AN EVALUATION OF KEYWORD SPOTTING PERFORMANCE UTILIZING FALSE ALARM REJECTION BASED ON PROSODIC INFORMATION

Masaki IDA , Ryuji YAMASAKI

Information Technology Research Center

OMRON Corporation

Shimo-kaiinji Nagaokakyo-city Kyoto, JAPAN 617-8510

{ida,ryuji}@ari.ncl.omron.co.jp

ABSTRACT

In this paper, we describe our effort in developing new method of false alarm rejection for keyword spotting type of speech recognition system. This false alarm rejection uses prosodic similarities, and works as posterior rescore basis.

In keyword spotting, there is always false alarm problem. Here, we propose a technique to reject those false alarms using prosodic features.

In Japanese, prosodic information is expressed in intonation form, while may of other languages is using stress accents. Therefore, it is easy to calculate prosodic information using fundamental frequency, so called F0, in our language. In our new keyword spotting engine, we get result by combining two scores. One is phonetic score calculated by front engine, and the other is pitch score calculated by post engine described in this paper.

We have accomplished 13%(point) improvement on keyword recognition accuracy using this method. We also have proposed robust modeling method for rejection using prosodic features.

1. INTRODUCTION

In recent days, there were many studies of human-machine interfaces basis on spoken dialogue. Those systems need to recognize and understand spontaneous utterances. However, most “spontaneous utterances” in real-world are out of grammar(syntax), having injections, hesitations, corrections, inversions, and/or omissions.It is impossible to describe every phenomenon with syntactic or statistic representations. In case of a system which handles limited task, though, it is able to understand user’s intentions only with understanding some important keywords out of their utterances. Particularly, on spoken dialogue systems, it is able to infer user’s intention in step by step basis through the dialogue with user. Therefore, word spotting technique which has feature that skips unnecessary segments and focus only important area, is very helpful and widely required in real life.

In word spotting, there is always false alarm problem. In this paper, we propose false alarm rejection method. This method re-scores the spotted segments, after front-end word spotter has decided, with use of prosodic information. In using prosodic information, since it is inevitable to have calculation error, a good modeling method which is both strict and flexibe to describe prosodic representation is needed. Here, we have described some prosodic modeling methods and their estimations in this paper.

2. OVERVIEW OF WORD SPOTTER

In this section, we describe the word spotter used on this paper which is using our own technique. Benefits of this engine are, reduction of calculation time and data-size with a feature vector extraction from LPC-cepstrum. This extraction method is called SPT(Successive Processing along Trajectory) method. It is based on LPC-Cepstrum using time normalization technique along vector trajectory length. Benefit of this technique is to have SPT feature vector with well-balanced phoneme density. 2 completely different merits, which are, strict description at extreme feature transformation segment like consonants, and rough description at constant segment like vowels, are realized at the same time, as a result of feature extraction density control. Fig.1 shows flowchart of our word spotting phoneme-base recognition engine. Fig.2 shows generation of SPT feature vectors. As it could be referred from the chart, consonants has shorter intervals compare to vowels. Roughly 50% cut on feature data is accomplished using SPT method.

3. FALSE ALARM REJECTION BASIS ON PROSODIC INFORMATION

In word spotter, there is always false alarm problem. This problem is due to occurence of phonetically similar sequence to the false keyword. Several studies has been reported concerning False Alarm Rejection Methods for Word Spotters. Use of high quality filler model to eliminate Non-recognizable-area, or, use of language model together with word spotter[1,2]. Those methods are all task-depended aproach, so, most important benefit of word spotter, which is task-independent robustness, is minimized. In this paper, we propose the use of prosodic information, which is independent from task, to reject false alarms.

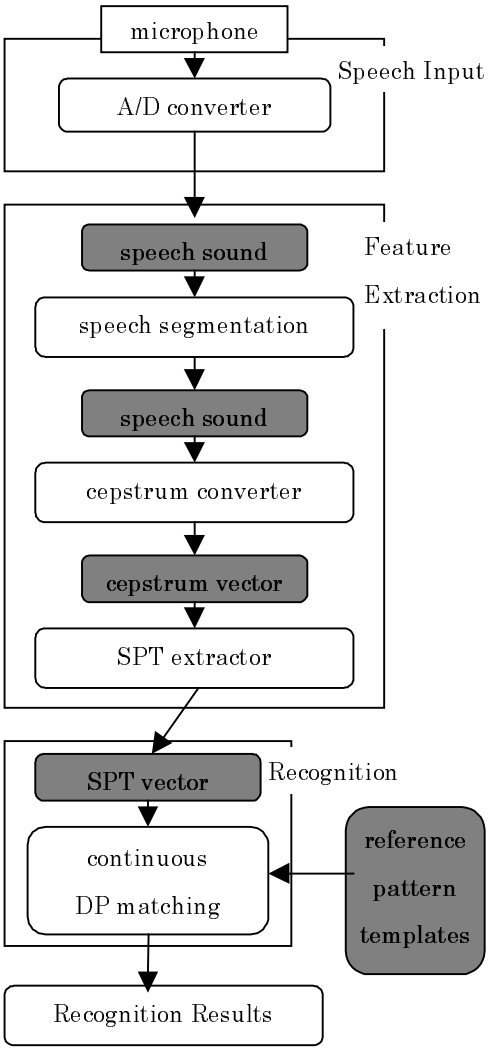


Fig.1 SPT Word-Spotter

When understanding speech, Accent information is very important element. In case of Japanese speech recognition, accent information is omitted in typical recognition engine. This is due to difference of accent type. Unlike most western languages which use Stress Accents, Japanese language uses Pitch Accent. In typical recognition engine, features are calculated using Cepstrum analysis. Cepstrum analysis divides speech signal into vocal tract information(represented as cepstrum vector) and vocal chords information(represented as Cepstrum residual signal). Normally, recognition engine uses only the former information(vocal tract). And Pitch accent information is contained in the later information(vocal chords). Eventually, in Japanese speech recognition, it is very effective to utilize prosodic information. Here are several work done before. Use of prosody information to segment phrase, as front end, and

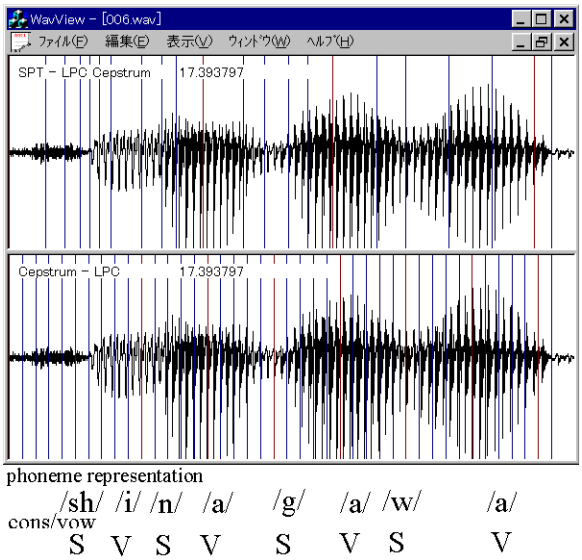


Fig.2 Feature Density between SPT and Cepstrum

sample speech of "Shi na ga wa"

upper : SPT (with constant **trajectory** intervals)

lower : Cepstrum (with constant **time** intervals)

use phrase boundary information that is obtained to the recognition[3]. Use of prosody information to understand phrase structure [4]. To identify vowel or consonant [5]. To distinguish language [6].

In this paper,unlike those described above, we have introduced prosody information to rescore the result of phonetic recognition. First, find keyword location (area) by phonetic recognition, and in second phase, recalculate(rescore) the matched keyword with predefined prosody pattern for the keyword. Integrated result of keyword spotter is calculated with combination of phonetic and prosodic recognition scores. There are few but some precedent approach similar to ours. [7] describe usage of similar recognition engine to word spotter to TV news program. In their approach, they use Dynamic Time Warp calculations (Dynamic Programing) to score prosodic matching. But in prosodic information, it is inevitable to have calculation error which causes vast differences when calculating distance score in matching procedure. They also select only one model as prosodic model. There is no doubt we need to consider speaker independency also in prosodic model, thus only one reference template is not sufficient.

Our goal is to realize gentle interface to any person based on human-machine dialogue using speaker independent speech recognition engine. In this term, we have also investigated prosodic information modeling methods which solve prosodic information extract error, and/or difference of speaker, for robustness. We have also suggested 2 modeling methods for prosodic information, and compared with DP matching model. Those models are described in section 4.3.

4.EXPERIMENTAL ESTIMATION AND RESULTS

4.1 Experimental Data

Experiment has been 2 tasks. They are recorded on 11025Hz sampling frequency, 16 bit A/D convert, Andorea headset microphone, and laboratory environment.

E @Tourist Information Terminal Command Task

This supposes spontaneous speech inputs to tourist information kiosk terminal. Target keywords include 18 sight location names in Kyoto. Evaluation set includes 3 speakers and 20 utterance each speaker. Examples as follows:

"Kinkakuji no haikanryo wo shiritai."

means "Tell me an admission fee of Kinkakuji."

keyword = Kinkakuji

E @Ticket Vending Machine Command Task

This supposes spontaneous command inputs of railway ticket vending machine with speech recognizer. Target keyword s include 50 station-name in Tokyo. Evaluation set includes 8 speakers, 200 utterances each speaker with following additional words. Examples for "Tokyo" as follows:

"Tokyo eki." means "Tokyo station."

"Tokyo made." means "To Tokyo."

"Tokyo onegai shimasu." means "Tokyo, please."

"Eeto Tokyo." means "Uh…, Tokyo."

4.2 Phonetic Recognition

Phonetic speech recognition has been used as front-end engine. Our phonetic recognition engine uses original feature SPT. This has been explained in section 2, in detail. Sampling rate of 11.025KHz, 23.2msec window and 15.9msec frame pitch for LPC-Cepstrum analysis. ¢ feature (cepsrtum) is not in use for this experiment. Fuzzy-C-Means clustering is selected to cluster multiple speakers' features for templates. Each word has 3 different templates to cover different features person to person. Original utterances for templates, are recorded by 15 male and 5 female for Ticket Vending Machine Command Task, and 10 male for Tourist Information Terminal Command Task.

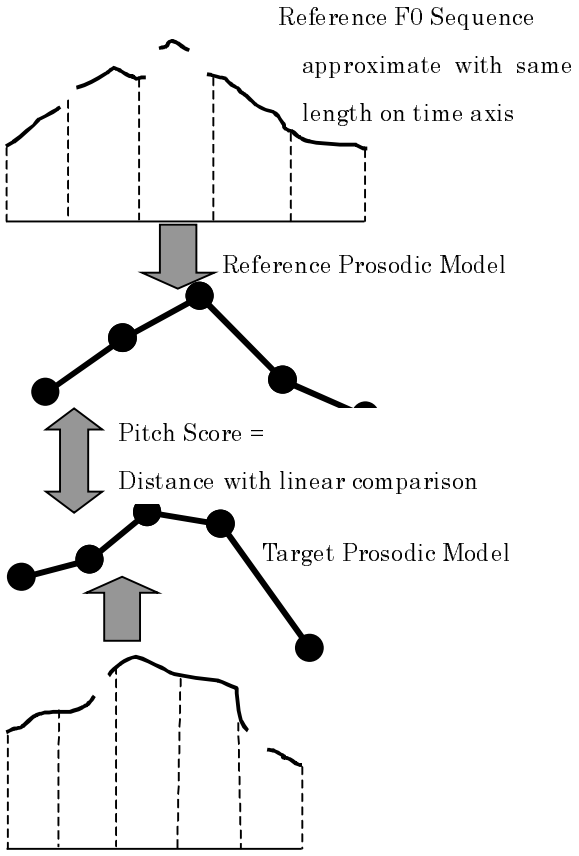


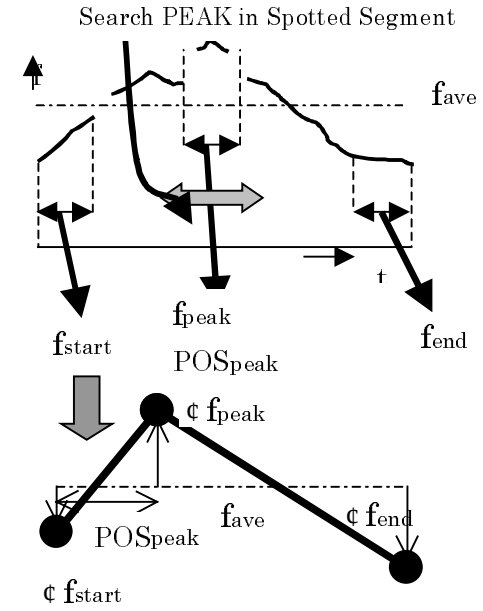
Fig.3 Polygonal Lines Approximation

4.3 Scoring based on Prosodic Information

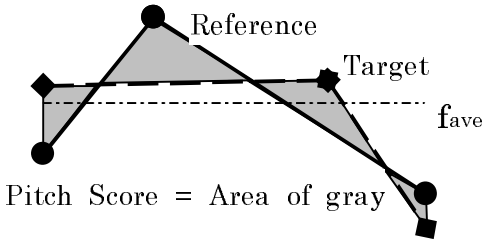
Prosodic Information is calculated by a tool based on Medan Technique, with frequencies of 10msec interval [8]. 1 representative speech data for each word has been selected and prosodic information for that particular data is taken for template for prosodic matching procedure. We have tested 3 different scoring calculations for matching procedure.

We applied prosodic information rescoring (1) Top 5 candidates, and (2) Candidates with less than 1.0 point difference of phonetic score compare to first available candidate. All candidates that do not fit above conditions are eliminated when rescoring.

(1) DP matching



(a) Components of Triangle Model



(b) Scoring of Triangle Model

Fig.4 Triangle Model

Use bare base frequency(F0) data and calculate distance using Dynamic Programming(DP) matching. Distance between template and feature data of target utterance is taken as score directly.

(2) Polygonal lines approximation (Fig.3)

For templates, average F0 data every 500msec is calculated and normalized. As a result, a set of several lines(2 to 5) of 500msec long is created to represent prosodic information for each word. Target utterance after phonic recognition, we already know matching boundary and candidates. Thus, split matching area data to the same number of which candidates templates has. After those calculations, distance between template and matching area of target utterance is calculated.

(3) Triangle model (Fig.4)

Calculate average F0 data of 3 points within target area. 3 points are (1) Average of first 5% of the area, (2) Average of 5% around F0 peak(either up or down), or if no peak appears within the area, 5% around middle place along time axis of the area, (3) Average of last 5% of the area. 5% is decided after experiment calculating 10% each or 20% each for average.

4.4 Experiment Result

Following table indicates result of Prosodic Information rescoring. Base line is bare phonetic speech recognition.

Table 1 Experimental Result

Task	Tourist [%]	Ticket [%]
Baseline	75.56	63.50
DP matching	88.30	62.88
Polygonal Lines	77.78	63.62
Triangle	78.89	63.56

We obtained at most 13%(points) improvement. We also confirmed effectiveness of prosodic information rescoring for speaker independent system and anti-grammar spontaneous speech input.

5. CONCLUSION AND FUTUREWORKS

We obtained at most 13%(points) improvement. We also confirmed effectiveness of prosodic information rescoring for speaker independent system and anti-grammar spontaneous speech input. DP matching model led best result, but with different condition, completely opposite result. DP matching model has problem in robustness. 2 other models that we have proposed in this paper, had less improvement, but stable result. In next step, we will modify prosodic information modeling methods to obtain better improvement maintaining stability.

References

1. Kawahara et al., “Word Spotting in Spontaneous Speech with Heuristic Language Model”, IEICE Trans., vol.J78-DII, No.7, 1995 (in Japanese)
2. Kawahara et al., “Robust Speech Understanding Based on A*-Admissible Phrase Spotting”, IEICE Trans., vol.J79-DII, No.7, 1996 (in Japanese)
3. Nakai et al., “Accent Phrase Segmentation Based on F0 Templates Using a Superpositional Prosodic Model”, IEICE Trans., vol.J80-D-II, No.10, 1997 (in Japanese)
4. Takeda and Ichikawa, “Analysis of prosodic features of prominence in spoken Japanese sentences”, Journal of ASJ, vol.47, No.6, 1991 (in Japanese)
5. Aoki et al., “An Investigation of False Alarm in Word Spotting -comparison of gross features of speech spectrumand filter bank-”, Tech. report of IEICE, EA96-31, 1996 (in Japanese)
6. Sasanuma and Itahashi, “Classification of Spoken Language by Fundamental Frequency”, Tech. report of IEICE, SP96-57,, 1996 (in Japanese)
7. Yamashita et al., “Keyword Spotting Using F0 Contour Information”, IEICE Trans., vol.J81-D-II, No.6, 1998 (in Japanese)
8. Tuerk and Robinson, “A New Frequency Shift Function for Reducing Inter-Speaker Variance”, Eurospeech, 1993