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 "Abu Riziq Rizik Abdur Rehman", only the word "Rizik" 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.

### Results

Figure 2 gives the results of testing 37 instances of Keywords in different carrier sentences on Keyword Spotter 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>
<thead>
<tr>
<th>Vocabulary Size</th>
<th>Number of Speakers</th>
<th>Total Utterances</th>
<th>Sampling Rate</th>
<th>Duration (Hours)</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location Names</td>
<td>49</td>
<td>300</td>
<td>1896</td>
<td>16k</td>
<td>8</td>
</tr>
<tr>
<td>District Names</td>
<td>19</td>
<td>600</td>
<td>22779</td>
<td>16k</td>
<td>8</td>
</tr>
<tr>
<td>Spontaneous Speech</td>
<td>12883</td>
<td>10</td>
<td>22550</td>
<td>16k</td>
<td>8</td>
</tr>
</tbody>
</table>

### Methodology

- **Keyword Spotter** is built using HTK toolkit, which models the keywords using Hidden Markov Models (HMM), whereas testing is performed through Julius decoder.
- **Phone Recognizer** is built using Sphinx toolkit, and it decodes the phonemes in the given utterance.
- The recorded utterance is passed through both the Keyword Spotter and Phone Recognizer to reduce false alarms.
- The objective is to spot keywords in unconstrained Urdu speech with high hit rate and minimal false alarms.

### Discussion

- **False alarms** in each dataset are 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.
- Figure 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 dear optimum value.
- The best compromise is obtained with **Language 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