Improvement of Rejection Performance of Keyword Spotting
Using Anti-Keywords Derived from Large Vocabulary
Considering Acoustical Similarity to Keywords

Makoto Yamada, Tsuneo Kato, Masaki Naito, and Hisashi Kawai
KDDI R&D Laboratories, Inc, Japan
{ma-yamada, tkato, Hisashi.Kawai}@kddilabs.jp

Abstract
This paper proposes an efficient anti-keyword derivation method to improve the rejection performance of keyword spotting. In this method, each anti-keyword is derived from the large vocabulary lexicon considering acoustic similarity to keywords, making use of the confusion matrix. Experimental results show that a 3% improvement of the rejection rate is obtained compared to conventional methods that do not have our anti-keyword models.

1. Introduction
Keyword spotting is an effective approach to speech recognition applications for detecting several to hundreds of keywords [1]-[4]. For such applications, a false alarm, misdetection of a keyword from an utterance without a keyword, is a significant problem. To reduce false alarms, background models (or garbage models) were introduced [1], [2]. As acoustic models for recognizing keywords are also used as background models in most cases, acoustic discrimination between keywords and non-keywords is not easy. Therefore various techniques have been proposed to alleviate this problem.

We have proposed a two-pass keyword spotting technique based on bilingual HMMs (Hidden Markov Models), which makes use of phonemic garbage models trained on a speech corpus of a language that is different - but acoustically similar - from the target one [5]. An advantage of this technique is that it does not need penalty tuning over garbage models, necessary for ordinary keyword spotting.

Another approach for alleviating false alarms is verifying whether the utterance truly contains a hypothetical keyword using anti-subword models [6], [7]. In [6], each subword in an utterance is verified based on confidence measures derived from subword models and anti-subword models. However, the threshold of the confidence measure is common for all keywords, but it is not practical to determine keyword-dependent thresholds from the viewpoint of computational cost.

Some verification methods using anti-keyword models have been proposed [8]-[10]. In [9], task-dependent discrimination training is conducted to train anti-keyword models using alternative hypotheses generated by a speech recognizer. Although this approach takes account of acoustic confusability between keywords and anti-keywords, an enormous amount of speech data is necessary to obtain accurate anti-keyword models.

In this paper, we propose a keyword spotting algorithm with anti-keywords derived from a large vocabulary based on their similarities, using the confusion matrix. The strengths of our algorithm are as follows: (1) enormous amounts of speech data are not necessary as the similarity is calculated based on the confusion matrix and (2) no need to worry about setting a threshold on the confidence measure as the performance changes gradually vis-à-vis the recognition score.

This paper is organized as follows: Section 2 describes the definition of acoustic similarity between a keyword and a non-keyword and a method for deriving anti-keywords from a large vocabulary based on acoustic similarity. Section 3 describes a method for verifying utterances by using anti-keywords. The results of experiments for evaluating performance of rejection and recognition are described in Section 4.

2. Definition of Word Similarity between Keyword and Non-keyword

2.1. Definition of word similarity between keyword and non-keyword
Figure 1 shows the grammar for keyword spotting in our system. Keywords and anti-keywords are connected in parallel between two garbage model networks.

A word similarity metric, which is a metric of acoustic similarity between subword (phoneme in this study) sequences of a keyword and those of a non-keyword, is computed based on the confusion matrix of the acoustic model of phonemes used for recognizing keywords by dynamic programming (DP) matching. Let $W$ represent the set of subwords. The acoustic word similarity $S(A,B)$ between words $A$ and $B$ is calculated as follows.

1. Derive confusion probabilities for subword $(phoneme) i, j$ from a confusion matrix. An element of the confusion matrix, $n(i,j)$, is the number of output $j$ caused by input $i$ in speech recognition. The set of subwords includes a null subword $\phi$ for the case when a subword is output from a null input (insertion) and the case when null output is observed.
for an arbitrary input (deletion). For subword $i$, the probability of a hit, $P_{\text{hit}}$, the probability of a substitution, $P_{\text{sub}}$, the probability of an insertion, $P_{\text{ins}}$, and the probability of a deletion, $P_{\text{del}}$, are defined as follows:

$P_{\text{hit}}(i) = \frac{n(i, i)}{\sum_{k \in W} n(i, k)}$  \hspace{1cm} (1)

$P_{\text{sub}}(i, j) = \frac{n(i, j)}{\sum_{k \in W} n(i, k)}$  \hspace{1cm} (2)

$P_{\text{del}}(i) = \frac{n(i, \phi)}{\sum_{k \in W} n(i, k)}$  \hspace{1cm} (3)

$P_{\text{ins}}(i) = \frac{n(\phi, i)}{\sum_{k \in W} n(k, i)}$  \hspace{1cm} (4)

2. Decompose keyword A and non-keyword B into sequences of subwords that are units of the confusion matrix.

$A = a_0, a_1, \ldots, a_m$ \hspace{1cm} $B = b_0, b_1, \ldots, b_n$ \hspace{1cm} $(m, n \geq 0)$  \hspace{1cm} (5)

3. The word similarity $S(A, B)$ is recursively calculated by the following equations (6)-(10), where $SIM()$ represents the similarity between the sub-sequence of keyword A and non-keyword B and $sim()$ represents similarity between subwords. The reference range for keyword A extends from the beginning to the $x_e$-th subword, while the reference range of non-keyword B extends from $y_s$-th subword to the $y_e$-th subword.

$S(A, B) = \max \{ SIM(x, y; A, B) \}$  \hspace{1cm} (6)

$SIM(x, y; A, B) = \max \left\{ \begin{array}{ll} SIM(x - 1, y; A, B) + \log P(a_x, b_y) & \text{if } x \leq y \leq y_e - 1 + 1 \\ SIM(x - 1, y - 1; A, B) + \log P(a_x, b_y) & \text{if } x < y \leq y_e - 1 \\ SIM(x - 1, y - 1; A, B) + \log \phi(b_y) & \text{if } 0 \leq x \leq y \leq y_e - 1 \\ 0 & \text{otherwise} \end{array} \right\}$  \hspace{1cm} (7)

where

$SIM(0, y; A, B) = \log P(a_0, b_y)$  \hspace{1cm} (8)

$P(a_i, b_j) = \begin{cases} \rho_{a_i}(i) & \text{if } i = j \\ \rho_{a_i}(i) & \text{if } j = \phi \\ \rho_{a_i}(i) & \text{if } i = \phi \\ \rho_{a_i}(i) & \text{if } i \neq j \end{cases}$  \hspace{1cm} (9)

$P(i, j) = \begin{cases} \rho_{a_i}(i) & \text{if } i = j \\ \rho_{a_i}(i) & \text{if } j = \phi \\ \rho_{a_i}(i) & \text{if } i = \phi \\ \rho_{a_i}(i) & \text{if } i \neq j \end{cases}$  \hspace{1cm} (10)

The term $c$ is a constant representing the minimum length of a subword sequence in calculating word similarity. The constant $c$ enables detection of similarity between similar subwords contained in a keyword and an anti-keyword candidate. By setting $x_e=m$, $y_s=0$, $y_e=n$, this function can be disabled.

Figure 2 Block Diagram of Derivation of Anti-keywords from Large Vocabulary Lexicon Based on the Acoustic Similarity Calculation.

Figure 3 Block Diagram of Keyword Spotting Using Proposed Anti-keyword Models

2.2. Derivation of anti-keywords from large vocabulary based on word similarity

We derive anti-keywords from a large vocabulary of non-keywords using word similarity $S(A, B)$. We assume that (1) a keyword A is set in advance, and (2) that a sufficiently large vocabulary is available for any non-keyword. Figure 2 shows the block diagram of our process for anti-keyword derivation from keywords, confusion matrix and large vocabulary of non-keywords.

If the calculated word similarity of a non-keyword exceeds a predetermined threshold, the non-keyword is added to the set of anti-keywords. In calculating the word similarity, as described in Section 2.1, the entire sequence of keyword and non-keyword as well as the partial sequence are considered. The word similarity $S(A, B)$ is calculated as the
maximum value of word similarities for the full and the partial sequences of keyword A and non-keyword B.

3. Utterance Verification with Anti-keywords

Figure 3 shows the block diagram of a keyword spotting system based on the proposed utterance verification method using anti-keyword models. The recognition lexicon consists of keywords and anti-keywords derived by the procedures shown in Section 2.

When an input speech is provided, a hypothesis with a maximum likelihood $\log P_k$ is first obtained from the HMM-based speech recognizer, in which keywords and anti-keywords are represented by context dependent Japanese HMMs, while garbage models are represented by context independent English HMMs [5]. We employ the English HMMs as their rejection performance is higher than that of Japanese HMMs [5]. An alternative hypothesis with the likelihood $\log P_b$ is obtained from the background network, where phoneme are connected, taking into account the Japanese syllable structure. The normalized score $S$ based on Japanese syllable structure is first obtained from the HMM-based speech recognizer, in which keywords and anti-keywords are modeled using gender-dependent, context-dependent HMM. The models are trained using a Japanese large corpus of landline and mobile telephone speech. The feature vectors are size 38 (12 MFCC + 12 ΔMFCC + 12 ΔΔMFCC + ΔEnergy + ΔΔEnergy). The garbage models are 28 context-independent, 3-state single Gaussian HMMs trained on MACROPHONE American English telephone speech corpus [5].

Table 1 shows the rejection rates for the 123 test utterances that do not contain keywords. The utterances are classified into 5 groups according to word similarity. The threshold for deriving anti-keywords varies from -15 to -5 by 5 for each group. When anti-keywords are not used, as can be seen in Table 1, the rejection rates decrease as the word similarity increases without anti-keywords (see the row “none”), and utterances with similarity larger than -5 are not rejected. In other words, non-keyword utterances cannot be discriminated from keyword utterances. In contrast, when anti-keywords are introduced, the rejection performance is improved as the threshold is lowered.

Table 1 Improvement of Rejection Performance for Various Thresholds for Anti-keywords Derivation

<table>
<thead>
<tr>
<th>similarity to keywords</th>
<th>-25</th>
<th>-20</th>
<th>-15</th>
<th>-10</th>
<th>-5</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>the proposed method</td>
<td>100</td>
<td>94</td>
<td>73</td>
<td>42</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>anti-keywords</td>
<td>100</td>
<td>94</td>
<td>73</td>
<td>42</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>none</td>
<td>100</td>
<td>94</td>
<td>73</td>
<td>42</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

4. Experiments

4.1. Effects of word similarity on the rejection performance

The purpose of this experiment is (1) to investigate the effect of word similarity between spoken non-keywords and keywords on the rejection performance and (2) to evaluate performance improvements when the proposed anti-keyword models are introduced.

The test set consists of Japanese cellular phone speech recorded in a clean environment. It contains 520 short utterances, 397 of which contain a keyword (viz. names of shops, stations, locations) while each utterance of the remaining 123 utterances contain at least one non-keyword. The lexicon consists of 100 keywords, while anti-keywords are derived from 914 non-keywords (the same task as for keywords). The threshold $th$ for acoustic word similarity was set to -5, -10 and -15. Entire sequences of non-keywords are used for anti-keywords in this experiment.

Common to all experiments in this paper, bilingual acoustic models are used as background models [5]. Keywords and anti-keywords are modeled using gender-dependent, context-dependent HMM. The models are trained using a Japanese large corpus of landline and mobile telephone speech. The feature vectors are size 38 (12 MFCC + 12 ΔMFCC + 12 ΔΔMFCC + ΔEnergy + ΔΔEnergy). The garbage models are 28 context-independent, 3-state single Gaussian HMMs trained on MACROPHONE American English telephone speech corpus [5].
5. Conclusions

In this paper, we proposed a derivation method of anti-keywords for improving the rejection performance of keyword spotting. We introduced anti-keywords consisting of non-keywords and parts of them, and word similarity as an acoustic measure of similarity between a keyword and a non-keyword. According to experimental results, we observed (1) the introduction of proposed anti-keywords can improve the rejection performance, and (2) the change of the rejection rate along the threshold $S_N$ for keyword verification is more moderate in the proposed method in comparison to the conventional method. These facts indicate that our method is effective for improving the rejection performance of keyword spotting and is robust vis-à-vis variations of keywords as it can automatically derive anti-keywords from large non-keyword vocabulary.

6. References


