Continuous Speech Recognition of Japanese Using Prosodic Word Boundaries Detected by Mora Transition Modeling of Fundamental Frequency Contours

Keikichi Hirose, Nobuaki Minematsu, Yohei Hashimoto and Koji Iwano***

Dept. of Frontier Informatics, School of Frontier Sciences, University of Tokyo
*Dept. of Inf. and Commu. Engineering, School of Inf. Science and Tech., University of Tokyo
Bunkyo-ku, Tokyo, 113-0033, Japan
hirose@gavo.t.u-tokyo.ac.jp, mine@gavo.t.u-tokyo.ac.jp, hashi@mlab.t.u-tokyo.ac.jp
**Dept. of Computer Science, School of Inf. Science and Engineering, Tokyo Inst. of Technology
Meguro-ku, Tokyo, 152-8552, Japan
iwano@furui.cs.titech.ac.jp

Abstract

An HMM-based method of detecting prosodic word boundaries was developed for Japanese continuous speech and was successfully integrated into a mora-basis continuous speech recognition system with two stages operating without and with prosodic information. The method is based on modeling the fundamental frequency (F_0) contour of input speech as transitions of mora-unit F_0 contours and operates after receiving mora boundary information form the 1st stage of the recognition system. The 1st and the 2nd stages use different mora bi-gram models as their language models: one trained not taking prosodic word boundary location into account and the other taking into account. Because of perplexity reduction of the model from the 1st to the 2nd stages, an improved recognition result can be obtained from the 2nd stage. In the current paper, the method is explained with experimental results. Issues of grammar scale factor for the boundary detection and N-best scheme for the speech recognition are also included. Improvements in mora recognition rates from the 1st to the 2nd stages were observable in both speaker-closed and –open experiments.

1. Introduction

As important features in human speech perception process, many researchers have tried to incorporate prosodic features into machine speech recognition process. There may be roughly two possible ways to use prosodic features in speech recognition process. One is to control acoustic features depending on the prosodic information, and the other is to detect prosodic events (prosodic boundaries, word accent types, speech acts, and so on) and to utilize them to control the speech recognition process. The first way has a major problem in speaker dependency and complexity of the effect of prosodic features on acoustic features [1], but will not be addressed here. The second way, to which the paper is related, also has two major problems: how to detect prosodic events accurately, and how to use them.

The most straightforward (but naive) method of the second way is to find out prosodic boundaries only with prosodic features and to use them to segment input speech prior to the speech recognition process. If the method properly works, it may largely increase recognition performance. Although several methods have been tried from this viewpoint, they did not work well. The major reasons are low boundary detection rates and large variations in boundary positioning by each speaker and for each utterance. The boundary detection rates are not improved so much by totally looking at various prosodic events, such as fundamental frequency (F_0) contour dips, phone duration lengthening, and so on, and/or by adopting statistical methods. The results suggest that prosodic features are not enough; segmental information should also be utilized for boundary detection. Since, in most continuous speech recognizers (decoders), two-pass algorithm is adopted, phoneme boundary information obtainable from the first pass can be cooperatively used to improve detection rates of prosodic boundaries. The second pass decoding process can be facilitated by the boundary information. The probabilistic factor of the prosodic boundary positioning may cause a hesitation in using prosodic information for speech recognition. However, we should note that the positioning is not a random process, and humans put boundaries only on possible locations, which correspond to some linguistic boundaries. A possible and good way is to use prosodic boundaries only when they are clearly found. A sophisticated answer to the problem was given as an efficient pruning during the decoding process [2].

Although introduction of stochastic language modeling realized a significant progress in continuous speech recognition, it includes a problem that the modeling is trained only for written texts. As outputs of human process of sound production, spoken sentences cannot be fully represented by written language grammars. They are largely related to prosodic features, and, therefore, prosodic information can be utilized to cope with the problem. The direct way is to construct a language model taking prosodic information into account. Using prosodic information in a statistical framework may be beneficial in avoiding the final recognition result to be seriously affected by the wrong information. The major difficulty along this line will be the collection of enough training corpora with prosodic information. This problem can be partly solved by finding prosodic boundary positioning features for a small speech corpus and by placing prosodic boundaries in the text corpus for language model training according to the features [3]. Preliminary results showed rather large test set perplexity reduction.

In the current paper, we summarize our work on mora transition modeling of F_0 contours, developed through the above considerations [4, 5]. Unlike segmental features, modeling of F_0 contours in frame units will not give a good

result. This is because prosodic features spread in wider ranges and should be treated in longer units. Taking into account that "mora" is the basic unit of Japanese pronunciation (mostly coinciding with a syllable) and its relative F_0 value is important for perceiving accent types and other prosodic events of Japanese, the modeling scheme was developed. Since the models are represented by state transitions time-aligned to mora boundaries (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, and it can be trained by a rather small sized speech corpus.

To detect prosodic boundaries, input speech is first segmented into mora unit based on the phoneme boundary information obtained by the first stage of the recognition process. Here, "stage" is used instead of "pass" in order to make it clear that the speech recognition system adopted in the current paper is different from ones for large vocabulary continuous speech as mentioned in section 3.1. Then, the F_0 contour of input speech is represented as a sequence of prosodic word F_0 contours, each of which is modeled as an HMM of mora F_0 transitions. Henceforth, this HMM is denoted as prosodic word model. Prosodic word F_0 contours are modeled separately for their accent types, and, therefore, accent type information is obtainable together with prosodic word boundary information. Here, "prosodic word" is a basic prosodic unit defined as a word or a word chunk corresponding to an accent component, which is also called as "accent phrase." A prosodic word mostly coincides with a "bunsetsu," a basic unit of Japanese language consisting of independent word(s) followed by particle(s), and, thus, its boundary information can be utilized to facilitate the recognition process.

Two versions of mora bi-gram are used in the 1st stage and the 2nd stage as language models. The 2nd stage uses the mora bi-gram trained after segmenting the text corpus into prosodic words, while the 1st stage uses the mora bi-gram trained before segmentation. Perplexity reduction from the model for the 1st stage to that for the 2nd stage leads to an improvement of final recognition results.

In the current paper, after a brief explanation on the modeling, the method of prosodic word boundary detection is first explained. Then, incorporation of the method into a continuous speech recognition system is introduced. Finally, grammar scale factor (weight of likelihood of prosodic word accent type bi-gram to that of prosodic word model) in the boundary detection process and N-best scheme in the recognition process are discussed with some experimental results.

2. Modeling and prosodic word boundary detection

2.1. Outlines

Figure 1 schematically shows the process of prosodic word boundary detection (and accent type recognition) using the mora F_0 contour transition models of prosodic word F_0 contours. In the method, mora F_0 contours obtained by segmenting sentence F_0 contours according to mora boundary locations 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) [6]. Prosodic word F_0 contours are grouped depending on their accent types and presence/absence of succeeding pauses, and each group is represented by a discrete HMM of mora F_0 transitions. When a mora F_0 contour is represented by a set of parameters such as F_0 slopes instead of codes, an HMM of continuous distribution can be introduced to model each group. The experimental results, however, showed no significant improvement, and use of continuous density HMM's is not regarded hereafter. The prosodic word models are matched against input utterances to obtain prosodic word sequences with their accent types. Since an input utterance is represented as a sequence of prosodic words, prosodic word boundaries can be detected. As for the grammar of the matching process, prosodic word bi-gram was trained and utilized. Here, we should note that the bi-gram used in the boundary detection process is somewhat different from that used in the recognition process; it is for prosodic contents (accent types), not for linguistic contents.

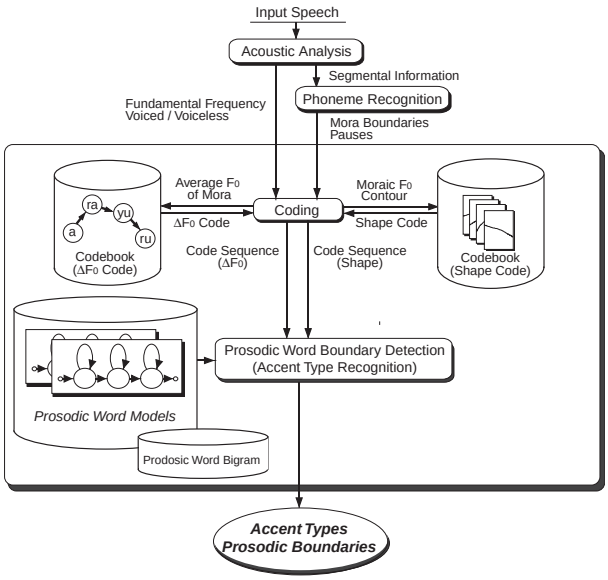

Figure 1: Prosodic word (accent phrase) boundary detection and accent type recognition based on the statistical modeling of mora F_0 contour transitions.

2.2. Shape Coding

Each mora F_0 contour may differ in length and in frequency range, and should be normalized prior to the 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.

After selecting mora F_0 contours without voiceless parts from the training data shown in section 2.6 (16434 samples for speaker MYI, and 18338 samples for speaker MHT), clustering was conducted using the LBG algorithm. 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. Through a preliminary experiment of boundary detection changing the codebook size from 4 to 256, 32 clusters were finally selected. They are henceforth 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. Assignment of codes 3 to 34 was done simply by the minimum distance basis. Voiceless periods, which might be included in mora F_0 contours of input speech, were excluded from the distance calculation. 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. Segments with code 1 should 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 form 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 on of two morae was voiceless mora.

These 36 codes were assigned to mora F_0 contours of input speech.

2.4. Prosodic Word Models

In the Tokyo dialect of Japanese, an m-mora word is uttered with one of m+1 accent types, which are usually denoted as type i (i=0~m) 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.

The following 7 models were trained in the discrete HMM framework using HTK software. Training was conducted by EM algorithm.

- T0 and T0-P models: for type 0 (or type n) prosodic words,
- T1 and T1-P models: for type1 prosodic words,
- TN and TN-P models: for types 2 to n-1 prosodic words,
- P model: 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. Figure 2 shows the HMM topologies, which were selected by 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.

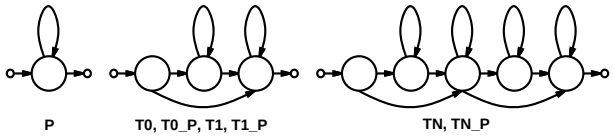

Figure 2: HMM topologies for prosodic word F_0 models.

2.5. Grammar for prosodic words

The prosodic word bi-gram was calculated using the same training data for the prosodic word models to be served as a grammar of prosodic word sequences in the boundary detection process. Weight of the language model (prosodic word bi-gram) likelihood to the acoustic model (prosodic word model) likelihood was set to 1 in sections 2 and 3.

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 both speakers' utterances are identical for training and for testing data. 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 [7] 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:

(a) T(MYI) for training and R(MYI) for testing.

(b) T(MHT) for training and R(MHT) for testing.

(c) T(MHT) for training and R(MYI) for testing.

(d) T(MYI) for training and R(MHT) 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 follows:

$$R_d = N_{cor}/N_{bou} \quad (1)$$

$$R_i = N_{ins}/N_{bou} \quad (2)$$

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 boundary detection results where mora boundary information is obtained through the forced alignment process using the same tri-phone set explained in section 3.

Table 1: Results of prosodic word boundary detection.

Experiment	R_d (%)	R_i (%)
(a) T(MYI) and R(MYI)	72.70	12.27
(b) T(MHT) and R(MHT)	75.38	12.31
(c) T(MHT) and R(MYI)	70.25	11.66
(b) T(MYI) and R(MHT)	73.85	14.77

3. Integration and continuous speech recognition

3.1. Outlines

The boundary detection method was integrated to a continuous speech recognition system as shown in Fig. 3. The system is different from widely used large vocabulary continuous speech recognition systems in that it is based on mora recognition and does not have a word lexicon. This is because the major aim of the current work is to clarify the effects of using prosodically obtainable word boundary information in speech recognition. In the system, mora recognition was conducted twice using differently trained mora bi-grams as the grammar [6]. The first stage operates with mora bi-gram trained in sentence unit (not taking word boundaries into account) and the resulting information on mora boundary location is fed to the process of prosodic word boundary detection. In the second stage, input speech is first segmented into prosodic words using the prosodic word boundary information thus obtained, and then mora recognition is re-conducted using mora bi-gram trained in the unit of prosodic word. 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:

- Mora dictionary consisting of all possible Japanese morae (125 morae). Pause period SP is also included.
- Phoneme HMMs selected from Japanese tri-phone models trained as "Basic Dictation Software for Japanese," developed under an IPA project [8].
- Two types of mora bi-gram as the language models as mentioned above: one obtained without taking prosodic word boundaries into account and the other obtained with taking them into account. In the latter model, at the beginning of a prosodic word, “boundary+mora” is counted, and, at the end, “mora+boundary” is counted. 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. Here, we should note that, if prosodic word boundaries are assumed after every 5th morae (average mora length of prosodic words), the perplexity reduction is not observable.

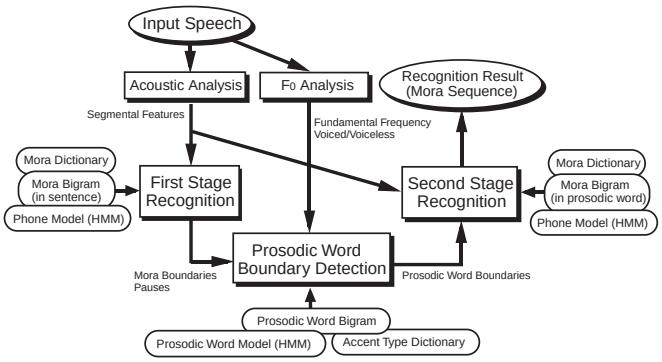

Figure 3: Two-stage speech recognition system with prosodic word boundary detection process.

Table 2: Conditions of acoustic analysis for speech recognition.

Sampling frequency	20 kHz
Analysis window	25 ms Hamming window
Frame shift	10 ms
Pre-emphasis coefficient	0.97
Feature parameters	12 MFCC + 12 ΔMFCC + Δpower
Filter banks	24 channels
Cepstram subtraction	For each utterance

3.2. Experimental results

Mora recognition experiments were conducted for cases (a) through (d) in section 2.6. Their results are shown in Fig. 4, where mora recognition rate C is defined as:

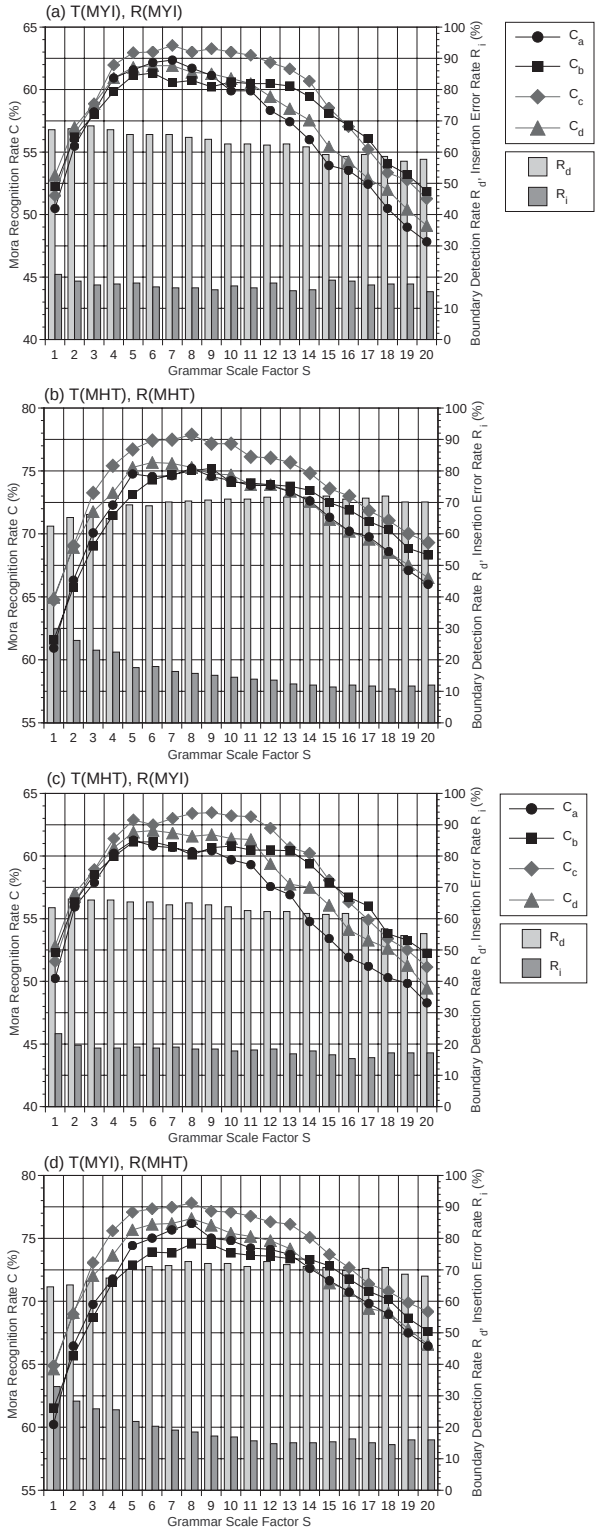

Figure 4: Mora recognition rates as functions of grammar scale factor before and after the second stage.

$$C = (N_{\text{all}} - N_{\text{del}} - N_{\text{sub}} - N_{\text{ins}})/N_{\text{all}} \qquad (3)$$

Here, N_{all}, N_{det}, N_{sub} and N_{ins} respectively represent total number of morae, number of deletions, number of substitutions and number of insertions. Suffices a, b, c, d to C indicate "after second pass," "after first pass," "after second pass when correct boundary information being supplied," "after second pass when boundary information of Table 1 being used," respectively. 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. Although different scale factors are possible for the first and second passes, they are set equal in the current experiments. If we set the factor S to 7, C_a outperforms by few percent from C_b 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_a has a value smaller than C_b.

3.3. Considerations on language model weighting and N-best scheme

Language model weight for boundary detection

Although, in the prosodic word boundary detection of the preceding sections, the grammar scale factor S_{acc} (weight of log-likelihood of the prosodic word bi-gram to that of the prosodic word model) was set to 1, it can be changed to obtain better results. Generally, if it is reduced, insertion errors come dominant, and, conversely, if it is increased, deletion errors come dominant. Recognition experiments were conducted as indicated in section 3 for cases (a) and (b) of section 2.6 (speaker-closed cases), and the mora recognition rates C_a's were obtained as listed in Table 3. The grammar scale factor S for the first and second passes of the recognition process was varied through the experiments to obtain the best results. Case (a) results and case (b) results in the table were obtained when S=5 and S=6, respectively.

Table 3: Mora recognition rates in for various prosodic word bi-gram weights.

Prosodic word bi-gram weight S_{acc}	Mora recognition rate (%)	
	(a) MYI	(b) MHT
0.1	69.8	75.0
1.0	70.9	75.3
3.0	71.9	75.9

3.4. Introduction of N-best scheme

In continuous speech recognition, it was reported that recognition results would be improved by taking N-bests into account [9]. In our experiment, 2-bests were taken into account for the 1st stage mora recognition results and also for prosodic word boundary detection results. Among 4 sentence candidates obtained as combinations of the above 2-bests, the one with largest score was selected. Here, the score is defined as sum of the likelihood of each constituting mora. Table 4. compares the mora recognition rates C_a's of the 1-best scheme and the N-best scheme of this section. The grammar scale factor for the prosodic word boundary

detection S_{acc} was set to 3. The figure clearly indicates improvements by the N-gram scheme.

Table 4: Comparison of mora recognition rates with and without N-best scheme for various mora bi-gram weights (grammar scale factors).

Mora bi-gram weight S	Mora recognition rate (%)			
	(a) MYI		(b) MHT	
	1-best	N-best	1-best	N-best
4	71.0	71.5	74.6	75.1
5	71.9	72.2	75.4	76.0
6	70.9	71.6	75.9	76.6
7	71.1	71.8	75.9	76.7
8	71.4	72.1	75.0	75.5
9	70.7	71.3	74.1	74.7

4. Conclusions

A method of prosodic word boundary detection is presented using a statistical modeling of mora transitions of prosodic word F_0 contours. The method is successfully integrated in a 2-stage continuous speech recognition process, where mora bi-grams are differently trained are used in the 1st stage and the 2nd stage; without prosodic word boundary information for the 1st stage, and with prosodic word boundary information for the 2nd stage. Issues of grammar scale factors for the prosodic boundary detection and N-gram selection are addressed. Finally, an improvement in mora recognition rate from 74.5% to 76.7 % is obtained due to the use of prosodic features.

5. References

[1] N. Minematsu, K. Tsuda, and K. Hirose, "Quantitative analysis of F0-induced variations of cepstrum coefficients," Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding, Red Bank (2001-10).

[2] S. Lee, K. Hirose and N. Minematsu, "Incorporation of prosodic module for large vocabulary continuous speech recognition," Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding, Red Bank (2001-10).

[3] M. Terao, N. Minematsu, K. Hirose, "Improvement of n-gram language model using prosodic information," Fall Meeting Report, Acoust. Soc. Japan, Vol.1, 2-1-21, pp.89-90 (2001-10). (in Japanese)

[4] 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).

[5] 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'98, Phoenix, Vol.1, pp.133-136 (1999-3).

[6] K. Hirose and K. Iwano, "Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition," Proc. IEEE ICASSP'2000, Istanbul, Vol.3, pp.1763-1766 (2000-6).

[7] J. J. Venditti, "Japanese ToBI labeling guidelines," Technical Report, Ohio-State University (1995).

[8] 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) Smith, J. O. and Abel, J. S., ``Bark and ERB Bilinear Transforms", *IEEE Trans. Speech and Audio Proc*., 7(6):697-708, 1999.

[9] R. Schwartz and S. Austin, "A comparison of several appropriate algorithms for finding multiple (N-best) sentence hypotheses," Proc. IEEE ICASSP'91, Vol.1, pp.701-704 (1991).