Performance Evaluation of Non-Keyword Modeling for Vocabulary-Independent Keyword Spotting

Young Kuk Kim¹, Hwa Jeon Song², and Hyung Soon Kim¹
¹Department of Electronics Engineering, Pusan National University, Geumjeong-gu, Busan 609-735, Korea
{ykukim, kimhs}@pusan.ac.kr
²Research Institute of Computer Information and Communications, Pusan National University, Geumjeong-gu, Busan 609-735, Korea
hwajeon@pusan.ac.kr

Abstract. In this paper, we develop a keyword spotting system using vocabulary-independent speech recognition technique, and investigate several non-keyword modeling methods to improve its performance. In order to overcome the weakness of conventional syllable model, we propose the syllable filler based on syllable information of keywords and syllable-like filler model. The former prohibits syllable filler network from taking the common syllables that keyword network has for better discrimination between keywords and filler. According to our experiments, syllable filler model using syllable information of keyword yields error reduction rate of 52%-54%. The latter constructs syllable filler network by concatenating the clustered CI phonemes classes. It leads to a 75 times faster decoding than conventional syllable filler while not requiring a large size of text corpus.

Keywords: vocabulary-independent keyword spotting, non-keyword modeling.

1 Introduction

The task of keyword spotting system is to detect a small set of keywords from a speech stream. In the past few decades a lot of effort has been funneled into developing keyword spotting system for applications where the detection of just a few words is enough for a transaction to take place like automated operator services, directory assistance, audio indexing, and surveillance of telephone conversations for security reasons.

In the literature typical keyword spotting methods are classified as follows: The first one is to use a grammar formed by the filler and keyword model [1, 2, 3]. The second one is to use large vocabulary continuous speech recognition (LVCSR) to produce a word string, and then search for keywords in that string [4, 5]. Last one is based on coarse-to-detail search algorithm which reduces the search space by using the approximation by extremely fast method for zooming into regions in speech where a given word has a good likelihood of occurring [6,7]. Among them, the most common approach to keyword spotting system design is to create a network of
keywords and complement it by filler or garbage models trained to account for the non-keyword speech and background noise.

Generally Gaussian mixture model (GMM), monophone-clustering, and syllable filler are widely used as filler models. Among them, syllable filler is known to be a very effective approach to express non-keyword speech segments. On the other hand, the more precise syllable fillers are, the more confusible with keywords they become. And it is computationally expensive to compute syllables fillers sharing HMM parameters with keywords. It is also noted that large text corpus is required to select highly frequent syllables for syllable filler.

In this paper, we propose the syllable filler based on syllable information of keyword to forbid syllable filler network to take the same syllables that keyword network has. We also suggest a syllable-like model. It creates syllable filler network by connecting the clustered CI phonemes so that we are free from a large number of text materials and obtain a significant reduction in computational complexity.

In next section, we describe our baseline vocabulary-independent keyword spotting system for evaluating our methods. Section 3 provides a description of non-keyword modeling methods. Section 4 shows experimental results, and conclusion is drawn in Section 5.

2 Vocabulary-Independent Keyword Spotting System

2.1 Baseline System Architecture

Vocabulary-independent keyword spotting system implemented in this paper is derived from the Korean vocabulary-independent speech recognition system. A schematic of the system is shown in Fig. 1. In this figure, keywords are represented as a concatenation of context-independent subword acoustic models. The optimized hidden Markov models (HMM) described in Section 2 are used for keywords. Non-keywords are modeled as a variety of filler models (GMM, monophone clustering, syllable) in Section 3.

The word spotter needs a finite state network where HMM models for keywords compete in the network with HMM models for non-keywords as shown in Fig. 2. In this figure NKW, KW, and Sil denote non-keyword, keyword, and silence model, respectively. Finite grammar network in Fig. 2 is designed for spotting only one keyword within a sentence. In our experiment we modified the grammar network to deal with one or two keywords in utterance.
2.1.1 Experimental Condition
Our baseline vocabulary-independent speech recognition system uses triphone with continuous mixture density HMM. Each HMM has three states and the number of mixtures per state varied from 1 to 19. The Korean phonetically optimized words (POW) DB [8] are used to construct the speaker-independent (SI) model using 37,992 words uttered by 20 male and female speakers. The speech data is sampled at 16 kHz and segmented into 20 ms frames at every 10 ms. We use 38 dimensional feature parameters including 12 MFCCs, their deltas, double deltas, delta energy, and double delta energy. The Korean phonetically balanced words (PBW) DB, provided by SiTEC (Speech Information Technology and Industry Promotion Center) in Korea, are used for evaluation. We perform the test experiments using 9,040 utterances spoken by the 10 male and female speakers with 452 different words.

2.1.2 Experimental Results
Baseline experiments according to mixture number are shown in table 1. According to table 1, 11 mixtures per state showed the best recognition performance of 97.36%. We used the optimized condition set in our vocabulary-independent keyword spotting experiments.
Table 1. Performance of vocabulary-independent speech recognition

<table>
<thead>
<tr>
<th>Mixture Number</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word accuracy (%)</td>
<td>94.36</td>
<td>96.59</td>
<td>97.21</td>
<td>97.29</td>
<td>97.35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mixture Number</th>
<th>11</th>
<th>13</th>
<th>15</th>
<th>17</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Accuracy (%)</td>
<td>97.36</td>
<td>97.19</td>
<td>96.95</td>
<td>96.86</td>
<td>96.60</td>
</tr>
</tbody>
</table>

3 Non-Keyword Modeling

Keyword spotting system uses filler models that are trained to account for the non-keyword speech and background noise. The performance of an HMM based keyword spotter heavily depends on the ability of the filler models to represent non-keyword speech segments, without rejecting the correct keywords. Since keywords in large vocabulary keyword spotting system are acoustically very similar to some non-keywords or out-of-vocabulary words, it is difficult to correctly spot keywords using conventional filler models. Therefore more sophisticated fillers to make a distinction between keywords and non-keywrods are needed. In following section, we examined various filler models based on GMM, monophone clustering, and syllable filler model.

3.1 GMM Filler Model

GMM [9] filler model is an implicit method for modeling non-keyword. This simple approach consists in training one single filler model on all training data with GMM. A Gaussian mixture density is a weighted sum of $M$ component densities given by

$$p(\tilde{x}, \lambda) = \sum_{i=1}^{M} c_i N_i(\tilde{x}, \mu_i, U_i)$$

where $\tilde{x}$ is a random vector, $c_i, i=i, \ldots, M,$ are the mixture weights, and $N_i$ is the $i$-th component density of Gaussian model $\lambda$ with mean vector $\mu_i$ and covariance $U_i$.

3.2 Monophone Clustering Filler Model [4]

In this case, we use the context-independent (CI) phoneme as filler model. To use all CI phoneme for filler has a problem of confusion between keywords and filler. Therefore filler model is made by merging the CI phoneme models into a small set of filler classes.
The clustering of these CI models is done in the following way: For each class pair starting from all the 45 CI phoneme classes, the distance of the combination of two classes is calculated and the class pair that minimizes the distance is merged into one class. This process is repeated until the desired number of filler classes is obtained. Each HMM of CI phonemes has three states and the number of mixtures per state is one. We examine Kullback-Leibler distance measure for computing dissimilarity between the two CI phonemes in pair.

3.2.1 Clustering with Kullback-Leibler Distance Measure
Kullback-Leibler distance measure [10] was calculated for each pair of phones. Kullback-Leibler distance between probability distributions of \( f(x) \) and \( g(x) \) is given by

\[
KL(f, g) = \int f(x) \log \left( \frac{f(x)}{g(x)} \right) dx
\]  

(2)

Assuming Gaussian probability distribution, the equation (2) is given by

\[
KL(f_x(x), f_y(x)) = \frac{1}{2} \left[ \ln \left( \frac{\sigma_x^2}{\sigma_y^2} \right) - \frac{\sigma_x^2 + (m_x - m_y)^2}{\sigma_y^2} + 1 \right]
\]

(3)

where \( m_x, \sigma_x^2, m_y, \sigma_y^2 \) are means and variances of \( f_x(x) \) and \( f_y(x) \).

Then the Kullback-Leibler distance between two CI phonemes is given by

\[
D_{KL}(p_i, p_j) = \sum_{s=1}^{N} \sum_{d=1}^{V} KL(f_{isd}(x), g_{jsd}(x))
\]

(4)

where \( N \) is the number of states, \( V \) is feature dimension, and \( f_{isd} \) is probability distribution of phoneme \( p_i \) with state \( s \) and \( d \)-th Gaussian component.

Since Kullback-Leibler distance has unsymmetrical characteristic, instead of equation (4), we use symmetric Kullback-Leibler distance shown in equation (5).

\[
D_{KL}(p_i, p_j) = \frac{1}{2} \left[ D_{KL}(p_i, p_j) + D_{KL}(p_j, p_i) \right]
\]

(5)
3.3 Syllable Filler Model

3.3.1 Conventional Syllable Filler [11]
In this section, we investigate syllables as filler model. In Korean, there are about 3,200 syllable units. Syllable filler is a more explicit filler model than any other filler model. But it is computationally expensive to use all syllables for filler. And it has a problem of confusion between keywords and syllable fillers. Therefore it is desirable to cover as many as possible of the unknown words with as less syllables as possible. We selected syllables for syllable filler from a large text corpus called Korean Information Base System 2 (KIBS 2) DB [12], provided by Korea terminology research center for language and knowledge engineering (KOTERM). KIBS 2 is composed of text materials from newscast, newspaper, and book.

Fig. 3 shows block diagram of syllable selection procedure for syllable filler. First, the various symbols, numbers, and other nonorthographic entities of text are converted into a common orthographic transcription suitable for subsequent phonetic conversion. Secondly, grapheme to phoneme conversion (GTP) generates syllable sequences from words. Finally we obtained 2,173 syllables from 13,471,753 syllables by removing repeated ones.

![Fig. 3. High frequency syllable generation](image)

Fig. 4 shows the coverage of the words due to the syllables sorted in descending frequency. In Fig. 4, 90% of the words in KIBS 2 DB can be covered by about 400 syllables. To obtain an optimumal set, we performed experiments accoring to changing the number of selected syllables.
3.3.2 Syllable Filler based on Syllable Information of Keyword
To use syllable filler is a very effective solution because there is no necessity to train a new HMM for filler model. But it happens that filler network has the same syllables that keyword network has. Hence it brings about a problem of confusion between keywords and filler.

In this section, we propose the syllable filler based on syllable information of keywords. In making a filler network, we remove the syllables that are found in both the keyword and the filler network. Consequently it leads to acoustically better discrimination between two kinds of words than the conventional syllable filler.

3.3.3 Syllable-like Filler
From Section 3.3.1, it can be seen that a large number of text corpus and time to analyze the text are needed for syllable filler model. In addition, it is computationally expensive because syllable filler shares HMM parameters which generally have multiple mixtures with keywords.

We proposed syllable-like filler model which doesn’t need a large number of text data, and expensive computation in decoding. Korean syllable structure is composed of consonant/vowel/consonant (CVC). We can make syllable model by concatenating CI phonemes. Instead of all CI phonemes, a small set of filler classes are used by using clustering method described in Section 3.2.
4 Experiments

4.1 Experimental Condition
We used phonetically optimized words (POW) DB [7] to construct the speaker independent (SI) model. According to our baseline experiments in 2.2.1, we obtained the best recognition performance with 11 mixtures per state. We also used sentence database called “CleanSent” from SiTEC(Speech Information Technology and Industry Promotion Center) for keyword model to deal with co-articulation effects in continuously spoken utterances. CleanSent DB is uttered by 100 different male and female speakers with 20,217 sentences.

The automated attendant database with 22 different department names, provided by ETRI, is used for evaluation. We perform the test experiments using 235 utterances which have one keyword and 315 utterances which have two keywords. Feature extraction procedure is the same as described in Section 2.2.1.

4.2 Experimental Results
Tables 2 and 3 show the performance of vocabulary-independent keyword spotting according to a variety of filler models. The number of keywords within a sentence is given in table 3 but not in table 2. In these tables, “POW” and “CleanSent” indicate performance using isolated words and continuously spoken sentences for training DB, respectively. “GMM” indicates optimized GMM filler with 5 mixtures, “Mono” means optimized monophone-clustered filler with 4 clusters, “Syllable” is conventional syllable filler with 60% coverage (88 syllable fillers), “Syllable-like” is syllable-like filler with 4, 10, and 6 monophone clusters for each phoneme in CVC structure. “Syllable*” indicates syllable filler based on syllable information of keyword as described in Section 3.3.2. “1 keyword” and “2 keywords” denote that there are one and two keywords within an utterance, respectively.

In table 2, “GMM” outperformed monophone clustered filler and “Syllable” yielded the best performance in “POW” DB. In the case of “CleanSent” we also get the best result with syllable. Performance of “Syllable-like” filler is comparable to that of conventional syllable filler without a large number of text corpuses. In case of “Syllable”, the number of Gaussians to compute likelihood is 4,536 (56 syllable × 9 states × 9 mixtures). On the other hand syllable-like model need just 60 Gaussians (20 monophone × 3 state × 1 Gaussian). Therefore we obtained 75 times faster in computing Gaussians. “Syllable*” showed a better performance than “Syllable” because common syllables found in both the keyword and filler network are removed from the syllable filler, and obtained the best performance with error reduction of 52%. Performance using “CleanSent” database outperformed “POW” DB because acoustic models from sentence database deal with co-articulation effects in continuously spoken utterances better than those from isolated word DB do.
### Table 2. Performance of keyword spotting according to filler model (GMM: Gaussian mixture model filler, Mono: monophone clustering filler, Syllable: syllable filler, Syllable-like: syllable-like filler model, syllable*: syllable filler based on syllable information of keyword)

<table>
<thead>
<tr>
<th>Training DB</th>
<th>Filler model</th>
<th>Keyword Accuracy (%)</th>
<th>Unknown Keyword Number within Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 keyword</td>
</tr>
<tr>
<td>POW (word)</td>
<td>GMM</td>
<td>5</td>
<td>71.49</td>
</tr>
<tr>
<td></td>
<td>Mono</td>
<td>C4</td>
<td>74.47</td>
</tr>
<tr>
<td></td>
<td>Syllable</td>
<td>60% (88)</td>
<td>77.87</td>
</tr>
<tr>
<td>CleanSent (sentence)</td>
<td>Syllable</td>
<td>50% (56)</td>
<td>79.57</td>
</tr>
<tr>
<td></td>
<td>Syllable-like</td>
<td>(4/10/6)</td>
<td>76.17</td>
</tr>
<tr>
<td></td>
<td>Syllable*</td>
<td>50% (56)</td>
<td>83.40</td>
</tr>
</tbody>
</table>

In contrast with table 2, the number of keywords within a sentence is known in table 3. Table 3 shows the same trend as in table 2. So syllable filler is better than other filler models. “Syllable*” yielded the best performance with 54% error reduction rate. And overall results in table 3 are better than in table 2.

### Table 3. Performance of keyword spotting according to filler model (GMM: Gaussian mixture model filler, Mono: monophone clustering filler, Syllable: syllable filler, Syllable-like: syllable-like filler model, syllable*: syllable filler based on syllable information of keyword)

<table>
<thead>
<tr>
<th>Training DB</th>
<th>Filler model</th>
<th>Keyword Accuracy (%)</th>
<th>Known Keyword Number within Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 keyword</td>
</tr>
<tr>
<td>POW (word)</td>
<td>GMM</td>
<td>5</td>
<td>92.34</td>
</tr>
<tr>
<td></td>
<td>Mono</td>
<td>C4</td>
<td>88.51</td>
</tr>
<tr>
<td></td>
<td>Syllable</td>
<td>60% (88)</td>
<td>95.32</td>
</tr>
<tr>
<td>CleanSent (sentence)</td>
<td>Syllable</td>
<td>50% (56)</td>
<td>92.77</td>
</tr>
<tr>
<td></td>
<td>Syllable-like</td>
<td>(4/10/6)</td>
<td>91.91</td>
</tr>
<tr>
<td></td>
<td>Syllable*</td>
<td>50% (56)</td>
<td>92.77</td>
</tr>
</tbody>
</table>
5 Conclusion

In this paper, we investigated a variety of filler models to improve the performance of vocabulary-independent keyword spotting system. Since keywords in large vocabulary keyword spotting system are acoustically very similar to some non-keywords or out-of-vocabulary words, it is difficult to correctly spot keywords using existing filler models. According to keyword spotting experiments, syllable filler model showed the best performance among the existing filler models. To obtain better performance than syllable filler model, we proposed its two modified versions.

The first one is the syllable filler based on syllable information of keyword, which prevents syllable filler network from having the common syllables that keyword network has. It yielded error reduction rate of 52%-54% for automated attendant database compared with conventional monophone-cluster-based filler model.

The second one is the syllable-like filler model, which make syllable filler network by means of clustered CI phonemes. Without requiring a large number of text corpuses, it obtained comparable performance to that of conventional syllable filler. In addition, it reduced the computational cost 1/75 compared with conventional syllable filler.

Acknowledgments. This paper was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea.

References


12. Korea terminology research center for language and knowledge engineering (KOTERM).: Korean Information Base System 2 (KIBS 2) DB (http://kibs.kaist.ac.kr/kibs/english/)