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

Introduction

Keyword Spotting (KWS) is a technique which is used to detect and decode only particular words in a continuous speech. It is extensively used in limited vocabulary ASR systems which are subject to out of vocabulary (OOV) words. For instance, in recorded utterance مجھے لاہور کا موسم جاننا ہے only the word لاہور is of our interest and needs to be spotted. This paper explores an HMM-based technique which models each keyword separately but uses a single model, called *filler* model, for non-keywords. The overall system accuracy is 94.59% for 8 keywords.

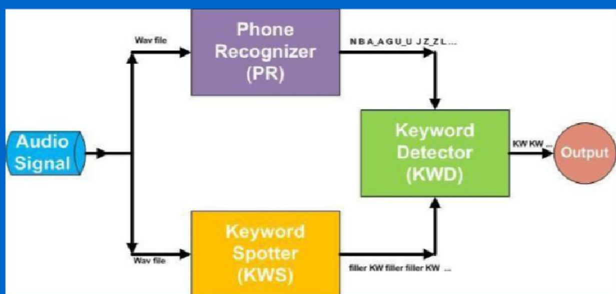

Figure 1 System Architecture

Methodology

- Keyword Spotter is built using HTK toolkit, which models the keywords using Hidden Markov Models (HMM), whereas testing is performed through Julius decoder.
- Figure 1 shows the system architecture. It consists of Keyword Spotter (KWS) and Phoneme Recognizer (PR).
- Each keyword has its unique HMM model but all non-keywords have a single model, called filler model.
- Phone Recognizer is built using Sphinx toolkit, and it decodes the phonemes in the given utterance.
- The recorded user utterance is passed through both the Keyword Spotter and Phoneme Recognizer to reduce false alarms.
- Keyword spotter gives a sequence of keywords and fillers (non-keywords), whereas the Phone Recognizer gives a sequence of phonemes.
- Fillers are discarded and keywords and phonemes are fed to keyword detector which compares the two sequences using a String Matching algorithm, and outputs only valid keywords.
- The objective is to spot keywords in unconstrained Urdu speech with high hit rate and minimal false alarms.

Results

Figure 2 gives the results of testing 37 instances of Keywords in different carrier sentences on Keyword Spotters built using 3 types of training data, which is presented in Table 1. Figures 3, 4 and 5 illustrate the effect of tweaking various system parameters on the hit rate and the false alarm rate of the keyword spotter.

Table 1 Training Datasets

	Vocabulary Size	Number of Speakers	Total Utterances	Sampling Rate	Duration (Hours)	Keywords
Location Names	49	300	1896	16k	0.5	8
District Names	19	600	22779	16k	2.7	8
Spontaneous Speech	12883	10	22550	16k	2.7	8

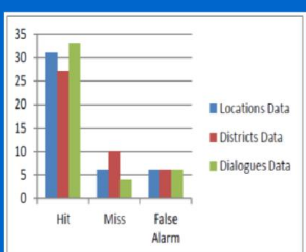

Figure 2 Performance Chart of KWS

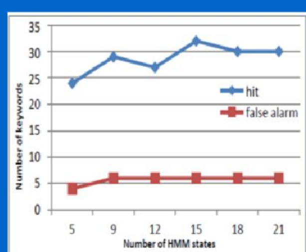

Figure 3 Effect of tweaking HMM states on hit rate & false alarm

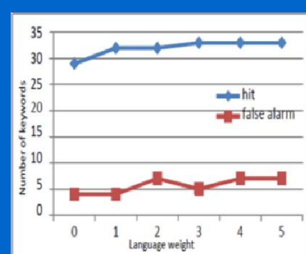

Figure 4 Effect of tweaking Language Weight on hit rate & false alarm

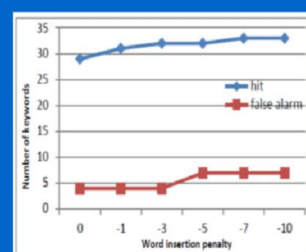

Figure 5 Effect of tweaking Word Insertion Penalty on hit rate & false alarm

Discussion

- Figure 2 describes the hit rate, miss rate and false alarm rate of three different training datasets, for 37 utterances of keywords in 82 testing sentences.
  - **False alarms** 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 dialogues speech, out of 37 utterances of keywords.
  - Best **hit rate** of 35 (94.59%) has been achieved on spontaneous speech.
- Figure 3 describes the effect of changing the number of states of HMMs of keywords on hit rate and false alarm.
  - The keywords consist of five to seven phonemes. Theoretically, 3 states are required to model each phoneme, which makes the ideal number of HMM states to be around 20 in a keyword of about six to seven phonemes.
  - But fig. 3 shows that **15 states** are sufficient to model a keyword and the accuracy actually drops after that.
- Figures 4 and 5 show the effect of tweaking decoding parameters, **language weight** and **word insertion penalty**, on the hit rate and the false alarms.
  - Both hit rate and false alarms increase with the increase in language weight and word insertion penalty, and there is no clear optimum value.
  - The best compromise is obtained with L**anguage Weight of 3** and **Word Insertion Penalty of -3**, which gives minimum false alarms and reasonably high hit rate.

Conclusions

- Out of 3 training datasets, best results (**94% hit rate**) are obtained on the spontaneous or dialogues speech dataset.
- 15 HMM states are enough to model a keyword with 5-7 phonemes.
- Increasing the value of decoding parameters like Language Weight and Word Insertion Penalty increase the hit rate as well as false alarm rate. So in absence of an optimal value, we have to select the values which give the best compromise i.e. minimum false alarm rate with as high hit rate as possible.

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.