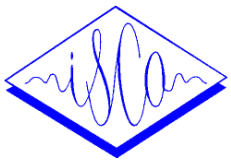
ISCA Archive
http://www.isca-speech.org/archive



3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

USE OF PROSODIC FEATURES IN THE RECOGNITION OF CONTINUOUS SPEECH

*Keikichi Hirose, Atsuhiro Sakurai and Hiroyuki Konno**

Department of Electronic Engineering, Faculty of Engineering, University of Tokyo,
Bunkyo-ku, Tokyo, 113 Japan *Currently with Sony Corp.

ABSTRACT

Two methods were proposed for the use of prosodic features in automatic speech recognition. One is to detect syntactic boundaries of input speech without information on the segmental level, which will be obtained by the ordinary speech recognition process. The other is to check the feasibility of recognition results. In the first method, both the microscopic and macroscopic features of fundamental frequency contours were taken into account, and 96 % of manually detectable boundaries were correctly extracted for the ATR continuous speech database. Several schemes were also proposed to reduce the insertion errors. As for the second method, a scheme of partial analysis-by-synthesis was developed, where fundamental frequency contours are generated using a functional model for the recognition hypotheses of the segmental level and are compared with the observed contour only for the part with recognition ambiguity. The hypothesis giving the best matching with the observation is the possible final recognition result. The proposed method was shown to be valid for recognition errors that include changes in the accent types and in the syntactic boundaries.

1. INTRODUCTION

It is widely admitted that prosodic features play an important role in the human process for speech recognition. For instance, in our recent perceptual experiments, the importance of fundamental frequency (henceforth, F_0) contours on word and sentence recognition was indicated [1]. It is, therefore, very important to make good use of prosodic features in automatic speech recognition.

In continuous speech recognition, information on the syntactic boundaries should be useful as the recognition constraints in the segmental feature level, and, therefore, can be utilized to reduce the searching area in finding the final recognition result. From this point of view, several methods have already been reported to detect syntactic boundaries of spoken sentences. As for the Japanese speech, methods include one based on detecting several boundary features in temporal patterns of F_0 and waveform amplitude [2], another based on finding dips in piecewise-linear F_0 contours and matching with boundary templates [3], and another one based on estimating boundary likelihoods [4]. Most of these methods were based on the macroscopic aspects of the prosodic features, and may include deletion errors, especially for high speech rates. We have proposed a method where both macroscopic and microscopic aspects of F_0 contours are taken into account [5]. Although temporal contours of (source) waveform power may also give us important information for the detection of syntactic boundaries, it was not included in the proposed method, for the following two reasons: 1) Since the power varies according to the recording condition and its temporal contour is largely affected by phoneme combination and noise level, it is rather hard to find an appropriate threshold to characterize a power dip as a syntactic boundary. 2) Since, especially in Japanese, any increase in F_0 is accompanied by an increase in source power, power information is more or less included in the F_0 contour. As for the durational information, pause lengths are useful to detect boundaries and to estimate their syntactic depths. In the current method, it was included in the microscopic features and was used for the detection of boundary candidates. Its use for boundary depth estimation was left for further studies.

Although it is important to increase the performance of the methods above for boundary detection, there should be some limitations in the conventional approaches, including ours. One reason is that it is rather hard to find a robust algorithm due to the rather large speaker-to-speaker and utterance-to-utterance variations of prosodic features. Another reason is that, in these approaches, syntactic boundaries are detected only by the prosodic features without referring to the recognition results based on the segmental features. To cope with these problems, a method is further proposed where the F_0 contour for the recognition candidate is generated using a functional model and is compared with the observed one [5]. For the comparison, a new scheme was introduced where the analysis-by-synthesis method is applied only for a small portion of the F_0 contour of the sentence. The degree of discrepancy of the generated contour from the observed one is calculated for each recognition hypothesis. The hypothesis giving the minimum discrepancy is the possible final recognition result. The proposed method was shown to be valid not only for the detection of recognition errors, causing changes in the syntactic boundaries (or for the reduction of syntactic ambiguities in the recognition results), but also for the detection of errors accompanied by changes in accent types. Comparison between the model-generated contour and the observed one was also conducted in another method to detect prosodic events, but it did not utilize the recognition results obtained by the segmental features [6].

2. DETECTION OF SYNTACTIC BOUNDARIES

2.1 Procedure of Boundary Detection

With the proposed method for the detection of syntactic boundaries, we can detect not only major boundaries like clause and phrase boundaries, but also minor boundaries like *bunsetsu* boundaries. Here, *bunsetsu* is a basic unit of Japanese syntax consisting of a content word with or without being followed by a string of function words.

Pitch extraction was first conducted for the input speech to obtain F_0 contours. A method based on the newly defined autocorrelation function with frame length proportional to the time lag was adopted for this purpose [7]. Then, a macroscopic F_0 contour was calculated as a piecewise linear contour from the original contour by the following steps:

1) Detect major dips in the amplitude envelope of the input speech waveform. Segment the F_0 contour at these dips and divide it into several periods.
2) For each period, find a recursive line with the least mean square error with respect to the original contour.
3) If the error averaged over the voiced part of the period exceeds a threshold (for the current experiment, 0.5 [(ln Hz)^2 sec]), proceed to the next step. If not, proceed to step 5).
4) Divide the period into two parts at the point with the maximum error, and return to step 2).
5) Proceed to the next input speech.

All dips in the macroscopic contour obtained by the procedure above were assigned to the boundary candidates and were denoted by G_i.

Further boundary candidates were detected from the original F_0 contour based on the following criteria. These candidates shall be denoted by L_i.

1) Positions where the derivative of F_0 contour changes its sign from negative to positive. With this criterion, dips in F_0 contour are detected.
2) Boundaries between unvoiced/silent regions and their preceding voiced regions when the derivative of F_0 contour is negative at the end of the preceding vowel region and is positive at the beginning of the succeeding vowel region.
3) Boundaries between unvoiced/silent regions and their preceding voiced regions when the derivative of the F_0 contour is negative at the end of the preceding voiced region, and the F_0 at the beginning of the succeeding voiced region is larger than that at the end of the preceding voiced region at least by a threshold (for the current experiment, 0.05 [ln Hz]).
4) Boundaries at the top and the end of unvoiced/silent regions equal to or longer than a threshold (say, 250 [ms]). Such regions are considered to be pauses.

Among the G_i and L_i candidates, the selection of the syntactic boundaries was conducted as follows:

1) From the G_i candidates, select 2nd step candidates which are also included in the L_i candidate group. Candidates with temporal discrepancy not exceeding a threshold (say, 45 [ms]) are considered to be identical.
2) From the second step candidates, select the initial syntactic boundary which is located apart from the beginning of the sentence at least by a threshold D.
3) From the second step candidates, select the i-th syntactic boundary which is located apart from the $(i-1)$-th syntactic boundary at least by the threshold D. This step is repeated until the end of the sentence.
4) From the G_i and L_i candidates not selected by step 3), select the syntactic boundaries located apart from all the selected boundaries at least by the threshold D.

2.2 Speech Material and Experimental Results

Experiments were conducted on the detection of syntactic boundaries to evaluate and improve the proposed method. The speech material used for the experiment was the continuous speech on the international conference registration (82 sentences) uttered by a male speaker (MAU), which is included in the ATR speech database. Its average speech rate is approximately 10 mora/sec, and it includes 279 syntactic boundaries.

An example of syntactic boundary detection is shown in Fig. 1. Vertical lines in the panel of waveform amplitude envelope indicate the dips used for the segmentation of F_0 contour to calculate the macroscopic contour. Among the detected boundaries b_1, b_2, b_3, b'_1, b_4 and b_5, b_2 was detected with one-mora temporal discrepancy from the corresponding segmental boundary. Although an additional post-processing is necessary, this kind of minor discrepancies can be corrected. Therefore, for the current experiment, boundaries with one-mora temporal discrepancies were treated as correctly detected. The boundary b'_1 is included only in the L_i candidate group, while the others are included in both the G_i and L_i candidate groups. The boundaries b_1, b_5 respectively indicate the beginning and the end of the sentence.

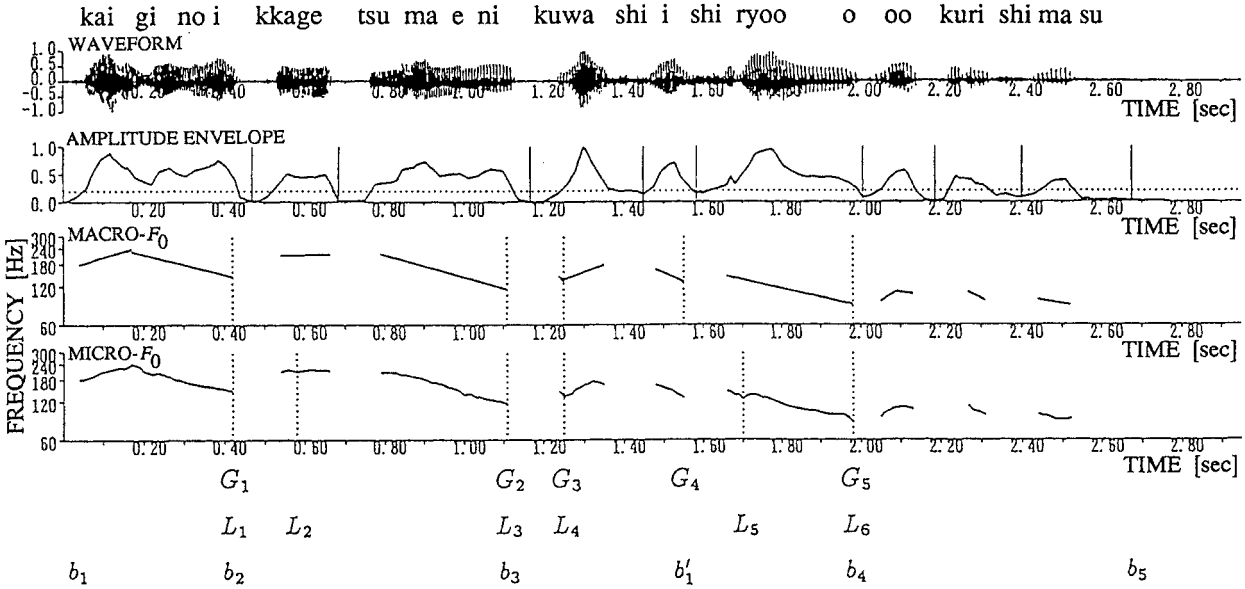


Fig. 1. An example of the syntactic boundary detection for the utterance "kaigino ikkagetsumaeni kuwashii shiryooo ookurishimasu." (Detailed materials will be sent out one month before the conference.)

In order to find an appropriate value for the threshold D, rates of deletion errors and insertion errors were calculated by changing D. As shown in Fig. 2, the rate of correct detection (defined by 1 minus the deletion error rate) decreases and the rate of insertion errors increases as D increases. Since the correct detection rate saturates when D is around 100 msec to 200 msec, for the current speech samples, 200 msec can be adopted as the most appropriate value for D. Using this value, the percentage of correctly detected boundaries to the boundaries of the total speech material (percentage rate of correct recognition) was 82.1 %, while the rate of insertion errors was 49.5 %. If the candidates L_i are neglected, these values decrease to 69.5 % and 19.4 %, respectively. This indicates the validity of the microscopic aspects of F_0 contours in decreasing deletion errors, but also indicates the defect of increasing insertion errors.

2.3 Reduction of Insertion Errors

In order to improve the performance of the method, the following modifications were applied to the proposed method taking the knowledge on F_0 contours into account:

1) The threshold distance D between two adjacent boundaries was decreased to 145 msec at the beginning of the sentence (distance between the beginning of the sentence and its succeeding boundary), while, at the end of the sentence (distance between the end of the sentence and its preceding boundary), it was increased to 300 msec.
2) If two candidates exist in a region not exceeding 120 msec, the one with smaller F_0 was selected. This modification was planned to reduce errors due to the left-to-right processing of the original method in selecting boundaries from candidates.

The following scheme was also adopted to decrease the insertion errors:

3) If an L_i candidate has an F_0 larger than the threshold function $L(t)$, it was excluded from the L_i candidate group. Figure 3 shows the threshold function which roughly represents the decreasing nature of the F_0 contour of a phrase.

As a result of the improvements above, the rate of insertion errors decreased to 30.5 % with no decrease in the rate of correct detection. The correct detection rate of 82.1 % does not seem so high, but we should note that, even by human experts, this rate can be increased only to at most 87 % because of the accent sandhi.

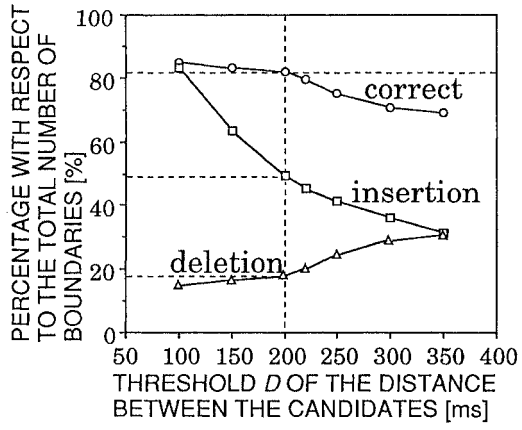


Fig. 2. Rates of correct recognition, insertion errors and deletion errors as functions of the threshold D between two adjacent syntactic boundaries.

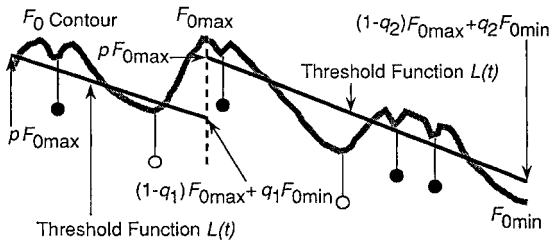


Fig. 3. Threshold function L(t) for an F_0 contour. Parameters p, q_1, q_2 were respectively set to 0.9, 0.6 and 0.85 for the current experiment. The boundary candidates indicated by filled circle were excluded from the L_i candidate group.

3. SELECTION OF THE RECOGNITION RESULT FROM SEVERAL CANDIDATES

3.1 Method of Partial Analysis-by-Synthesis

In order to generate an appropriate F_0 contour for the recognition candidate in segmental level, a functional model was adopted, where the F_0 contour in logarithmic frequency scale is represented by the sum of phrase and accent components [8]. These components are generated from impulse-like phrase commands and step-wise accent commands. A hill-climbing method, known as the analysis-by-synthesis (AbS) method, is widely used to find a combination of model parameters yielding the contour that best fits to the observed one. For the current purpose of the evaluation of recognition candidates, however, the best fitting search over the entire sentence will obscure the mismatching around the part with recognition errors. From this point of view, a scheme of partial AbS was developed where the best fitting search was conducted on the limited portion with recognition ambiguity. The base line value F_b and the parameters for the preceding phrase commands were set to those obtained by the command estimation process for the entire utterance [9]. Since unlimited search of parameter values may yield a good fitting for every candidate, the following constraints were put on the model parameters:

T_0 (onset of phrase command): 80 ± 20 msec before the corresponding segmental boundary,
T_1 (onset of accent command): 70 ± 50 msec before the corresponding segmental boundary,
T_2 (offset of accent command): 100 ± 50 msec before the corresponding segmental boundary,
α (natural angular frequency for the phrase control mechanism): 3.0 with ± 20 % tolerance,
β (natural angular frequency for the accent control mechanism): 20.0 with ± 20 % tolerance,
A_p (magnitude of phrase command): ± 20 % tolerance from the initial value,
A_a (amplitude of accent command): ± 20 % tolerance from the initial value.

3.2 Experiments and Results

With the scheme of partial AbS, the degree of discrepancy between the model-generated contour and the observed contour was calculated for each recognition candidate. The candidate giving the minimum discrepancy was selected as the final recognition result. In order to show the validity of the method in finding out correct syntactic boundaries, a small experiment was conducted for the following two speech samples uttered by a male speaker (HK):

S1: /umigameno maeni hirogaru/ (Stretching in front of a

turtle.)
S2: /keQsekishita kuniNno tamedesu/ (It is for the nine who were absent.)

These utterances can easily be wrongly recognized as follows:

S1': /umiga menomaeni hirogaru/ (The sea is stretching out in front of our eyes.)
S2': /keQsekishi kakuniNno tamedesu/ (Being absent. This is for the confirmation.)

The second example includes a recognition error (/ta/ ⇒ /ka/) in the phonemic level, while the first example includes no errors. For the speech sample S1, the partial AbS was conducted for the part /menomaeni/ with the following two hypotheses:

Hypothesis 1: An additional phrase command at the beginning of the part.
Hypothesis 2: No additional phrase command.

The hypotheses 1 and 2 correspond respectively to the wrong recognition and to the correct recognition in the speech sample S1. For the speech sample S2, the partial AbS was also conducted for the parts /takuniNno/ and /kakuniNno/ with the hypotheses 1 and 2 above. The discrepancies between the observed contours and the model-generated contours are shown in Fig. 4. The smaller discrepancy for hypothesis 2 of speech sample S1 indicates that the final recognition result should be /umigameno maeni hirogaru/. As for the speech sample S2, the discrepancies for the two hypotheses are rather large, though hypothesis 1 has a smaller value. For this case, the discrepancy was further calculated for the following hypothesis:

Hypothesis 3: An additional phrase command inside of the part.

As shown in Fig. 4, the discrepancy takes a small value for hypothesis 3, indicating the correct recognition being /keQsekishita kuniNno tamedesu/.

The proposed method is also valid in detecting recognition errors causing changes in accent types. A small experiment was also conducted to show the validity. Four short sentence utterances were recorded for each of the following cases:

Case 1: Recognition error causing the accent type change from type B to type A.
Case 2: Recognition error causing the accent type change from type C to type A.
Case 3: Recognition error causing the accent type change from type A to type B.

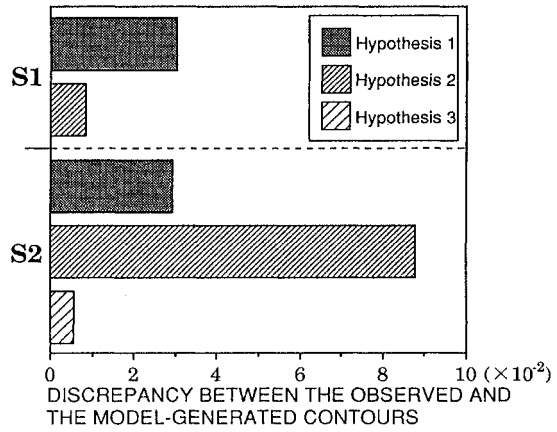


Fig. 4. Discrepancies between the observed and the model-generated contours (partial AbS errors) for samples S1 and S2 with hypotheses 1, 2 and 3.

Case 4: Recognition error causing the accent type change from type A to type C.

Here, types A, B, C respectively denote accent types without a rapid downfall in F_0, with a rapid downfall at the end of the first mora, with a rapid downfall at the end of the second mora or of one of the succeeding ones. For instance, /oHkuno gaikotsuo mita/ (Saw many skeletons.) of the example of case 2 can be recognized wrongly as /oHkuno gaikokuo mita/ (Saw many countries.). The partial AbS was conducted for the underlined part. The partial AbS error was calculated for each hypothesis with the recognition error and normalized by the error for the corresponding hypothesis of correct recognition. The normalized errors exceeded one for all the speech samples, indicating the validity of the proposed method. The values, however, are rather small for the cases 1 and 2, and may lead to a misjudgement. Fine alignment in the restrictions on the model parameters is necessary to cope with the problem.

4. CONCLUSION

Two methods were proposed for the use of prosodic features in continuous speech recognition. One is for the detection of syntactic boundaries of input speech and the other is for the selection of the correct recognition result out of several recognition candidates. Although their validity were experimentally shown, the following studies are also planned: 1) to reduce insertion errors in the former scheme using the F_0 contour model, 2) to incorporate the concept of boundary likelihood, 3) to increase the performance of the partial AbS method, and 4) to incorporate the developed methods in recognition systems.

REFERENCES

[1] N. Minematsu and K. Hirose, "Role of prosodic features in the human process of speech perception," Proc. ICSLP 94, 21.8 (1994-9).
[2] Y. Suzuki, Y. Sekiguchi and M. Shigenaga, "Detection of phrase boundaries using prosodics for continuous speech recognition," Trans. IEICE, Vol.J72-D-II, No.10, pp.1609-1617 (1989-10).
[3] E. Oohira, A. Komatsu and A. Ichikawa, "Structure inference algorithm of conversational speech sentence using prosodic information," Trans. IEICE, Vol.J72-A, No.1, pp.23-31 (1989-1).
[4] S. Okawa, T. Endo, T. Kobayashi and K. Shirai, "Phrase recognition in conversational speech using prosodic and phonemic information," IEICE Trans. Inf. & Syst., Vol.E76-D, No.1, pp.44-50 (1993-1).
[5] H. Konno and K. Hirose, "Detection of syntactic boundaries using prosodic information," Technical Report of IEICE, SP93-122, pp.31-38 (1994-1).
[6] E. Geoffrois, "A pith contour analysis guided by prosodic event detection," Proc. EUROSPEECH 93, Vol.2, pp.793-796 (1993-9).
[7] K. Hirose, H. Fujisaki and S. Seto, "A scheme for pitch extraction of speech using autocorrelation function with frame length proportional to the time lag," Proc. ICASSP 92, Vol.I, pp.149-152 (1992-3).
[8] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," J. Acoust. Soc. Jpn. (E), Vol.5, No.4, pp.233-242 (1984-10).
[9] H. Fujisaki, K. Hirose, H. Fujisaki and S. Seto, "A study on automatic extraction of characteristic parameters of fundamental frequency contours," Rec. Fall Meeting, Acoust. Soc. Jpn., 2-6-18, pp.255-256 (1992-3).