Urdu Keyword Spotting System using HMM

Saad Irtza
EED, UET, Lahore
Saadirtza786@gmail.com

Khawer Rehman
CLE, KICS, UET, Lahore
k.rehman163@gmail.com

Dr Sarmad Hussain
CLE, KICS, UET, Lahore
sarmad.hussain@kics.edu.pk

ABSTRACT

This paper reports the development of Urdu keyword spotting system (KWS). The approach in the development of KWS is based on filler models to account for non-keywords speech intervals. An impact of using different training datasets to develop filler models has been explored. In addition, a phoneme recognizer (PR) based on all phone model automatic speech recognition system (ASR) has been developed on keywords. Training and decoding parameters of KWS system have been tweaked to get the optimum performance. In the end, KWS and PR systems are integrated and string matching algorithm has been used to improve the performance of Urdu keyword spotter system. The overall system accuracy is 94.59% on the data set used.

Keywords: Automatic speech recognition (ASR), Keyword spotting system (KWS), Out of vocabulary words (OOV).

1. INTRODUCTION

Automatic Speech Recognition is a key component in applications e.g. speech document retrieval (SDR) and human-computer interaction via voice commands. Keyword Spotting (KWS) is a technique which is used to decode only particular words from a continuous speech (Tejedor, 2006). It is extensively used in large vocabulary ASR systems which are subjected to out of vocabulary (OOV) words. Generally in dialogue systems, users speak some extra words other than exact query [11] therefore these ASRs often encounter out of vocabulary OOV words. KWS is used to spot the desired words in continuous speech. For instance in weather mobile service, the user is instructed to speak the desired district name to acquire its weather report, but in some cases the users speak complete sentences e.g. "مجھے وہلارا ک مسوم ہنا کرنا ہے-" ("I need to find the weather of Lahore"). In such cases KWS must spot "Lahore" in the input string. A good keyword spotter should identify all the keywords and minimize the false alarms i.e. not decode non-keyword parts of speech as keywords. This paper reviews some relevant work done on KWS, followed by the experimental details and the results of the current work on as system being developed for Urdu.

2. LITERATURE REVIEW

Different techniqueshave been deployed for keyword spotting. KWS has been developed on five keywords of Urdu using word boundary detection [12]. Training and testing data set consists of isolated words of 7500 and 3200 utterances respectively. The accuracy of system has been found to be 98.1%. Sliding Model Method [8] has been used to develop KWS which isimplemented withDistance time Warping (DTW) and HMM [10].In this method,feature vectors (512 point FFT) are extracted from speech and acoustic vectors have been prepared for each training sample. In the decoding process, sliding window is used to find the distance between acoustic vectors of input speech file and acoustic vectors of keyword. A 20,000 vocabulary size has been used in this system. Testing data set consists of 100 utterances from 14 male and female speakers. The word error rate has been found to be 10.6%. In recent years mostly HMM based keyword spotting techniques are used [1][2][6][7][11]. In [1][6][7] keyword spotting using filler model is implemented. In filler model technique, non-keywords are modeled as fillers while keywords are modeled [7]. Filler model can be modeled on word level or phoneme level [1][2]. Different filler models results in different hit and false alarm rate [1]. Bengali KWS has been developed on 12 keywords using filler modeling approach. Training data set consists of 350 utterances of keywords and subset of TIMIT English speech corpus has been used to develop filler model. Test data set consists of 240 speech utterances. The overall accuracy of system has been found to be 95.83%. The performance of KWS has been improved by using phoneme recognition in the first stage and in second stage, search for keywords using phone lattice [4][5], edit distance algorithm [3] or string searching algorithm [8]. These methods report good accuracy with low false alarm rate but required large amount of training data which should cover all vocabulary and are also computational very expensive [4]. Spanish KWS on 80 keywords has been developed using Albazyin database. Confidence Measure method for keywords is implemented to decrease false alarms. The hit and false alarm rate of the system has been found to be 84.33% and 41.44% respectively.

3. METHODOLOGY

The performance of dialogue system degrades because of OOV words [11]. The objective is to develop a KWS to address the OOV words and to spot keywords in unconstrained Urdu speech with high hit rate and minimal false alarms. Filler modeling technique has been implemented to detect eight locations names of district Lahore. Figure 1 shows the system architecture. It consists of Keyword Spotter (KWS) and Phoneme Recognizer (PR). KWS is implemented by using filler modelling. All phone model is implemented in PR. Speech input is processed by KWS and PR processes. The output of KWS is stream of OOV words and keywords while PR outputs string of phones. Keyword Detector (KWD) measures the confidence score of keywords spotted by KWS in phonemes string decoded by PR.

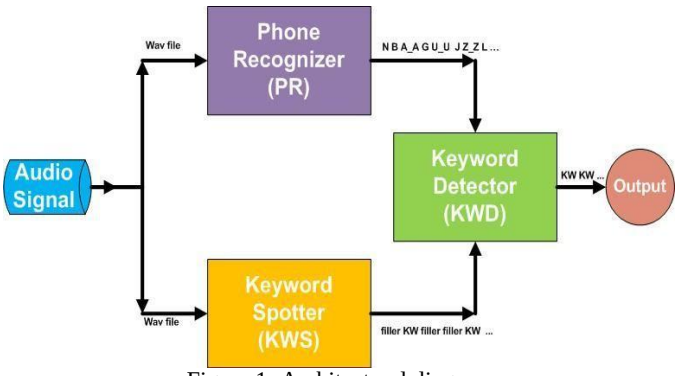

Figure 1: Architectural diagram

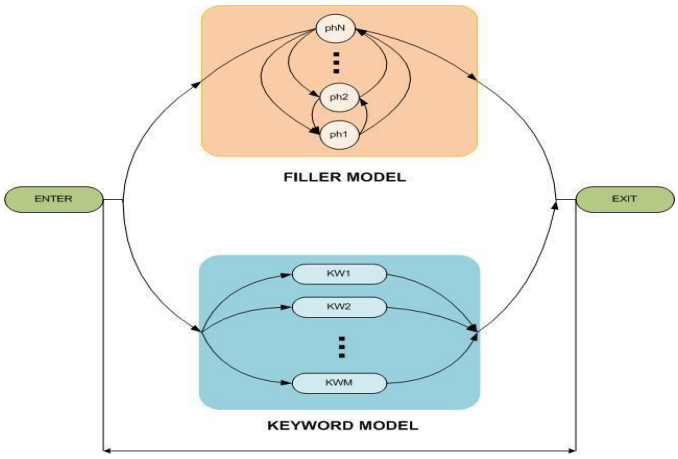

Figure 2: HMM model of KWS

Keyword Spotter is based on Hidden Markov Model (HMM). Acoustic model has been developed using HTK toolkit and Julius is used to test the performance of acoustic model. In KWS, all the non-keywords are modeled as fillers and transcribed at phoneme level. Keyword models are for isolated words and transcribed at word level. The HMM model of KWS is shown in Figure 2.

Phone Recognizer (PR) is implemented by using CMU Sphinx toolkit. The tri-phone basd acoustic model has been developed. Training data used in PR is same as that has been used for training of keywords. Keyword Detector compares the outputs of KWS and PR. It validates the presences of keyword by measuring its confidence in output of PR. For confidence measuring Bitap algorithm [2] is used.

In the first experiment, three different training data sets has been used to model the filler words. The datasets used are: 1) 49 location names of Lahore district, 2) 19 district names of Pakistan, 3) continuous spontaneous speech with general Urdu vocabulary coverage. Table-1 describes the detail of training datasets. In experiment 2, different training and decoding parameters have been tweaked. The tweaking includes: 1) number of states of HMM of keyword, 2) language weight, 3) word insertion penalty. The best performance training dataset has been used in this experiment to model filler words.

	Training Datasets of KWS			Training Datasets of PR
	Location names	District names	Spontaneous speech	Location names
Vocabulary size	49	19	12,883	49
Number of Speakers	300	600	10	300
Total Utterances	1896	22779	22550	1896
Sampling rate	16KHz	16KHz	16KHz	16KHz
Duration (Hours)	0.5	2.7	2.7	0.5
Acoustic model	All phone	All phone	All phone	All phone
Keywords	8	8	8	-

Table 1: Training datasets

Table 2 describes the data used to test the system trained on the different training data sets.

Keywords vocabulary size	8
Number of Speakers	10
Total Utterances	82
Sampling rate	16KHz
Duration (minutes)	80
Sentence templates	8
Language weight	15
Word insertion penalty	-10

Table 2: Testing datasets

4. EXPERIMENTAL RESULTS

Table 3 and Figure 3 give the recognition results of KWS systems developed. The overall accuracy is higher when the training set is from the same domain, but highest when general Urdu corpus is used with larger amount of training data, giving 94.59% accuracy.

	Training datasets					
	Location names		District names		Spontaneous speech	
Keywords	37	-	37	-	37	-
Hits	31	83.78%	27	72.97%	35	94.59%
Misses	6	16.2%	10	27%	2	5.4%
False Alarms	6	16.2%	6	16.2%	6	16.2%

Table 3: Recognition results of KWS

Table 4 describes the recognition results of PR.

Test utterances	Language weight	Accuracy (%)
180	5	97.8

Table 4: Recognition results of PR

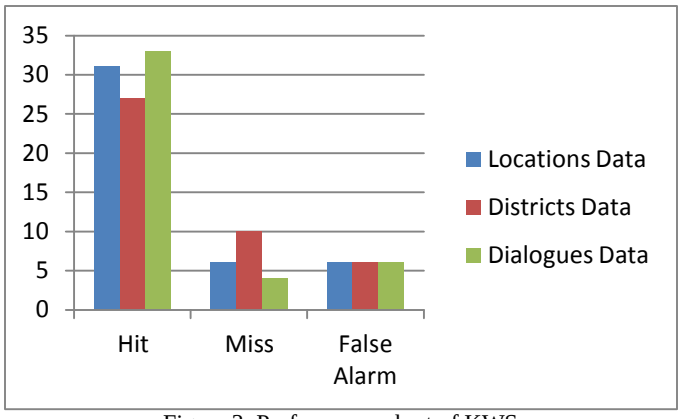

Figure 3: Performance chart of KWS

Figure 4 shows the effect of varying number of states of keywords on hit rate and false alarm.

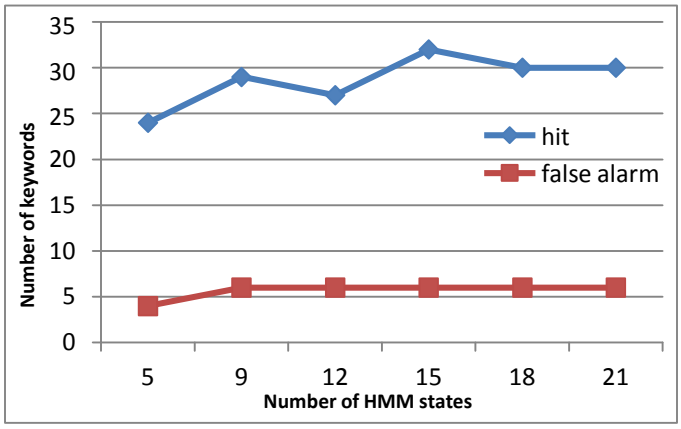

Figure 4: Effect of tweaking HMM states on hit rate and false alarm

Figure 5 shows the effect of varying language weight on hit rate and false alarm.

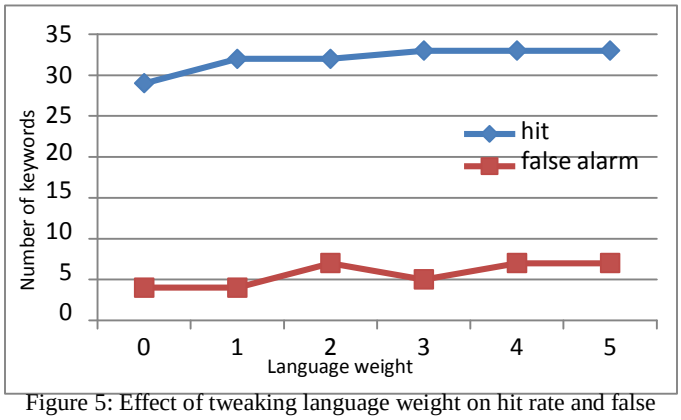

Figure 5: Effect of tweaking language weight on hit rate and false alarm

Figure 6 shows the effect of varying word insertion penalty on hit rate and false alarm.

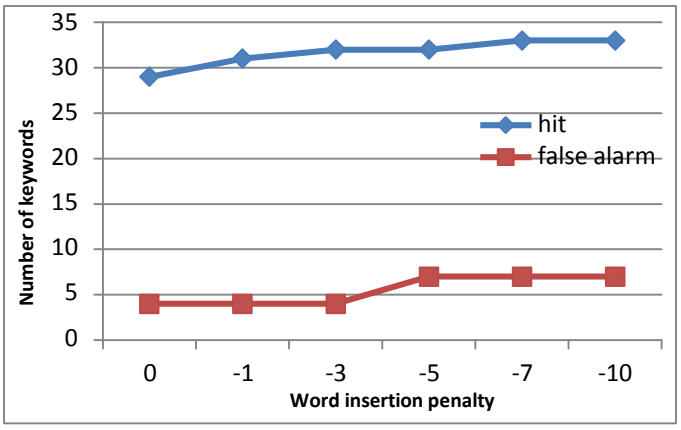

Figure 6: Effect of tweaking word insertion penalty on hit rate and false alarm

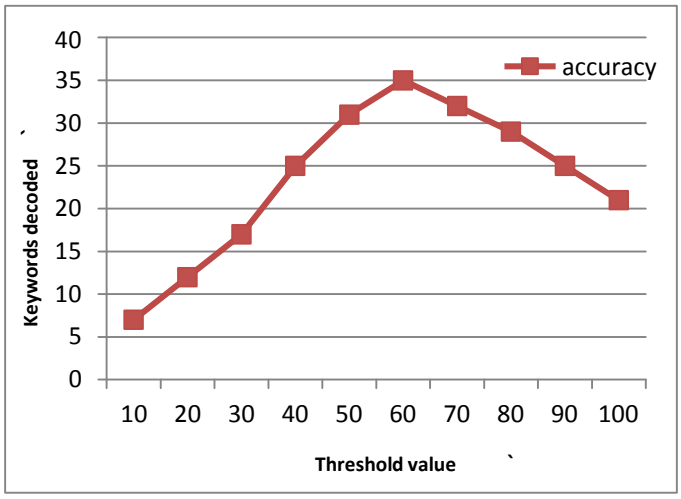

Figure 7: Effect of tweaking threshold value on hit rate

5. DISCUSSION

Table 3 describes the hit rate, miss rate and false alarmon three different datasets. False alarm in each dataset is same. Miss rate is maximum i.e. 10 (27%) on location names dataset and minimum i.e. 2 (5.4%) on spontaneous speech out of 37 utterances of keywords in 82 sentences. Best hit rate of 35 (94.59%) has been achieved on spontaneous speech. Figure 4 describes the effect of changing the number of states of keywords on hit rate and false alarm. In all phone model, 5 number of states have been used for all phonemes. The keywords consist of five to seven phonemes. It has been explored how many states are required to model each keyword. Figure 4 shows that 15 number of states are sufficient to model a keyword that consist of five to seven phonemes.

In the second experiment, the acoustic model has been developed on optimum value of number of states of keywords. Figures 5 and 6 shows the effect of tweaking decoding parameters i.e. language weight and word insertion penalty. Figures 5 and 6 show that hit rate and false alarms will increase with the increase in language weight and word insertion penalty. Language weight and word insertion penalty has been

selected such that hit rate is maximized. The false alarms have been reduced by tweaking the KWD module. The keyword will be considered correct if the output of bitap algorithm is equal to minimum threshold value. The threshold value is tweaked to minimize the false alarm without effecting hit rate. Figure 7 shows the effect on tweaking the threshold value on hit rate. The optimum value of threshold comes out to be 60% for all keywords. False alarm has been reduced from 16.2% to 5.4%.

6. CONCLUSION

It is concluded from this experiment that to increase the performance of ASR system states of words or phonemes should be tweaked in training process. Decoding parameters have significant effect on performance of keyword spotter system. The performance of string matching algorithm also effect the accuracy of keyword spotter.

7. REFERENCES

[1] Tejedor, Javier, and José Colás. "Spanish keyword spotting system based on filler models, pseudo N-gram language model and a confidence measure." *Proceedings of IV Jornadas de TecnologıadelHabla* (2006): 255-260.

[2] S.Das and P.C Ching, "Speaker Dependent Bengali Keyword spotting in unconstrained English Speech", A Project report, Indian Institute of Technology Guwahati, India, 2005

[3] K. Audhkhasi and A. Verma, "Keyword search using modified minimum edit distance measure," Proc. ICASSP-2007, vol. 4, pp. 929-932, 2007.

[4]Lin, Hui, Alex Stupakov, and Jeff A. Bilmes. "Spoken keyword spotting via multi-lattice alignment." *INTERSPEECH*. 2008.

[5] Lin, Hui, Alex Stupakov, and Jeff Bilmes. "Improving multi-lattice alignment based spoken keyword spotting." *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009.

[6]Li, Weifeng, AudeBillard, and HervéBourlard. "Keyword Detection for Spontaneous Speech." *Image and Signal Processing, 2009. CISP'09. 2nd International Congress on*. IEEE, 2009.

[7]Szöke, Igor, et al. "Phoneme based acoustics keyword spotting in informal continuous speech." *Text, Speech and Dialogue*. Springer Berlin Heidelberg, 2005.

[8]Silaghi, Marius-Calin. "Spotting Subsequences Matching an HMM Using the Average Observation Probability Criteria with Application to Keyword Spotting." *Proceedings of the National Conference on Artificial Intelligence*. Vol. 20. No. 3. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[9] Kim, Joo-Gon, Ho-Youl Jung, and Hyun-Yeol Chung. "A keyword spotting approach based on pseudo N-gram language model." *9th Conference Speech and Computer*. 2004.

[10]Nitta, Tsuneo, et al. "Key-word spotting using phonetic distinctive features extracted from output of an LVCSR engine." *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*. 2003.

[11]Wilpon, Jay G., et al. "Automatic recognition of keywords in unconstrained speech using hidden Markov models." *Acoustics, Speech and Signal Processing, IEEE Transactions on* 38.11 (1990): 1870-1878.

[12]Juang, B. H. "Recent developments in speech recognition under adverse conditions." *First International Conference on Spoken Language Processing*. 1990.