REPRESENTING PROSODIC WORDS USING STATISTICAL MODELS OF MORAIC TRANSITION OF FUNDAMENTAL FREQUENCY CONTOURS OF JAPANESE

Koji Iwano and Keikichi Hirose
{iwano, hirose}@gavo.t.u-tokyo.ac.jp

Department of Information and Communication Engineering
School of Engineering, University of Tokyo
Bunkyo-ku, Tokyo, 113-8656, Japan

ABSTRACT

We have formerly proposed a statistical model of moraic transitions of fundamental frequency (F_0) contours and showed its effectiveness for prosodic boundary detection and accent type recognition. This model represented F_0 contours of prosodic words to simultaneously detect and recognize prosodic word boundaries and accent types. This paper proposes a method where prosodic word F_0 contours are modeled separately according to their accent types and presence/absence of succeeding pauses. An utterance is regarded as a sequence of prosodic words under a simple grammar. Each moraic F_0 contour is represented by a pair of codes; the original shape code and the newly introduced delta code representing the degree of F_0 change between the mora in question and its preceding mora. Compared with earlier results, the boundary detection rate inproves from 87.7 % to 91.5 %. Accent type recognition rate reached 76.0 % (type 1 accent discrimination).

1. INTRODUCTION

Prosodic features of speech are unutilized (with the exception of durational modeling) in most current speech recognition systems. One exception will be the system developed under the well-known Verbmobil project that uses prosodic features, for instance, to determine whether the input utterance is declarative or interrogative [1]. However, usage is limited to a small part of the recognition process. As clear from the human process of speech perception, more positive use of prosodic features is necessary for further advancement of recognition systems.

Although several methods have been developed for the use of prosodic features, their effects were rather limited. This is because most methods attempted to detect prosodic events (such as syntactic boundaries with F_0 contour dips) from prosodic features 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],[3]. Although modeling in frame units is possible, it dose not give good results. Supra-segmentals should be treated in longer periods. We have developed a moraic transition model 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 accent-type perception. Since this modeling is time-aligned to segmental boundaries, it can be easily incorporated into phoneme-based speech recognition processes.

We already applied this modeling for syntactic boundary (prosodic word boundary) detection of Japanese sentences [2],[3] and accent type recognition of Japanese 4-mora words [3]. Although favorable results were obtained indicating the potential ability of the modeling to represent prosodic events, further improvements were required. One problem is that, by modeling the boundaries straightforward, the period of observation was decided rather heuristically. The other problem is that only the shape was accounted when representing a moraic F_0 contour by a code.

From this viewpoint, we newly developed a method to detect prosodic word boundaries and accent types simultaneously. When representing moraic F_0 contours with codes, there average values were also took into account. In the current paper, after a brief explanation of the new modeling, we will present some experimental results.

2. STATISTICAL MODELING OF F_0 CONTOURS

2.1. Outlines

The proposed 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. The prosodic word models were 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. In the current modeling, 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 adjacent (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.

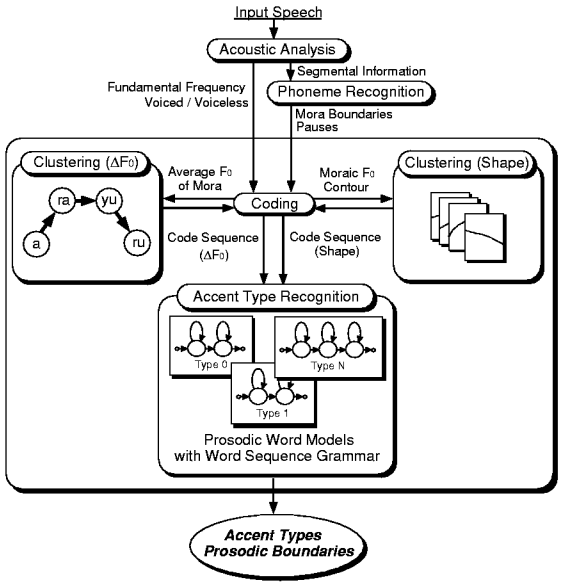

Figure 1: Method of prosodic word boundary detection and accent type recognition based on the modeling of F_0 contours in moraic units.

2.2. Shape Coding

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 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.

Shape codes were decided by clustering 983 moraic F_0 contours without voiceless part. These contours were selected from ATR continuous speech corpus; 85 sentence utterances by a male announcer (speaker MYI) on task SD (a pile of sentences with no context to each other). The clustering scheme was that based on the single linkage method and the leader method [4]. As the result, 9 clusters were obtained and named as codes 3 to 11 as shown in Table 1. Two additional codes 1 and 2 were also prepared respectively for pauses and voiceless mora.

These 11 codes was 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 was assigned based on the minimum distances between its moraic F_0 contours and averaged F_0 contours of the clusters. Different from the case of clustering, a moraic F_0 contour may include voiceless regions. Such regions were excluded from the distance calculation.

Table 1: Distribution of the shape codes in the training data. "Convex #1" means the peak of convex locates between the middle to the right edge of the contour, while "Convex #2" means the peak of convex locates between the middle to the left edge of the contour.

Code Number	Feature of the Shape	Number of Mora
1	Pause	2,307
2	Voiceless	1,999
3	Flat	2,476
4	Slightly Rising	1,183
5	Rising	385
6	Sharply Rising	338
7	Slightly Falling	4,867
8	Falling	2,353
9	Sharply Falling	1,703
10	Convex #1	330
11	Convex #2	332

2.3. ΔF_0 Coding

Clustering was conducted by selecting pairs of moraic F_0 contours adjacent to each other form the same corpus as used in the shape code clustering. Only pairs of voiced mora 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 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 as shown in Table 2. Codes 1 and 11 were used to represent the distances exceeding 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 as 0.
2. For a voiceless mora, its average F_0 is calculated as the interpolation between the av-

Table 2: Distribution of the ΔF_0 codes in the training data.

Code Number	ΔF_0	Number of Mora
1	Negative(falling)	1,526
2	·	93
3	·	429
4	·	1,517
5	·	4,256
6	Zero(no change)	6,180
7	·	1,757
8	·	847
9	·	260
10	·	104
11	Positive(rising)	1,304

erage F_0 of its preceding voiced (or pause) mora and that of its succeeding voiced (or pause) mora.

2.4. Prosodic Word Modeling

Discrete HMMs with left to right configuration in HTK software were adopted to model the prosodic words. The training and the recognition were done by EM algorithm and Viterbi algorithm respectively.

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

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 $b_j(o_t)$ of generating observation o_t is given by:

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

where $P_{js}(o_{st})$ 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.

Training data of these models contained 503 sentences; 3,365 prosodic words (include 658 pauses) from the same corpus used in the code clustering.

2.5. 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)." Bigram was constructed from the same training data for prosodic word models.

3. EXPERIMENTS

Testing data was 50 sentences included in the training data (closed condition experiments). The total number of prosodic words in the testing data was 326 including 70 pauses. Code weightings γ_s and γ_r were both set to 1.0.

3.1. Accent Type Recognition

Accent type recognition rates R, A are defined as:

$$R = \frac{N-D-S}{N} \quad (2)$$

$$A = \frac{N-D-S-I}{N} \quad (3)$$

where N, D, S and I respectively denote the numbers of total prosodic words, deletions, substitutions and insertions. Table 3 shows R and A for the following 2 cases:

1. Type 0 and type 1 accent recognition: to tell type 0, type 1 or others.
2. Type 1 accent recognition: to tell type 1 or others.

Table 3: Result of accent type recognition

Recognition Outputs	Grammar	Recognition Rate (%)	
		R	A
Type 0,Type 1, Others	Constraint	76.77	62.63
	Bigram	71.46	66.67
Type 1, Others	Constraint	84.34	72.47
	Bigram	80.05	76.01

3.2. Prosodic Word Boundary Detection

Boundary detection rates C, C_B and C_N are defined as:

$$C = \frac{H_B + H_N}{N_B + N_N} \quad (4)$$

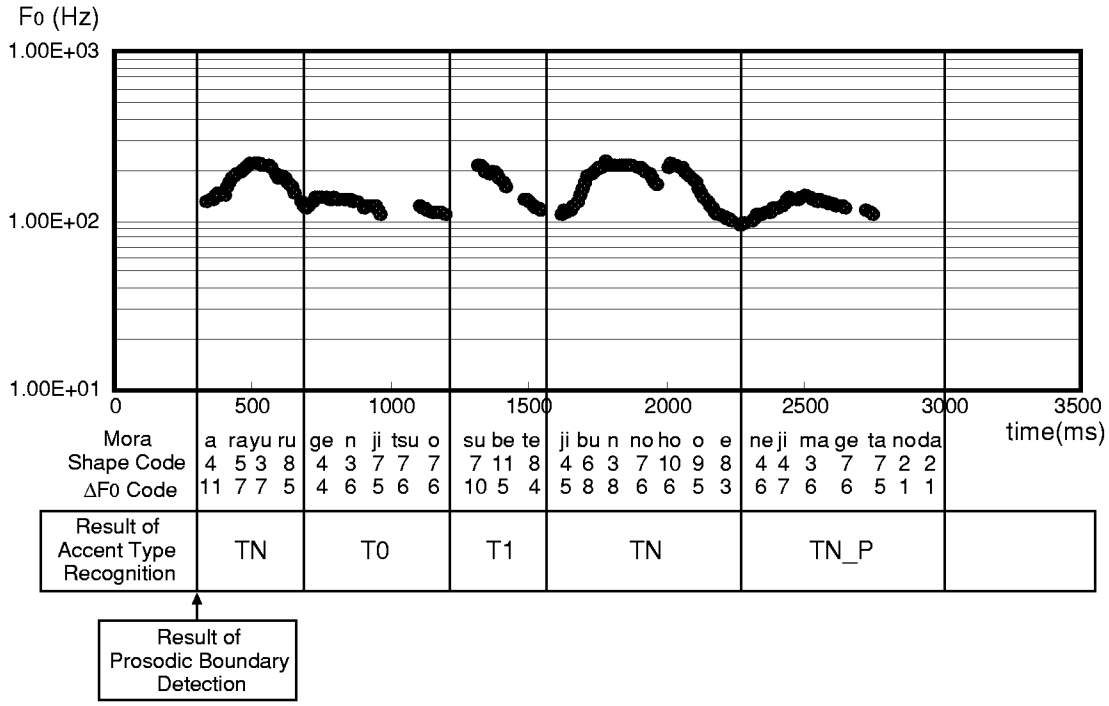

Figure 2: An example of accent type recognition and prosodic word boundary detection. This figure shows a case of correct recognition/detection.

$$C_B = \frac{H_B}{N_B} \quad (5)$$

$$C_N = \frac{H_N}{N_N} \quad (6)$$

where N_B, N_N, H_B and H_N respectively denote the numbers of total prosodic word boundaries, mora boundaries not prosodic word boundaries, mora boundaries correctly judged as prosodic word boundaries and mora boundaries correctly judged as not prosodic word boundaries. Table 4 shows these prosodic boundary detection rates for the two types of grammars. It also shows the detection rates by the former method [2],[3] for comparison. The results clearly shows the improvements of the proposed method. Figure 2 shows an example of accent type recognition and prosodic word boundary detection.

Table 4: 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. DISCUSSION AND CONCLUSION

A method of prosodic word modeling was presented where both the accent type recognition and prosodic word boundary detection were conducted simultaneously. Although favorable results were obtained, the experiments may include some problems. In the ATR corpus, only accent phrase boundaries are labeled, which are assumed to be identical to the prosodic word boundaries in the current experiments. This is actually not the case and degrades the detection performance. From this viewpoint, we are now planning to construct speech database with prosodic labeling. We are also planning to apply the developed method to speech recognition.

5. REFERENCES

1. C. Lieske, J. Bos, M. Emele, B. Gambäck and C. J. Rupp, "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, Vol.1, 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*, Seattle, Vol.1, pp.25-28 (1998-5).
4. J. A. Hartigan, Clustering Algorithms, John Wiley & Sons, New York (1975).