2011 IEEE International Workshop on Machine Learning for Signal Processing
September 18-21, 2011, Beijing, China

FALSE ALARM REDUCTION BY IMPROVED FILLER MODEL AND POST-PROCESSING IN SPEECH KEYWORD SPOTTING

Amirhossein Tavanaei, Hossein Sameti, Seyyed Hamidreza Mohammadi

Department of Computer Engineering,
Sharif University of Technology, Tehran, Iran
tavanaei@ce.sharif.edu, sameti@sharif.edu, shmohammadi@ce.sharif.edu

ABSTRACT

This paper proposes four methods for improving the performance of keyword spotting (KWS) systems. Keyword models are usually created by concatenating the phoneme HMMs and garbage models consist of all phonemes HMMs. We present the results of investigations involving the use of skips in states of keyword HMMs and we focus on improving the hit ratio; then for false alarm reduction in KWS we model the words that are similar to keywords and we create HMMs for highly frequent words. These models help to improve the performance of the filler model. Two post-processing steps based on phoneme and word probabilities are used on the results of KWS to reduce the false alarms. We evaluate the performance of the improved keyword spotting in FarsDat corpus and compare the approaches. The presented techniques depict better performances than the popular KWS systems.

Index Terms**—** Keyword spotting, Hit ratio, False alarm, Keyword model, Filler model, False alarm reduction

1. INTRODUCTION

The aim of the keyword spotting system is to detect the keywords from speech stream. Several approaches are used for this purpose, such as HMM-based acoustic modeling, LVCSR and phoneme lattice KWS [1].

Feature vectors of speech frames are obtained by different methods such as Mel-Scale, Bark-scale, and LPC cepstral coefficients. The Mel Frequency Discrete Wavelet Transform (MFDWT) feature extraction is also used for phoneme recognition [2]. In this paper we use MFCC features for speech recognition. The HMM is the most popular classifier in speech recognition [3] and we use it in this research for both the keyword and filler models.

The purpose of KWS is to increase the number of true hits and decrease the number of false alarms. When the hit ratio is improved, the system tends to recognize more words as keywords; so the probability of mistakes (False alarm rate) is increased. Recently researchers try to achieve high hit ratio, and control the number of false alarms. Keyword HMMs can be created by whole word training or concatenating the phoneme HMMs. KWS system needs the filler model to detect the non-keywords. Filler model can consist of all phoneme models or GMMs, syllable models, and other speech utterances such as silence HMM [4, 5]. This is the basis of most KWS systems and many examples of them have been designed and evaluated [6, 7 and 8]. If we model all non-keywords as the filler models, the hit ratio and false alarm rates are improved; but this kind of filler model is not conceivable because it needs great time and space. Filler models have very important effect on the false alarm rate and by improving the filler models; the gratifying false alarm rates are achieved. Filler models are also used for ASR to reject the out-of-vocabulary utterances [9]. In [10] a new model addition to the basic filler model is used that models the syllables according to the database and keywords. Tejedor and Colas (2006 [11]) used language model and confidence measure to increase the hit ratio and reduce the false alarm rate. Post-processing on the results of keyword spotting can improve the performance of the system. The probability of the detected keyword is used to reduce the false alarms [12].

In this paper, first, the keyword HMM models with diverse skips in states for keyword models are studied and a delighted hit ratio is achieved. Second we apply a new filler model to reduce the false alarms. Finally two post-processing methods on the results of KWS system are used to reduce the false alarms.

The rest of this paper is on network grammar of keyword spotting (Section 2), we then present the strategy of keyword model creation in Section 3, the filler model and methods of false alarm reduction are presented in Section 4. Finally our experimental results are reported in Section 5 and conclusions are given in Section 6.

2. NETWORK GRAMMAR OF KEYWORD SPOTTING SYSTEM

Keyword spotting system recognizes a set of keywords and rejects other sections of the speech stream. There are two HMM models: keyword and filler models. For detecting the keywords in sequence of frames, the Viterbi algorithm is used. Fig. 1 depicts the network grammar of KWS system. Parameter "*kwp*" is the probability of keyword occurrence

978-1-4577-1623-2/11/$26.00 ©2011 IEEE

and parameter "*fap*" is the probability of non-keyword occurrence.

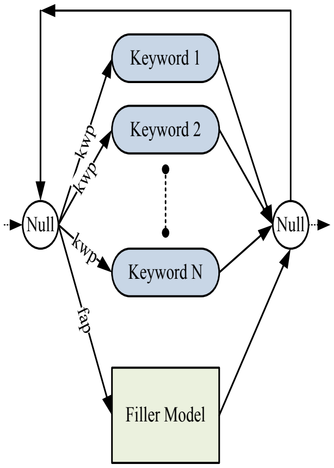

Figure 1: *Network grammar of KWS*

3. KEYWORD MODEL

In this paper the keyword HMM is created by concatenating the phoneme HMMs. This method is called phoneme-based keyword modeling. Assume we use the 3-states HMM for phoneme models and for every pronunciation of each keyword. Simple implementation of keyword models are achieved by simple concatenation of phoneme models as shown in Fig. 2. Other keyword models are obtained by changing the transition probabilities of keyword HMMs. In Fig. 3 one-state skip is shown and Fig. 4 presents one-state and two-states skip. When we concatenate the phoneme HMMs, the number of states in keyword is increased. So increasing the skip parameter in HMM improves the performance of finding the keywords.

4. FILLER MODEL

We use HMMs of all phonemes as the filler model. Rest of this section our developed filler model is described.

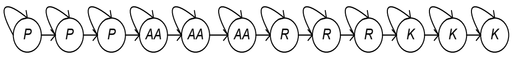

Figure 2: *Simple concatenation of phonemes in keyword HMM of "PARK"*

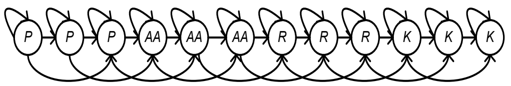

Figure 3: *One-state skip in keyword HMM of "PARK"*

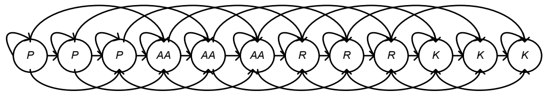

Figure 4: One-state and *Two-states skip in keyword HMM of "PARK"*

4.1. Improved filler model

The words in speech stream can be very similar. For instance the pronunciations of "talk" and "walk" are similar. Generally, there are words with similar pronunciations to keywords. Furthermore some verbs and pronouns are used frequently in speech. We use HMMs of these words to help the filler model. Suppose keyword spotting system wants to detect the "Talk" and "Earth" as keywords; a part of the filler model is presented in Fig. 5.

4.2. Post-processing filler model

The false alarm reduction is proposed to be applies after the keyword decoding. The Viterbi algorithm for KWS represents 4 records for any detected word as follow:

- Detected word (W)
- Start time
- End time
- Log probability of the word (P)

The log probability of a true detected word is higher than the log probability of false alarms. We present two methods to find the false alarms.

4.2.1. Average word probability

Log probability of each detected word is compared with the average log probability of the keyword. Eq.1 computes the average log probability of each keyword.

$$Ave(w) = \frac{\sum_{w \in \text{detected word}} P(w)}{Number\ of\ detected\ word} \qquad (1)$$

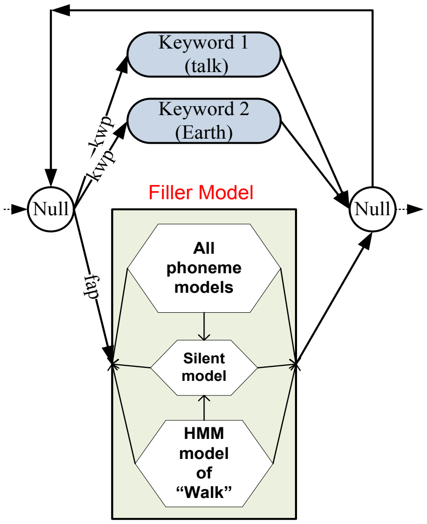

Figure 5: *A part of the improved filler model*

4.2.2. Average phoneme probability

Each true detected phoneme in phoneme recognition step of the KWS system has an average and a minimum log probability depicted by *AvePh(ph)* and *minPh(ph)* respectively. Average log probability of each keyword (*Ave(w)*) is obtained by Eq. 2.

$$Ave(w)=\sum_{ph\in w} AvePh(ph) \tag{2}$$

So we need to compare the average log probability of the keyword with the log probability of each detected word. Eq. 3 depicts this object.

$$\begin{cases} P(w)\geq Ave(w)-k^2(w) & True\,Hit \\ P(w)< Ave(w)-k^2(w) & False\,Alarm \end{cases} \tag{3}$$

Where *k2(w)* is threshold of each word that is obtained as the difference between minimum and average log probabilities of each phoneme (Eq. 4). So false alarms are reduced due to their log probabilities.

$$k^2(w)=\sum_{ph\in w}\left(AvePh(ph)-minPh(ph)\right) \tag{4}$$

5. EXPERIMENTS AND RESULTS

In this paper, the FarsDat corpus was used. The FarsDat contains Persian read speech recordings of 100 speakers. We use 75% of the database to train the system and the other 25% for testing. 13 MFCCs, signal energy, delta coefficients and acceleration of coefficients are extracted for each frame of speech stream. So feature vectors have 39 dimensions. The test data consist of 1.5 hours of wave files with 360 keyword occurrences.

Phoneme HMMs are 5-state left to right models with 16 Gaussian Mixture Models (GMM). Keyword HMMs have simple structure with one-state and two-state skips. Table 1 shows the KWS accuracy rates versus the number of false alarms obtained by three types of keyword models. It is concluded from Table 1 that keyword models with skip of one state give better performance (with 81.30% hit ratio and 54 false alarms) than other types of keyword models. Keyword models with skip of two states give higher hit ratio than the others, but the number of false alarms is not desirable. So we use keyword HMMs with skip of one state.

Table 1. *The KWS accuracy rates based on three types of keyword models*

Keyword Model Structure	Hit Ratio	Number of False Alarms	False Rejection Rate
Without state skip	67.19%	23	32.81%
1-state skip	**81.30%**	**54**	**18.70%**
2-states skip	83.49%	86	16.51%

Based on this KWS system the improved filler model, described in Section 4.1, is used. Table 2 presents the results of KWS with improved filler model. By comparing Table 1 and Table 2, it is concluded that the improved filler model reduces the number of false alarms. In the final step we reduce the false alarms by post-processing of the keyword spotting results. Tables 3 and 4 show the hit ratio and the number of false alarms after using the two methods of post-processing on keyword spotting based on the improved filler model. It is clear from the results that false alarms are reduced. Fig. 6 depicts the ROC curve of the hit ratio versus the number of false alarms after post-processing. It is observed that average phoneme probability post-processing method gives better results in KWS.

Table 2. *KWS accuracy rates based on improved filler model*

Keyword Model Structure	Hit Ratio	Number of False Alarms	False Rejection Rate
Without state skip	69.47%	15	30.53%
1-state skip	**80.43%**	**30**	**19.57%**
2-states skip	84.42%	67	15.58%

Table 3. *KWS accuracy rates after using the post-processing by average word probability method*

Keyword Model Structure	Hit Ratio	Number of False Alarms	False Rejection Rate
Without state skip	69.25%	13	30.75%
1-state skip	**81.30%**	**28**	**18.70%**
2-states skip	83.11%	58	16.89%

Table 4. *KWS accuracy rates after using the post-processing by average phoneme probability method*

Keyword Model Structure	Hit Ratio	Number of False Alarms	False Rejection Rate
Without state skip	69.27%	13	30.73%
1-state skip	**81.27%**	**27**	**18.73%**
2-states skip	83.11%	56	16.89%

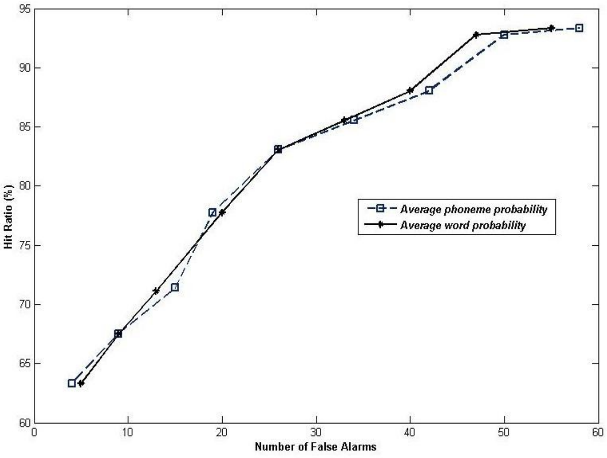

Figure 6: *The ROC curve of KWS after post processing*

For better presentations of the KWS results based on simple keyword structure, keyword HMMs with skip of one state and improved filler model, the ROC curve obtained by changing the keyword occurrence parameters kwp is exhibited in Fig. 7. From Fig. 7 it is clear that KWS based on keyword HMMs with skip of one state and improved filler model have gratifying performance. We use postprocessing based on average phoneme probability on the results of KWS described above. Fig. 8 compares the KWS results with and without post-processing.

Results shown in tables and figures prove that the KWS system based on keyword HMMs with skip of one state and the improved filler model with post-processing by average phoneme probability gives better performance than the other methods.

In order to compare our results with prevalent methods, we implemented a good method of using post-processing step for false alarm reduction introduced by SMIDL LUBOS and TRMAL JAN ([12]). In that work, score comparison post-processing is used. The authors of [12] used the fact that false alarm has a higher score (= is less probable) than a true keyword. Their algorithm compares an average word score "WS" to the average buffer score "BS" corresponding to the word time boundaries (Eq. 5) and 80.64% Figure-Of-Merit (FOM%) was obtained (FOM is computed by average hit ratio per FA/KW/hour from 0 to 10). They used anti-keyword model that is similar to our improved filler model. The anti-keywords are created from keywords by substitution of each character with the closest character. We implemented the score comparison and antikeyword modeling on the FarsDat database and compared their performance with our KWS system. In Table 5 the comparisons of the accuracy rates are presented.

$$BS(w) = \frac{\sum_{t=start(w)}^{end(w)} BUFF(t)}{end(w) - start(w) + 1} \qquad (5)$$

Where

$$BUFF(t) = \min_{w \in all\, words} Score\, of\; w : t \in < start(w); end(w) > \qquad (6)$$

Table 5. *Comparison of implemented method of [12] on FarsDat and our KWS method*

Post-Processing method	Hit Ratio (%)	Number of False Alarms
Score Comparison	79.42	30
SC+Anti-Keyword	78.96	28
APP + IFM	**81.27**	**27**

6. CONCLUSIONS

This paper described the improved keyword spotting. We changed the skips of states in keyword HMMs and compared their results. Keyword HMMs with one-state skip had higher hit ratios than the other keyword HMMs. For false alarm reduction, three methods were used. The first method modeled the words with similar pronunciations to keywords and created HMMs for the pronouns and highly frequent verbs. These models helped the phoneme loop filler model to reduce the false alarms. After obtaining the KWS results, two post-processing methods based on average phoneme probability and average word probability were used to decrease the number of false alarms.

The keyword spotting system based on the improved filler models and post-processing reduced the false alarms to 27 compared to the general KWS with 57 false alarms while the hit ratio of our keyword spotting system was reduced just by 0.03%.

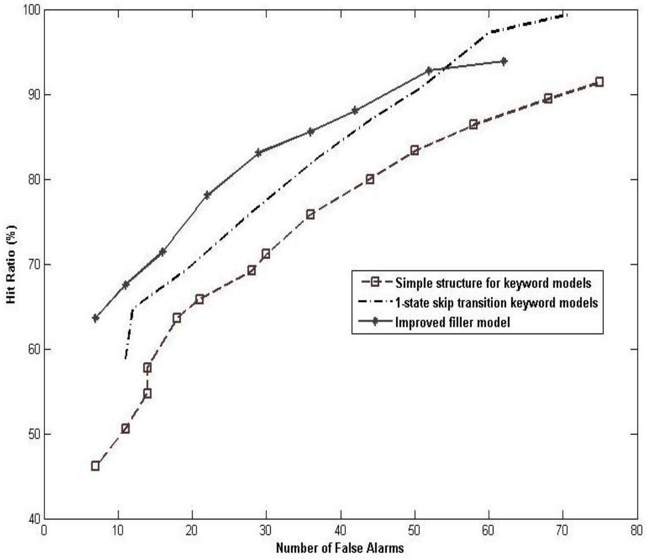

Figure 7: *ROC curve based KW HMM structures and improved filler model*

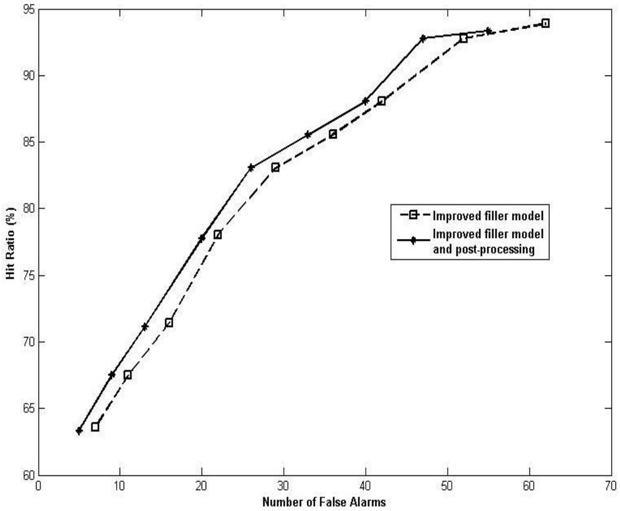

Figure 8: *KWS results with post-processing and without post-processing*

7. REFERENCES

[1] I. Szőke, P. Schwarz, L. Burget, M. Fapšo, M. Karafiát, J. Černocký, and P. Matějka, “Comparison of Keyword Spotting Approaches for Informal Continuous Speech,” Interspeech Conference, Lisbon, pp. 633-636, 2005.

[2] A. H. Tavanaei, M. T. Manzuri, and H. Sameti, “Mel-Scaled Discrete Wavelet Transform and Dynamic Features for the Persian Phoneme Recognition,” International Symposium on Artificial Intelligence and Signal Processing (AISP) 2011. pp.138-140, Tehran, Iran, June 15-16 2011.

[3] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE, vol.77, no.2, pp. 257-286, Feb 1989.

[4] Y. K. Kim, H. J. Song, and H. S. Kim, “Performance Evaluation of Non-Keyword Modeling for Vocabulary-Independent Keyword Spotting,” ISCSLP, Kent Ridge Singapore, pp. 420-430, 2006.

[5] R. E. Meliani, and D. O'Shaughnessy, “New Efficient Fillers for Unlimited Word Recognition and Keyword Spotting,” ICSLP, Philadelphia, PA , USA, pp. 590-593, 1996.

[6] R.C. Rose, and D. B. Paul, “A Hidden Markov Model based Keyword Recognition System,” ICASSP, Albuquerque, pp. 129-132 vol.1, Apr 1990.

[7] K. Knill, and S. Young, “Speaker Dependent Keyword Spotting for Accessing Stored Speech,” Tech. Rep. CUED/F-INFENG/TR 193, Cambridge University Engineering Department, 1994.

[8] A. Jansen, and P. Niyogi, “An Experimental Evaluation of Keyword-Filler Hidden Markov Models,” Tech. Rep. TR-2009-02, U. Chicago, Apr 2009.

[9] M. E. Dunnachie, P. W. Shields, D. H. Crowford, and M. Davies, “Filler Models for Automatic Speech Recognition Created from Hidden Markov Models using the K-Means Algorithm,” EUSIPCO, pp. 544-548, 2009.

[10] S. Tangruamsub, P. Punyabukkana, and A. Suchato, “Thai Speech Keyword Spotting using Heterogeneous Acoustic Modeling,” IEEE International Conference, pp. 253-260, March 2007.

[11] J. Tejedor, and J. Colas, “Spanish keyword spotting system based on filler models, pseudo N-gram language model and a confidence measure,” in Proc of IV Jornadas de Tecnologia del Habla, Zaragoza, pp. 255-260, 2006.

[12] S. Lubos, and T. Jan, “Keyword Spotting Result Postprocessing to Reduce False Alarms,” Recent Advances in Signals and Systems, WSEAS Press, Budapest, pp. 49-52, 2009.