ACCENT TYPE RECOGNITION AND SYNTACTIC BOUNDARY DETECTION OF JAPANESE USING STATISTICAL MODELING OF MORAIC TRANSITIONS OF FUNDAMENTAL FREQUENCY CONTOURS

Keikichi Hirose and Koji Iwano

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

ABSTRACT

Experiments on accent type recognition and syntactic boundary detection of Japanese speech were conducted based on the statistical modeling of voice fundamental frequency contours formerly proposed by the authors. In the proposed modeling, fundamental frequency contours are segmented into moraic units to generate moraic contours, which are further represented by discrete codes. After modeling the accent types and syntactic boundaries, their recognition/detection was done for ATR speech corpus. As for the accent type recognition, 4-mora words were used for the training and testing, and recognition rates around 74 % were obtained for speaker open experiments. As for the syntactic boundary detection, detectability of accent phrase boundaries was tested for sentence speech. Although the experiments were conducted only for the closed condition due to availability of speech corpus, the result indicated the usefulness of separating the boundary model into two depending on whether the boundary is accompanied by a pause or not.

1. INTRODUCTION

Although introduction of statistical modeling largely improved the performance of speech recognition, the modeling was mostly on segmental features. As it is clear from the consideration on the human process of speech perception, further advancement in speech recognition requires a new scheme to make a good use of prosodic features. Many efforts were already devoted to extract syntactic boundaries and structures from prosodic features, but the achieved results were not so satisfactory. One possible reason is that, in these efforts, they are trying to detect syntactic boundaries only from the prosodic features. From this point of view, we formerly developed a method to evaluate recognition candidates by comparing model-generated fundamental frequency contour (F_0 contour) for each candidate and that of actually observed [1]. Although this method was proved to be effective in detecting recognition errors accompanied by accent type change and/or syntactic boundary shifts, it suffered from the variations in F_0 contours. To cope with this problem, we formerly developed a statistical modeling of F_0 contours, and showed its validity in detecting syntactic boundaries [2]. Assuming that information on syllable boundaries is obtained during the phoneme recognition process, in the proposed method, statistical modeling of F_0 contours are conducted after segmenting them into moraic unit; a basic unit of Japanese pronunciation mostly coinciding with a syllable.

The proposed modeling is considered to be also valid for the accent type recognition of Japanese words. Since each Japanese word is uttered with a definite accent pattern, and this pattern sometimes plays a dominant role in distinguishing word meaning, the total speech recognition performance considered to be largely increased with accent type information. In the current paper, after the brief explanation of the modeling, we will present some results on accent type recognition experiments. The paper also includes the experimental results on the improvements of syntactic boundary detection.

2. STATISTICAL MODELING OF F_0 CONTOURS

2.1. Outlines

Although statistical modeling has already been introduced in several works to represent prosodic features [3], most of them fell into a mistake of representing frame to frame transitions as in the case of phoneme recognition, and could not realize satisfactory results. Prosodic features should be treated in longer units. Observation sequence in prosodic features is a possible answer for this issue [4], but it is not always clear how we can select it. In the case of Japanese, we have a rhythm that each mora is uttered with a similar duration, and the relative F_0 value of each mora is known to be important to perceive prosodic features. With this consideration, we proposed a modeling of prosodic features in moraic unit. Different from the work on the observation sequence [4], phrase final lengthening is not used here, because it is not always clearly observable in Japanese. Since, as compared to the case of frame unit, the number of morae in a sentence is very small and, therefor, the number of varieties in transition is very limited, the proposed modeling supposed to show a good performance even when only small sized training data are obtainable.

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 recognition. Then, a discrete code is assigned to each moraic F_0 contour. Finally, the obtained code sequence is matched against statistical models of accent types or syntactic boundaries.

Discrete hidden Markov models with left to right configuration in HTK software were adopted to model the accent types and syntactic boundaries. The training and the recognition were done by EM algorithm and Viterbi algorithm respectively. In the previous paper for syntactic boundary detection [2], the best result was obtained when the number of states coincides with the number of morae under the observation. Therefore, in the present paper, these two numbers were made equal throughout all the experiments; no self transition allowed for every state.

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. Clustering and Coding

In order to assign a discrete code to each moraic F_0 contour of the training and testing data, a clustering was first conducted for 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 MHT) 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 [5]. 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 morae.

One of these 11 codes is assigned to each moraic F_0 contour of input speech as follows:

1) A pause period is divided into 100 ms segments 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 V % of the whole length of the mora. In the current experiments, V was fixed to 10.
3) For other morae, one of the codes 3 to 11 was assigned based on the minimum distances between 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: Feature of the average F_0 contour of each cluster and the result of classification for the training data.

Cluster Number	F_0 Contour Feature	Number of Mora
1	pause	4377
2	voiceless	1577
3	flat	4241
4	slightly rising	1480
5	rising	522
6	sharply rising	422
7	slightly falling	4214
8	falling	2201
9	sharply falling	852
10	flat then rising	280
11	flat then falling	169

3. ACCENT TYPE RECOGNITION

3.1. Word Accent Types

In the Tokyo dialect of 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 morae. 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 first mora. Type 0 accent shows no apparent downfall in its F_0 contour. In Figure 1, F_0 contours are schematically shown as high and low patterns for 5 accent types possible for a 4-mora word.

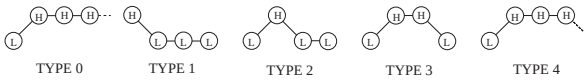

Figure 1: Binary description of F_0 contour for each of 5 accent types of 4-mora words of the Tokyo dialect.

3.2. Experiments

Fundamental frequency contours extracted from the utterances using pitch extraction scheme based on the autocorrelation of LPC residual [6] were first segmented into moraic units. Although the segmental boundary information should be given from the phoneme recognizer, in the current experiments, that attached to the corpus was used instead. When no information on mora boundary is supplied, as in the case of long vowels, a mora boundary was assumed to be locating at the center. Each moraic F_0 contour thus obtained was normalized and classified into one of 11 clusters as explained already. The same procedure was also adopted for the experiments on syntactic boundary detection in section 4.

Table 2: Numbers of word speech samples used for the experiments on accent type recognition.

		Accent Type				Total
		Type 0	Type 1	Type 2	Type 3	
Male Speaker	MHT	99	51	50	48	248
	MKT	95	51	50	49	245
	MAU	98	51	50	48	247
Female Speaker	FKN	99	50	50	49	248
	FAF	100	51	50	47	248
	FFS	100	51	50	48	249

Experiments were conducted using ATR speech corpus of Japanese 4-mora words uttered by 3 male and 3 female announcers. Although 5 accent types are possible for 4-mora words as explained above, type 4 accent was excluded from the experiments. This is because type 4 accent has an F_0 contour similar to that of type 0 accent when uttered in isolation. Table 2 shows numbers of word samples used for the training and testing. For each accent type, a 4-state model without self transition was trained using speech material by 4 speakers (speakers MHT, MKT, FKN and FAF). Then the accent type recognition was conducted for utterances of other 2 speakers not included in the training.

Tables 3 and 4 show the recognition results as confusion matrices. Although rather high recognition rates were obtained for type 1 accent, they were rather low for types 0 and 3 accents. One possible reason is that the normalization process (DC value elimination in moraic contours) before the coding makes flat contours to correspond to the same code regardless of their absolute F_0 values. A scheme may be necessary to incorporate F_0 absolute values in the recognition process.

Table 3: Result of accent type recognition for male speaker MAU.

Accent Type	Accent Type as Recognized			
	Type 0	Type 1	Type 2	Type 3
Type 0	69.9%	8.0%	15.1%	7.0%
Type 1	0.0%	98.0%	2.0%	0.0%
Type 2	2.0%	12.0%	84.0%	2.0%
Type 3	12.5%	10.4%	27.0%	50.0%
Total	74.5%			

Table 4: Result of accent type recognition for female speaker FFS.

Accent Type	Accent Type as Recognized			
	Type 0	Type 1	Type 2	Type 3
Type 0	92.3%	2.0%	4.0%	1.0%
Type 1	3.9%	90.3%	1.9%	3.9%
Type 2	20.0%	10.0%	66.0%	4.0%
Type 3	70.8%	6.2%	0.0%	23.0%
Total	73.5%			

In the current experiment, the coding of moraic F_0 contours was conducted using the clusters shown in Table 1. Since these clusters were obtained for continuous speech, a better result will be obtainable by re-clustering moraic F_0 contours using the training data in Table 2.

4. SYNTACTIC BOUNDARY DETECTION

In the former experiments, *bunsetsu* (a basic linguistic unit peculiar to Japanese defined as "a word chunk consisting of a content word optionally followed by a function word or a string of function words") boundary detection was conducted for various cases (including speaker open and task open cases) using ATR speech corpus of continuous speech [2]. Although fairly good results were obtained, they included the following two problems: 1) *bunsetsu* being a unit of written language and its boundaries not strictly coinciding with prosodic boundaries of spoken language, 2) rather low detection rates for boundaries without a pause. As for the first problem, accent phrase boundaries were detected instead. Accent phrase is roughly corresponds to *bunsetsu*, but defined as a unit which is usually uttered with one accent component. As for the second problem, on the other hand, syntactic boundaries were modeled separately for cases with and without pauses.

4.1. Modeling of Syntactic Boundaries

In the current experiments, the period of observation is fixed to 2 morae before and after (totally 4 morae) the boundary in question, which showed the best results in the previous experiments [2]. Several models were arranged depending on the existence of an accent phrase boundary before the boundary in question and depending on the existence of a pause at the boundary in question. Experiments on boundary detection were then conducted for the following model combinations:

Combination 1: X-B-X, X-N-X

Combination 2: X-B(P)-X, X-B(NP)-X, X-N-X

Combination 3: X-B-X, B-N-N, N-N-B, B-N-B, N-N-N

Combination 4: X-B(P)-X, X-B(NP)-X, B-N-N, N-N-B, B-N-B, N-N-N

In each model, the first/third symbol indicates whether an accent phrase boundary exists before/after the boundary in question (B) or not (N). The second letter indicates whether the boundary in question is an accent phrase

Table 5: Various detection rates of accent phrase boundaries for the 4 model combinations.

Detection Rate	Combination			
	1	2	3	4
C	83.82 %	74.80 %	88.33 %	82.58 %
C_B	78.53 %	91.10 %	63.80 %	82.20 %
C_N	85.01 %	71.13 %	93.85 %	82.67 %
C_B(NP)	66.50 %	87.37 %	44.66 %	73.79 %

boundary (B) or not (N). Symbol X means "not concern." Symbols P and NP in parentheses mean the boundary in question being accompanied by a pause or not, respectively.

4.2. Experiments

Experiments on accent phrase boundary detection were conducted using ATR continuous speech corpus of text reading. Since speech corpus with labeling of accent phrase boundaries is available only for utterances by male speaker MYI, clustering was again conducted using his speech of the same 85 sentences in section 2.3. The result was very similar to that shown in Table 2 for speaker MHT; indicating the consistency of the clustering.

For the model training, 503 utterances of speaker MYI on task SD were used. Depending on the labeling attached, these utterances contain 3365 accent phrase boundaries and 14908 non-accent-phrase boundaries. As for the test, utterances of 25 sentences were selected from the training data (closed condition experiments).

Table 5 shows the total detection rate C in various model combinations. Here, the total detection rate C is defined as: $C = (H_B + H_N)/(B + N)$, where B and N respectively denote numbers of accent phrase boundaries and non-accent-phrase boundaries, while H_B and H_N denote those correctly detected. The table also includes $C_B = H_B/B$ and $C_N = H_N/N$. Detection rate C_B was also calculated for accent phrase boundaries not accompanied by pauses, and was denoted by C_B(NP) in the table. By dividing the accent boundary model into two cases, one with a pause and the other without, C_B was sharply increased. However, C_N decreased simultaneously, causing a slight decrease in the total detection rate C. Although the results can be evaluated in several ways, combination 4 yielded the best performance from the stand point of rather high C_B and C_N with values similar to each other.

Experiments were also conducted on *bunsetsu* boundary detection for comparison. For every combination, however, the introduction of accent phrase boundary showed only a slight improvement. This result may be ascribable to the fact that the labeling of accent phrase boundaries of the corpus was not done based on the observed prosodic features.

5. CONCLUSION

Accent type recognition and syntactic boundary detection were conducted using statistical modeling of moraic transition of F_0 contours. The results showed the modeling was useful for the both purposes, but also showed necessity of further improvements. Currently, a new idea is tested for the improvement of clustering and classification scheme. We are also planning to use the corpus with prosodic labeling such as J-ToBI.

6. REFERENCES

[1] K. Hirose and A. Sakurai, "Detection of syntactic boundaries by partial analysis-by-synthesis of fundamental frequency contours," Proc. IEEE ICASSP, Atlanta, pp.809-812 (1996-5).

[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] A. Ljolje and F. Fallside, "Recognition of isolated prosodic patterns using hidden Markov models," Computer Speech and Language, vol.2, pp.27-33 (1987).

[4] K. Ross and M. Ostendorf, "A dynamical system model for recognizing intonation patterns," Proc. EUROSPEECH'95, Madrid, pp.993-996 (1995-9).

[5] J. A. Hartigan, Clustering Algorithms, John Wiley & Sons, New York (1975).

[6] K. Hirose, H. Fujisaki and N. Seto, "A scheme for pitch extraction of speech using autocorrelation function with frame length proportional to the time lag," Proc. IEEE ICASSP, San Francisco, pp.149-152 (1992-3).