Performance Improvement in Keyword Spotting for Telephony Services

M. Assadi
Master graduated student in Computer Engineering,
AmirKabir University of Technology
Tehran, Iran

M. M. Homayounpour
Professor, Computer Engineering Department
AmirKabir University of Technology
Tehran, Iran

ABSTRACT

In this paper, a new hybrid approach is presented for keyword spotting. The proposed Method is based on Hidden Markov Mode (HMM) and is performed in two stages. In the first stage by using phoneme models, a series of candidate keyword(s) is recognized. In the second stage, word models are used to decide on acceptance or rejection of each candidate keyword. Two different methods are presented in the second stage to improve the spotting performance of the first stage. In the first method, we make a decision to accept or reject each candidate keyword using the similarity between candidate word and the corresponding word model. In the second method, the similarity values between candidate keyword with HMM models of keywords and some HMM models of out of vocabulary words are calculated. These similarity values form a feature vector and are given to a SVM classifier to make the final decision on the correctness of the decision made in the first step. The proposed method was evaluated on two evaluation datasets. Comparing the result obtained from the proposed method and the results obtained by the one stage keyword spotting using the filler models (i.e. the first method on the second step), 5.6% of improvement on the first test set and 4.5% of improvement on the second test set were obtained. By implementation and evaluation of the second method in the second stage, an improvement of 10.3% was achieved using the second dataset.

General Terms

Speech processing, Speech recognition, Keyword spotting, Pattern Recognition.

Keywords

Keyword spotting, Speech recognition, Confidence measure, Hidden Markov Model (HMM), Support Vector Machine (SVM)

1. INTRODUCTION

The keyword spotting system analyzes a given spoken content and searches every speech segment in which one of pre-defined keywords is uttered [1]. Most of works on keyword spotting are based on Hidden Markov models [2, 3]. Besides, there are some methods that are independent of Hidden Markov models [4]. Keywords are important for carrying out the basic concepts of speech; and the meaning of speech, is known by identify them [5]. The problem of detecting a limited number of keywords can be solved in three major ways. The most obvious approach is to use a large vocabulary continuous speech recognition system to produce a word string, and then to search for the keyword in this word string. Theoretically, this is the best way, but there are problems with out-of-vocabulary words, false starts, hesitations, repetition, and other irregularities. The second presented approach is based on analyzing the output of the phonetics decoder – acoustic base form (ABS). The third approach combines the filler model with the confidence measure approach [20]. In this paper, in the first stage for keyword spotting, we introduce and evaluate several filler models in order to decide if a keyword was or was not spoken.

A standard keyword spotting system consists of two stages: model training and keyword spotting [21]. The model training stage aims at constructing two kinds of speech models, respectively called keyword models and garbage models. A keyword model indicates acoustic characteristics of the corresponding keyword, which are estimated from a set of keyword utterances. On the other hand, a garbage model, also known as a filler model, is used to absorb non-keyword segments [31]. The most keyword spotters use a set of Hidden Markov Models (HMM) to represent the non-keyword portions. A widely used approach is to split the training data into keyword and non-key-word data. The keywords are represented by HMMs trained using the keyword speech and the garbage models are trained using the non-keyword speech. The main disadvantage of this method is the task dependency [12,13]. Another approach is to use a common set of acoustic models for both keywords and garbage models. However, this method faces a major problem. In a keyword spotter, the garbage models are usually connected to allow any sequence. Therefore, the keywords are also included in these sequences. When the same training data are used for keyword and garbage models, the garbage models also cover the keywords. In order to overcome these problems, a new method has been proposed in [13] for modeling the non-keyword intervals. In that method, the garbage models are phonemic HMMs trained using a speech corpus of a language other than—but acoustically similar to—the target language.

In a method used for garbage model by Yapanel, several filler models have been used for keywords spotting [14]. Yapanel showed that using less garbage models decreases the efficiency of keyword spotting. But false acceptance rate increases when the number of garbage models increases. Besides, using less garbage model decreases the computation time. The suitable number of garbage models balances false acceptance and false rejection rates.

In this paper, our proposed method is a two stages method. In the first stage several filler models are used and a series of candidate keyword(s) is recognized. In the second stage, two different methods are presented to decide on acceptance or rejection of each candidate keyword. Our method is independent of non-keywords dictionary. Therefore it is not necessary to train the whole system for a new keyword and just the model of that word is needed. A non-keyword is any speech segment other than the keywords.

2. SYSTEM COMPONENTS

This section briefly describes the various components of our system for keyword-spotting.

2.1 Classification by Support Vector Model (SVM)

Support Vector Machines represent a new approach to pattern classification developed from the theory of structural risk minimization [17]. It is a machine learning technique proposed by Vapnik (1995) to solve two class classification problems. This method is defined over a vector space, and the problem is to find the best separator surface that assigns the data to the classes. The training set is optimally separable if it is classified without error and if the distance between the nearest training vectors to the hyper plane is maximum. This method has valuable capacity that improves it among others methods. For example there are no problems with local minimums in the training phase and also it constructs a classifier with maximum generalization [15]. In one of the recommended methods in this paper, SVM is used to improve and confirm the results of keyword spotting obtained from first stage.

2.2 Proposed techniques for improvement of keyword spotting

The method proposed in this paper is performed in two stages. The first stage is a phoneme based keyword spotting system that only consists of keyword models and is used to identify normal sentences in order to obtain some mapping that may be correct or not. It means that when the system encounters a word that is not in the grammar, it will be recognized as a filler model and will be removed or may automatically be mapped incorrectly as a keyword. Therefore the output of this system may consist of some correct recognized keywords or incorrect recognized words. After that in the second stage, word models are used to decide on acceptance or rejection of each candidate keyword.

2.2.1 Stage one: keyword spotting system based on phoneme

In this stage, all keywords which may occur in the input speech and different filler models used for implementation of this stage are defined. Different filler models are:

- Group 1 filler model: a general filler model is trained by all Persian phonemes and used as a filler model named as Out of Vocabulary Model or briefly OOV1.
- Group2 filler models: in this group, Persian phonemes are classified in several phonetic groups and for each phonetic group a model is trained. These models are used as filler models, named as OOV2.
- Group 3 filler models: these filler models include two general filler models where one of them is trained by vowel phonemes and the other is trained by consonant phonemes, named as OOV3.
- Group 4 filler model: As proposed in [13], speech data from some common languages such as English, French, Germany, Japanese, Chinese, etc. is classified in 7 phonetic groups and a filler model is trained for each group. These filler models are named LanGrp filler models, named Langrp.

Each of the aforementioned models is also combined with the Persian phoneme models to consist new groups of filler models. In this case a “ph” is added to the name of filler model group (as can be seen in Figures 1 and 2).

2.2.2. Stage two: improving the first stage results

The second stage is a keyword spotting system based on the word models to improve the first stage results. Two different methods are presented in this paper.

2.2.2.1. First method: comparing with word model

In this method after the first stage, feature vectors of speech segment which have been recognized as a keyword, will be compared to Hidden Markov Model of that keyword and likelihood as a similarity measure will be calculated. Finally the likelihood will be compared to a decision threshold in order to accept or reject the candidate keyword. Before implementation of this stage, the decision threshold for each keyword is determined in a training phase.

2.2.2.2. Second method: using SVM to accept each keyword

Classification of the keywords which have similar pronunciation to others keywords or other non-keywords and revising them for more assessment can improve the accuracy of the keyword detection [16]. According to this point, in the second method, a keyword spotting correcting method, based on support vector machine (SVM) is proposed. The SVM as a method for pattern classification is based on the lowest structural risk theory [3]. In this method, the similarity values between speech segment of candidate keyword with HMM models of some similar keywords and some HMM models of some out of vocabulary words are calculated. Some anti-word models are needed for each keyword [17]. An anti-word model is a non-keyword model which has the most similarity to an assumed word. A non-keyword dictionary is needed in order to make anti-word model of each keyword. In our paper, confidence measure of each candidate keyword is calculated using *n* keywords more similar to candidate keyword and *m* anti-words more similar to the candidate keyword using a speech recognizer based on the Hidden Markov Model.

For implementing the method, a cohort list is specified for each word of the dictionary. Cohort words are those words that have a similar pronunciation to the target word [19]. This list includes both similar keywords and similar anti-words whose HMM models have a good similarity with the speech segment of the candidate keyword. Then, a vector is made that consists of the similarity between target keyword and each model existing in the cohort list of a given keyword. This vector will be given as an input to SVM. The acceptance/rejection decision for a word is based on the confidence score which is provided by SVM classifier. The decision is performed separately for each word in the vocabulary.

3. System evaluation

To evaluate the performance of the proposed method, two evaluation measures [17] are used including the False

Acceptance Rate (FAR), and False Rejection Rate (FRR), defines as:

$$FAR = \frac{Total\ False\ Acceptance}{Total\ False\ Attempts} \quad (1)$$

$$FRR = \frac{Total\ False\ Rejection}{Total\ True\ Attempts} \quad (2)$$

According to equation 1 and 2, Detection Rate (DR) is calculated as follows:

$$DR = 100 - \frac{FAR + FRR}{2} * 100 \quad (4)$$

Where:
Total False Acceptance is Total Number of Non-Keywords recognized as keyword.
Total False Rejection is Total Number of Keywords recognized as Non-keyword.

3.1 Datasets

In order to evaluate the methods explained in this paper, four datasets were used as follows:

- TPersianDat dataset which has been recorded and collected in the Laboratory for Intelligent Multimedia Processing (LIMP), at computer engineering department, Amirkabir University of techmology. This dataset has been segmented and phonetically labeled.
- 40 speech files of telephony FarsDat dataset. This dataset has been used only for a better training of phonetical models [18].
- Telephony bank dataset. This dataset consists of 60 sentences which have been recorded and collected in LIMP laboratory.
- OGI dataset. This dataset consists of speech from10 languages including Farsi language that has been used for training in one of the filler models.

The speech signal has been digitized at 8 kHz sampling rate. Then the pre-emphasized acoustic waveform has been segmented into 30ms frames every 10ms. A Hamming window has been applied to each frame and 12 Mel frequency Cepstral Coefficients (MFCC) have been computed. Then the energy and delta (first order derivatives) and delta-delta (second order derivatives) and delta-delta-delta (third order derivatives) MFCCs have also been calculated. All these feature coefficients are used in each acoustic feature vector for HMM model training and test.

3.2 Results

Stage one: keyword spotting using phoneme models

100 keywords of telephony bank dataset have been used in this experiment. There is no dictionary of non-keywords and all filler models introduced above have been used. Figure shows the results.

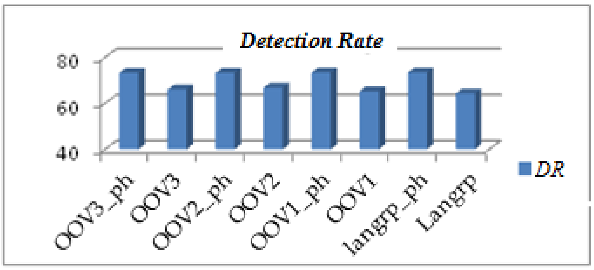

Figure 1: Comparison of keyword spotting Detection Rate versus various filler models

As it is observed in Figure , the lowest detection rate is referred to those cases that Langrp and OOV1 filler models are used to identify non-keyword models. This is due to high level of false acceptance rate. In fact these filler models are not able to recognize and separate non-keywords exactly and can only recognize a small group of them. Of course by using these models along with single phoneme models, performance has increased significantly (comparing Langrp and Langrp_ph). Also, in Figure it can be seen that the maximum detection rate of keyword spotting in this stage is related to filler model of phoneme groups of foreign languages along with Persian phoneme models (Langrp_ph model) with the detection rate of 71.35%. Based on the results, using TPersianDat dataset with 28 keywords and Langrp filler models, the minimum performance is obtained due to high false acceptance rate.

Stage Two: implementing the first method in the second stage

According to the experiments of the stage 1, to improve the results, that filler model group is used that identify more candidate keywords. So, it seems that using OOV1 and Langrp filler models is an appropriate choice because of high level of false acceptance, so and a large number of chosen keyword candidates. For a more complete analysis and assessment of the performance of the system, all the filler models even those with low performance were used. Figure shows the detection rate (DR) obtained in stage 1 and in stage 2 using the TPersian dataset.

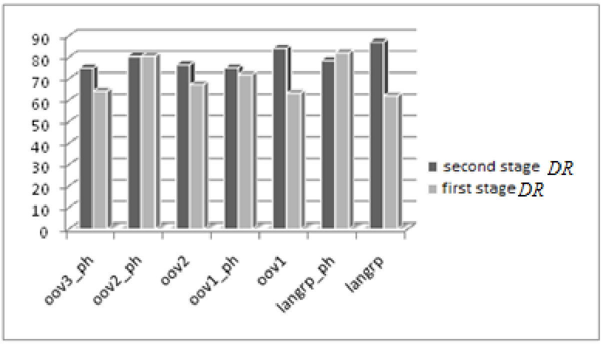

Figure 2: comparison of Detection Rate of keyword spotting in tow step on TPersinaDat

According to the Figure , a considerable improvement in the keyword spotting performance can be observed for all filler models. As expected, the two-stage method presents the maximum detection rate (DR) if the Langrp (filler models obtained using phonetic groups of several languages), i.e. group 4 filler models, are used as non-keywords models. But the detection rate is not as high as the detection rate obtained on telephony bank dataset. It can be explained by this fact that when the number of recognized keywords is low (as it is in the telephony bank dataset), phoneme keywords spotting and filler models have a better performance.

Stage Two: improving the first stage results using second method (using SVM)

Various parameters affect the performance of these methods. These parameters are the number of keywords defined in the cohort list, the number of defined Anti-words in the cohort list. Table 1 shows the effect of the number of words in the cohort list on the Detection Rate. In this experiment only one anti-word has been defined for each keyword in a cohort list.

Table 1: Keyword spotting performance vs. the number of words in the cohort list

Number of keyword in cohort list	Number of anti-words	Detection Rate
1	1	71.0
2	1	78.8
3	1	80.8
4	1	80.6
5	1	79.2
6	1	74.5
7	1	75.0

According to Table 1, when the number of anti-words is kept 1, the best efficiency is obtained when three keywords are included in a cohort list. In another experiment, the effect of the number of anti-words existing in cohort list has been assessed and it has been shown that the best efficiency is obtained when three anti-words are included in a list of cohorts.

In the experiments on telephone bank dataset, 100 keywords have been identified. For evaluation performance of this method on various numbers of keywords, we repeated the experiments on the same dataset but for just 50 keywords. According to the results, three keywords and three anti-words are appropriate number of keywords and non-keywords in a cohort list.

The noticeable point in this experiment is that efficiency decreases when the number of keywords decreases. It is due to the method complication. In fact, when filler models are used in systems with limited number of keywords, good detection rate is achieved. Using inappropriate filler models and growing false acceptance error in order to improve the performance of support vector machine, will not affect the detection rate. This method is served to improve the word spotting performance when a dataset including a large number of keywords is used. Experiments show that in this case, using support vector machine has high effect on detection rate improvement. Table 2 presents the results of using this method and its comparison with two former methods are shown.

As can be seen, maximum detection rate was obtained in identification of 100 keywords in telephony bank dataset in two-stage method using SVM. But the false acceptance error rate confidence measure method is used is lower than the performance for two other methods. But totally the efficiency of using SVM is prominent due to the increased number of false acceptance rate in confidence measure method.

Table 2: Keyword spotting performance for recognition of 100 keywords vs. various methods on telephony bank dataset

Method	FAR	FRR	Detection Rate
SVM Classification	32.9	3.9	81.6
Threshold	10.1	38.3	75.8
One-stage filler model	36.2	21.1	71.3

4. Conclusion

According to the results, using the confidence measure implementation, performance improvements of 4.5% on telephony bank dataset and 5.6% on TPersianDat dataset were achieved. The performance improvement on telephony bank dataset was 10.3% when support vector machine technique was used. Since in the first stage of our proposed method, those words that are highly suspected to be non-keywords are omitted, and even those speech parts which have been similar to a keyword have not been rejected and have been kept to be checked afterwards, so false rejection rate has decreased. Usually false acceptance rate increases by decreasing the false rejection rate, but in our method, which uses a phoneme-based speech recognition system, if a non-keyword speech is similar to a word, it is considered as a keyword candidate and is reassessed in the second stage of the algorithm. This reduces the false acceptance error.

There is less false acceptance rate by using SVM, compared to the one-stage common method, when the optimal filler models are used. Using SVM, more information is used for calculation of confidence measure and better results are achieved. Of course in the proposed method, misclassification error is also considered as false rejection error. In fact SVM can very well recognize a keyword and distinguish between a keyword and other similar words and non- keywords. Therefore in the cases where a keyword is recognized instead of another keyword (misclassification), that keyword is considered as a false keyword and is rejected.

There are many ways for growing the efficiency of method that can be experimented and assessed as future works including increasing the training data for keyword model training, focusing on making better models for keywords and their combination and using common suffixes and prefixes of words for training a set of filler models.

References:

[1] Jeong-Sik Park, Daejeon, South Korea, International Journal of Multimedia and Ubiquitous Engineering Vol. 7, No. 2, April, 2012. Confidence Measure for Utterance Verification in Keyword Spotting System

[2] H. Ketabdar, J. Vepa, S. Bengio, and H. Boulard, Proceedings of Interspeech, Pittsburgh, Pennsylvania, 2006. Posterior based keyword spotting with a priori thresholds,

[3] Y. B. Ayed, D. Fohr, J. P. Haton, and G. Chollet, Proceedings of International Conference on Audio, Speech and Signal Processing, Montreal, Canada, 2004. Confidence measure for keyword spotting using support vector machines

[4] J.Keshet, D. Grangier, and S. Bengio, Workshop on Non-Linear Speech Processing NOLISP, 2007. Discriminative keyword spotting

[5] E. Gouws, K. Wolvaardt, N. Kleynhans, and E. Barnard, Proceedings of the 15th Annual Symposium of the Pattern Recognition Association of South Africa, p. 169, 2004. Appropriate baseline values for HMM-based speech recognition

[6] Jeong-Sik Park, Daejeon, South Korea, International Journal of Multimedia and Ubiquitous Engineering, April 2012. Confidence Measure for Utterance Verification in Keyword Spotting System

[7] S. Veisi, MSc thesis, Sharif University, 2006. Recognition of Out of Vocabulary Words in order to improve the Performance of Speech Recognition Systems (in Farsi)

[8] Matsushita, M., Nishizaki, H., Utsuro, T. Kodama, Y., Nakagawa, S., Toyohashi University of Technology, Japan; Kyoto University, Japan, Eurospeech, 2003.Evaluating Multiple LVCSR Model Combination in NTCIR-3 Speech-Driven Web Retrieval Task.

[9] Davis, S. B., Mermelstein, P., IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-28 pp. 357-366, August 1980. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences.

[10] Young, S., IEEE Signal Processing Magazine, September 1996. A Review of Large-Vocabulary Continuous-Speech Recognition

[11] Digalakis V., Murveit, H., International Conference on Acoustics, Speech, and Signal Processing, Adelaide, Australia, pp. 537-540, 1994. Genones: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer.

[12] Yu, P., Chen, K., Ma, C., Seide, F., IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 5, no. 13, pp. 635–643, 2005.Vocabulary independent indexing of spontaneous speech.

[13] Heracleous, P., Shimizu, T., Speech Communication, pp. 373–386, 2005. A novel approach for modeling non-keyword intervals in a keyword spotter exploiting acoustic similarities of languages.

[14] Yapanel, U., Ph.D. thesis, Istambul Teknik University, 1997. Garbage modeling techniques for a Turkish keyword spotting system.

[15] Clarkson, P., Moreno, P., Acoustics, Speech and Signal Processing, pp. 2:585–588, 2000. The use of support vector machines for phonetic classification.

[16] Mori, S., Nishimura, M., Itoh, N. IBM Japan Ltd., Japan, Eurospeech, Geneva, Switzerland, 2003.Language Model Adaptation Using Word Clustering.

[17] Shilei, H., Xiang, X., Jingming, K., Department of Electronic Engineering, Beijing Institute of Technology, Beijing, P.R.China, INTERSPEECH, 2006. Improving the Performance of Out-of-vocabulary Word Rejection by Using Support Vector Machines.

[18] Bijankhan, M., et al (1994), Proc. Australian Conference On Speech Science and Technology. Vol 2, pp. 826-830, 1994. FARSDAT – The Speech database Of Farsi Spoken Language.

[19] K. Thambiratnam, S.Sridharan, EuroSpeech, 2003. Isolated word verification using Cohort Word-level Verification.

[20] Lobus smidle, Josef V. Psutka, INTERSPEECH, ICSLP, 2006. Comparison of Keyword Spotting Methods for Searching in Speech.

[21] Z. Chenyan, L. Shuqin and S. Chengli, Proceedings of the 8th International Conference on Signal Processing, 2006.

[22] Study of Design and Implementation of Speech Keyword Recognition System based on Streaming Media.