

A METHOD OF REPRESENTING FUNDAMENTAL FREQUENCY CONTOURS OF JAPANESE USING STATISTICAL MODELS OF MORaic TRANSITION

Keikichi Hirose and Kouji 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

A statistical modeling of voice fundamental frequency contours was proposed for the purpose of developing effective ways to utilize prosodic features in speech recognition. In view of the fact that prosodic features should be treated in longer units, the proposed modeling represents the transition in moraic units. A fundamental frequency contour was first segmented into moraic units and then each moraic contour was represented by a code depending on the shape. After modeling fundamental frequency contours for the portions of several morae around boundaries in question based on HMM scheme, experiments on syntactic boundary detection were conducted. Detection rate reached to 89.2 % for the closed condition experiment and was around 85 % for the open (speaker and topic) condition experiment. Experiments on accent type recognition were also conducted yielding around 74 % of correct recognition for the speaker independent cases.

1. INTRODUCTION

As it is clear from the consideration on the human process of speech perception, prosodic features should be utilized in speech recognition for its further advancement. Although, from this point of view, a number of methods have already been reported to detect syntactic boundaries using prosodic features, they were mostly intended to be used in a supplementary way before the main recognition process. Therefore, syntactic boundaries were detected only by the prosodic features, which may limit the detection ability.

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 hypothesis and that of actually observed [1]. Although this method was proved to be effective in detecting syntactic boundaries [2], it suffered from the variations in F_0 contours. Statistical modeling such as hidden Markov modeling is known as one of the best ways to cope with variations in acoustic features, and it was already introduced in several works to represent prosodic features [3]. However, in these works, an HMM is mostly constructed so that its transitions correspond to those of frame to frame of input speech. Since prosodic features

are known as supra-segmental features, they should be treated with longer units. A statistical method was already developed, where observation sequence was introduced for the statistical modeling of prosodic boundaries [4], but it is not always clear. 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.

From this point of view, assuming that mora boundary information is obtainable during the ordinary recognition process, we newly developed a scheme of representing moraic transitions of F_0 contours by statistical modeling. 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. Experiments on syntactic boundary detection together with those on accent type recognition were conducted to evaluate the modeling. Since, as compared to the case of frame unit, the number of morae in a sentence is very small and, therefore, 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.

2. STATISTICAL MODELING OF MORaic TRANSITION

Figure 1 shows the method of syntactic boundary detection based on the proposed modeling. 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 syntactic boundaries (or those for items to be detected or recognized). As for the statistical modeling, discrete HMM of HTK software was utilized.

2.1. 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

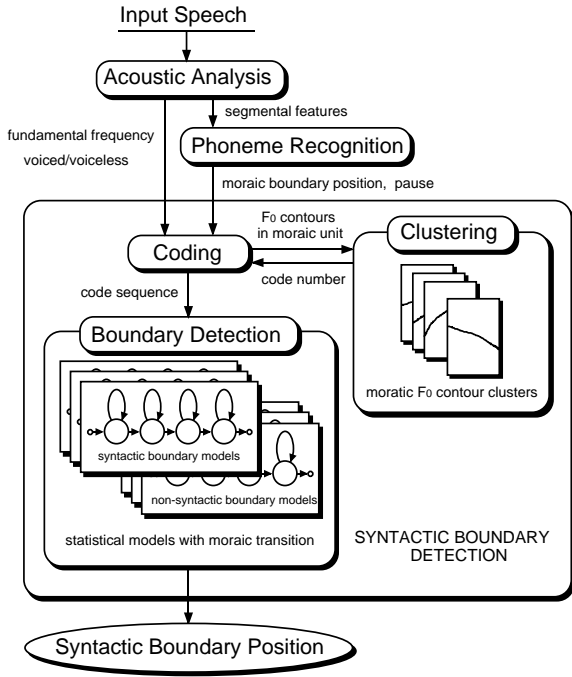


Figure 1. Method of syntactic boundary detection based on the proposed modeling of F_0 contours in moraic units.

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.

2.2. 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 utterances of 85 sentences by a male announcer, a portion of the training data consisting of 503 sentences (see section 3). 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 later in Table 1. Two additional codes 1 and 2 were also prepared respectively for pauses and voiceless morae.

If a pause period is included in speech samples, it 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. 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. 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.

2.3. Modeling Syntactic Boundaries

Discrete HMMs with left to right configuration were adopted to represent syntactic boundaries. Period of observation was varied and checked from the viewpoint of performance of syntactic boundary detection. Concretely, it was varied from 1 to 4 morae before and after the mora boundary in question. Although the proposed HMMs can separately represent prosodic boundaries of different levels (such as, prosodic word boundary, prosodic phrase boundary, and so on), in the current paper, they are merged to a group of prosodic word boundaries and their detection will be discussed.

In the case of Japanese, a prosodic word boundary is roughly corresponding to a *bunsestu* boundary, where *bunsestu* is a basic linguistic unit peculiar to Japanese and is defined as “a word chunk consisting of a content word optionally followed by a function word or a string of function words.” Since no ample amount of database with information of prosodic word boundaries is obtainable in Japanese, these two types of boundaries were assumed to be the same in the experiment, though this assumption is not exactly the case and may degrade the performance of boundary detection.

The following HMMs were arranged to model a mora boundary being a *bunsestu* boundary:

- B1:** *bunsestu* before the boundary directly relating to the *bunsestu* immediately after the boundary,
- B2:** *bunsestu* before the boundary directly relating to the second *bunsestu* after the boundary,
- B3:** *bunsestu* before the boundary directly relating to the third *bunsestu* after the boundary,
- B:** combination of B1, B2 and B3.

On the other hand, to represent non-*bunsestu* boundary, the following HMMs were arranged:

- N1:** a mora boundary before the boundary in question being a *bunsestu* boundary,
- N2:** a mora boundary after the boundary in question being a *bunsestu* boundary,
- N3:** no *bunsestu* boundary existing in the period of observation,
- N:** combination of N1, N2 and N3.

3. EXPERIMENTS

Experiments on the *bunsestu* boundary detection (that is to tell a mora boundary being *bunsestu* boundary or non-*bunsestu* boundary) were conducted using ATR continuous speech corpus of text reading. For the HMM training, 503 utterances of speaker MHT on task SD (a pile of sentences with no context to each other) were used. These utterances contain 3425 *bunsestu* boundaries and 16910 non-*bunsestu* boundaries, totally 20335 mora boundaries. As for the test, utterances of 50 sentences were selected from the training data for the closed condition experiments, and utterances of the same sentences by a male

(speaker MTK) and a female (speaker FKN) announcers were used for the speaker independent experiments. In 50 sentences, 320 *bunsestu* boundaries are included. Utterances on conference registration (task SC with 40 sentences, 284 *bunsestu* boundaries) by the same three speakers were also used for the task-open experiments.

Fundamental frequency contours, extracted from the utterances using the pitch extraction scheme based on the auto-correlation of LPC residual with frame length proportional to the time lag [6], were first segmented into moraic units. Although the segmental boundary information should be given from the phoneme recognizer in the total method shown in Fig.1, 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. Table 1 summarizes the result of classification for the training data together with the features of the average contour of each cluster.

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

Figure 2 shows the total detection rate C for the closed condition experiment in various HMM conditions. Here, the total detection rate C is defined as:

$$C = (H_B + H_N)/(B + N),$$

where B , N , H_B and H_N respectively denote numbers of *bunsestu* boundaries, non-*bunsestu* boundaries, *bunsestu* boundaries detected and non-*bunsestu* boundaries correctly judged. The best result (89.2 %) was obtained for the case of “B N1 N2 N3” combination of “2-2, 4 states.” This result implies the modeling need not be “hidden” in the current framework of moraic F_0 contours. The *bunsestu* boundary detection rate defined by $C_B = H_B/B$ was 74.1 % for this case, which was not the best result. The best condition should be decided as a compromise of these two indices. Figure 3 shows C for each combination of the three speakers and two tasks. For speaker and topic open conditions, C of around 85 % was obtained. An example of boundary detection is shown in Fig.4, where the last boundary was failed to be detected in the correct position(one deletion error in H_B and one insertion error in H_N). Since *bunsestu* boundaries are

not necessarily appear in the prosodic features, the above results may underestimate the performance of the boundary detection ability of the proposed modeling. Further experiments are planned from this viewpoint.

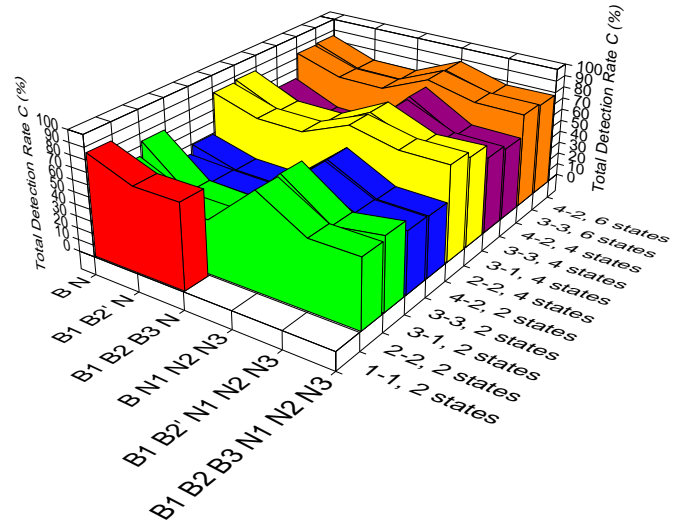


Figure 2. Total detection rate C for various HMM conditions. “4-2, 4 states” indicates that the period of observation is 4 morae before and 2 morae after the boundary and the number of HMM states is 4.

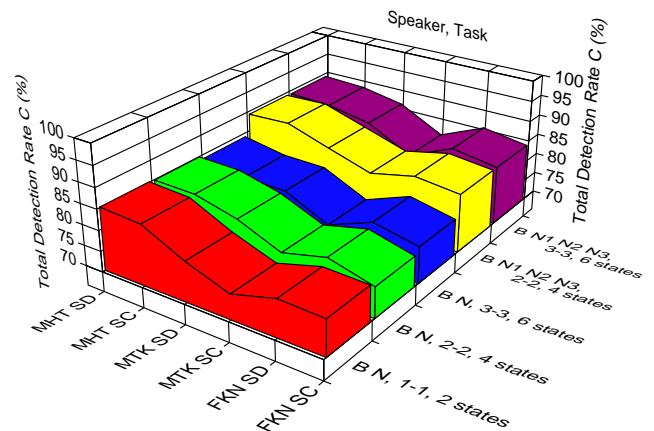


Figure 3. Total detection rate C for various combinations of speakers and tasks.

4. ACCENT TYPE RECOGNITION

The proposed modeling can also be used for the accent type recognition. Experiments were conducted using ATR speech corpus of Japanese 4-mora words. Although 5 accent types are possible for 4-mora words in standard Japanese, the type 0 and type 4 accents show similar F_0 contours when uttered in isolation. Therefore, the experiments were conducted on the recognition of types 0 to 3 accents. For each of these four accent types, 4-state discrete HMM was trained using speech material by

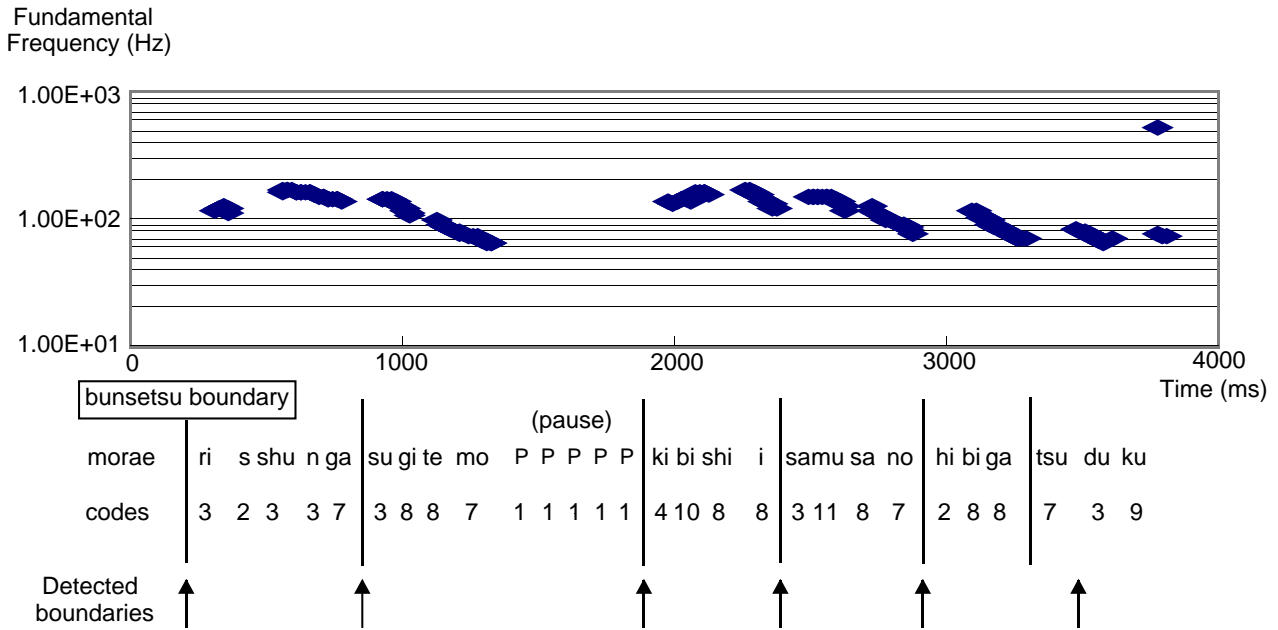


Figure 4. An example of boundary detection for an utterance of task SD by speaker MHT. Boundary modeling of “B N1 N2 N3” combination and HMM of “2-2, 4 states” were used.

4 speakers (2 male speakers MHT, MTK and 2 female speakers FKN, FAF). Totally, 989 samples were used for the training. Recognition experiments were conducted for utterances of speakers MHT and FKN, which were included in the training data. Recognition rates reached 89.5 % and 79.3 %, respectively. Recognition experiments were also conducted for the case of speaker open. Around 74 % of correct recognition was obtained for a male speaker and for a female speaker. Recognition errors from type 3 to type 0 occurred rather frequently for the female speakers. In the current experiment, the coding of moraic F_0 contours was conducted using the same clusters shown in Table 1. Since these clusters were obtained for continuous speech of a speaker, a better result will be obtainable by re-clustering moraic F_0 contours for the word training data.

5. CONCLUSIONS

A method was developed for the statistical modeling of moraic transition of F_0 contours, and its validity was shown by the experiments on *bunsetsu* boundary detection and accent type recognition. Since F_0 contours are modeled in moraic units, the proposed modeling can be combined with the phoneme recognition process rather easily. Construction of the total method of syntactic boundary detection in Fig. 1 is under the way. Further research is planned to improve the clustering and classification scheme.

REFERENCES

- [1] K. Hirose, A. Sakurai and H. Konno, “Use of prosodic features in the recognition of continuous speech,” Proc. ICSLP, Yokohama, pp.1123-1126 (1994-5).
- [2] 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).
- [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).