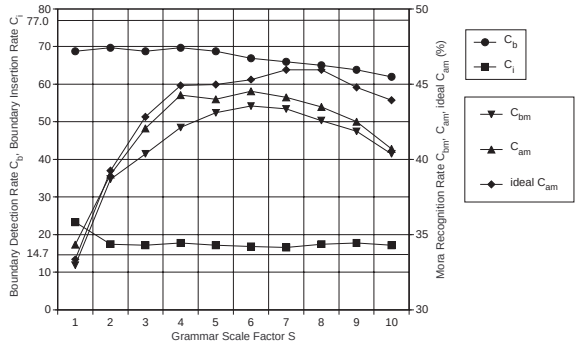

Figure 3: Result of mora recognition

Here, N_{mora}, N_{det}, N_{subst} and N_{ins} respectively represent total number of morae, number of deletions, number of substitutions and number of insertions. Ideal C_{am} denotes the mora recognition rate when the correct prosodic word boundary information is obtainable. Horizontal axis of the figure is the grammar (mora bi-gram) scale factor S which means the log-likelihood being multiplied by S before combining it with acoustic likelihood. Improvements from C_{bm} to C_{am} are observable, indicating the validity of the proposed method in speech recognition. The figure also shows the results of prosodic word boundary detection. Insertion error rate C_i is defined as:

$$C_i = \frac{H_i}{N_b} \quad (6)$$

where H_i indicates number of insertion errors. The horizontal bars in the figure at 77.0 % and 14.7 % respectively show the boundary detection rate and insertion error rate when the correct mora boundary information is obtainable. Different form the case in section 3, boundaries detected inside the ±100 ms region from the correct position were assumed to be correct.

5. CONCLUSION

A method of prosodic word boundary detection was presented, where prosodic word F_0 contours were modeled using the statistical modeling of moraic transitions. The method was integrated into a continuous speech recognition scheme and evaluated from the viewpoint of mora recognition rates. Although favorable results were obtained, the experiments may include a problem on that accent phrase boundaries labeled in the ATR corpus are not strictly coincide with prosodic word boundaries. They are temporarily assumed to be identical in the experiments. From this viewpoint, we are now planning to construct speech database with prosodic word labeling. We are also planning to develop a scheme to decrease the search space in speech recognition, using the method.

6. REFERENCES

[1] C. Lieske, et. al., "Giving prosody a meaning," *Proc. EUROSPEECH'97*, Rhodes, Vol.3, pp.1431-1434 (1997-9).

[2] K. Hirose and K. Iwano, "A method of representing fundamental frequency contours of Japanese using statistical models of moraic transition," *Proc. EUROSPEECH'97*, Rhodes, pp.311-314 (1997-9).

[3] K. Hirose and K. Iwano, "Accent type recognition and syntactic boundary detection of Japanese using statistical modeling of moraic transitions of fundamental frequency contours," *Proc. IEEE ICASSP'98*, Seattle, Vol.1, pp.25-28 (1998-5).

[4] K. Iwano and K. Hirose, "Representing prosodic words using statistical models of moraic transition of fundamental frequency contours," *Proc. ICSLP'98*, Sydney, to be published (1998-12).

[5] H. Fujisaki, K. Hirose and N. Takahashi, "Manifestation of linguistic information in the voice fundamental frequency contours of spoken Japanese," *IEICE Trans. Fundamentals*, Vol.E76-A, No.11, pp.1919-1926 (1993-11).

[6] K. Takeda et. al., "Common platform of Japanese large vocabulary continuous speech recognition research: construction of acoustic model," *Information Processing Society of Japan, Research Reports*, 97-SLP-18-3 (1997-10). (in Japanese)