*International Symposium on Chinese Spoken Language Processing (ISCSLP), pp.281-284, 15-18 Dec. 2004, Hong Kong*

# A TWO-STEP KEYWORD SPOTTING METHOD BASED ON CONTEXT-DEPENDENT *A POSTERIORI* PROBABILITY

*Thomas Fang Zheng, Jing Li, Zhanjiang Song*, and Mingxing Xu*

Center for Speech Technology, State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science & Technology, Tsinghua University, Beijing, 100084, China
{fzheng, lij, xumx}@cst.cs.tsinghua.edu.cn, http://cst.cs.tsinghua.edu.cn/

* Beijing d-Ear Technologies Co., Ltd., zjsong@d-Ear.com, http://www.d-Ear.com

## ABSTRACT

Keyword weighting plays an important role in traditional keyword spotting (KWS) systems: it helps detect keyword candidates in an utterance so that they will not be missed. However, if the keywords are over-weighted, there will be a high number of false alarms, which will slow down the system and might introduce rejection errors; on the other hand, if the keywords are inefficiently weighted, the detection rate is not guaranteed. It is difficult to make a compromise with regard to keyword weighting. A two-step KWS method based on context-dependent *a posteriori* probability (CDAPP) is proposed in this paper as a way to solve this problem. The first step adopts a continuous speech recognition method, to generate a sequence of acoustic symbols for the second step, which performs a fuzzy keyword search. Preliminary experiments show that the proposed strategy is a promising one that needs additional investigation.

## 1. INTRODUCTION

Among automatic speech recognition technologies, keyword spotting (KWS) is one of the most practical, with large market potential including Call Centers and Command-and-Control. In these kinds of applications, KWS technology provides a flexible and convenient way for users to access information or control something.

The most popular KWS methods are often either template matching based [1] or keyword-filler network based [2][3]. A confidence measure (CM) is a very important part in such systems for post-processing to remove false alarms.

No matter which of the above-mentioned methods is used, keywords are usually weighted so as to guarantee a high keyword detection rate. While this contributes to a higher keyword detection rate, it also encourages a high false alarm rate, a major problem in KWS systems. By using CM techniques in the searching procedure, which we refer to as CM Look-Ahead, researchers are trying to find an efficient solution to pruning impossible keyword hypotheses ahead of time. However, this runs the risk of false pruning.

By studying the keyword spotting errors in our previous KWS systems, we can summarize the types of errors into the following categories:

- Speaker: the speaker does not utter the keyword clearly or correctly. This is difficult to solve using automatic technology.
- Environment: the background is noisy. Noise reduction or speech enhancement methods can provide an efficient solution.
- Keyword Weighting: if a keyword class is over-weighted, the false alarm rate will be high; otherwise the keyword detection rate will be lower. It is difficult to find a compromise.
- Confidence Measure: there is not currently a satisfying CM method for use in either the searching procedure or the post-processing phase.

Based on the above facts, a two-step KWS method is proposed and evaluated in this paper. The motivation here is to eliminate the adverse effects of keyword weighting, which is often determined empirically and experimentally to a great extent. In our method, the first step produces a sequence of acoustic symbols (phones, syllables, Chinese initials/finals (IFs), etc.) and the second step allows a sufficient search of keywords in the sequence provided by the first step.

This paper is organized as follows. In Section 2, the framework of the two-step KWS method is described in detail. Experimental results are given in Section 3. In Section 4 we analyze the experimental results and point out future research plans.

## 2. TWO-STEP KWS FRAMEWORK

### 2.1. Application background

Our experiments were based on the *d-Ear Attendant* [5] application, in which a caller speaks out whom he/she wants to call (where the name of the person to call can be embedded in a sentence), and the system detects the person's name and make a connection. The system assumes that the caller will be cooperative, or in other

words, that each utterance will contain at most one name. The *d-Ear Attendant* is a Chinese language KWS system; therefore in this paper we will use the set of Chinese initials and finals as our acoustic modeling units.

### 2.2. Framework description

The proposed method divides the KWS task into two steps: a continuous speech recognition step (performed without using any language model), and a fuzzy keyword search step. The details of each step are given below.

#### *2.2.1. Continuous speech recognition*

A continuous Chinese speech recognizer without any language model is used for the first step. The output is a sequence of Chinese IFs. Unlike several other methods, we do not generate an IF lattice, because we will never use it for later processing.

#### *2.2.2. Fuzzy keyword search*

The goal of the second step is to identify keyword candidates in the sequence of Chinese IFs obtained in Step 1. In order to keep from being influenced by keyword weighting, and to improve robustness (especially when there are mispronounced or unclearly pronounced IFs), a fuzzy keyword search algorithm is proposed based on a context-dependent *a posteriori* probability (CDAPP).

Suppose the IF sequence observed in Step 1 is defined as $O=(o_1,\cdots,o_n,\cdots)$ , where $o_n$ is the *n*-th acoustic symbol (in our case Chinese IF) in the sequence. Then let $K_i=(k_{i1},...,k_{iN_i})$ be the *i*-th keyword consisting of $N_i$ Chinese IFs. Ignoring insertion and deletion errors, the probability of matching $K_i$ with $O_n$, a subsequence of $O$ starting from the *n*-th IF, is proposed to be calculated as

$$\begin{aligned} P(K_i \mid O_n) &= P(k_{i1},\cdots,k_{iN_i} \mid O_n) \\ &= \prod_m P(k_{im} \mid O_n; k_{i1},\cdots,k_{i,m-1}) \\ &= \prod_m P(k_{im} \mid o_1^{(n)},\cdots,o_m^{(n)}; k_{i1},\cdots,k_{i,m-1}) \end{aligned}$$

where $o_m^{(n)} = o_{n+m-1}$ $(n,m \in Z)$.

If we assume that the calculation of the probability of the current IF in a keyword is only affected by at most its first order left context -- the Chinese IF to the left of the current position being matched -- the equation can be then rewritten as

$$P(K_i \mid O_n) \approx \prod_m P(k_{im} \mid o_{m-1}^{(n)}, o_m^{(n)}, k_{i,m-1})$$

or

$$P(K_i \mid O_n) \approx \prod_m \frac{P(k_{i,m-1}, k_{im} \mid o_{m-1}^{(n)}, o_m^{(n)})}{P(k_{i,m-1} \mid o_{m-1}^{(n)}, o_m^{(n)})}$$

Related to this equation, both $P(c_{m-1}, c_m \mid o_{m-1}, o_m)$ and $P(c_{m-1} \mid o_{m-1}, o_m)$ can be estimated from a training set, where $c_m$ is a canonical Chinese IF at the current position (which can be obtained from the transcription), and $o_m$ and $o_{m-1}$ are the observed IFs at the current position and its left position (which can be obtained from an existing speech recognizer, i.e. the same one used in Step 1).

The problem of sparseness must be solved when estimating the values of $P(c_{m-1}, c_m \mid o_{m-1}, o_m)$ and $P(c_{m-1} \mid o_{m-1}, o_m)$ . For the sake of simplicity, a straightforward method is used in this paper: assigning a small, predefined value to any zero probabilities.

For comparison purposes, three cases of CDAPP were considered, as in the following equations where *CI* stands for context-independent, *LCD* for left context-dependent, and *LCD2* for a variation of the *LCD* case.

$$P(K_i \mid O_n) \approx \prod_m P(k_{im} \mid o_m^{(n)}) \qquad (CI)$$

$$P(K_i \mid O_n) \approx \prod_m P(k_{im} \mid o_{m-1}^{(n)}, o_m^{(n)}) \qquad (LCD)$$

and

$$P(K_i \mid O_n) \approx \prod_m P(k_{im} \mid o_{m-1}^{(n)}, o_m^{(n)}, k_{i,m-1}) \quad (LCD2)$$

## 3. EXPERIMENTS

### 3.1. Databases

The application background is as described in Section 2.1. Experiments were performed across databases collected from different telephone channels at an 8 kHz sampling rate. An extended context-dependent Chinese initial/final set [6] was used for speech recognition units (acoustic symbols), with each unit modeled by a 3-state HMM using HTK [4].

Keywords were generally person names made up of 2 or 3 Chinese characters (syllables). A very small number of keywords consisted of 5 characters (syllables) such as *Prof. Fang Di-Tang* (or *Fang*1 *Di*4 *Tang*2 *Jiao*4 *Shou*4 in Pinyin). Each name in the list had a unique pronunciation, except two names that shared the same toneless Chinese syllable string.

The acoustic model was trained from a read-style training set consisting of more than 50,000 sentences uttered by approximately 400 speakers, where the content of the sentences were independent of the *d-Ear Attendant* domain. One third of the data were down-sampled from the data taken from the *863 Database* [7] at a 16 kHz sampling rate through the telephone channel. This part of the data contains both syllable

transcriptions and time boundary information. The other two thirds of the data were collected directly from the telephone channel with syllable transcriptions, but without time boundary information.

The testing set contained 600 sentences: 300 spontaneously uttered sentences and 300 sentences each containing a single name. It was different from the training set, especially in speaking style. More specifically, the testing set was closer to real-world use, where some people like to directly speak out a single name, while other people prefer to speak in a natural way (i.e. in full sentence). The person names spoken in the 600 sentences were taken from a list of 110 different names.

### 3.2. Baseline system

The baseline system chosen was the one with the best results as described in [5]. In the baseline system extended template matching is used to direct keyword spotting.

The traditional template matching (TM) method achieves satisfactory results when input utterances match quite well with the templates, but the performance degrades for mismatched utterances. On the other hand, traditional KWS based on a loop network of keywords and fillers enables flexible speech input; however the keyword detection rate is lower. The extended template matching (ETM) method takes advantage of the said two methods (see Figure 1) while at the same time adopting an online filler modeling method [5].

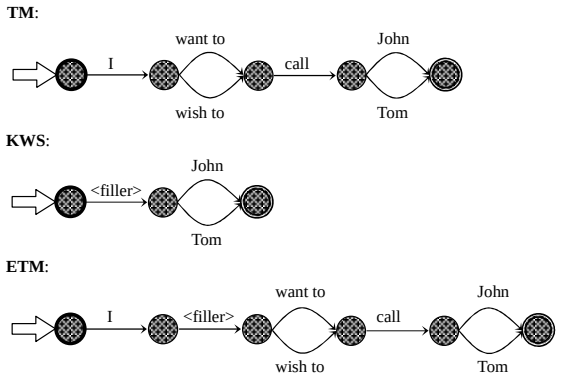


*Figure 1*: Illustration of extended template matching (ETM) compared with traditional template matching (TM) and KWS when the input is "I *often* want to call John".

### 3.3. Accuracy comparison

Experimental results on the testing set are listed in Table 1.

*Table 1:* Experimental Results

| Methods | Cor(%) |
|---|---|
| Baseline | 87.0 |
| CI | 80.3 |
| LCD | 80.3 |
| LCD2 | 88.0 |

## 4. DISCUSSION AND SUMMARY

### 4.1. Discussion

This paper reported preliminary experimental results done using a context-dependent *a posteriori* probability (CDAPP) method for keyword spotting, and the experimental results are encouraging. From the framework of this method, we can see that it has the following advantages:

- The ambivalent keyword weighting problem has been eliminated. Because there are no keyword weights, there is no need to worry about weight adjustment.
- Compared with traditional KWS methods, the proposed CDAPP method has a smaller search space in the fuzzy keyword search step. This is because what is being searched is a sequence of acoustic symbols with a much smaller length in IFs than the total length in frames. This enables more pattern matching and comparison attempts, which helps detect unclear or ill-pronounced keywords with the aid of CDAPP.
- The proposed CDAPP method does not rely on an acoustic model as strongly as other methods do, since the CDAPP probabilities can be used as an acoustic regulator. Even if the recognizer performs poorly, the proposed method will "*pull*" those incorrectly recognized IFs back as correctly as possible. Of course, the acoustic model cannot be too poor; it should be above average.

### 4.2. Next steps

The preliminary experiments show that the proposed method is a promising one. It is expected that the performance of KWS systems using the CDAPP method will be greatly improved, particularly because there are many other methods that might further improve results if integrated into the proposed method.

- *Using the acoustic modeling score*. In the fuzzy keyword search step, the criterion we use is a context-dependent probability, $P(K_i \mid O_n)$. A very important part, the acoustic score $P(O_n)$ obtained from the first step, is ignored.

- *Integrating language modeling*. In our experiments, only a pure acoustic model was used to decode the input utterance. It is well known that the language model plays a very important role in automatic speech recognition. Obviously by integrating it into the first step, a more accurate sequence of acoustic symbols will be obtained, thereby improving the overall KWS performance.
- *Dealing with insertion and deletion errors*. Based on previous KWS experiments, substitution errors account for around 87.6% of all errors, while insertion errors account for 7.7% and deletion errors for 4.7%. This is the reason that we ignore insertion and deletion errors in the second step. However, this portion of errors is not small in percentage. Some rule-based methods or search strategies should be helpful in improving the KWS detection rate.
- *Considering higher order context*. According to the experimental results presented in this paper, where *CI* considers a one-IF width context, the higher order context information, i.e. the two-IF width context (*LCD2*), outperforms the lower order context information. It is expected that even better results can be obtained using higher order context information, provided that the zero probability estimation problem is well solved.
- *Using back-off methods for probability re-estimation*. In our CDAPP calculation, we assign a small, predefined value to those zero probabilities, which is very approximate. Back-off algorithms have been proven efficient and effective in language modeling. The situation here is similar, so it is expected that back-off methods will work well here as well. Probability re-estimation is a very important part in the proposed method.
- *Using a confidence measure*. This is very important for reducing the false alarm rate.

## 5. REFERENCES


[1] Higgins, A. L.; Wohlford, Robert E., "Keyword Recognition Using Template Concatenation," *ICASSP'1985*, 3: 1233-1236, 1985

[2] Manos, A. S., "A Study on Out-of-Vocabulary Word Modeling for a Segment-Based Keyword Spotting System," M.S. Thesis, Massachusetts Institute of Technology, 1996.

[3] Rose, R.-C.; Paul, D.-B.. "A hidden Markov model based keyword recognition system," Proceedings of *ICASSP'1990*. v.1, pp. 129-132, Albuquerque, NM. 1990

[4] Young, S.; Everman, G.; Kershaw, D.; Moore, G.; Odell, J., Ollason; D., Valtchev, V.; Woodland, P. "The HTK Book 3.1". Cambridge: Entropic, 2001.

[5] Zhang, G.-L.; Sun, H.; Zheng, F.; and Wu, W.-H. "Robust Speech Recognition Directed by Extended Template Matching in Dialogue System," *The 5th World Congress on Intelligent Control and Automation (WCICA)*, June 14-18, 2004, Hangzhou, China

[6] Zhang, J.-Y.; Zheng, F.; Li, J. "Improved Context-Dependent Acoustic Modeling for Continuous Chinese Speech Recognition," Proceedings of *EuroSpeech2001*, pp. 1617-1620. Aalborg, Danmark. 2001

[7] Zheng, F., Song, Z.-J, and Xu, M.-X, "EASYTALK: A large-vocabulary speaker-independent Chinese dictation machine," *EuroSpeech'1999*, Vol.2, pp.819-822, Budapest, Hungary, 1999