

# USE OF PROSODIC FEATURES IN SPEECH RECOGNITION

*Keikichi Hirose\**

*Kouji Iwano\**

*Atsuhiko Sakurai\*\**

\*Dept. of Information and Communication Engineering  
School of Engineering, Univ. of Tokyo  
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

\*\*TRDC, Texas Instruments, Tsukuba, 305 Japan

## ABSTRACT

Two methods were proposed for the use of prosodic features in speech recognition: one to detect major syntactic (phrase) boundaries as the initial phase of speech recognition, and the other to check the feasibility of the results of ordinary recognition process from the viewpoint of prosodic features. In the first method, fundamental frequency contours were assumed as waveforms as functions of time and were low-pass filtered to suppress accent components in the contours. Then the derivative of filtered contour was used to detect phrase boundaries. An experiment was conducted on the ATR continuous speech database, showing that the method managed to detect about 77% of manually detectable phrase boundaries. The second method is based on generating fundamental frequency contours for recognition candidates using a speech synthesis scheme and comparing them with the observed contour. The candidate giving the best matched contour to the observed contour should be the final recognition result. The method was shown to be valid in detecting recognition errors accompanied by changes in accent types or/and in syntactic boundaries. The method was then evaluated in its performance for the detection of phrase boundaries. Allowing 1-mora discrepancies, the detection rate reached 92% for the ATR database, which was further improved to 97% by a simple speaker adaptation method.

## 1. INTRODUCTION

Prosodic features of speech, as major features characterizing spoken language, are closely related with various kinds of linguistic and non-linguistic information, such as word meaning, syntactic structure, discourse structure, speaker's intention and emotion, and so on. In human speech communication, therefore, they are playing an important role in the transmission of information. In current speech recognition methods, however, their use is rather limited even for linguistic information. Although the hidden Markov modeling has been successfully introduced in the speech recognition yielding rather good results only with segmental features, prosodic features also need to be incorporated for the further improvement.

Different from the case of segmental features, use of segmental features in speech recognition should be supplementary. Since prosodic and linguistic features belong to two different aspects of language, respectively spoken and written language, they do not bear a tight relationship. For instance, a major syntactic boundary (in written language) does not necessarily correspond to a major prosodic boundary (in spoken language). Therefore, in incorporating prosodic features in the speech recognition process, this

factor should be taken into consideration.

In continuous speech recognition, unless the perplexity of the recognition task is small, a lot of computation is required for the searching process in the linguistic level, still not yielding a good result. Information on syntactic structures is thought to be effective in order to improve recognition performance, when utilized as the constraints for the searching process. From this point of view, several works have been conducted on the prosodic segmentation. In an early attempt for conversational speech of Japanese, a pattern matching method was applied to fundamental frequency contours ( $F_0$  contours) to find out syntactic structures[1, 2]. Simultaneous use of microscopic and macroscopic features of  $F_0$  contours was shown to yield a good result for syntactic boundary detection[3].

Although the above methods can detect syntactic boundaries rather well, they can give us only a limited information on the type of boundaries, which is considered to be very useful for speech recognition. This is because, they are mostly based on a rude assumption that a deeper dip in  $F_0$  contours will correspond a deeper syntactic boundary. An  $F_0$  contour model should be utilized which can qualitatively classify the syntactic boundaries. The superpositional model for  $F_0$  contour generation[4] is considered to be appropriate for the purpose, where the classification is possible by the existence/absence and magnitude of phrase command at a boundary. This model was already utilized to make constraints in the matching process of  $F_0$  contours[5], and in an attempt to detect prosodic events by means of a left-to-right algorithm[6]. However, the most direct use of this model should be the analysis-by-synthesis method, where the model parameters were searched in the process of adjusting the model generated contours to the observed contour. The major problem of this idea is that the analysis-by-synthesis process requires a good set of initial parameter values, which is usually rather hard to be given without manual aid. Since the undulations due to phrase components are less dominant in  $F_0$  contours as compared to those due to accent components, automatic estimation of parameter values is especially hard for the phrase components, though they are useful to classify the syntactic boundaries as mentioned above.

From this viewpoint, we have developed a method to detect phrase components (and thus phrase boundaries) by suppressing the accent components using a low-pass filtering[7]. This method will be explained in section 2.

In most of the above methods, prosodic features were planned to be used in a supplement process before the main recognition process. Therefore, syntactic boundaries were detected only by the prosodic features, without referring to the recognition results obtainable from the main process. Accurate boundary detection using only of prosod-

ic features may not be always possible at least in the current technologies. This consideration led us to a new method to use prosodic features during the main process of recognition[3, 8]. Section 3 will be allocated for the detailed explanation of the method.

## 2. DETECTION OF PHRASE BOUNDARIES BY LOW-PASS FILTERING OF FUNDAMENTAL FREQUENCY CONTOURS

The method consists basically of the following two steps: step for pre-processing and that for event and boundary detection. The pre-processing consists of filtering the  $F_0$  contour using a simple low-pass filter and calculating the derivative of the filtered curve. The second step, event and boundary detection, consists of detecting temporal events that denote the occurrence of a phrase component and calculating the phrase boundary positions in the speech waveform.

### 2.1. Pre-Processing

After the pitch extraction process based on the calculation of an autocorrelation function with the frame length proportional to the time lag[9], a continuous  $F_0$  contour is obtained by linearly interpolating unvoiced segments of the contour.

The Butterworth filter was selected for the low-pass filtering due to its relatively flat phase characteristics. In order to satisfactorily eliminate the accent component of the  $F_0$  contour, the cut-off frequency should be on the order of 1 to 2 Hz, as reported by Ström[10]. However, if the cut-off frequency were set within this range, the phrase component would also be greatly affected by the filtering process, making the process of phrase boundary detection more difficult. Here, we opted to select a higher cut-off frequency, leaving vestiges of accent components on the  $F_0$  contour. The characteristics of the filter utilized in the experiments are: 3rd Butterworth filter with pass-band frequency of 2 Hz and stop-band frequency of 7 Hz.

As the final process of the pre-processing, the derivative of the filtered curve is calculated by taking the slope of the minimum square error line across the 3 successive points..

### 2.2. Event and Boundary Detection

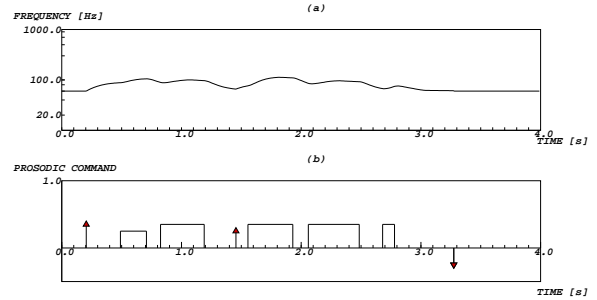
In principle, the derivative of an  $F_0$  contour with no accent component would be a constantly negative curve, becoming positive only at the verge of a new phrase command. Therefore, some of the events that could be used as indicators of new phrase components in an  $F_0$  contour are:

- (1) Negative-to-positive transitions of the derivative (zero-crossings);
- (2) A threshold involving the value of the derivative curve;
- (3) A threshold involving the average of the derivative over a moving window instead of the instantaneous value.

In order to find the most suitable event for the purpose of our method, we performed experiments using model-generated  $F_0$  contours containing an inner phrase boundary (a phrase boundary not located at the beginning of the sentence), varying the magnitude of the phrase command and the amplitude of the subsequent accent command[4]. We generated  $F_0$  contours for the following utterance: "sochirano kokusaikaigini / roNbuNo tookooshitaito omouNdesuga." (I'd like to submit a paper to this international conference.) The slash '/' represents the syllabic position of

the phrase boundary in question. The  $F_0$  contours for the utterance above were generated using rules for speech synthesis[11], and both the magnitude of the inner phrase command (phrase command corresponding to the phrase boundary in question) and the amplitude of the accent command were varied in the interval from 0.1 to 0.6.

Figure 1 illustrates the case when the magnitude of the inner phrase command and the amplitude of the subsequent accent command are 0.35. The generated curve is shown in (a), and the  $F_0$  model commands, in (b). The vertical line in (a) represents the segmental position of the phrase boundary. It should be noted that no other command besides the phrase command in question and the following accent command was varied.

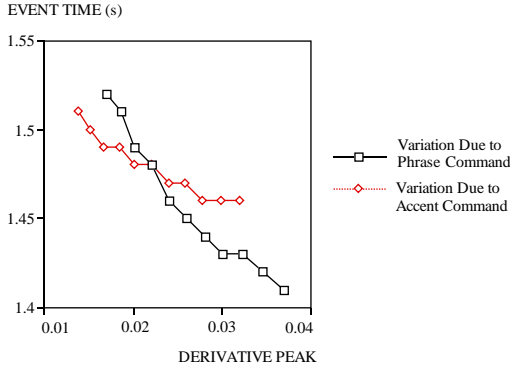


**Figure 1.** Example of model-generated  $F_0$  contour with the inner phrase command magnitude 0.35 and the subsequent accent command amplitude 0.35. The corresponding phrase boundary (vertical bar) is located at 1.5 s.

For each case, events (1), (2) and (3) were detected and plotted against the derivative peak corresponding to the boundary in question. The value of the derivative peak is searched within the interval beginning 200 ms before the event, ending 400 ms after. From the results, it was noted that event (3) is the most suitable to detect the position of the phrase command, since the relation between event (3) and the derivative peak was almost linear within the observed range of variation of the magnitudes of the phrase command. Moreover, the variation of the magnitudes of the derivative peak due to the variation of the amplitude of the accent command, when plotted against event (3), showed a similar behavior as when the magnitude of the phrase command was varied, which means that the influence of accent components can be easily handled using the same approach as for phrase components. The obtained curve can be seen in Figure 2.

In the current method, therefore, phrase boundaries are determined using event (3) (when the average of the derivative of the filtered  $F_0$  contour over a moving window exceeds a threshold) and the derivative peak at the boundary in question.

Using the average over a moving window instead of the instantaneous value seems logical in the present method. Since accent components are not eliminated completely, local rises corresponding to onsets of accent commands can result in negative-to-positive transitions of the derivative curve even when no phrase command is present. By using an averaging process, though, the algorithm becomes sensitive only to significant rises of the  $F_0$  contour. The sensitivity of the search process can be controlled by the width of the moving window: a narrow window results in a highly sensitive algorithm, and a wide window makes the trigger occur only at long and substantial rises of the derivative. Here,



**Figure 2.** Derivative peak near the event as a function of the event time.

after realizing the filtering and derivation process on a set of 25  $F_0$  contours (the same  $F_0$  contours will be later utilized in evaluation experiments), and based on various observed values of widths of derivative peaks, we adopted a window width of 300 ms. As to the threshold, a phrase boundary is detected whenever the mean value of the derivative peak in the window exceeds  $7 \times 10^{-3} \text{ s}^{-1}$ .

Once the event is detected, the relation between the instant of occurrence of the event (*event time*) and the actual position of the phrase boundary (*boundary position*) should be investigated. From Figure 2, we derived a linear relationship between the *event time* and the *derivative peak*, allowing us to find the deviation of *event time* from the actual phrase boundary.

$$B = ET + 4.77 \times DP - 5.64 \times 10^2 \text{ [s]},$$

where  $B$  represents the location of phrase boundary,  $ET$  is the *event time* and  $DP$  is the *derivative peak*.

### 2.3. Experimental Results

A set of experiments were realized using  $F_0$  contours extracted from the ATR continuous speech corpus on conference registration[12]. We selected 25 speech samples, which were uttered by the male speaker MAU with an approximate speech rate of 10 morae/s. We extracted their  $F_0$  contours and determined the phrase commands of their underlying models of section 3.1 by the analysis-by-synthesis[4]. Here, we did not deal with phrase boundaries occurring after pauses longer than 500 ms, which can be more easily detected using the temporal information of the pause.

Next, we applied the proposed method of automatic detection of phrase boundaries to the same set of  $F_0$  contours in order to compare the results. The results of the experiments are summarized in Table 1. From the table, it can be seen that the method was able to find approximately 77% of the manually detectable phrase boundaries, with an insertion rate comparable to the deletion rate.

**Table 1.** Experimental result showing numbers of detected boundaries and insertions.

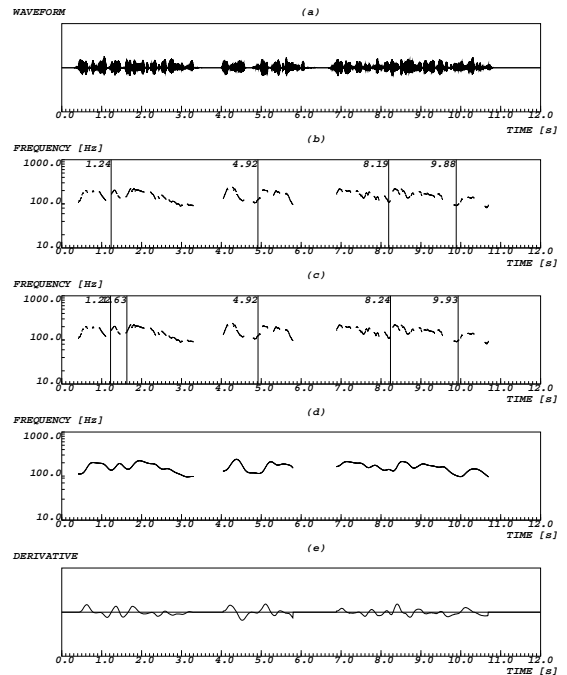
Inner Phrase Boundaries	56
Detected Boundaries	43
Insertions	14

In the next set of experiments, we evaluated the method as to the deviation of the detected boundaries with respect to the actual boundaries in the speech waveform. We selected, among the phrase boundaries used in the previous experiments, 40 phrase boundaries that actually correspond to major syntactic boundaries, and determined the deviation of the detected boundaries with respect to their correct positions, in terms of number of morae. The results are described in Table 2.

**Table 2.** Experimental result with respect to the deviation in the detected boundary position.

Inner Phrase Boundaries		40	
Detected Boundaries	No Deviation	15	35
	1-mora Deviation	18	
	2-morae Deviation	2	

Figure 3 shows an example of phrase boundary detection using the proposed method. The content of the utterance is: "daimokutoshitewa kanari koohaNna kotoga kakaretearuN-desuga yoosuruni jidootsuuyakudeNwadesuka sorenikaNsuru gijutsudeshitara naNdemo yoroshiitoui kaishakude yoroshiiwake desune." (As to the title, they cover a wide variety of topics, but in short, if the theme has any relation with automatic interpreting telephone, or any related technological issues, then it will be fine, won't it?)



**Figure 3.** An example of phrase boundary detection. (a) speech waveform, (b) correct phrase boundaries (vertical lines), (c) detected phrase boundaries, (d) filtered  $F_0$  contour and (e) derivative of filtered contour.

### 2.4. Considerations

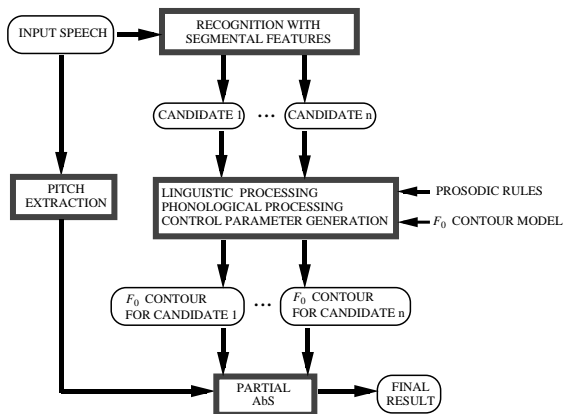
In the present method, the experiments showed that eliminating insertions is an important issue that remains to be solved. (In the example shown in Figure 3, for instance,

there is an insertion at 1.63 s.) It was also noted that most of the insertions occurred when two phrase boundaries were detected close to each other (within an interval of less than 1.0 s.)

The main reason for the insertion errors is the fact that the accent components are not completely compressed by the low-pass filter. Lowering the cut-off frequency would reduce insertion errors, but with the serious expense of lowering the detection rate. In order to deal with this problem, it is possible to use the partial analysis-by-synthesis scheme in the next section, by setting various recognition hypotheses with and without a phrase command.

### 3. METHOD OF POSTERIOR USE OF PROSODIC FEATURES

The proposed method is based on generating  $F_0$  contours for recognition candidates obtained by the ordinary recognition process, and then comparing them with the observed contour. The candidate with smallest mismatch should be the correct recognition result. Generation of  $F_0$  contours are conducted using prosodic rules for speech synthesis, which are a simplified version of those formerly constructed for a text-to-speech conversion system[11]. Because of considerably large utterance-to-utterance or speaker-to-speaker variations in the observed  $F_0$  contours, a mere comparison between the generated contour and the observed contour could yield a large distance even for the correct recognition candidate. Therefore, before calculating the distance, the generated contour is adjusted in a limited extent to the observed contour by a scheme of partial analysis-by-synthesis. Figure 4 shows the total configuration of the method.



**Figure 4.** Total configuration of the method for finding correct recognition result from several candidates.

The proposed method is considered to be valid to detect recognition errors, to ensure recognition results and, thus, to realize an effective search in continuous speech recognition. The method can also be used in conjunction with the above methods for the detection of syntactic boundaries: to check if an extracted boundary is correct or to identify the type of boundary.

#### 3.1. Superpositional Model of $F_0$ Contours

The prosodic rules are based on a functional model for  $F_0$  contour generation[4]. This model represents an  $F_0$  contour in logarithmic scale of frequency as a superposition of phrase and accent components on a baseline. The phrase components and the accent components are respectively

considered to be generated from impulselike phrase commands and step-wise accent commands, which are known to have good correspondences respectively with the syntactic structure and the lexical accent. The generation processes of phrase and accent components from their corresponding commands are represented by critically damped second order linear systems.

#### 3.2. Partial Analysis-by-Synthesis

The method of analysis-by-synthesis based on the "hill-climbing" algorithm is widely used to find a combination of model parameter values yielding the contour that best fits the observed one. Although, in the case of  $F_0$  contours, the best fitting search is usually done over the entire utterance unit (prosodic sentence) delimited by respiratory pauses, for the current purpose of evaluating recognition candidates, this procedure should be difficult if the unit includes several portions with recognition ambiguity. Even if it is possible, it will obscure the mismatch due to recognition errors. From this point of view, a new scheme of partial analysis-by-synthesis was developed, where the best fitting search was conducted only on the limited portion with recognition ambiguity. The distance between generated and observed contours is given as the analysis-by-synthesis error per a frame averaged over the voiced part of the portion.

In order to start the analysis-by-synthesis process, a set of initial values are required for model parameters. It is given by the prosodic rules for speech synthesis, as mentioned already. Table 3 shows the command values assigned to the phrase and accent symbols, which will serve as the initial values for the analysis-by-synthesis process[11]. Since, in Standard Japanese, word accent of type 0 without rapid downfall in its  $F_0$  contour shows rather different features in prosodic rules as compared to other accent types with rapid downfall, accent symbols are prepared differently for type 0 accent (FH, FM and FL) and others (DH, DM and DL). The table also shows the initial positions of commands with respect to the voice onset of corresponding syllable. Initial values for the natural angular frequency of phrase control mechanism and that of accent control mechanism are set respectively to  $3.0 \text{ s}^{-1}$  and  $20.0 \text{ s}^{-1}$ . Value of the baseline component was decided so that the model-generated contour had the same average (on logarithmic frequency) with the observed contour.

Although, in the scheme of partial analysis-by-synthesis, the best fitting search is conducted only on the limited portion, it may possibly be affected by the phrase components generated prior to the portion. Therefore, proper assignment of the preceding phrase components is important for the performance of the method. With the prosodic rules, symbol P1 is usually assigned at the top of a prosodic sentence, viz., immediately after a respiratory pause. When the prosodic sentence starts with a conjunction word, symbol P1 is changed to symbol P2 with additional symbol P1 after the word. At syntactic boundaries in a sentence, symbol P2 or P3 will be further assigned. Although in the original prosodic rules, selection of P2 or P3 is done with the information of the depth of syntactic boundary, for the proposed scheme, only the number of morae from the adjacent phrase command is counted. If more than two phrase commands are assigned before the portion of partial analysis-by-synthesis, they cannot be searched separately by the scheme. In the proposed scheme, only the closest command to the portion is included in the searching process and the other commands are left unchanged. Since a phrase component comes to almost zero in several morae

**Table 3.** Command values and positions assigned to the phrase and accent symbols in the prosodic rules. These will serve as the initial parameter values for the process of analysis-by-synthesis.

Type	Symbol	Command Magnitude/Amplitude	Position with Respect to Voice Onset (ms)
Phrase Symbol	P1	0.35	-210
	P2	0.25	-80
	P3	0.15	-80
	P0	(reset)	-80
	Accent Symbol	FH	0.50
	FM	0.25	-70
	FL	0.10	-70
	DH	0.50	-70
	DM	0.35	-70
	DL	0.15	-70
	A0	(reset)	-70

due its declining feature, this simplification is considered to affect the result only by little.

In the original analysis-by-synthesis method, search of parameter values is conducted within a wider range of parameter space. This process may possibly yield similar contours for different recognition candidates and, therefore, may give the best fitting even for a wrong candidate. To cope with this problem, the searching space need to be limited to a smaller range. For the current experiment, all of the command positions were searched in the range between  $-20$  ms and  $+20$  ms from the initial values, while the command magnitudes/amplitudes and the natural angular frequencies were searched in the range of  $\pm 20\%$  of the initial values.

### 3.3. Detection of Recognition Errors

The proposed method is considered to be valid for the detection of recognition errors causing changes in the accent types or/and in syntactic boundaries[3, 8]. In order to show this point, several experiments have been conducted. As for the accent type changes, utterances of four short sentences were recorded for each of the following cases:

**Case 1:** Recognition error causing accent type change from type N to type 0,

**Case 2:** Recognition error causing accent type change from type 1 to type 0,

**Case 3:** Recognition error causing accent type change from type 0 to type N,

**Case 4:** Recognition error causing accent type change from type 0 to type 1.

Here, types 1 and N respectively denote accent types with a rapid downfall in the  $F_0$  contour at the end of the first mora and with a rapid downfall at the end of the second mora or of one of the succeeding ones. The term "type N" was defined temporarily in this paper to denote accent types other than types 0 and 1. For each of cases 1 to 4, sentences U1, U2, U3 and U4 in Table 4 were adopted in the experiment. For each of these sentences, a phoneme recognition error was assumed in one of consonants of the

underlined prosodic word, producing a different sentence (such as U1', U2', U3' and U4'), and making the accent type of the word change according to one of the cases 1 to 4. For instance, "ookuno ga'ikotsuo mita" (U1 of case 2) was assumed to be wrongly recognized as "ookuno gaikokuo mita" (U1' of case 2) with accent type change from type 1 to type 0. The partial analysis-by-synthesis was conducted on the underlined portions. Figure 5 shows the results of the experiment for utterances of a male speaker of the Tokyo dialect. For every utterance, a smaller error was obtained for the correct result, indicating the validity of the proposed method. However, in several utterances, the errors were rather large for correct results, and, conversely, were rather small for wrong results. Fine alignment in the restrictions on the model parameters should be further necessary.

**Table 4.** Sentences used for the experiment on the detection of recognition errors accompanied by the changes in accent type.

Case 1	U1	"higa <u>toppu'ri</u> kureta" (The sun set completely.)	
	U1'	"higa <u>tokkuri</u> kureta" (The sun set 'tokkuri.') (semantically incorrect)	
	U2	"ishani <u>kaka'tte</u> iru" (I'm under a doctor's care.)	
	U2'	"ishani <u>katatte</u> iru" (I'm talking to a doctor.)	
	U3	"anokowa <u>uchi'wao</u> motteita" (She had a fan.)	
	U3'	"anokowa <u>ukiwao</u> motteita" (She had a swim ring.)	
	U4	"sorewa <u>fuko'odato</u> omou" (I think it is unhappy.)	
	U4'	"sorewa <u>futoodato</u> omou" (I think it is unfair.)	
	Case 2	U1	"ookuno <u>ga'ikotsuo</u> mita" (I saw many skeletons.)
		U1'	"ookuno <u>gaikokuo</u> mita" (I saw many foreign countries.)
U2		"kareo <u>ka'nkokuni</u> maneita" (I invited him to Korea.)	
U2'		"kareo <u>kantokuni</u> maneita" (I invited him as a supervisor.)	
U3		"tookuni <u>go'oruga</u> mieta" (I saw the goal far away.)	
U3'		"tookuni <u>booruga</u> mieta" (I saw the ball far away.)	
U4		"ichiban <u>ko'kuna</u> yarikatada" (It is the most cruel way.)	
U4'		"ichiban <u>kotsuna</u> yarikatada" (It is the most 'kotsuna' way.) (semantically incorrect)	
Case 3	U1	"ishani <u>katatte</u> iru" (I'm talking to a doctor.)	
	U1'	"ishani <u>kaka'tte</u> iru" (I'm under a doctor's care.)	
	U2	"anokowa <u>ukiwao</u> motteita" (She had a swim ring.)	
	U2'	"anokowa <u>uchi'wao</u> motteita" (She had a fan.)	
	U3	"sorewa <u>futoodato</u> omou" (I think it is unfair.)	
	U3'	"sorewa <u>fuko'odato</u> omou" (I think it is unfair.)	
	U4	"hisokani <u>kitaio</u> joseru" (To expect secretly.)	
	U4'	"hisokani <u>kika'io</u> joseru" (To bring a machine closer in secret.)	
Case 4	U1	"ookuno <u>gaikokuo</u> mita" (I saw many foreign countries.)	
	U1'	"ookuno <u>ga'ikotsuo</u> mita" (I saw many skeletons.)	
	U2	"kareo <u>kantokuni</u> maneita" (I invited him as a supervisor.)	
	U2'	"kareo <u>ka'nkokuni</u> maneita" (I invited him to Korea.)	
	U3	"kanojono <u>koppuni</u> tsugu" (To pour into her cup.)	
	U3'	"kanojono <u>to'ppuni</u> tsugu" (To be second to her.)	
	U4	"tookuni <u>booruga</u> mieta" (I saw the ball far away.)	
	U4'	"tookuni <u>go'oruga</u> mieta" (I saw the goal far away.)	

As for the syntactic boundary changes, an experiment was conducted for the following two speech samples:

**S1:** "umigameno maeni hirogaru" (Stretching in front of a turtle.)

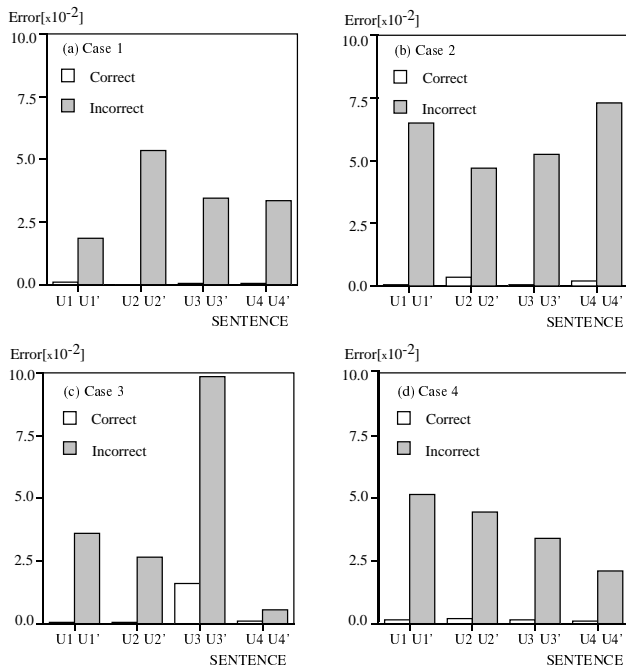
**S2:** "kessekishita kuniNno tamedesu" (It is for the nine who were absent.)

Due to an error in detecting morpheme boundaries (S1) or due to a phoneme recognition error of /ta/  $\Rightarrow$  /ka/ (S2), these utterances can be wrongly recognized as follows:

**S1':** "umiga menomaeni hirogaru" (The sea is stretching out in front of our eyes.)

**S2':** "kessekishi kakuniNno tamedesu" (Being absent. This is for the confirmation.)

The portion subject to partial analysis-by-synthesis was chosen so as to begin at the earliest syntactic boundary in question, ending 5 morae later. In the current cases, the



**Figure 5.** Partial analysis-by-synthesis errors for the utterances of cases 1 to 4 with correct and incorrect hypotheses on the accent types.

portion "menomaeni" was selected for S1 and the portion "takuniNno" for S2. According to the prosodic rules, additional phrase components (phrase components generated by symbols P2 or P3) occur at  $F_0$  contours corresponding to major syntactic boundaries (phrase boundaries). In the experiment, the following three cases were assumed as the possible hypotheses for the additional phrase component:

**H1:** a phrase command (of additional phrase component) immediately before the portion,

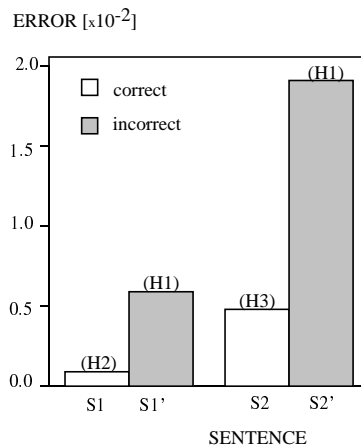
**H2:** no phrase command around the portion,

**H3:** a phrase command inside of the portion, viz., between "umigameno" and "maeni" for S1 and between "kessekishita" and "kuniNno" for S2.

Hypothesis H1 corresponds to the results S1' and S2' of the incorrect recognition. Although both of hypotheses H2 and H3 were assumed as the  $F_0$  contours for the correct recognition, hypothesis H1 agreed with the prosodic rules for S1, while hypothesis H2 agreed with those for S2. Distances between observed contours and model-generated contours are shown as errors of partial analysis-by-synthesis in Figure 6. In both samples, smaller distances were observed for the correct recognition, viz., hypothesis H2 for S1 and hypothesis H3 for S2, indicating that the final recognition results can be correctly selected from several candidates using prosodic features.

### 3.4. Detection of Phrase Boundaries

Although the proposed method need to be finally evaluated after being incorporated in segmental-based recognition systems, its performance was tested in the detection of phrase boundaries. This is because information on phrase boundaries is very useful as the constraints in the recognition process, but their correct detection is sometimes quite difficult using only of prosodic features. Assuming that phrase boundary positions had been shifted by one or two



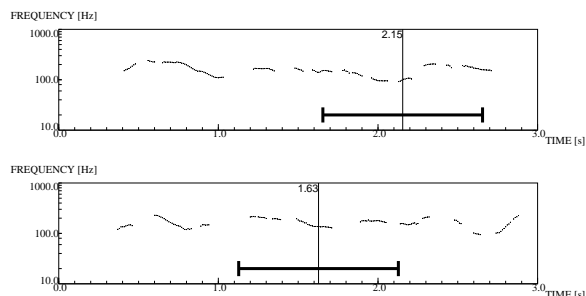
**Figure 6.** Partial analysis-by-synthesis errors for samples S1 and S2 with hypotheses of correct and incorrect recognition.

morae due to recognition errors (one of the hardest conditions for the method) the proposed method was evaluated as to whether it could detect such shifts[13]. The evaluation was conducted using the same ATR continuous speech corpus in section 2.3. First, major phrase boundaries were selected manually from the written text of the corpus, and, then, for each selected boundary, the existence of a phrase command was checked for the observed  $F_0$  contours by the original analysis-by-synthesis process. The experiment was conducted for phrase boundaries actually accompanied by phrase commands. We excluded phrase boundaries with long pauses (longer than 350 ms for the current experiment). Consequently, 37 phrase boundaries were selected for the experiment.

Unlike the previous section, the portion to be subject to the partial analysis-by-synthesis was automatically set as the period of 1 s with the initial position of the command for correct recognition at the center. Figure 7 shows the positions for the following two speech samples:

**Q1:** "koozabangooo shiteeshiteitadakereba / jidootekini hikiotosaremasu." (If the banking account is specified, the charge will be automatically subtracted.)

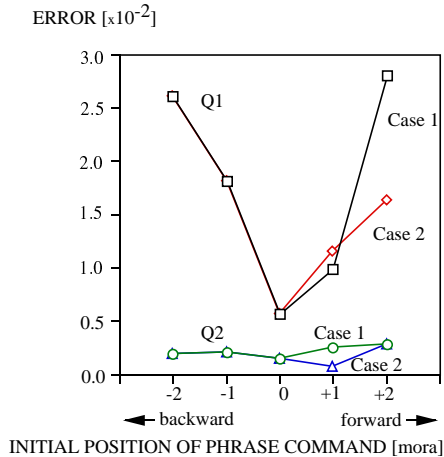
**Q2:** "mochiron happyoonotokimo / nippongode yoroshiinodesune?" (Naturally, we can make the presentation also in Japanese, can't we?)



**Figure 7.** Portions of partial analysis-by-synthesis for two sentence speech samples Q1 (upper) and Q2 (lower), indicated by thick horizontal bars. The vertical lines in the  $F_0$  contour indicate the locations of phrase boundaries.

The slashes '/' indicate the location of phrase boundaries. Besides the case of correct recognition, the partial analysis-by-synthesis was conducted after shifting the initial position of the phrase command backward and forward, by one and two morae. When shifting the phrase command forward, we had to note some peculiarities of Japanese accentuation. In standard Japanese, an  $n$ -mora word is uttered only in one out of  $n + 1$  accent types, although the  $F_0$  contour could be formed by  $2^n$  combinations of high and low constituent morae. As a result, if the first mora of a word has a high  $F_0$  contour, the following morae should have a low  $F_0$  contour. This accent type is denoted by type 1, as already mentioned in section 3.3. On the contrary, if the first mora has a low  $F_0$  contour, the second mora must have a high  $F_0$  contour. Therefore, if the accent type of the first word in a phrase is originally non-type 1, after a one-mora forward shift of the phrase command, it can still remain non-type 1 with a one-mora forward shift of the onset of the accent command (case 1) or it should be transformed into type 1 with no shift in the onset of the accent command (case 2). When the original accent type was of type 1, it was left unchanged, with a one-mora forward shift on both the onset and end of the accent command.

Figure 8 shows the results of the partial analysis-by-synthesis for the speech samples Q1 and Q2. The horizontal axis of the figure indicates the positions of assumed phrase boundaries represented by the number of morae with respect to the correct boundary location. The results for these two samples indicate two extreme cases: the first one, when the boundary is detected correctly at the right position and the second one, when the correct detection is quite difficult. The exact detection of phrase boundaries became difficult when the portion of partial analysis-by-synthesis included long voiceless parts and/or the magnitude of the phrase command was small.



**Figure 8.** Partial analysis-by-synthesis errors for the sentence speech samples Q1 and Q2 as functions of the initial position of phrase command. Two hypothesis were considered when phrase boundary was shifted forwards: case 1 and case 2.

In all, out of 37 phrase boundaries for the experiment, 15 boundaries were correctly detected without deviation, and, allowing the maximum deviation of 1 mora, 34 boundaries were detected. The results showed that the proposed method is valid in detecting recognition errors causing more than 2-mora deviations in the phrase boundary.

By inspecting the  $F_0$  contours that were adjusted by the

partial analysis-by-synthesis, the value of the bias level was found out excessively high. This is an indication that the initial values of the phrase and accent commands were too small for the speech samples used in the experiment. To cope with this problem, partial analysis-by-synthesis was conducted again after increasing initial amplitudes for accent commands. Concretely, each command amplitude was multiplied by a factor of 1.24, based on the results of original analysis-by-synthesis for several utterances of the speaker. The number of correct detection with maximum deviation of 1-mora increased to 36, while that without deviation decreased to 14. The results are summarized in Table 5.

**Table 5.** Results of phrase boundary detection with the proposed method and the conventional methods 1 and 2.

	Proposed Method		Method 1	Method 2
	Original Results	With Speaker Adaptation		
Correct Search	15/37 (41%)	14/37 (38%)	15/37 (41%)	29/37 (78%)
Correct Search Admitting 1-Mora Errors	34/37 (92%)	36/37 (97%)	31/37 (84%)	36/37 (97%)

Although, because of the complete difference in the methodology (use recognition results or not), the fair comparison with other conventional methods for boundary detection is difficult, experiments of phrase boundary detection were conducted using the same database by the method (method 1) based on low-pass filtering (method discussed in section 2) and the method (method 2) based on the simultaneous use of macroscopic and microscopic aspects of  $F_0$  contours[3]. The results were also included in Table 6, showing that the performance of the proposed method is better than that of method 1, and the same with that of method 2, when allowing 1-mora deviation. When the deviation is not allowed, method 2 gave the best result. However, we should note that method 2 detects a lot of minor prosodic events besides the phrase boundaries.

#### 4. CONCLUSIONS

In searching possible ways of utilizing prosodic features in speech recognition, two methods were proposed: one to detect phrase boundaries based on low-pass filtering of  $F_0$  contours, and the other to select the correct recognition result among plural recognition candidates. Although the basic validity was shown for both methods, further studies are still necessary and planned. As for the first method, it should be combined with a method of detecting accent commands to produce a good estimation of model parameters, which can be used as initial parameters for starting full-automatic analysis-by-synthesis process. On the other hand, as for the second method, further investigations are necessary on the following points:

- (1) to find out better way to cope with the  $F_0$  contour variations,
- (2) to establish a criterion to relate the partial analysis-by-synthesis errors to the boundary likelihood,
- (3) to incorporate the method in recognition systems.

## REFERENCES

- [1] A. Komatsu, E. Ohhira, A. Ichikawa and H. Endoh, "Prosodic aids to structural analysis of conversational speech," *Proc. IEEE ICASSP 86*, 42.15, pp.2283-2286 (1986-4).
- [2] E. Ohhira, 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). (in Japanese)
- [3] K. Hirose, A. Sakurai and H. Konno, "Use of prosodic features in the recognition of continuous speech," *Proc. ICSLP 94*, S20-12, pp.1123-1126 (1994-9).
- [4] 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).
- [5] M. Nakai, H. Singer, Y. Sagisaka and H. Shimodaira, "Automatic prosodic segmentation by  $F_0$  clustering using superpositional modeling," *Proc. IEEE ICASSP 95*, Vol.1, pp.624-627 (1995-5).
- [6] E. Geoffrois, "A pitch contour analysis guided by prosodic event detection," *Proc. EUROSPEECH 93*, 24.4, pp.793-797 (1993-9).
- [7] A. Sakurai and K. Hirose, "Detection of phrase boundaries by low-pass filtering of fundamental frequency contours," *Proc. ICSLP 96*, to be published (1996-10).
- [8] K. Hirose, "Disambiguating recognition results by prosodic features," *Computing Prosody*, ed. Y. Sagisaka et. al., Springer-Verlag, New York, to be published (1996).
- [9] 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 92*, Vol.1, pp.149-152 (1992-3).
- [10] V. Ström, "Detection of accents, phrase boundaries and sentence modality in German with prosodic features," *Proc. EUROSPEECH 95*, THpmiC.4, pp.2039-2041 (1995-9).
- [11] K. Hirose and H. Fujisaki, "A system for the synthesis of high-quality speech from texts on general weather conditions," *IEICE Trans. Fundamentals*, Vol.E76-A, No.11, pp.1971-1980 (1993-11).
- [12] K. Takeda, Y. Sagisaka, S. Katagiri, M. Abe and H. Kawabata, "Speech database user's manual," ATR Technical Report (1988-5).
- [13] K. Hirose and A. Sakurai, "Detection of syntactic boundaries by partial analysis-by-synthesis of fundamental frequency contours," *Proc. IEEE ICASSP 96*, Vol.4, pp.809-812 (1996-5).