

DETECTION OF PROSODIC WORD BOUNDARIES BY STATISTICAL MODELING OF MORA TRANSITIONS OF FUNDAMENTAL FREQUENCY CONTOURS AND ITS USE FOR CONTINUOUS SPEECH RECOGNITION

Keikichi Hirose and Koji Iwano***

* Department of Frontier Informatics, School of Frontier Sciences

** Department of Information Engineering, School of Engineering

University of Tokyo

Bunkyo-ku, Tokyo, 113-0033, Japan

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

ABSTRACT

We have been developing a reliable method of prosodic word boundary detection for Japanese continuous speech based on the statistical modeling of mora transitions of fundamental frequency contours of prosodic words. Modifications in the codebook sizes and in the HMM topologies improved the boundary detection performance. When using mora boundary information obtainable from the phoneme recognition process, the detection rates were reached around 73 % with 12.5 % insertion errors for speaker-open experiments. This method was then integrated to a continuous speech recognition system with un-limited vocabulary. The integrated system conducts recognition process in two stages: first stage to detect mora boundaries without prosodic information and second stage to increase mora recognition rate using prosodic word boundary information. Slight improvements in mora recognition rates were observed both in speaker-closed and -open experiments.

1. INTRODUCTION

In view of the importance of prosodic features in the human process of speech perception, a rather large number of works have been conducted to develop methods to utilize prosodic features in machine speech recognition process. However, the usage is limited to a small part of the process, such as to detect interrogation from the sentence final rise of fundamental frequency (F_0) contours, though it is known that prosodic boundary information may facilitate the recognition process and improve the recognition performance. In the works related to the Verbmobil project, prosodic boundaries were used to constrict search space [1], but they were limited to major boundaries, which might be accompanied by pauses. This situation is mainly due to the rather low boundary detection rates for minor boundaries not accompanied by pauses. One major reason for the low detection rates is that the methods attempted to detect prosodic boundaries without taking temporal relation between prosodic features and segmental boundary information well into account. The segmental boundary information was utilized to measure the segmental duration, i.e. to detect phrase final lengthening.

From this point of view, we have been developing methods to locate prosodic boundaries using temporal relationship between prosodic events appeared on F_0 contours and segmental boundaries, which were assumed to be obtainable through the phoneme

recognition process. One such method is the statistical modeling of F_0 contour transitions in mora units [2][3]. Different from the case of segmental features, modeling in frame units will not give a good result. This is because prosodic features are those of supra-segmental 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 modeling scheme. Since the models are time-aligned to segmental boundaries, they can be rather easily incorporated into phoneme-based speech recognition process. The modeling in mora unit further has several advantages over frame-based modeling; it can be robust to F_0 contour fluctuations at consonants, and it can be trained by a rather small sized speech corpus.

We already have applied this modeling scheme to the modeling of F_0 contours of prosodic words and succeeded to simultaneously detect their boundaries and recognize their accent types with rather high accuracy [4][5]. Here, a prosodic word is defined as a word or a word chunk corresponding to an accent component, which is also called as an accent phrase. We then integrated the developed method of prosodic word boundary detection into continuous speech recognition system, and obtained a favorable result through a preliminary experiment [6]. However, several considerations are still needed on the coding of mora F_0 contours, HMM topology, and so on.

In the current paper, after a brief explanation on the modeling, improvements on the prosodic word boundary detection method will be explained with several experimental results. Results on the continuous speech recognition experiments for speaker-closed and -open cases will be also given.

2. STATISTICAL MODELING OF F_0 CONTOURS AND PROSODIC WORD BOUNDARY DETECTION

2.1. Outlines

A sentence F_0 contour in logarithmic frequency scale is segmented into mora units using mora boundary information obtained by the phoneme recognition process. The segmented F_0 contours are denoted by mora F_0 contours, and, in the current modeling, they are represented by pairs of 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). The developed method mod-

els prosodic words differently according to their accent types and presence/absence of succeeding pauses. The prosodic word models are 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.

2.2. Shape Coding

Each mora F_0 contour may differ in length and in frequency range, and should be normalized before shape coding. Currently, normalization was conducted simply by shifting the average value of a mora 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.

Clustering was conducted by selecting mora F_0 contours without voiceless parts from the training data shown in section 2.6. Distance between two normalized mora F_0 contours was calculated as the difference in logarithmic F_0 values for corresponding points averaged over the whole period of mora F_0 contours. Although in the original method, 9 clusters were adopted [4], they were increased to 32 using the LBG algorithm for better performance. They were called as codes 3 to 34. Two additional codes 1 and 2 were also prepared for pauses and voiceless morae, respectively. Here, voiceless mora is defined as that whose voiced portion does not exceed 20 % of the whole length. These 34 codes were assigned to mora F_0 contours of input speech. In order to take pause length into account, a pause was divided into 100 ms segments and code 1 was assigned to all of them. Code 1 was also assigned to the last segment in a pause, which might be shorter than 100 ms. These segments with code 1 will be denoted as pause morae hereafter for ease of explanation. Also, morae with codes 3 to 34 will be denoted as voiced morae.

2.3. ΔF_0 Codes

Clustering for ΔF_0 codes was conducted by selecting pairs of voiced morae adjacent to each other from the same corpus as used in the shape code clustering. After calculating average $\log F_0$ for the voiced portion of each mora, differences between the averages of the first to the second morae were calculated for all the selected pairs. Then, the LBG algorithm was used to obtain 32 clusters, to which codes 5 to 36 were assigned. Codes 1 to 4 were reserved to represent pairs of morae when one or both of morae were voiceless (or pause) morae as follows:

Code 1: both the first and second morae were pause morae.

Code 2: only the second mora was pause mora.

Code 3: only the first mora was pause mora.

Code 4: at least one of two morae was voiceless mora.

2.4. Prosodic Word Models

In the 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. Letter “ i ” indicates the dominant downfall in F_0 contour occurring at the end of i th mora. Type 0 accent shows no apparent downfall.

The following 7 models were trained in the discrete HMM framework.

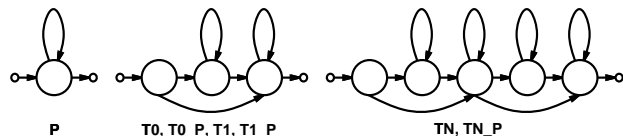


Figure 1: HMM topologies for prosodic word F_0 models.

T0 and T0-P models: for type 0 (or type n) prosodic words,

T1 and T1-P models: for type 1 prosodic words,

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

P model: for pauses.

Here, “-P” indicates that the model is 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. Figure 1 shows the HMM topologies, which were modified from the original ones taking the F_0 contour features of Japanese into consideration. A double code-book scheme was adopted to assign a pair of shape and ΔF_0 codes to each mora F_0 contours. The stream weights for shape codes and ΔF_0 codes were set to 1 for the current experiments.

2.5. Grammar for prosodic words

Prosodic word bi-gram was calculated using the same training data for the prosodic word models to serve as grammar of prosodic word sequences.

2.6. Detection of prosodic word boundaries

In order to conduct boundary detection experiments in speaker-closed and -open conditions, utterances of two male 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 both speakers are identical. As already mentioned, training data (T) were used not only to train prosodic word models, but also to cluster shape and ΔF_0 codes, and to calculate prosodic word bi-gram. Since prosodic labels necessary for the experiments, such as lexical accent types and prosodic word boundaries, are not included in the data by speaker MHT, they are converted from tone and break indices of J-ToBI labels 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.

Mora boundaries were detected by the forced alignment using tri-phone HMMs explained later in section 3.1. The following four combinations of the training and testing data were selected for the boundary detection experiments:

Table 1: Results of prosodic word boundary detection.

Experiment	R_d (%)	R_i (%)
(a) MYI for training and testing	72.7	12.3
(b) MHT for training and testing	75.4	12.3
(c) MHT for training, MYI for testing	70.3	11.7
(d) MYI for training, MHT for testing	73.9	14.8

Table 2: Conditions of acoustic analysis for speech recognition.

Sampling frequency	20 kHz
Analysis window	Hamming window with 25 ms
Frame shift	10 ms
Pre-emphasis coefficient	0.97
Feature parameters	12MFCC+12 Δ MFCC+ Δ power
Filter-banks	24 channels
Cepstral subtraction	Each utterance

- (a) T(MYI) for training and R(MYI) for testing.
- (b) T(MHT) for training and R(MHT) for testing.
- (c) T(MYI) for training and R(MHT) for testing.
- (d) T(MHT) for training and R(MYI) for testing.

Cases (a) and (b) are for speaker-closed experiments, while cases (c) and (d) for speaker-open experiments.

Detection rate R_d and insertion error rate R_i for prosodic word boundaries are respectively defined as $R_d = N_{cor}/N_{bou}$ and $R_i = N_{ins}/N_{bou}$. Here, N_{bou} , N_{cor} and N_{ins} indicate the numbers of total prosodic word boundaries in the test data, boundaries detected inside the ± 100 ms region from the correct position and insertion errors, respectively.

Table 1 shows the results. Although the performances are similar to those obtained by the original modeling, superiority of the revised modeling is clear if we take it into account that the mora boundary information attached to the database was used in the former experiments [5].

3. CONTINUOUS SPEECH RECOGNITION

3.1. Outlines

The boundary detection method was integrated to a continuous speech recognition system 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 integrated system shown in Figure 2, mora recognition was 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 using HTK software Ver.2.1. Conditions of acoustic analysis are summarized in Table 2.

The following items were arranged for the both stages:

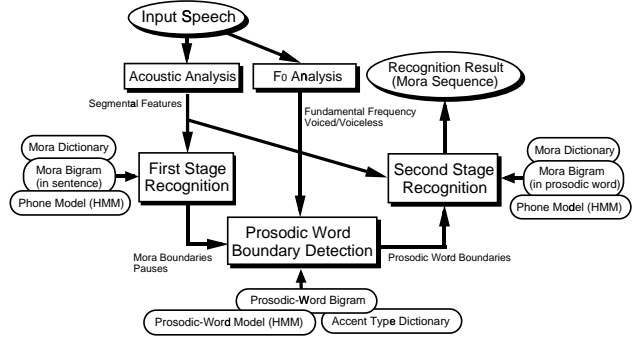


Figure 2: Integrated speech recognition system.

1. Mora dictionary consisting of all possible Japanese morae (125 morae). Pause period SP is also included.
2. Phone HMMs selected from Japanese tri-phone models trained as “Basic Dictation Software for Japanese,” developed under an IPA project [7].
3. Two types of mora bi-gram as language modeling: one obtained without taking prosodic word boundaries into account and the other obtained with taking them into account. The former one was used in the first stage of the recognition 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. Mora bi-gram perplexities were around 40 to 42 for the first stage, while they were around 29 for the second stage. The perplexity reduction from the first stage to the second stage indicates the possibility of better recognition results when prosodic word boundary information is used.

3.2. Experimental results

Mora recognition experiments were conducted for cases (a) to (d) in section 2.6. Their results are shown in Figure 3, where mora recognition rates before and after the second stage C_{bm} and C_{am} are defined as:

$$C_{bm}, C_{am} = (N_{mora} - N_{del} - N_{subst} - N_{ins}) / N_{mora} \quad (1)$$

Here, N_{mora} , N_{del} , 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 that the log-likelihood is multiplied by the factor S before combining it with acoustic likelihood. If we set the factor S to 7, C_{am} outperforms by few percent from C_{bm} for cases (a) and (d), indicating the validity of the proposed method in speech recognition. In the cases (b) and (c), improvements in the recognition rates are not clear, but the use of prosodic word boundary still has no negative effect on speech recognition. When the factor S is increased, the recognition rates decreases and C_{am} has a value smaller than C_{bm} . This indicates that a certain level in recognition rate (of the first stage) is required to obtain a positive effect from the prosodic word boundary information. The figure also shows the results of prosodic word boundary detection.

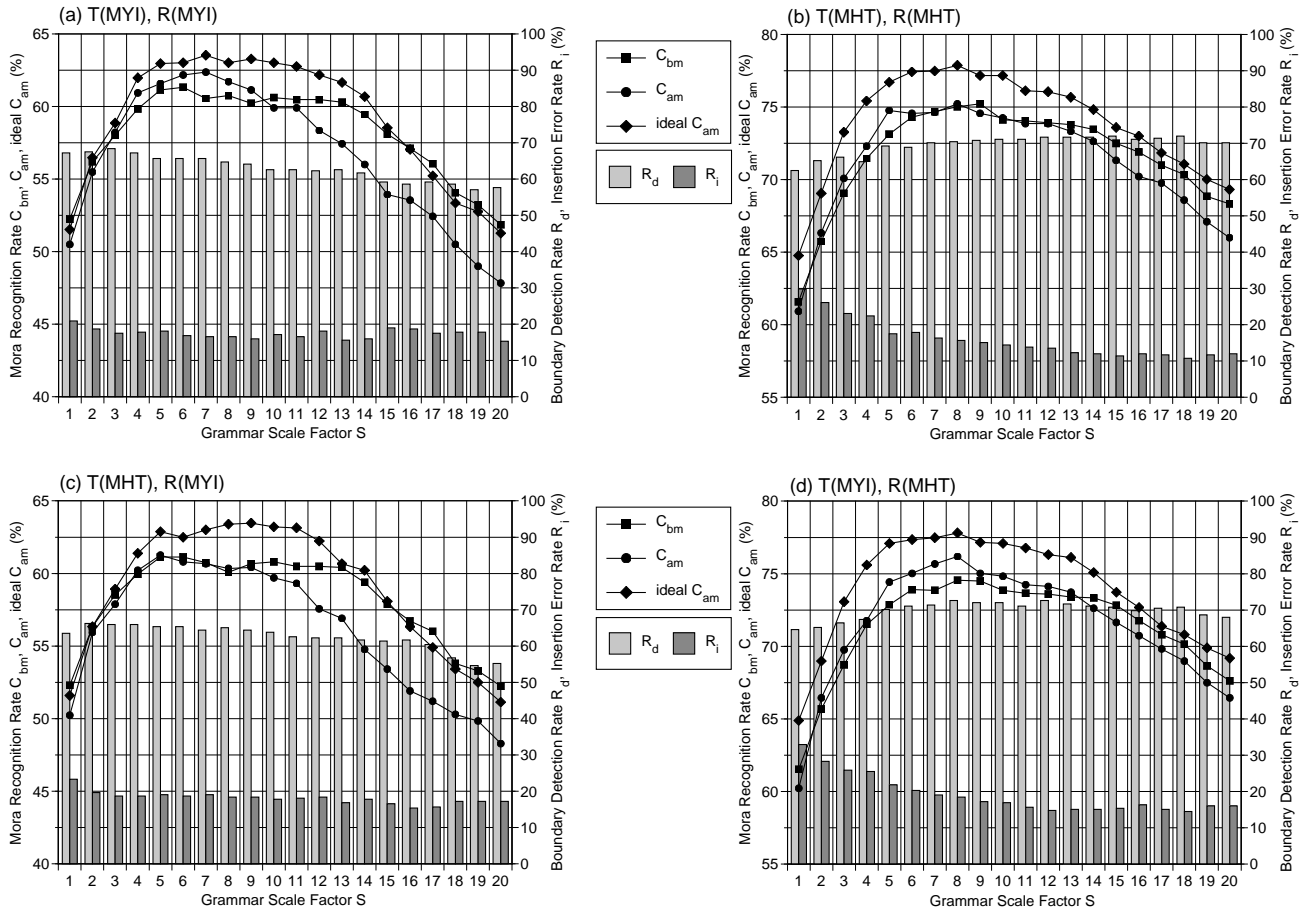


Figure 3: Mora recognition results for four cases.

4. CONCLUSION

Experiments on prosodic word boundary detection were conducted using a statistical modeling of mora transitions of prosodic word F_0 contours. By increasing the codebook sizes and modifying the HMM topologies, detection rates better than those by our original modeling were obtained. This boundary detection method was further combined with a continuous speech recognition scheme and evaluated from the viewpoint of mora recognition rates. Although favorable results were obtained, the experimental results indicated that the original phoneme recognition rate should exceed a certain level (say 60 %) in order that the prosodic boundary information improved the recognition rate. In view of the lack of speech corpus with prosodic word boundary locations, we are now planning to construct such a corpus.

5. REFERENCES

- [1] H. Niemann et al., "Using prosodic cues in spoken language systems," *Proc. SPECOMWorkshop*, St. Petersburg, pp.17-28 (1998).
- [2] K. Hirose and K. Iwano, "A method of representing

fundamental frequency contours of Japanese using statistical models of moraic transition," *Proc. EUROSPEECH'97*, 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*, 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*, pp.599-602 (1998-12).
- [5] K. Iwano, "Prosodic word boundary detection using mora transition modeling of fundamental frequency contours -Speaker independent experiments-", *Proc. EUROSPEECH'99*, pp.231-234 (1999-9).
- [6] K. Iwano and K. Hirose, "Prosodic word boundary detection using statistical modeling of moraic fundamental frequency contours and its use for continuous speech recognition," *Proc. IEEE ICASSP'99*, pp.133-136 (1999-3).
- [7] K. Takeda et. Al., "Common platform of Japanese large vocabulary continuous speech recognition research: construction of acoustic model," *Information Processing Society of Japan, SIG Notes*, 97-SLP-18-3 (1997). (in Japanese)