This full-text paper was peer-reviewed and accepted to be presented at the IEEE WiSPNET 2017 conference.

Continuous Hindi Speech Recognition Model Based on Kaldi ASR Toolkit

Prashant Upadhyaya,[1] Omar Farooq,[2] Musiur Raza Abidi and Yash Vardhan Varshney
Department of Electronics Enginerring, Aligarh Muslim University, Aligarh 202002
Email: [1]upadhyaya.prashant@gmail.com [2]omarfarooq70@gmail.com

***Abstract*—In this paper, continuous Hindi speech recognition model using Kaldi toolkit is presented. For recognition, MFCC and PLP features are extracted from 1000 phonetically balanced Hindi sentence from AMUAV corpus. Acoustic modeling was performed using GMM-HMM and decoding is performed on so called HCLG which is construted from Weight Finite State Transducers (WFSTs). Performance of both monophone and triphone model using N-gram language model is reported which is computed in term of word error rate (WER). A significant reduction in word error rate (WER) was observed using the triphone model. Further, it was found that MFCC feature provide higher recognition accuracy than PLP feature. Goal is to show the performance of Hindi language using present state-of-the-art (Kaldi) system.**

***Index Terms*—Kaldi ASR, Weight Finite State Transducers, MFCC, Speech recognition.**

I. Introduction

Speech is the most intuitive form of communication among people. Though, it has shown the significant improvement in speech recognition based application but still there is question for how fast and reliable is the system model. Thus, success of ASR depend upon how accurate the system model which is measure in term of speech recognition accuracy. Although, there are numbers of various toolkit available for developing speech recognition based application. Some of the popular toolkit are HTK [1], Julius [2], Sphinx-4 [3], RWTH [4] and Kaldi [5] ASR toolkit. Recently, Kaldi is one of the most popular and state-of-the-art toolkit for the researcher working in the speech recognition area.

Kaldi is an open-source toolkit for speech recognition written in C++. The advantage of speech recognition based application developed using Kaldi produces high-quality lattices and are sufficiently fast for real-time recognition [5], [6]. Kaldi toolkit is actively maintained, and is distributed under the permissive Apache 2.0 license [7]. For compiling, OpenFST (Finite State Transducer) toolkit is used. Internal structure of Kaldi toolkit is shown in Fig. 1.

Here in this paper, continuous large vocabulary Hindi speech recognition using Kaldi toolkit is presented. The reason for selecting Hindi language for building speech recognition based application is due to its more popularity. Hindi language is the fourth most spoken language followed by Mandarin, Spanish and English [8]. Application such as smart phone, laptop and many IVRS based application use Hindi speech as an interface for controlling or accessing these applications. Thus, allowing to access the technology more freely. Some work related to Hindi speech recognition is reported in [9–12].

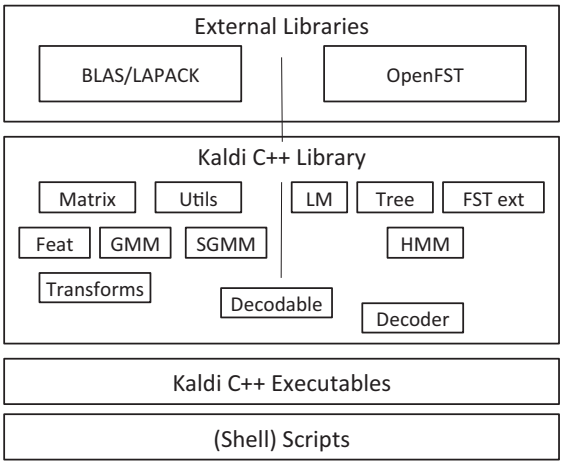

Fig. 1. Internal structure of Kaldi ASR toolkit [5].

II. Acoustic and Language Modeling

Acoustic modelling (AM) is the heart of every speech recognition model. It search for the most probable sequence of words $w^\star$ given the acoustic observations $\boldsymbol{O}$ as described in Eq. (1).

$$w^\star = \arg\max_i \{P(w_i|\boldsymbol{O})\} \tag{1}$$

where w_i is the i'th vocabulary word. Using Bayes' Rule gives

$$P(w_i|\boldsymbol{O}) = \frac{P(\boldsymbol{O}|w_i)P(w_i)}{P(\boldsymbol{O})}. \tag{2}$$

Therefore, for a given set of prior probabilities $P(w_i)$, the most probable spoken word depends only on the likelihood $P(\boldsymbol{O}|w_i)$ so Eq. (2) can be reduce to

$$P(w_i|\boldsymbol{O}) = \arg\max_i \{P(\boldsymbol{O}|w_i)P(w_i)\}. \tag{3}$$

The task of acoustic modelling is to estimate the parameters θ of a model so that the probability $P(\boldsymbol{O}|w_i)$ is as accurate as possible. Similarly, the LM represents the probability $P(w_i)$ [1]. Fig. 2 represent the complete structure of statistical speech recognition system using Kaldi ASR.

Kaldi use an FST-based framework, therefore any language model can be used which support FST. One can easily implement N-gram model using the IRSTLM or SRILM toolkit which are include in their recipe [5].

978-1-5090-4442-9/17/$31.00

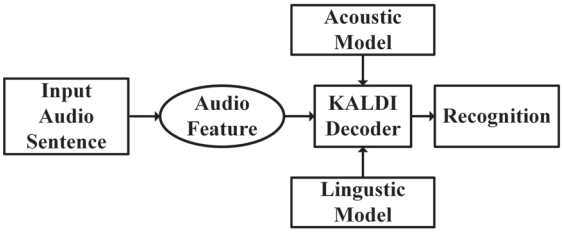

Fig. 2. Automatic speech recognition model using Kaldi toolkit.

Training and decoding algorithm in Kaldi use Weight Finite State Transducers (WFSTs). The weight FST provide the well studied graph operation which can effectively used for acoustic modelling. It use the "pdf-ids" (assign numeric value to decoding graph correspond to context-dependent states). Since the different phone share the same pdf-ids, therefore it "transistion-id" is used which encode the pdf-ids of phone member and use arc(transition) within the topology specify for that phone [5], [6]. Thus, decoding is performed on so called decoding graph HCLG which is constructed from simple FST graphs as given in Eq. (4) [5–7].

$$HCLG = H \circ C \circ L \circ G. \tag{4}$$

The symbol $\circ$ represents an associative binary operation of composition on FST. Here, G is an acceptor that encodes the grammar or language model, L represents the lexicon (its input symbols are phones and output symbols are words), C represents the relationship between context-dependent phones on input and phones on output and H contains the HMM definitions, that take as input id number of PDF and return context-dependent phones.

Next section deal with the data preparation that are mandatory for Kaldi ASR. Thus, maintain the meta-data of each speakers which are used for training and testing the acoustic and language models.

III. Data Preparation for Kaldi ASR

This section deal with the step by step procedure for creating simple ASR using your own set of data using Kaldi toolkit. For our experiment AMUAV database is chosen which consists of 100 speakers. Each speakers utterance the 10 short Hindi sentence from which two sentence are common to each speaker. Thus, 1000 continuous Hindi speech database is prepared which is phonetically balanced. To train the model 900 sentences are chosen and rest are used for testing.

Finally, acoustic meta-data of each speakers is to be created which are used for training and testing the acoustic models. Data preparation is divided into acoustic data and language data. Meta-data use for acoustic data is given below which are mandatory for Kaldi ASR:

a.) $spk2gender \Rightarrow \langle$ speaker ID $\rangle\langle$ gender $\rangle$
This file informs about speakers gender. Speaker ID is a unique name of each speaker (sometime also referred as recording ID).

b.) $wav.scp \Rightarrow \langle$ uterrance ID $\rangle\langle$ path of the recorded.wav $\rangle$
It provide the path of recorded audio files sentence along with speakers ID.

c.) $text \Rightarrow \langle$ uterrance ID $\rangle\langle$ transcription $\rangle$
This file contains every utterance matched with its text transcription.

d.) $utt2spk \Rightarrow \langle$ uterrance ID $\rangle\langle$ speaker ID $\rangle$
This has the mapping of the utterance of particular speaker.

e.) $corpus.txt \Rightarrow \langle$ transcription $\rangle$
It contain all the utterance transcription that are use for building the model.

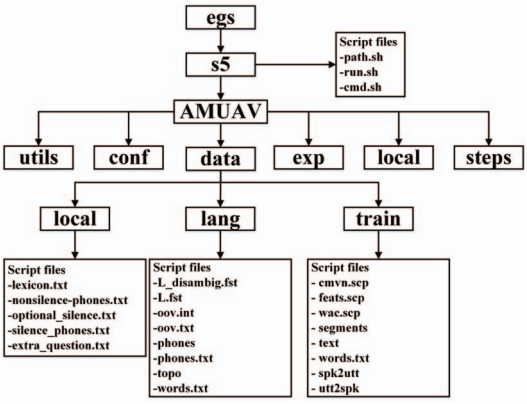

Fig. 3. Kaldi directories structure for AMUAV corpus.

Meta-data for preparing language data is given below which are mandatory for Kaldi ASR:

a.) $lexicon.txt \Rightarrow \langle$ word $\rangle\langle$ phone 1 $\rangle\langle$ phone 2 $\rangle$.
This contain the phone transcriptions of every word.

b.) $nonsilence\_phones.txt \Rightarrow \langle$ phone $\rangle$.
This contain all the phones that are used for preparing the database.

c.) $silence\_phones.txt \Rightarrow \langle$ phone $\rangle$.
This contain the silence and short pause phone.

Complete Kaldi directories structure for AMUAV data preparation is created in Kaldi-trunk (main Kaldi directory) as shown in Fig. 3.

IV. Feature Extraction

Most important feature extraction technique for speech recognition are Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP). Here in this paper we have extracted both MFCC and PLP features. Thus, MFCC and PLP transformations are applied on a sampled and quantized audio signal. Here only MFCC feature extraction process is described. The feature is extracted by applying 25 ms window shifted by 10 ms. The audio signal was sampled at 16 kHz. Therefore, 16000*0.025 = 400 samples in one window are reduce to 13 static cepstral coefficients. To include the temporal evolution of MFCC, additional feature Δ and $\Delta - \Delta$ values is computed. Finally, these feature vectors are concatenated to form as a single modality, i.e., $d_a = 39$. Complete steps involved in MFCC features extraction process is shown in

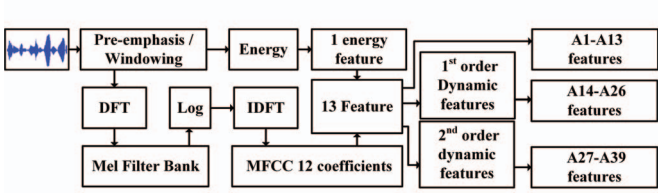

Fig. 4. MFCC feature extraction process.

TABLE I
VOCABULARY SIZE OF AMUAV DATABASE.

Word	100 speaker	Training speaker	Testing speaker
Vocabulary size	2007	1917	419
Unique vocabulary	2007	1678	329

TABLE II
WER PERFORMANCE FOR MFCC AND PLP FEATURE USING MONO (MONOPHONE TRAINING) MODEL.

Feature	2-gram	3-gram	4-gram
MFCC	17.82	16.09	16.34
PLP	19.68	17.95	18.44

Fig. 4. Finally, cepstral mean and variance normalization (CMVN) per speaker is computed on the extracted features.

V. EXPERIMENTAL APPROACH

The machine configuration on which experiment was conducted has Ubuntu 16.04 LTS (64-bit operating system), Processor Intel Core 2 Duo with 2.20 GHz.

Experimental results on AMUAV corpus is reported which consists of 1000 phonetically balance Hindi sentences which is spoken by 100 speakers. The vocabulary size of the AMUAV database is 2007 (unique word). Total number of word in AMUAV dataset are 10664. Number of words present in training and testing are shown in Table I.

Context-dependent triphone system with simple GMM-HMM model was developed. The features are MFCCs and PLP with per-speaker cepstral mean subtraction. Since Kaldi use FST-based framework, so SRILM toolkit was used to build LM model from the raw text. For experimental purpose N-gram model (i.e., $N = 2,3$ and 4) was used for recognition.

Performance is measured in term of word error rate (WER) defined as

$$WER(\%) = \frac{(D + S + I)}{N} * 100(\%) \quad (5)$$

where N is the number of word used in test, D is number of deletions, S is number of substitutions and I is the number of insertion error.

Table II shows the performance of MFCC and PLP feature using monophone training model using 2-gram, 3-gram and 4-gram LM model. As seen form the Table II MFCC feature gives improvement over PLP feature. Best recognition rate was achieved at 3-gram LM model. Increasing of LM model to 4-gram degrade the performance of speech recognition model.

TABLE III
WER PERFORMANCE FOR MFCC AND PLP FEATURE USING TRI1 (TRIPHONE TRAINING).

Feature	2-gram	3-gram	4-gram
MFCC	14.36	15.97	15.97
PLP	16.21	16.34	16.21

Table III shows the performance of MFCC and PLP feature using triphone training model using 2-gram, 3-gram and 4-gram LM model. As seen form the Table III MFCC feature gives improvement over PLP feature. Best recognition rate was achieved at 2-gram LM model. Increasing of LM model from 2-gram to 4-gram degrade the performance of speech recognition using triphone model.

Best recognition is obtained for triphone modelling, due to its context dependency. On the otherhand, WER obtained using monophone model is high due to its insufficient variation of phones with respect to left context and right context [1].

VI. CONCLUSION

In this paper, continuous Hindi speech recognition using AMUAV corpus is reported using Kaldi toolkit. Two feature were selected for recognition and it was shown that MFCC feature outperformed the PLP feature. Also, the triphone model give the best accuracy. Further, speech recognition by varying the LM model from 2-gram to 4-gram is also reported. It was clearly shown from the results, that increasing the LM model will increase the complexity, thus it can degrade the performance of speech recognition model. To our knowledge this is the first work on Kaldi toolkit for continuous Hindi speech. In future, our task is to found the new robust feature which can further increase the robustness of speech recognition model. Also, use of deep neural network can also increase the performance of ASR system.

REFERENCES

[1] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book (for version 3.4)*, Cambridge University Engineering Department, 2009.

[2] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," in *EUROSPEECH*, 2001, pp. 1691–1694.

[3] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, *Sphinx-4: A flexible open source framework for speech recognition*, Sun Microsystems Inc., Technical Report SML1 TR2004-0811, 2004.

[4] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Loof, R. Schluter, and H. Ney, "The RWTH Aachen University open source speech recognition system," in *INTERSPEECH*, 2009, pp. 2111–2114.

[5] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, and J. Silovsky, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (No. EPFL-CONF-192584)*, IEEE Signal Processing Society, 2011.

[6] D. Povey, M. Hannemann, G. Boulianne, L. Burget, A. Ghoshal, M. Janda, M. Karafit, S. Kombrink, P. Motlek, Y. Qian, and K. Riedhammer, "Generating exact lattices in the WFST framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4213–4216.

[7] Kaldi Home Page (kaldi-asr.org).

[8] http://www.internationalphoneticalphabet.org.
[9] V. Chourasia, K. Samudravijaya, M. Ingle, and M. Chandwani, "Hindi speech recognition under noisy conditions," *International Journal of Acoustic Society India*, pp. 41–46, 2007.
[10] O. Farooq, S. Datta, and A. Vyas, "Robust isolated Hindi digit recognition using wavelet based de-noising for speech enhancement," *Journal of Acoustical Society of India*, vol. 33, no. 1–4, pp. 386–389, 2005.
[11] A. Mishra, M. Chandra, M. Biswas and S. Sharan, "Robust features for connected hindi digits recognition," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 4, no. 2, pp. 79–90, 2011.
[12] P. Upadhyaya, O. Farooq, M. R. Abidi, and P. Varshney, "Comparative study of visual feature for bimodal hindi speech recognition," *In Archives of Acoustics*, vol. 40, no. 4, pp. 609–619, 2015.