WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Speech Keyword Spotting (KWS) using KALDI toolkit

Talware Samyak Manohar[1], Vipul Singh[2], Chinmay Sharma[3], Hrithik Tawania[4]

(B201033EC)
(B200944EC)
(B200998EC)
(B200105EC)
BACHELOR OF TECHNOLOGY IN ELECTRONICS AND COMMUNICATION ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY, CALICUT NIT CAMPUS P.O., CALICUT
KERALA, INDIA 673601.

Introduction :

Keyboard and mouse, although are popular medium but not very convenient as it requires a certain amount of skill for effective usage. Current computer interfaces also assume a certain level of literacy from the user. It also expects the user to have certain level of proficiency in English. Speech interface can help us tackle these problems. In this paper, we discuss about the survey done in Hindi language for building large vocabulary speech recognition systems.

Speech Keyword Spotting (KWS)

Keyword search (KWS), also known as spoken term detection (STD), allows for searching through extensive spoken collections like lectures, meeting recordings, call centre conversations, or web videos. A keyword spotting system suggests potential matches for search terms specified by the user in an audio collection.
Many KWS systems use automatic speech recognition (ASR) systems to transcribe input speech into word sequences before searching. ASR-free KWS has been studied [9, 10, 11]. KWS helps diagnose ASR performance beyond word error rate.
To enable good KWS systems, accurate estimation of timing and confidence scores for words is required. A high recall of words in the ASR lattice outputs is also im- portant. However, unlike hybrid ASR, which is decoded using methods based on the weighted finite state transducer (WFST) framework and provides a rich lattice where timing information and confidence scores can be easily obtained, ASR results do not naturally come with timing or confidence per word. This work will address and evaluate the impact of these challenges on KWS performance.

Literature Survey :

Overview of KWS Technology

The advancements in speech processing via Keyword Spotting (KWS) technology plays a critical role by enabling the recognition of particular keywords or phrases in audio streams. A thorough introduction of KWS is given by Sahbi and Duraiswami (2006), who also address its challenges, applications, and changing trends. Applica- tions for KWS can be found in security monitoring, content indexing, voice-activated gadgets, and transcription systems.

Language-Specific Challenges in KWS

In contrast to languages like English, there is comparatively less annotated speech data available for KWS tasks in Hindi, despite the language being spoken widely. Large-scale speech datasets must be gathered, annotated, and curated in order to facilitate the development and assessment of KWS systems for Hindi, as there are insufficient labelled speech corpora available. Other challenges include cross-lingual interference, dialectal variations, phonetic complexity.

Speech Processing for Hindi

There have been significant developments in speaker adaptation, end-to-end tech- niques, acoustic and language modelling, and Hindi speech processing with Kaldi. Customised language and acoustic models for Hindi are made possible by Kaldi's tools, and speaker adaptation methods maximise recognition for specific speakers.
These developments enhance recognition accuracy and efficiency for Hindi speech applications, underscoring Kaldi's pivotal role in advancing Hindi speech processing capabilities.

Datasets and Resources

To further KWS research in Hindi, scholars and practitioners have made sure that reliable evaluation frameworks and high-quality annotated datasets are available. These resources are invaluable for KWS system validation, training, and benchmark- ing, which helps to create dependable and efficient speech processing solutions for Hindi language applications.

Feature Extraction and Model Architectures

K. Kumar and V. B. Surya's, "MFCC Based Features for Speaker Recognition in Hindi," (2018) paper investigate the effectiveness of Mel-Frequency Cepstral Coef- ficients (MFCCs) for Hindi speaker recognition. In particular, their work explores feature extraction methods tailored to the phonetic peculiarities of Hindi speech with the goal of finding the best feature representations for Keyword Spotting (KWS) tasks.

Evaluation Metrics and Benchmarks

N. Jaitly et al. propose evaluation metrics and benchmarks for large-scale acous- tic indexing tasks in their 2014 IEEE Transactions on Audio, Speech, and Language Processing paper "Large-Scale Acoustic Indexing for Audio Streams". Their research illuminates Keyword Spotting (KWS) performance evaluation methods.

Jaitly et al. simplify KWS system performance evaluation and comparison across research studies and implementations by using standardised evaluation metrics and benchmark datasets. Their contributions improve KWS research reliability and re- producibility and provide a framework for assessing KWS system efficacy and effi- ciency in real-world applications. Jaitly et al.'s paper is essential for understanding KWS evaluation methodologies and bench-marking practices in a project report lit- erature survey.

Future Directions and Challenges

Multilingualism, domain adaptation, and real-time processing are among the main challenges could be faced in future KWS works. Their study illuminates the changing landscape of KWS research and emphasises the need for advanced methods and tech- nologies to address these challenges. Pradhan and Rao's work guides KWS research
by identifying future directions and potential obstacles, providing crucial consider- ations for researchers and practitioners. Their paper provides valuable insights on KWS research's progress and challenges in a project report literature survey.

Problem Definition

The objective is to accurately detect predefined keywords or phrases within spoken Hindi audio streams. The system should be capable of robustly identifying target keywords amidst various acoustic environments, speaker variations, and linguistic nuances inherent in Hindi speech.

The KWS system must leverage Kaldi's tools and techniques for acoustic modeling, language modeling, and speaker adaptation, tailored specifically for Hindi phonetics and language structure.

The ultimate goal is to enable efficient and reliable keyword spotting in Hindi, fa- cilitating applications in speech transcription, content indexing, and voice-controlled devices.

Major Contributions

The following contributions collectively advance the state-of-the-art in speech key- word spotting for Hindi, enabling the development of efficient and reliable KWS systems tailored for Hindi language applications.

- **Development of a Customized KWS System**

The main contribution is the creation of a customised Keyword Spotting (KWS) system that uses Kaldi and is specifically made for the Hindi language. With the help of linguistic and acoustic models tailored for Hindi phonetics and language structure, this system makes it possible to precisely identify keywords in Hindi speech data.

- **Creation of Annotated Datasets**

The development of annotated speech datasets for Hindi KWS tasks is another noteworthy contribution. These datasets make it easier to train, validate, and assess the KWS system because they contain labelled audio recordings with predefined keywords.

- **Optimization of Feature Extraction**

In order to improve the representation of Hindi speech signals, research ef- forts are concentrated on optimising feature extraction techniques, such as Mel-Frequency Cepstral Coefficients (MFCCs). The KWS system performs better at keyword detection as a result of this optimisation.

- **Evaluation Framework Development**

Contributions include the creation of a thorough framework for evaluating the effectiveness of the Hindi KWS system. Scoring and calculating Word Error Rate (WER) are a part of this framework, which enables impartial and thorough system performance evaluation.

- **Robustness and Scalability Enhancements:**

Work is focused on making the KWS system more resilient and scalable to accommodate a range of acoustic conditions, speaker variations, and practical uses. This covers methods for scalability to large-scale datasets, speaker adap- tation, and domain adaptation.

Work Done :

A comprehensive system for Keyword Spotting (KWS) in the Hindi language, which is combined with Automatic Speech Recognition (ASR) using the Kaldi toolkit. The system being suggested utilizes the advanced automatic speech recognition (ASR) capabilities of Kaldi to improve the process of identifying and retrieving keywords in spoken Hindi text. The essential elements encompassed in this context are audio modelling, language modelling, and efficient decoding algorithms, which collectively contribute to the achievement of precise and timely keyword detection.The integra- tion of Kaldi's Automatic Speech Recognition (ASR) and Keyword Spotting (KWS) modules enhances the development of spoken language processing technologies for the Hindi language. This integration paves the way for enhanced human-computer interaction and improved information retrieval in multilingual scenarios.

ASR(Automatic speech recognition) System

Data Collection

The first step here was to compile 150 grammatically rich sentences in Hindi having maximum possible unique words and trying to capture distinct situations .

Obviously the more the number of unique sentences (and words), the better would be the performance of the recognition system.

Below are the steps for KALDI format data.

- **wav.scp** file in your train folder and save it.
- **text** file and save it.
- **utt2spk**: create file on <filename> <speakerID> pattern and save it
- **spk2utt**: Sentences spoken by each speaker.
- **spk**: create file list on <lang name speakerid> pattern.
- **utt**: create file lists on <unique utterance id> pattern and save it. Repeat same steps for your **test** folder.

Next we store the lexicon.txt that we had prepared earlier in the folder data/local/dict

In the same folder , we also create the following files :

- **Non silence phones.txt** : this contains all the phones present in the lexicon.
- **Optional silence.txt**: for our scenario , this only contains the silence sil.
- **Silence phones.txt** : contains the sounds that do not contain acoustic information, but are present like noises(also called fillers)

Language Model Preparation

Here we are working with N-gram language model, copy below script to your folder and replace the path , we have taken the n gram=2 which mean that we are building bi-gram language model.

- **Feature Extraction**: In this step we extract MFCC features of each utterance (audio). Open your terminal and run below command.
- **make mfcc.sh** used for computing MFCC coefficients.
- **nj**- number of jobs, you can set it to according to your cpu.
- **data/train**: folder path for which you want to compute MFCC.
- **exp/make mfcc/train**: log file stored in this directory.
- **mfcc**: directory name where we store extracted feature.

Acoustic Model Preparation a)Monophone-HMM (WER-3.44 %)

- Train-steps/train mono.sh - - nj 10 data/train data/lang bigram exp/mono
- Combine-utils/mkgraph.sh - - mono data/lang bigram exp/mono exp/mono/graph
- Decode-steps/decode.sh - -nj 5 exp/mono/graph data/test exp/mono/decode

Triphone-HMM (WER-2.80 %)

- **Train**-steps/train deltas.sh 2000 16000 delta/train data/lang bigram exp/mono ali exp tri
- **Combine**-utils/mkgrapgh.sh data/lang bigram exp/tri exp/tri/graph
- **Decode**-steps/decode.sh - - nj 5 exp/tri/graph data/test exp/tri/decode

DNN-HMM (WER-2.30 %)

- **Train**-steps/nnet2/train tanh.sh - -initial-learning- rate 0.015 - -final-learning-rate 0.002 - -num-hidden-layers 1 - -minibatch-size 128 - -hidden-layer-dim 256 - -num-jobs-nnet 10 - -num-epochs 5 data/train data/lang bigram exp/tri ali exp/DNN
- **Decode**-steps/nnet2/decode.sh –nj 4 exp/tri/graph data/test exp/DNN/decode.

KWS(Keyword Spotting) System

Keyword Spotting is performed by scripts/do all kws search.sh which sets up the keyword search task, creating and specifying the necessary data resources required for the keyword search. This script also has the optional function to score the search results.

Running KWS :

- **Usage**: do all kws search.sh [options] <in:wav list> <in:kwd list><in:lang> <out:dir work> <out:kwd results>

Parameters :

- wav list : list of full file paths to audio files to be searched.
- kwd list : list of keyword phrases to search. Each keyword phrase should be on a new line. Keyword phrases can consist of single or multiple words.
- lang : [hindi] - specifies audio language and determines which decoding model is used.
- dir work : directory in which required resources are created.
- kwd results : file to which results will be written.

Final Output

Scoring KWS

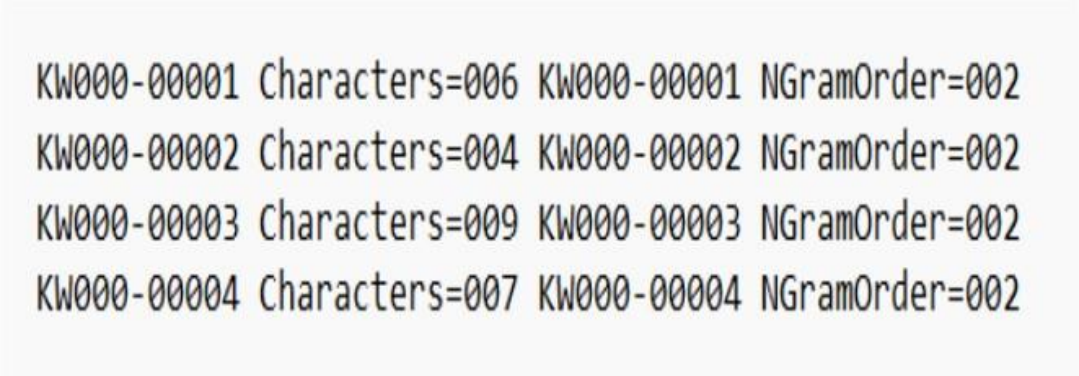
Scoring of the keyword search results is optional, and disabled by default. If enabled via the optional argument –skip-scoring false, the keyword search results are evaluated by the F4DE scorer from NIST.

Parameters

One of the following optional arguments must be provided:

- **–aligned-ctm aligned ctm** : Accurately aligned ctm of reference files, corresponding to audio files. [Suggested option for optimal results].
- **–txt-list txt list** : List of file paths to reference transcriptions, corresponding to audio files. Transcript text must be on a single line. A ctm with approximate time alignments will be created from the transcription texts.

```
<kwslist kwlist_filename="" language="hindi" system_id="">
     <detected_kwlist kwid="KW000-00001" kwd_text="saapha" search_time="1" oov_count="0">
        <kw file="hindi_parliament_clip" channel="1" tbeg="0.45" dur="0.51" score="1" decision="YES"/>
     </detected_kwlist>
<kwslist kwlist_filename="" language="hindi" system_id="">
     <detected_kwlist kwid="KW000-00002" kwd_text="jaba" search_time="1" oov_count="0">
        <kw file="hindi_parliament_clip" channel="1" tbeg="2.41" dur="2.51" score="1" decision="YES"/>
     </detected_kwlist>
<kwslist kwlist_filename="" language="hindi" system_id="">
     <detected_kwlist kwid="KW000-00003" kwd_text="manxsuxya" search_time="1" oov_count="1">
        <kw file="hindi_parliament_clip" channel="1" tbeg="0.35" dur="0.39" score="0.85" decision="YES"/>
     </detected_kwlist>
<kwslist kwlist_filename="" language="hindi" system_id="">
     <detected_kwlist kwid="KW000-00004" kwd_text="pustaka" search_time="1" oov_count="1">
        <kw file="hindi_parliament_clip" channel="1" tbeg="3.13" dur="3.24" score="0.30" decision="YES"/>
     </detected_kwlist>
</kwslist>
```



Conclusion :

- The main contribution is the creation of a customised Keyword Spotting (KWS) system that uses Kaldi and is specifically made for the Hindi language. With the help of linguistic and acoustic models tailored for Hindi phonetics and language structure, this system makes it possible to precisely identify keywords in Hindi speech data.
- The development of annotated speech datasets for Hindi KWS tasks is another noteworthy contribution. These datasets make it easier to train, validate, and assess the KWS system because they contain labelled audio recordings with predefined keywords.
- In order to improve the representation of Hindi speech signals, research efforts are concentrated on optimising feature extraction techniques, such as
- Mel-Frequency Cepstral Coefficients (MFCCs). The KWS system performs better at keyword detection as a result of this optimisation.
- Contributions include the creation of a thorough framework for evaluating the effectiveness of the Hindi KWS system. Scoring and calculating Word Error Rate (WER) are a part of this framework, which enables impartial and thorough system performance evaluation.

BIBLIOGRAPHY :

1. F. Tariq, M. Khandaker, K.-K. Wong, M. Imran, M. Bennis, and M. Debbah, “A speculative study on 6G,” *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 118-125, Aug. 2020,
2. Xinran Fang, Wei Feng, Te Wei, Yunfei Chen, and Cheng Xiang Wang, “5G Embraces Satellites for 6G Ubiquitous IoT: Basic Models for Integrated Satellite Terrestrial Networks,” *IEEE Internet of Things J.*, vol. 8, no. 18, September 15, 2021. 387, 1975.
3. M. J. F. Gales, K. M. Knill, and A. Ragni, “Unicode-based graphemic sys- tems for limited resource languages,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2015, pp. 5186–5190. (https://kaldi-asr.org/doc/kws.html)
4. J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddington, “Results of the 2006 spoken term detection evaluation,” in Proceedings of the ACM SIGIR Conference, vol. 7, 2007, pp. 51–57.
5. S. Wegmann, A. Faria, A. Janin, K. Riedhammer, and N. Morgan, “The tao of atwv: Probing the mysteries of keyword search performance,” in 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Dec 2013, pp.