# Keyword Spotting Using F0 Contour Matching

*Yoichi Yamashita*
Dep. of Computer Science, Ritsumeikan University
1-1-1, Noji-Higashi, Kusatsu-shi, Shiga, 525-77 Japan
yama@cs.ritsumei.ac.jp

*Riichiro Mizoguchi*
I.S.I.R., Osaka University
8-1, Mihogaoka, Ibaraki-shi, Osaka, 567 Japan
miz@ei.sanken.osaka-u.ac.jp

## ABSTRACT

This paper describes keyword spotting using prosodic information as well as phonemic information. A Japanese word has its own F0 contour based on the lexical accent type and the F0 contour is preserved in sentences. Prosodic dissimilarity between a keyword and input speech is measured by DP matching of F0 contours. Phonemic score is calculated by a conventional HMM technique. A total score based on these two measures is used for detecting keywords. The F0 contour of the keyword is smoothed by using an F0 model. Evaluation test was carried out on recorded speech of a TV news program. The introduction of prosodic information reduces false alarms by 30% or 50% for wide ranges of the detection rate.

## 1. INTRODUCTION

Keyword spotting is an important technique for indexing applications, such as TV news speech classification into one of specified topics, and so on. Speech recognition based on the spotting is also a promising approach to understanding spontaneous speech that includes many non-grammatical sentences. In general, phonemic similarity between input speech and a keyword is used for detecting the keyword. If a speech segment, which is not a keyword, is similar to a keyword in a sense of the segmental spectral measure, it can be incorrectly detected as the keyword, that is a false alarm. Reduction of false alarms is a crucial issue in the keyword spotting.

Some false alarms in the conventional keyword spotting have different prosodic patterns from that of the keyword although the phoneme sequences are similar to that of the keyword. A typical case of such false alarms is detection of a segment across a phrase boundary. In this case, there is often a dip in the F0 contour, and it is completely different from that of one word. In other false alarms, a detected word has a different type of the lexical accent from the keyword and its F0 contour also differs from that of the keyword. These observations of conventional spotting results imply that the use of F0 information can improve performance of the keyword spotting.

Prosody can be used to improve continuous speech recognition. Particularly, phrase boundary information is very useful to disambiguate syntactic analysis or semantic interpretation. Many studies have tried prediction of break indices from syntactic information in order to evaluate sentence hypotheses prosodically[1][2] [3]. A method of the keyword spotting using word boundary information derived from F0 has been also proposed[4]. These works address the relationship between phrase boundaries and syntactic information, and they do not use the relationship between the F0 contour and the hypothesized word sequence. In Japanese, a word has an inherent lexical accent which is classified into $n$ types for $n$-mora words. The accent type generates characteristic patterns of F0 in sentences. The F0 pattern across a word or phrase boundary is also very different from that in the middle of one word. The similarity between F0 contours can be useful information to reduce false alarms in the keyword spotting. F0 contour matching between a keyword template and input speech is an alternative method of speech recognition using prosody. It has following advantages over the phrase boundary detection for the keyword spotting.

1. It may differentiate words with different accent types even if they are phonemically similar to each other. The phrase boundary detection does not contribute to differentiate them.
2. It gives a better F0 representation of long compounds. The long compound often has small dips of the F0 pattern which cause incorrect detections of the phrase boundary. In the F0 contour matching, the keyword F0 template represents the whole F0 contour and it can avoid this problem.
3. It does not need global processing. The phrase boundaries are global characteristics of F0 pattern of the sentence. It is very difficult to identify them in local information. The F0 contour matching can be realized by using local information in a sentence. This characteristic is convenient to the word spotting and the speech recognition based on the left-to right processing.

In this paper, we describe a method of the keyword spotting using matching score of F0 contours as well as segmental spectrum information in order to reduce false alarms without degradation of the detection rate.

 

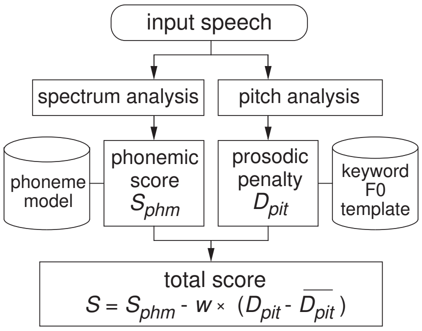


Figure 1. Flow of keyword spotting using prosodic score.

## 2. METHOD

The keywords are spotted based on the following score.

$$S = S_{phm} - w \times (D_{pit} - \overline{D_{pit}}) \qquad (1)$$

$S_{phm}$ is a phonemic score which represents spectral distance between an input speech segment and a keyword, and it is calculated by a conventional HMM technique. $D_{pit}$ is a prosodic penalty evaluated in terms of F0 contour distance between an input speech segment and a template prepared for each keyword, and $\overline{D_{pit}}$ is an average of $D_{pit}$ for all keywords found in speech data. The $w$ is a scaling factor, and $w = 0$ gives a conventional keyword spotting without F0 information. If a segment in input speech has a larger score than a threshold, it is detected as a keyword. A schematic flow of this process is depicted in Figure 1.

### 2.1. Phonemic Score

The HMM recognizer calculates two phonemic likelihoods for an input whole utterance using linguistic constraints, shown in Figure 2. The first likelihood, $L_{syl}$, is a phoneme probability in logarithm for the most probable syllable sequence. An utterance is recognized under a linguistic constraint of a syllable model, shown in Figure 2(a), which allows any syllable concatenation including pause. Another likelihood, $L_{kw}$, is given by a keyword model, shown in Figure 2(b) which hypothesizes existence of a keyword. The keyword model represents a keyword embedded in preceded and followed background models described by the syllable model. The keyword model is prepared for each keyword whereas the syllable model is commonly used for all keywords.

The phoneme recognition result of the second recognition always includes a keyword. If the input speech actually contains a keyword, the first recognition result includes a very similar phoneme sequence to the keyword and $L_{kw}$ is very close to $L_{syl}$. If the input speech does not contain a keyword, the keyword model forces the keyword to match with a part of the input speech and it decreases $L_{kw}$. The phonemic score for detecting a keyword, $S_{phm}$, can be defined as

$$S_{phm} = \frac{L_{kw} - L_{syl}}{keyword\ length}. \qquad (2)$$

Difference of two likelihoods is normalized by a keyword length which is evaluated in terms of a phoneme length of the keyword. In general, $L_{kw}$ is less than $L_{syl}$ because a strong constraint is given for the second recognition. Therefore, $S_{phm}$ is always negative.

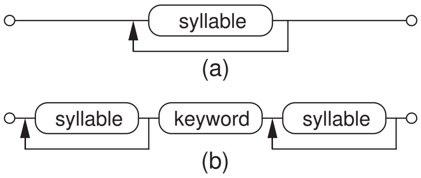


Figure 2. Linguistic constraints for phonemic scoring.

We used the HTK tools provided by Entropic to obtain the phonemic scores. The feature vector is composed of 12 melcepstrum, 12 delta-melcepstrum, and delta-energy. Phonemes are divided into 26 categories. Each phoneme model has 5 states with 4 mixtures and 3 loops. The phoneme models were trained with 13.6 hour phonetically balanced sentences by 64 speakers in the ASJ Continuous Speech Corpus for Research. This speech recognizer uses a phoneme bigram trained with 503 phonetically balanced sentences. The accuracy of the recognizer is 54% for the first 10 news utterances in speech data mentioned later.

### 2.2. Prosodic Penalty

The prosodic penalty, $D_{pit}$, is defined as the dissimilarity calculated in terms of the DP matching distance between F0 contours of input speech and a keyword template. The preprocess removes pitch analysis errors from raw F0 patterns by smoothing for the keyword template and by filtering for the input speech before the F0 contour matching.

The F0 model proposed by Fujisaki[5] smoothes and interpolates the F0 contour of the keyword in order to obtain a continuous F0 template. We manually selected an utterance containing the most typical F0 pattern for each keyword from news speech data. The F0 pattern of the whole sentence is fitted and smoothed by the F0 model and the keyword segment was elicited from the sentence. Thus, a keyword template is prepared for each keyword.

For the input speech, pitch analysis errors are removed by the heuristic rules based on the size and relationship of clusters in the F0 pattern. For example, the size of consecutive voiced frames is less than 4, they are regarded as the pitch analysis error.

The frame distance in the DP matching, $d(i, j)$, is defined as

$$\begin{aligned} d(i,j) &= |\log p_k(i) - \log p_i(j)| \quad (p_i(j) \neq 0) \\ &= 0 \qquad\qquad\qquad\qquad\quad\; (p_i(j) = 0). \end{aligned} \qquad (3)$$

The $p_k(i)$ and $p_i(i)$ is the F0 frequency of $i$-th frame for the keyword template and the input speech segment, respectively. Note that $p_k(i)$ never takes 0 because the F0 contour for the keyword is interpolated by the F0 model.

The process of the DP matching iteratively calculates the prosodic penalty, $D_{pit}$, to equalize the F0 averages for the keyword and the input speech.

Step-1: Carry out the DP matching between two F0 contours using the frame distance defined above.
Step-2: Calculate an average pitch for the voiced frames in the input speech. Get the frame-to-frame mapping and calculate an average pitch in the keyword template using only frames which correspond to voiced frames in the input speech.
Step-3: Shift pitch values for the keyword template by the difference of two average pitches.
Step-4: Go to step-1 if the difference of average pitches is larger than the threshold. Otherwise, the last DP score is $D_{pit}$.

## 3. SPEECH DATA

Speech data was collected from a TV news program. Recorded speech was automatically divided into 610 speech fragments, called utterances, based on pause longer than 180msec. The TV news speech includes very noisy utterances, such as interview overlapped by environmental noise and narration with background music, and so on. These noisy utterances were manually removed from speech material for experiments. The total number of clean utterances are 514, and the total duration is 26.2 minutes. They includes several non-professional speakers, such as interviewees in the street, unclearly speaking politicians, and so on, as well as professional news announcers.

Twenty words frequently appeared in this news speech, such as '*dairishomei* (representative signature)', '*gakko-kyushoku* (school lunch)' and so on, were manually selected as keyword. Average mora length of 20 keywords is 8.8. The occurrence of the keywords in the clean utterances is 235 in total and 27.0 [/hour/keyword] in average.

## 4. EVALUATION

Performance of the keyword spotting is evaluated by the detection rate and the false alarm (FA) rate defined as follows.

$$detection\ rate[\%] = \frac{N_{correct}}{N_{keyword}} \times 100$$

$$FA\ rate[/hour/keyword] = \frac{N_{FA}}{DUR_{total} \times KW}$$

$N_{correct}$ and $N_{FA}$ is the number of correct detections and false alarms in a spotting result, respectively. $N_{keyword}$, $KW$, and $DUR_{total}$ is is the total occurrence of the keywords in speech data (235), the kind of the keywords (20), and the total duration of speech data in hour (0.44=26.2/60), respectively.

We performed a simple spotting using the HTK tool. All possibility of the keyword existence is not investigated. First, the keyword spotting without F0 information detects a segment which is the most likely to a keyword for each utterance. Thus, 10280 (=514x20) segments was selected as test speech. Secondly, the F0 contour matching is applied to these segments. When an utterance contains two or more same keywords, only the most probable one is used for evaluation and the others are ignored.

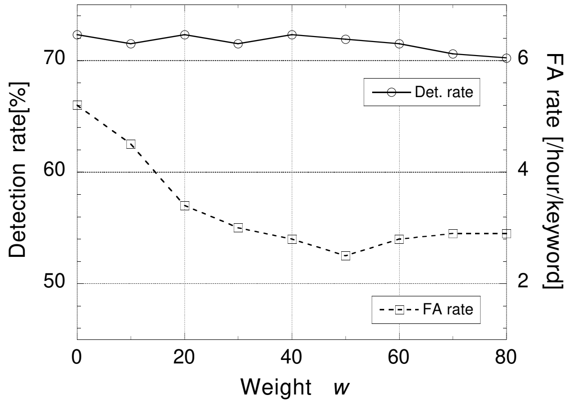


Figure 3. Effect of prosodic information on the false alarm reduction.

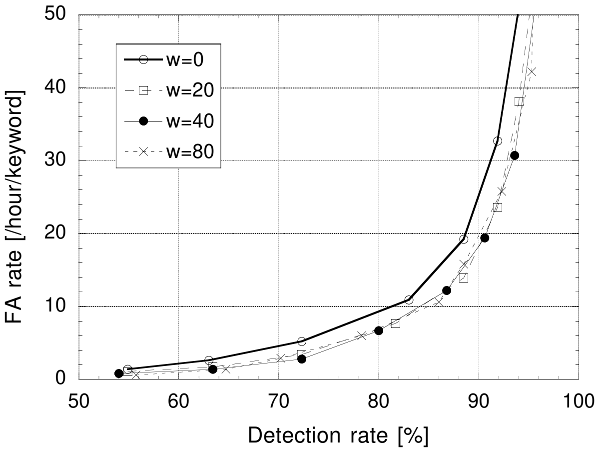


Figure 4. Spotting performance using prosodic information.

Figure 3 shows the detection rate and the FA rate for changing the scaling factor $w$ when the spotting threshold is 7.0. In a case of the spotting without the F0 information, which are plotted on the leftmost, the detection rate is 72.3%, and the FA rate is 5.2. The introduction of the F0 contour matching decreases the FA rate to less than the half with few degradation of the detection rate.

Figure 4 shows spotting results for various spotting thresholds. The threshold for the spotting score controls the number of spotted keywords. The small threshold can detect most of keywords but increases the false alarms, and vice versa. The solid bold line represents the baseline performance without the F0 information. The introduction of the F0 contour information can reduce the false alarms for all range of the detection rate. From this figure, it reduces 30% or 50% of false alarms for 80% or 70% of the detection rate, respectively. The false alarm reduction is

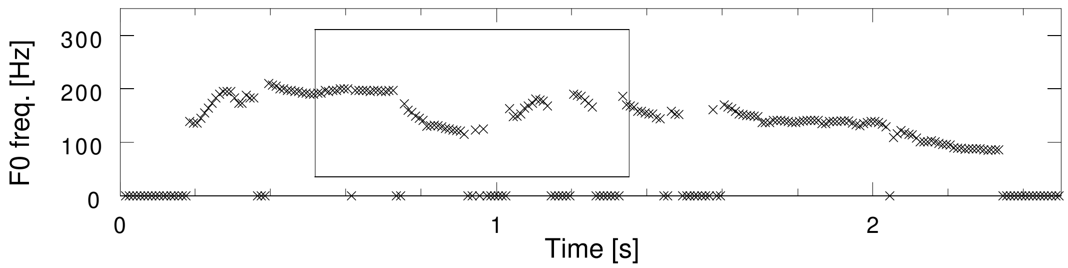


Figure 6. An example of F0 pattern ('*iNtanetto no kyusokuna fukyu* (rapid spread of internet)').

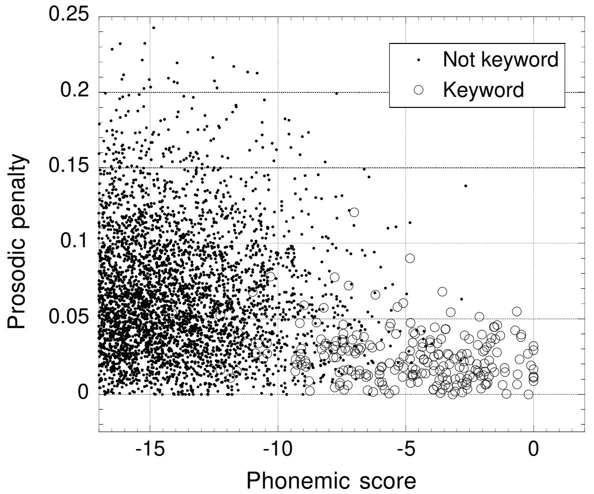


Figure 5. Distribution of the phonemic score and the prosodic penalty.

salient for lower detection rates.

Figure 5 shows the distribution of the phonemic score and the prosodic penalty of spotted segments. Correct keywords are plotted by a white circle and false alarms are plotted by a small dot. In a conventional spotting without F0 information, a vertical line is a boundary between keywords and non-keywords because an input speech segment is spotted regardless of a prosodic penalty. The use of prosodic penalty makes this boundary tilt to the right according to the $w$. Some non-keyword data are distributed in upper areas of the keyword data. They are successfully removed from spotted keywords by the prosodic penalty. However, too large $w$ incorrectly judges the data with a very small prosodic penalty and a small phonemic score.

Figure 6 shows a successful example of false alarm reduction by F0 contour matching. It depicts F0 pattern for an utterance of '*iNtanetto no kyusokuna fukyu* (rapid spread of internet)'. The area in a rectangle is incorrectly detected as the keyword '*gakko-kyushoku* (school lunch)' without F0 contour information because a partial phoneme sequence '*nettonokyusoku*' in the utterance is similar to the keyword. This false alarm was suppressed by the F0 contour matching. The F0 pattern of this false alarm is very different from that of the keyword '*gakko-kyushoku*' which is shown in Figure 7.

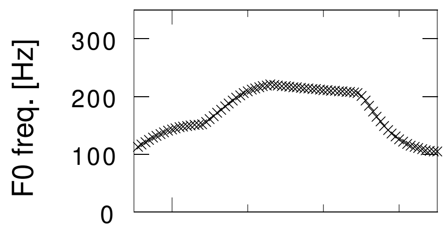


Figure 7. An example of keyword F0 template ('*gakko-kyushoku* (school lunch)').

## 5. CONCLUSION

This paper describes the keyword spotting using F0 contour information as well as segmental spectral similarity. An evaluation test shows that the introduction of prosodic information reduces false alarms by 30% or 50% for wide ranges of the detection rate.

The F0 contour of a keyword is represented by a template which is prepared for each keyword. The F0 contour has a lot of varieties according to speakers and context in a sentence or a story. Some keywords had a little bit different pattern from the template. To cope with diversity of the F0 pattern, a multi-template method based on clustering of F0 contours or modeling relationship between F0 contours and lexical and contextual information is necessary.

## REFERENCES

[1] M. Ostendorf, C.W. Wightman, N.M. Veilleux: "Parse scoring with prosodic information: an analysis/synthesis approach", *Computer Speech and Language*, 7, pp.193-210 (1993).

[2] R. Kompe, et al.: "Prosodic Scoring of Word Hypotheses Graphs", *Proc. of ICASSP '95*, pp.1333-1336 (1995).

[3] A.J. Hunt: "Syntactic Influence on Prosodic Phrasing in the Framework of the Link Grammar", *Proc. of Eurospeech'95*, pp.997-1000 (1995).

[4] T. Hanazawa, Y. Abe, K. Nakajima: "Phrase Spotting Using Pitch Pattern Information", *Proc. of Eurospeech '95*, pp.2137-2140 (1995).

[5] H. Fujisaki and K. Hirose: "Analysis of voice fundamental frequency contours for declarative sentence of Japanese", J.A.S.J.(E), 5, 4, pp.233-242 (1984).