Int. j. inf. tecnol. (August 2023) 15(6):3079–3092
https://doi.org/10.1007/s41870-023-01339-y

ORIGINAL RESEARCH

Isolated words recognition of Adi, a low-resource indigenous language of Arunachal Pradesh

Sajal Sasmal[1] · Yang Saring[1]

Received: 28 December 2022 / Accepted: 2 June 2023 / Published online: 22 June 2023

Abstract This paper describes a method of the automatic recognition system for isolated words of Adi Language using the 'Kaldi' toolkit. Adi is an extremely low-resource endangered tribal language of Arunachal Pradesh, India. In this research, a speech corpus of 21 native Adi speakers of Arunachal Pradesh has been employed. The corpus consists of 2088 unique Adi words with 14,490 utterances. Mel frequency cepstral coefficients (MFCC) features were extracted from recorded Adi speech samples. In this recognition system, various speech models like Monophone, Triphone (tri1, tri2, tri3), and Sub Space Gaussian Mixture Model (SGMM) were considered. The system's performance was measured using word error rate (WER) and word recognition accuracy (WRA). In the monophone model, the WRA was found to be 72.63%. In triphone models, the recognition efficacy of tri1, tri2, and tri3 was improved to 84.76, 87.54 and 91.06%, respectively. The SGMM model gave the least WER of 7.95%. The authors also investigated overall phone alignment and occurrences. WER also calculated separately for individual speakers with different models. This proposed model may be helpful in the real world to build different speech recognition applications of man-machine interfaces like interactive voice response systems (IVRS), voice security systems, banking by telephone, voice dialing, database access services, and home appliances control using the Adi language.

Keywords Low-resource indigenous language · Adi · Arunachal Pradesh · Word recognition · WER · Kaldi

1 Introduction

Speech is the most popular and easiest way of human interaction. Sometimes it is too tricky due to different languages and accents. In this era, modern technology should be language-independent to enjoy its benefits. As per a study on open voice data in Indian languages in 2020, only 10–12% of Indians are comfortable in English, i.e., approximately 1253 million Indians are comfortable only in their mother tongue [1]. In India, there are almost 1369 classified languages; more than 120 languages have more than 10,000 speakers [2]. The top automatic speech recognition (ASR) [3, 4] based companies like Amazon Alexa, Apple Siri, Google Assistance, Microsoft- Cortana digital assistant, etc. [5, 6] are interested in working only on those languages which are commercially beneficial. Google-Home works on merely 13 languages throughout the world, only Hindi from India, whereas Microsoft has ASR systems on Hindi, Marathi, Gujarati, Telugu, and Tamil.

Presently India has approximately 1.18 billion mobile connections. Among them, 700 million are regular Internet users. Smartphone users' growth rate is 25 million per quarter [1]. So, we need to implement ASR in all native languages to access the internet content efficiently and save the low-resource languages in this globalization. Some research has already been initiated to develop speech recognition technologies in some resourceful Indian languages. Kumar et al. [7] worked on Hindi ASR system in noisy environment using hybrid feature extraction technique like perceptual linear predictive (PLP) and MFCC. Babhulgaonkar et al. [8] implemented generalized back-off

✉ Sajal Sasmal
sajal.sasmal@gmail.com

Yang Saring
ys.nitap@gmail.com

[1] Department of Electronics and Communication Engineering, National Institute of Technology Arunachal Pradesh, Jote, India

and factored language model by combining of linguistic features to predict upcoming words for an ASR of Hindi. Guchhait et al. [9] build a Kaldi based ASR system of Bengali language. In [10], the authors designed a speech recognizer of Marathi using the HTK toolkit of 910 sentences from 6 speakers. An ASR system was developed to transcribe Telugu TV news automatically by Reddy et al. [11]. Lokesh et al. [12] implemented a Tamil ASR based on a bidirectional recurrent neural network. An HMM-based speech and isolated word recognition system was developed by Ananthi et al. [13]. The research was carried out with the data collected from 20 different speakers, and the system obtained 87.42% recognition accuracy. An isolated spoken digit recognition of the Assamese language using HMM was implemented by Bharali et al. [14]. This digit recognition model used the speech corpus of 10 native Assamese speakers. In [15], an isolated spoken Marathi words recognition system using HMM has been executed. A small speech database of 20 phonetically rich Marathi words of ten native speakers was used in this research. The recognition system got a maximum accuracy of 86.50%. A convolutional neural network (CNN) based automatic speech recognition system for isolated words was designed by Slívová et al. [16]. A word-level recognition system for the Hindi language using the Kaldi toolkit was created by Sri et al. [17]. The various acoustic models, Monophone, Triphone, and SGMM were designed in this work. A GMM and MFCC-based isolated spoken numerals recognition system for the Bengali language was presented by Paul et al. [18] with a corpus of 1000 audio samples. The recognition system achieved 91.7% prediction efficiency. An isolated Arabic words recognition model has been proposed based on the hidden Markov model (HMM) in [19].

With the help of ASR technology, it is easy to recognize and translate the spoken language into text form by machine. According to the Atlas of the World's Languages in Danger report UNESCO 2017, there are 33 endangered languages of Arunachal Pradesh, including Adi. The year 2019 is declared the Year of Indigenous Languages by United Nations to illustrate awareness of languages worldwide that are in jeopardy of disappearing.

As per Census 2011, a total of 2,48,834 speakers of Adi are found in Arunachal Pradesh, mainly spread over the west, east, and upper Siang districts [2] of Arunachal Pradesh. Adi language originated from the Tibeto-Burman family, typically associated with the Sino-Tibetan family [20].

The main challenges to working with Adi language are

1. No speech data are available on the internet or any other digital media, as Adi is a very Low-Resource Indigenous Language of Arunachal Pradesh.
2. Adi has adopted modified Roman script for writing in the Adi language, which is still being developed; therefore, it is a challenge to represent Adi words in proper phonetic transcripts.
3. Most Adi native speakers cannot read modified Roman scripts of Adi words, making data collection more complex.
4. The Adi tribes are spread in the different mountainous areas of Arunachal Pradesh, making the work more difficult.

Lalrempuii worked on the Morphology of Adi [21]. In [22], spectral and formant studies of Adi consonants have been investigated. A speech recognition system of Adi language proposed in [23]. In this work, the authors endeavored to develop an automatic isolated words recognition system for the Adi language using Hidden Markov Model-Gaussian Mixture Model (GMM-HMM) [24, 25] and SGMM [26].

The significant contributions of this research are:

1. The Corpus consists of 2088 unique isolated words of Adi Language that have never been endeavored before.
2. For the first time, this research shows the phonetic transcriptions of 2088 Adi words with proper phoneme sequences.
3. This proposed Adi words recognition model may become an opening step to building an ASR system in Adi.
4. This work demonstrates excellent recognition accuracy on monophone, three different triphones [27], and SGMM models with the help of the Kaldi toolkit.
5. The authors investigated overall phone alignment and occurrences and calculated WER separately for individual speakers with different models.

The paper is organized as follows: Section 2 describes the model building using Kaldi; the system configuration is discussed in Sect. 3; feature extraction is illustrated in Sect. 4. Section 5 explains language model creation; training and decoding are described in Sect. 6. After that, experimental results and discussion are given in Sect. 7. Section 8 concludes the current research work. Finally, Sect. 9 emphasizes future work of the present research.

Table 1 List of Adi consonants

	Alveolar	Bilabial	Velar	Palatal	Glottal
Stops (uv)	t	p	k		
(v)	d	b	g		
Fricatives(uv)	s				h
Nasals (v)	n	m	ŋ (ng)	ñ (ny)	
Roll (v)	r				
Approximant(v)				y	
Affricates (v)	j				
Lateral Approximant(v)	l				

v voiced, *uv* voiceless/unvoiced

Table 2 List of Adi vowels

		Front (Unrounded)	Central (Unrounded)	Back (Rounded)
Open	(short)		/a/ [a]	
	(Long)		/aa/ [a:]	
Open-mid	(short)			/o/ [ɔ]
	(Long)			/oo/ [ɔː]
Close	(short)	/i/ [i]	/í/ [ɨ]	/u/ [u]
	(Long)	/ii/ [i:]	/íí/ [ɨː]	/uu/ [u:]
Close-mid	(short)	/e/ [e]		
	(Long)	/ee/ [e:]		
Mid	(short)		/é/ [ə]	
	(Long)		/éé/ [əː]	

2 Model building using Kaldi Toolkit

In this research, 16 consonants, 14 vowels (7 short and 7 long), 19 diphthongs, and a triphthong of the Adi language are enlisted [21]. The Adi consonants are shown in Table 1.

In Adi, the duration of the vowels is extremely noteworthy as the meaning of the same word is determined by the length of vowel phonemes, whether short or long. Generally, a short vowel may be replaced by an equivalent long vowel to give the different meanings of a word. The five phonetic properties present in the Adi are alveolar, bilabial, glottal, palatal, and velar. Also, this under-resourced language's six different manners of articulation are affricates, fricatives, glides, liquids, nasals, and stops. Table 2 shows the list of Adi vowels. Some Diphthongs of this language are /aé/ [aə], /ai/ [ai], /ía/ [ɨa], /ao/ [aɔ], /oa/ [ɔa], and sole Triphthong is 'uai.' Dental, labio-dental fricative, aspirated and retroflex sounds are absent in Adi. The fricatives such as [h] and [s] can be interchanged in Adi words without any disparity in the meaning.

Kaldi is a highly efficient open-source speech recognition toolkit built by Johns Hopkins University [28]. In Fig. 1, the structural design of the automatic word recognition model has been illustrated. Different speech samples of Adi words have been recorded from 21 native Adi speakers of Arunachal Pradesh. Then MFCC features are extracted from the recorded speech samples. The extracted MFCC feature vectors are fed into the Decoder, which decodes and classifies the features using the Acoustic model, Lexicon (phonetic transcript of the Adi), and language model.

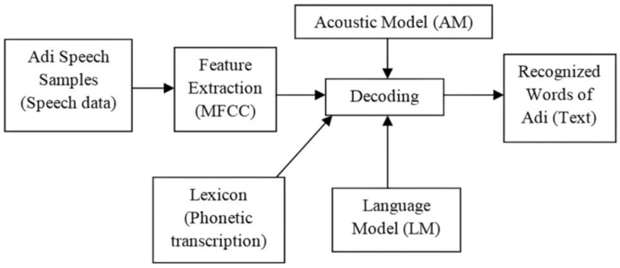
Acoustic Model (AM)
Adi Speech Samples (Speech data)
Feature Extraction (MFCC)
Decoding
Recognized Words of Adi (Text)
Lexicon (Phonetic transcription)
Language Model (LM)

Fig. 1 The architecture of automatic words recognizing model of Adi

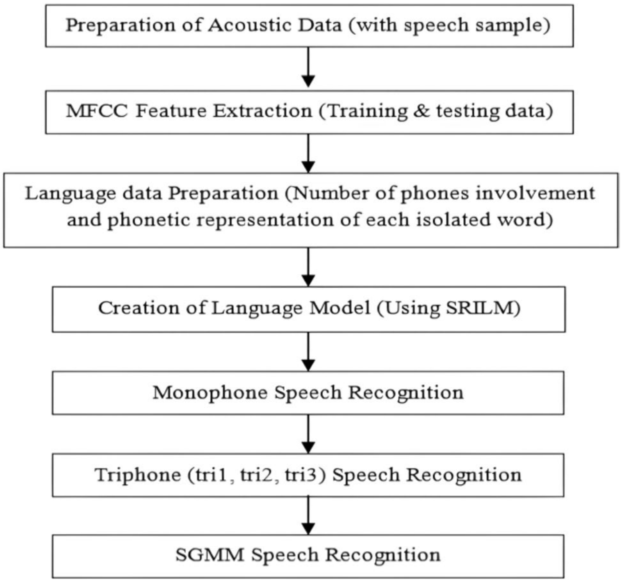
Preparation of Acoustic Data (with speech sample)
MFCC Feature Extraction (Training & testing data)
Language data Preparation (Number of phones involvement and phonetic representation of each isolated word)
Creation of Language Model (Using SRILM)
Monophone Speech Recognition
Triphone (tri1, tri2, tri3) Speech Recognition
SGMM Speech Recognition

Fig. 2 Step by step process of words recognizing model

Figure 2 describes the step-by-step process of the automatic word recognizing system. Acoustic samples are recorded and stored in waveform audio (.wav) file format. In this work, the sampling rate of audio data was fixed at 16 kHz with the bit rate of 129k signed PCM encoding and a 16-bit mono channel. The language data was prepared with 50 phonemes and every isolated 2088 Adi word's phonetic representation. SRILM toolkit has been used to construct a language model. Finally, the ASR system efficiently recognizes isolated words of the Adi language using monophone, triphones, and SGMM speech models.

Figure 3 shows the directory structure of our proposed model in the Kaldi toolkit, where different directories and indispensable files were created. Kaldi toolkit's 'egs' is the

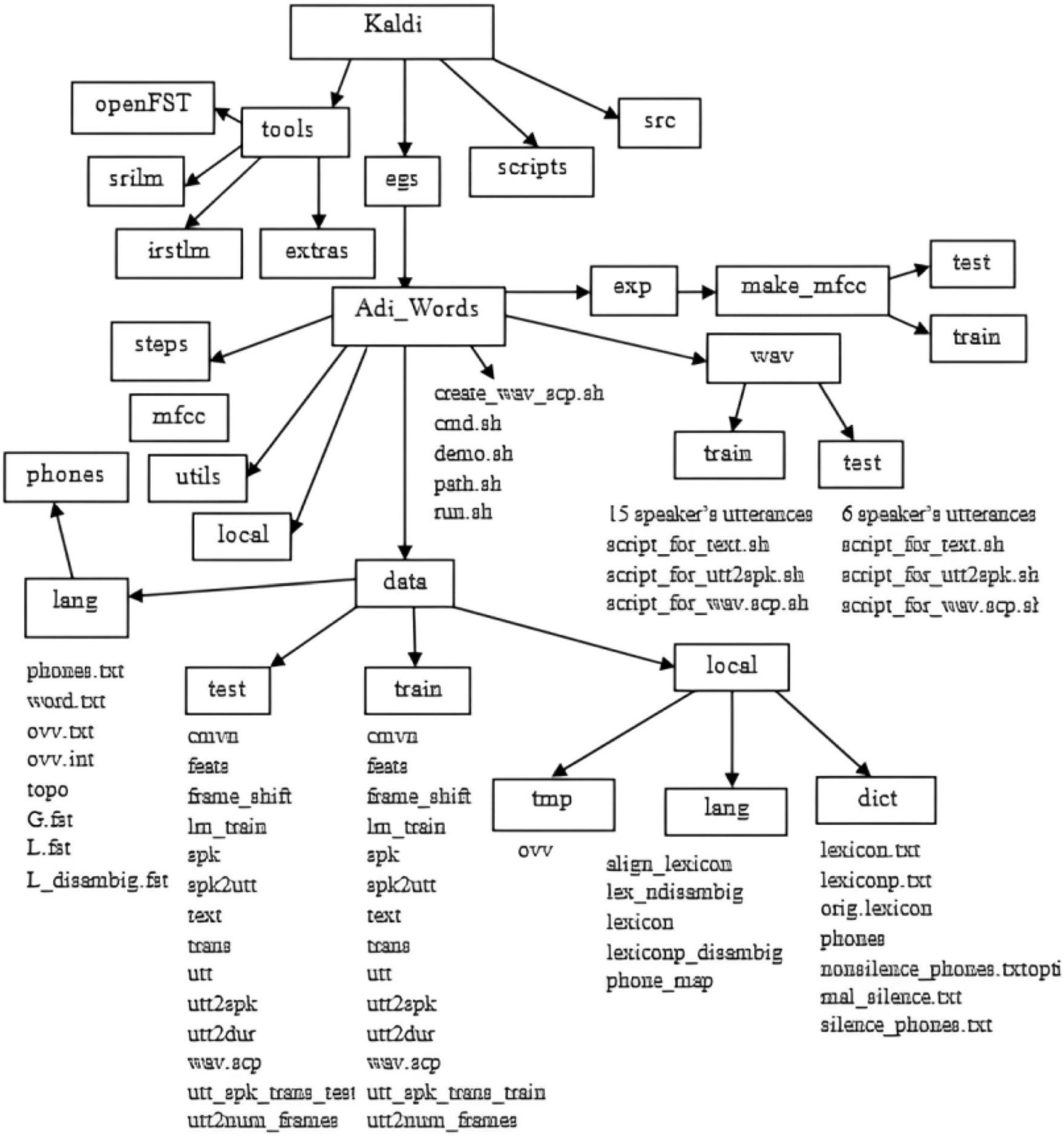

Fig. 3 The directory topology of the recommended model in Kaldi

most important directory under which 'Adi_Words' and a few sub-directories are designed to place all the stuff associated with the recommended model.

2.1 Speech dataset

The speech samples from native Adi speakers of Arunachal Pradesh were collected with the help of voice recording tools and were stored by the authors. The specifications of data for this research are as follows.

(i) Bit Rate: 256k.
(ii) Encoding: PCM.
(iii) Channels: Mono, 16-bit.
(iv) Sampling Rate: 16 kHz.
(v) Format: waveform audio file (.wav).

In this isolated word recognition system, the speech corpus consists of 21 Adi speakers (9 male and 12 female) from Arunachal Pradesh. The corpus consists of 2088 unique Adi words, and 14,490-word utterances are present in this dataset.

Table 3 List of files obligatory in acoustic data

File name	File details
wav.scp	It includes the location of every acoustic file to link all utterances with an associated acoustic file used in this model in .wav format. For example < uterranceID > < full path to audio file > AD001F0001 /home/sajal/kaldi/egs/0adi/ADI_Reco/wav/AD001F0001.wav
text	Transcription of each word utterance includes in the 'text' file. It is an utterance-by-utterance transcript of the Adi word corpus.
utt2spk	This file connects each utterance with the correlated speaker. The outline of this file is < uterranceID > < speakerID>. e.g.AD001F0651 001 F.
spk2utt	This file holds the speakers to utterance mapping.e.g.001 F AD001F0001 AD001F0002 … AD001F0533. It keeps the record of the first female speaker with her recorded utterance IDs.
corpus.txt	It includes every single utterance transcription that may take place in this recognition system. In this corpus, 21 native speakers uttered 2088 unique Adi words.

Table 4 List of files requisite in language data

File name	File details
lexicon.txt	It shows all unique Adi words and their pronunciations with the help of the Adi Phoneme Set developed for this finding. The 'phone transcription' of each Adi word is revealed here. e.g. aapui a: p eu [< word > < phone 1 > < phone 2 > < phone 3>].
nonsilence_phones.txt	This file encloses a list of each phone which is not silent and present in this model. This word recognition system employs 50 non-silence Adi phones.
silence_phones.txt	This file embraces silent phones. Only 'sil' and 'spn' were employed in this model.
optional_silence.txt	It contains a single phone which can optionally appear among words. Here merely 'sil' phone is used.
phone.txt	This file has all silence, and non-silence phones with four different categories Begin, End, Internal, and Singleton.
word.txt	This file keeps the documentation of all words present in this model.
ovv.txt	It keeps the record of Out-of-vocabulary (OOV) words.
utt_spk_trans_train	It is responsible for mapping utterance ID, speakers ID, and corresponding transcripts.
wav	This file connects the utterance ID and the location of every acoustic file.

2.2 Acoustic data

The acoustic data section of the model is organized with a few crucial files like spk2gender, wav.scp, utterance transcripts, utt2spk, and corpus.txt. 'spk2gender' contains gender details of every Adi speaker engaged in this ASR system. Location of speech samples for every utterance is placed in 'wav.scp' and utterance transcriptions in a 'text' file. 'utt2spk' is created for mapping speakers to utterances, whereas the transcript of the corpus is put together in corpus.txt file. The files essential to communicate between acoustic data and the Kaldi toolkit are shown in Table 3.

Table 5 Phoneme sequences of some Adi Words in Lexicon

Adi Word	Phoneme sequence
eeríng	e: r í ŋ
aadak	a: d a: k
bisulangka	b i s u l a: ŋ k a:
buik	b eu k
domapeka	d ɔ m a: p a k a:
ngaét	ŋ aə t
geartuidung	g e: r t eu d u ŋ
doato	d ɔa t ɔ
luingkoem	l eu ŋ k ɔ a m
oletyea	ɔ l a t y e:

2.3 Language data

The language data comprising Lexicon, silence, and non-silence phone information are listed in Table 4. The files 'silence.txt' and 'nonsilence.txt' include silence and non-silence phones.

A total of 50 non-silence phonemes of the Adi language are enlisted for this research. Some of the phones in this

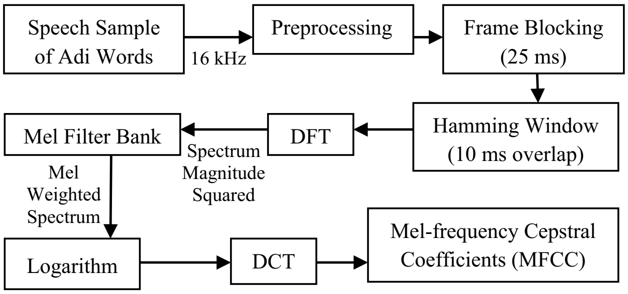
Speech Sample of Adi Words
16 kHz
Preprocessing
Frame Blocking (25 ms)
Hamming Window (10 ms overlap)
DFT
Spectrum Magnitude Squared
Mel Filter Bank
Mel Weighted Spectrum
Logarithm
DCT
Mel-frequency Cepstral Coefficients (MFCC)

Fig. 4 MFCC featuresextraction

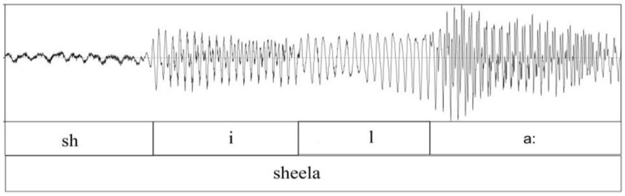

Fig. 5 Monophone alignment for 'sheela'

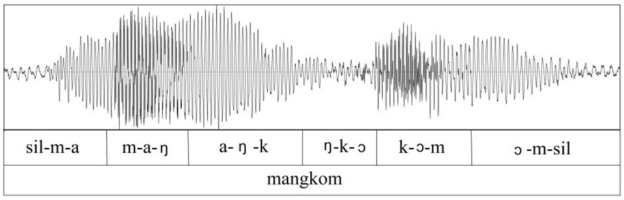

Fig. 6 Triphone alignment for 'mangkom'

word recognition system are a, a:, e, e:, i, i:, b, eu, k, l, m, n, ŋ, ɔ, p, s, t, z, ɔa, aə. 'sil' and 'spn' are the two silence phones apply in this ASR model. The phonetic representation of some Adi words is given in Table 5.

In this corpus of the Adi language, 21 native speakers uttered 2088 unique isolated words with 14,490 utterances of words. The speech files are stored in .wav format.

3 System configuration

The Adi language's automatic word recognition system was intended using Ubuntu 20.04 LTS (a 64-bit operating system). The system utilized is Lenovo IdeaPad Laptop associated with the Intel 10th Gen i5-10300 H Processor (2.5 GHz (Base)–4.5 GHz (Max), 4 Cores, 8 MB Cache), 8 GB RAM DDR4-3200, 512 GB SSD with NVIDIA GTX 1650 4GB GDDR6 dedicated GPU.

4 Feature extraction

The prime objective of feature extraction is to compute the feature vectors that symbolize the input speech as a sequence of observations. MFCC is a proficient and well-known speech feature extraction technique.

The steps for MFCC feature extraction are shown in Fig. 4 with the help of a block diagram. The MFCC features were taken out using a 25 ms Hamming windowing technique with a 10ms overlap. The recording sampling rate is 16 kHz, indicating 16,000 samples for every second of audio samples. So a sole window incorporates 400 samples abridged to 13 cepstral coefficients. Then, another 13 delta and an extra 13 delta-delta coefficients are computed for every frame, i.e., an MFCC feature vector of 39-dimensional is employed in this work. First, the analog speech sample is converted into a digital gesture, followed by the pre-emphasis method. The motive behind splitting the speech utterances into frames of tiny duration is that the speech is non-stationary, and its temporal characteristics change too fast as every 10 ms–100 ms, the vocal tract changes its shape to produce different sounds. So, by captivating a small frame dimension, one can assume that the speech utterances will be stationary, and their uniqueness will not differ much inside the frame. A 10-millisecond frameshift is selected to track the continuity of the audio signal and not miss out on any sudden changes in the edges of the frames. Then discrete fourier transform (DFT) is applied to extract frequency domain information from the time domain. The fourier transform output is first squared and applied to a Mel filter bank to get the information into the power spectrum. The logarithm of the output for the Mel filter bank is taken to eradicate the acoustic variants, which are not significant for this ASR model. Lastly, the outcomes of the log Mel spectrum are again converted into the time domain with the help of discrete cosine transform (DCT) [29], and cepstral coefficients (MFCC) are obtained as the final output.

Table 6 Silence and non-silence phones occurrences for all models

ASR Model	Silence (sil) at utterance begin %	Optional-silence (sil) at utterance end (%)	Overall non-silence (%)	Overall silence (sil) occupancy (%)
Mono	97.70	49.71	92.8	28.0
Tri-1	97.5	74.71	92.4	39.8
Tri-2	97.7	73.54	92.4	39.4
Tri-3	97.2	71.28	92.5	38.7
SGMM	97.4	71.64	92.5	38.5

5 Language model creation

The Kaldi toolkit employs a structure that counts on finite state transducers (FST). Typical ARPA format of language models modified into the .fst format using OpenFst. According to the Kaldi toolkit, .fst file format is required to compile this ASR model. In the language model L.fst, G.fst and L_disambig.fst files are created. L.fst or the Phonetic Dictionary FST is made with phonetic symbols in the input and word symbols in the output to represent the lexicon in .fst format. G.fst stands for the language model FST, used for grammar, where L_disambig.fst is the phonetic dictionary with disambiguation Symbols FST. The SRI language modelling toolkit (SRILM) is utilized to implement a statistical language model in this speech recognition system. An ARPA trigram model file is built to track the probabilities of each phone engaged in this ASR system.

6 Training and decoding

The speech corpus comprises 14,490-word utterances spoken by 21 native Adi speakers with 50 non-silence phones. Word error rates (WER) are computed to validate the ASR system for different recognition models. So, each utterance is divided into equal alignments for phone time marking and mapped each division to the particular phoneme symbol in the sequence. Each phoneme's probability density function (PDF) can be segmented and classified in this approach. All word utterances are split into individual units of speech known as a phoneme, as revealed in Table 5.

Table 7 Phones alignment and occurrences (Utterance begin/end) for tri3 model

Utterance begin/end	Phone	Phone occurrences (%)	Duration (median,mean, 95-percentile) frames
Begin	Sil	97.2	(57,62.4,121)
Begin	Nonsilence	2.8	(4,12.2,18)
Begin	a:_B	0.7	(3,6.2,4)
Begin	n_B	0.6	(4,6.5,7)
Begin	s_B	0.6	(3,5.0,4)
End	Sil	71.3	(206,217.7,517)
End	Nonsilence	28.7	(3,57.1,269)
End	ŋ_E	11.8	(6,55.5,226)
End	e:_E	5.1	(3,89.7,269)
End	ɔ_E	3.1	(3,4.6,11)
End	a:_E	2.7	(3,4.8,8)
End	e_E	2.1	(3,4.7,10)
End	u_E	1.9	(111,204.5,474)
End	i_E	0.7	(3,3.2,3)
End	a_E	0.6	(3,3.0,3)

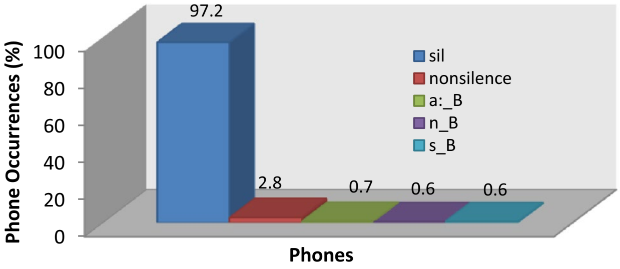

Fig. 7 Phones occurrences (utterance begin)

In the monophone model, every phoneme was compared separately, i.e., a single phone was taken into account independently. All neighboring phones were ignored at the time of training. Here in Fig. 5, the waveform and phonetic mapping have been shown for the word 'sheela.' The phonetic transcription of 'sheela' is 'sh-i-l-aa(a:).' So four non-silence phonemes are 'sh,' 'i', 'l,' and 'a:.' The total utterance duration for the word 'sheela' is 387.37 ms. Among these, 'sh' phone utterance duration is 89.4 ms,' i' uttered for 96.03 ms, and 'l' duration 84.67 ms. The end phone 'a:' exists for a slightly longer duration of 117.27 ms as it is a long vowel. A monophone model trained speech data with individual phone utterances in the first experiment. A word utterance does not simply depend on its phone sequences, but there is a strong influence of the phonemes that come left and right of each phoneme. So, combination with surrounding phonemes effectively improves the overall system recognition efficiency. The waveform and phonetic representation for the Adi word 'mangkom' ('Instead' in English) are in Fig. 6. The phonetic transcription of 'mangkom' is 'm-a-ng(ŋ)-k-o-m .' The triphone model embraces three successive phonemes to decide the probability of a specific phoneme. In this model, a total of 50 Adi phones are involved. So one hundred fifty (50*3) possible HMM states can be assigned, one for every triphone. In the training section, decision trees were formed using the collection of triphones. In this research, the triphone model comprises three training models: tri1, tri2, and tri3. The tri1 training model uses MFCC features and delta + delta-delta (Δ + Δ Δ) features.

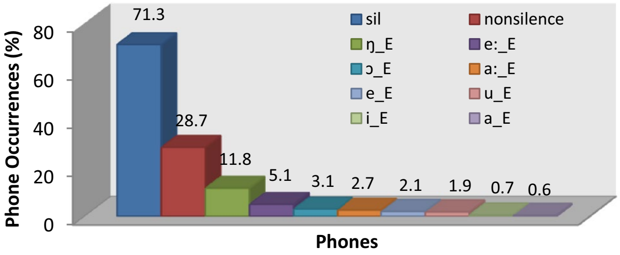

Fig. 8 Phones occurrences (utterance end)

Table 8 Overall phones alignment and occurrences for tri3 model

Phone	Phone occurrences (%)	Frame duration (median,mean, 95-percentile)
nonsilence	92.5	(8,13.1,30)
Sil	7.5	(59,102.3,351)
a:_I	6.3	(9,10.6,23)
ɔ_I	5.2	(8,9.0,20)
a_I	4.1	(8,10.0,26)
u_I	3.8	(5,46.4,334)
l_I	3.5	(7,11.1,32)
d_I	2.8	(6,7.7,17)
n_I	2.8	(5,7.0,15)
k_I	2.7	(6,7.8,17)
eu_I	2.7	(8,11.7,33)
i_I	2.4	(7,9.0,23)
p_I	2.2	(7,8.2,15)
k_B	2.1	(8,9.0,19)
ŋ_E	2.0	(8,24.1,122)
t_I	1.8	(7,8.0,15)
a:_E	1.6	(13,15.6,33)
d_B	1.5	(7,9.0,20)
ŋ_I	1.5	(7,19.3,57)
e:_I	1.4	(5,12.9,40)
b_B	1.4	(7,9.1,18)
y_I	1.3	(7,32.4,160)
a_B	1.3	(8,10.9,28)
r_I	1.3	(5,8.5,22)
m_B	1.2	(8,11.1,32)
s_B	1.2	(15,12.9,23)
a_E	1.2	(11,13.9,35)
eu_E	1.2	(12,14.9,36)
g_I	1.1	(7,10.0,30)
s_I	1.1	(13,11.8,21)
i_B	1.0	(11,11.7,26)
n_B	1.0	(4,5.7,12)
l_B	0.9	(8,13.2,40)
ɔa _B	0.9	(5,11.9,46)
t_B	0.9	(5,6.5,14)
i_E	0.9	(11,13.5,30)
k_E	0.8	(7,9.0,20)
p_B	0.8	(8,8.6,16)
ɔa_I	0.7	(9,11.6,29)
n_E	0.6	(4,7.8,20)
u_E	0.5	(5,37.1,214)
y_B	0.5	(9,21.9,45)

Tri2 employs linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT). LDA is usually responsible for decreasing the feature space and creating specific conversions for every speaker. In the tri3 model, speaker adaptive training (SAT) is included with LDA and MLLT [30]. Noise and speaker normalization are achieved using data metamorphose for every speaker. The SGMM is similar to a GMM system that inquires all HMM states to share a similar GMM configuration with an equal number of Gaussians in every state. The SGMM arrangement has sub-states without any speaker adaptive framework. The advantage of SGMM is that it is comparatively firm than monophone and triphone models as it has less number of parameters than actual speech states. This superiority benefits the ASR system's recognition efficiency with limited training data. A simple SGMM model with no sub-states performs better than the best GMM system, and after adding the sub-states model, WER further improved. There are two parts to a speech recognition system: acoustic modeling and decoding. The acoustic model transforms audio information into a phonetic attribute. The main goal is to precisely recognize phonemes and give outputs as a posteriorgram which signifies posterior probabilities for phonemes at every speech frame. The decoding model identifies features in the most probable word from a specific speaker's utterance. A decoding graph [28] is constructed as a weighted finite state machine with the assistance of a lexicon, acoustic, and language model to envisage the most probable word. A decoding graph can be symbolized as HCLG = H∘C∘L∘G. 'H' stands for Hidden Markov Model, 'C' represents context-dependency, 'L' denotes Lexicon, and 'G' designates language model or grammar.

7 Results and discussion

In this ASR system, HMM-GMM monophone model, HMM-GMM triphone (tri1, tri2, tri3) model, and SGMM, a total of five models are used for isolated word recognition of the Adi language. All non-silence and silence phones length, as well as occurrences, are analyzed for all models. Table 6 illustrates that for the monophone model, silence (sil) accounts for 97.7% of phone occurrences at the time of utterance begin, and optional silence 'sil' is seen only 49.71% of the time at utterance end. In tri1, tri2, tri3 and SGMM models the silence accounts 97.5, 97.7, 97.2 and 97.4% respectively at the time of utterances begin and 74.71, 73.54, 71.28 and 71.64% respectively at the time of utterance end.

7.1 Phones alignment Analyzation (Tri-3 model)

Phones alignment analysis of a speech recognition model is extremely imperative to know the actual silence and nonsilence phone occurrences involved in the recognition model of a specific language. The optional-silence 'sil' is

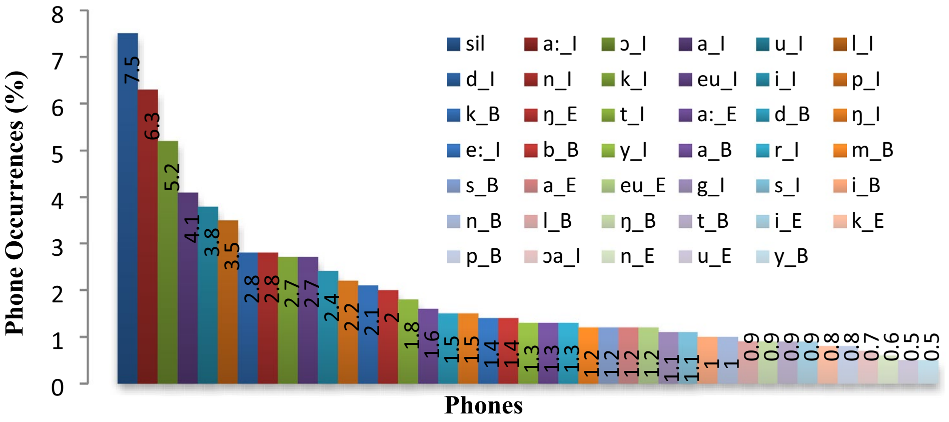

Fig. 9 Overall occurrences of each phone

seen only 71.28% of the time at utterance end in the tri-3 model. Table 7 shows the phone alignment and phone occurrences at the time of utterance begin and utterance end for the tri3 model. The phone's duration showed in frames using 'median, mean, 95-percentile'. In this mentioned model, at utterance begin, the silence (sil) accounts for 97.2% of phone occurrences, i.e., nonsilence accounts for 2.8%. Among these nonsilence phone occurrences at utterance begin, three phone a:_B, n_B, and s_B account for 0.7, 0.6 and 0.6%, respectively. At the end of the utterance, the silence (sil) accounts for 71.3% and nonsilence for 28.7%; mainly eight nonsilence phones 'ŋ_E', 'e:_E', ' ɔ_E', 'a:_E', 'e_E', 'u_E', 'i_E' and 'a_E' are observed at utterance end. Phone's occurrences at the time of utterance begin, and utterance end has been exhibited in Figs. 7 and 8 respectively.

4140 Adi words are considered for testing in this model. Table 8, overall phone alignment and occurrences are enlisted for the tri-3 model. At the time of decoding, all 4140 Adi words are phonetically divided. In widespread occurrences of each phone in the tri-3 model, the silence 'sil' is perceived only 7.5% with the duration of (59,102.3,351) frames considering 'median, mean and 95-percentile' and nonsilence phone occurs 92.5% with the duration of (8,13.1,30) frames. Forty nonsilence phone occurrences are illustrated in Table 8, and Fig. 9 shows three different phone categories: begin, end, and internal. Here a:_I (internal) confirms the highest overall occurrences with 6.3%, whereas u_E (end) and y_B (begin) assert the minor occurrences of 0.5% for each phone. So this type of analysis can provide a clear idea about the occurrences of every phone for a particular speech dataset and can also bestow a prediction for phone generation probability of a specific language.

The optional-silence phone 'sil' occupies 38.7% of overall frames. Limiting the stats to the 62.1% of frames not covered by an utterance-[begin/end] phone, optional-silence sil occupies 5.7% of frames. Utterance-internal optional-silences sil comprise 2.2% of utterance-internal phones, with duration (median, mean, 95-percentile) of (22,34.1,101). In word boundary analysis, the silence can appear in five different ways in the speech samples: nonwords (sil), begin (sil_B), end (sil_E), internal (sil_I), and singleton (sil_S). Every nonsilence phone has four categories: begin, end, internal, and singleton. For example, 'ŋ' nonsilence phone categorization are ŋ _B (begin), ŋ_E (end), ŋ_I (internal), and ŋ_S (singleton).

7.2 WER and WRA

This isolated word recognition system is trained with 15 Adi speakers (6 male and 9 female) and tested with six speakers (three male and three female). The word error rate (WER) and word recognition accuracy (WRA) can determine the efficiency of any speech recognition system. The WER is the least edit distance between the ASR output and the definite transcriptions. Word error rate is typically illustrated in Eqs. (1) and (2) as

$$WER = \frac{Deletions(D) + Substitutions(S) + Insertions(I)}{Total\ number\ of\ the\ words\ (N)} \tag{1}$$

$$WER\ (\%) = \frac{Deletions\ (D) + Substitutions\ (S) + Insertions\ (I)}{Deletions\ (D) + Substitutions\ (S) + Correct\ Words(C)} \times 100 \tag{2}$$

Where N specifies the total number of words, D symbolizes deletion error, S represents the numeral of substitutions error, and I stand for insertion error. The word recognition accuracy (WRA) is conferred in Eqs. (3) and (4) as,

$$\mathrm{WRA} = (1 - \mathrm{WER}) = \frac{\mathrm{N - D - S - I}}{\mathrm{N}} = \frac{\mathrm{C - I}}{\mathrm{N}} \tag{3}$$

$$WRA\ (\%)\ = 100 - WER\ (\%) \tag{4}$$

In Table 9; Fig. 10, eight recognition results for each model are enlisted. The Monophone, tri1, tri2, tri3, and

Table 9 Performance analysis of different ASR models

ASR Model	Insertion	Deletion	Substitution	WER (%)	WRA (%)
Mono	**92**	**384**	**657**	**27.37**	**72.63**
	83	401	672	27.92	72.08
	101	322	742	28.14	71.86
	78	409	710	28.91	71.09
	66	277	899	30.00	70.00
	85	195	1030	31.64	68.36
	107	372	845	31.98	68.02
	194	400	780	33.19	66.81
Tri-1	**3**	**377**	**251**	**15.24**	**84.76**
	3	333	302	15.41	84.59
	2	336	317	15.82	84.18
	7	352	298	15.87	84.13
	4	405	274	16.50	83.50
	9	341	360	17.15	82.85
	15	423	284	17.44	82.56
	1	443	293	17.80	82.20
Tri-2	**10**	**285**	**221**	**12.46**	**87.54**
	10	197	310	12.49	87.51
	10	164	368	13.09	86.91
	12	204	383	14.47	85.53
	9	251	342	14.54	85.46
	10	173	431	14.83	85.17
	7	174	446	15.14	84.86
	4	224	421	15.68	84.32
Tri-3	**15**	**195**	**160**	**8.94**	**91.06**
	15	195	163	9.01	90.99
	17	191	179	9.35	90.65
	15	203	173	9.44	90.56
	16	209	181	9.81	90.19
	13	233	161	9.83	90.17
	19	197	196	9.95	90.05
	19	194	202	10.02	89.98
SGMM	**3**	**177**	**149**	**7.95**	**92.05**
	0	154	179	8.04	91.96
	1	173	198	8.99	91.01
	2	141	232	9.06	90.94
	1	142	198	8.24	91.76
	1	126	374	12.10	87.90
	1	114	417	12.85	87.15
	1	180	380	13.55	86.45

The best WRAs are shown in bold in the first row of each model

SGMM model has 33 different recognition outputs. 4140-word utterances of the Adi are considered for every recognition consequence to analyze the performance of different models. Table 9 shows eight different recognition outputs for each model. The best WERs and WRAs are illustrated in the table's first row of each model. After overall analysis, the Monophone model was the least efficient, whereas SGMM showed the best recognition performance. The best recognition output of the Monophone model led to 92 insertions, 384 deletions, and 657 substitution errors. So, only 3007 words were correctly recognized among 4140 words used for testing. The SGMM model shows three insertions, 177 deletions, and 149 substitution errors, i.e., 3811-word utterances are properly recognized among 4140 words. The best WRAs are shown in bold in the first row of each model.

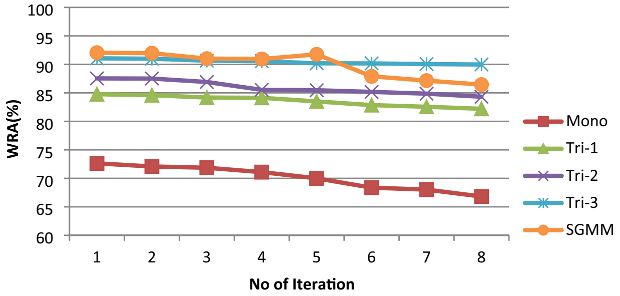

Fig. 10 Recognition accuracy of different ASR models

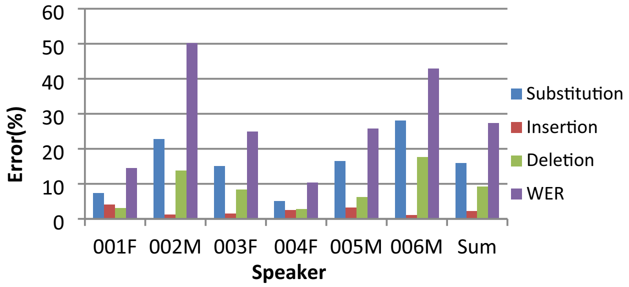

Fig. 11 Monophone model performance analysis for individual speaker

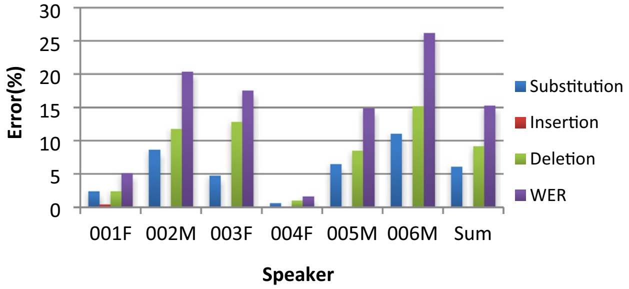

Fig. 12 Tri-1 model performance analysis for individual speaker

In this work, the authors applied 10,350-word utterances (recordings of 15 speakers) for training and 4140-word

Table 10 WER analysis of Monophone model for individual speaker

Speaker	No of words uttered	Correctly recognize		Substitution error		Insertion error		Deletion error		Error (WER)	
		Words	%	Words	%	Words	%	Words	%	Words	%
001 F	764	684	89.53	56	7.33	31	4.06	24	3.14	111	14.53
002 M	580	368	63.45	132	22.76	7	1.21	80	13.79	219	50.17
003 F	657	503	76.56	99	15.07	10	1.52	55	8.37	164	24.96
004 F	515	475	92.23	26	5.05	13	2.52	14	2.72	53	10.29
005 M	651	504	77.42	107	16.44	21	3.23	40	6.24	168	25.81
006 M	973	565	58.07	237	28.06	10	1.03	171	17.57	418	42.96
Sum	**4140**	**3099**	**74.86**	**657**	**15.87**	**92**	**2.22**	**384**	**9.28**	**1133**	**27.37**

Bold values show the overall performance of different ASR models

Table 11 WER analysis of Tri-1 model for individual speaker

Speaker	No of words uttered	Correctly recognize		Substitutions error		Insertion error		Deletion error		Error (WER)	
		Words	%	Words	%	Words	%	Words	%	Words	%
001 F	764	728	95.29	18	2.36	3	0.39	18	2.36	39	5.10
002 M	580	462	79.66	50	8.62	0	0	68	11.72	118	20.34
003 F	657	542	82.50	31	4.72	0	0	84	12.79	115	17.50
004 F	515	507	98.45	3	0.58	0	0	5	0.97	8	1.55
005 M	651	554	85.10	42	6.45	0	0	55	8.45	97	14.90
006 M	973	719	73.90	107	11.00	0	0	147	15.11	254	26.10
Sum	**4140**	**3512**	**84.83**	**251**	**6.06**	**3**	**0.07**	**377**	**9.11**	**631**	**15.24**

Bold values show the overall performance of different ASR models

Table 12 WER analysis of Tri-2 model for individual speaker

Speaker	No of words uttered	Correctly recognize		Substitutions error		Insertion error		Deletion error		Error (WER)	
		Words	%	Words	%	Words	%	Words	%	Words	%
001 F	764	717	93.85	30	3.93	8	1.05	17	2.23	55	7.20
002 M	580	481	82.93	40	6.90	0	0	59	10.17	99	17.07
003 F	657	566	86.15	36	5.48	2	0.30	55	8.37	93	14.16
004 F	515	513	99.61	0	0	0	0	2	0.39	2	0.39
005 M	651	571	87.71	48	7.37	0	0	32	4.92	80	12.29
006 M	973	786	80.78	67	6.89	0	0	120	12.33	187	19.22
Sum	**4140**	**3634**	**87.78**	**221**	**5.34**	**10**	**0.24**	**285**	**6.88**	**516**	**12.46**

Bold values show the overall performance of different ASR models

Table 13 WER analysis of Tri-3 model for individual speaker

Speaker	No of words uttered	Correctly recognize		Substitutions error		Insertion error		Deletion error		Error (WER)	
		Words	%	Words	%	Words	%	Words	%	Words	%
001 F	764	736	96.34	22	2.88	7	0.92	6	0.79	35	4.58
002 M	580	523	90.17	25	4.31	0	0	32	5.52	57	9.83
003 F	657	583	88.74	33	5.02	2	0.30	41	6.24	76	11.57
004 F	515	514	99.81	0	0	1	0.19	1	0.19	2	0.39
005 M	651	588	90.32	31	4.76	4	0.61	32	4.92	67	10.29
006 M	973	841	86.43	49	5.04	1	0.10	83	8.53	133	13.67
Sum	**4140**	**3785**	**91.43**	**160**	**3.86**	**15**	**0.36**	**195**	**4.71**	**370**	**8.94**

Bold values show the overall performance of different ASR models

Table 14 WER analysis of SGMM for individual speaker

Speaker	No of words uttered	Correctly recognize		Substitutions error		Insertion error		Deletion error		Error (WER)	
		Words	%	Words	%	Words	%	Words	%	Words	%
001 F	764	759	99.35	5	0.65	0	0	0	0	5	0.65
002 M	580	539	92.93	22	3.79	0	0	19	3.28	41	7.07
003 F	657	578	87.98	27	4.11	0	0	52	7.91	79	12.02
004 F	515	512	99.42	3	0.58	0	0	0	0	3	0.58
005 M	651	586	90.02	30	4.61	0	0	35	5.38	65	9.98
006 M	973	840	86.33	62	6.37	3	0.31	71	7.30	136	13.98
Sum	**4140**	**3814**	**92.13**	**149**	**3.60**	**3**	**0.07**	**177**	**4.28**	**329**	**7.95**

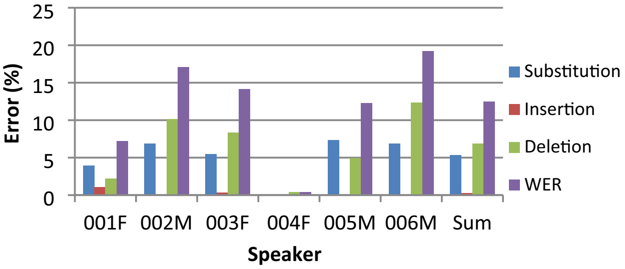

Fig. 13 Tri-2 model performance analysis for individual speaker

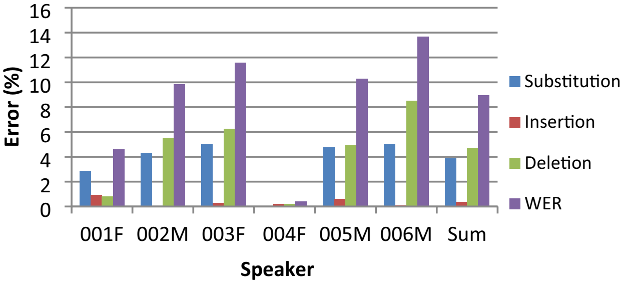

Fig. 14 Tri-3 model performance analysis for individual speaker

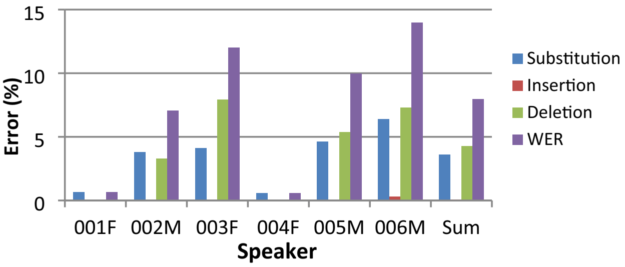

Fig. 15 SGMM performance analysis for individual speaker

utterances (recordings of 6 speakers) for testing, i.e., 71.43% of data used for training and 28.57% for testing. In the monophone model, WRA is 72.63%, and WER is 27.37%. In triphone models, WER was decreased, and the recognition accuracy of tri, tri2, and tri3 became 84.76, 87.54 and 91.06%, respectively. The SGMM model offered the lowest WER of 7.95%. So considering the recognition outputs of each model, Fig. 10 shows SGMM has the highest recognition accuracy, followed by Tri-3, Tri-2, Tri-1, and Monophone models. Although SGMM offers the highest recognition outputs for decoding results, some recognition outputs of the Tri-3 model performed better than SGMM.

7.3 WER for individual speakers

Among 21 speakers, six speakers' speech samples have been used for testing. The WER can be calculated separately for different speakers with different models. The WER of individual speakers are given in Tables 10, 11, 12, 13 and 14, considering substitution, insertion, and deletion errors with the number of uttered words and correctly recognized words for Monophone, Triphone, and SGMM models. Whereas, Figs. 11, 12, 13, 14 and 15 demonstrates the performance analysis for individual speakers for the corresponding five models.

8 Conclusion

This research signifies the first initiative to build an ASR system for a low-resource Arunachali endangered tribal language, 'Adi.' In this work, an isolated word recognition system has been developed on the Kaldi toolkit using speech samples from native Adi speakers. MFCC features were extracted from the speech samples and were used in the ASR system. Out of 14490-word utterances in the dataset with 2088 unique Adi words, 10350-word utterances (recordings of 15 speakers) were used for training, and 4140-word utterances (recordings of 6 speakers) were used for testing, i.e., 71.43% of the dataset was used for training, and 28.57% of the dataset was used for testing. The monophone model achieved a WRA of 72.63% and a WER of 27.37%. In triphone models, the WER decreased as the system complexity increased; the recognition accuracy achieved using the tri1 model, tri2 model, and tri3 model are 84.76, 87.54, and 91.06%, respectively. The SGMM model offered the lowest WER of 7.95%. Further, it was also observed that WER varies from speaker to speaker. In this work lowest WER of 0.58% was observed with speaker' 004F'.

9 Future scope

Future works for the current research may include testing of the models with a larger speech corpus of different age group and dialects. Additionally, noisy data may be applied in these present models to make this Adi ASR system more robust and efficient. Development of various ASR applications in the upcoming future will be helpful to preserve this low recourse tribal language in this digitalized era.

Acknowledgements Authors are extremely thankful to all the 21 native Adi speakers of Arunachal Pradesh for sharing their audio samples to make the speech corpus.

Funding No funding has been received for this particular research work.

Data Availability The custom code dataset are available from thecorresponding author on request.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

References

1. Sharma G (2020) A Study on open voice data in indian languages. Deutsche Gesellschaft fürInternationale Zusammenarbeit (GIZ) GmbH. https://toolkit-digitalisierung.de/app/uploads/2021/02/Study-on-Open-Voice-Data-in-Indian-Languages_GIZ-BizAugmentor.pdf. Accessed 12 Feb 2022
2. Office of the Registrar General, India (2018) Language – India, States and Union Territories. https://censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf. Accessed 15 Dec 2021
3. Yu D, Deng L (2016) Automatic speech recognition. Springer, Berlin
4. Pillai LG, Mubarak DMN (2021) A stacked auto-encoder with scaled conjugate gradient algorithm for Malayalam ASR. Int J Inf Tecnol 13:1473–1479. https://doi.org/10.1007/s41870-020-00573-y
5. López G, Quesada L, Guerrero LA (2018) Alexa vs. Siri vs. Cortana vs. Google Assistant: a comparison of Speech-Based natural user interfaces. Advances in human factors and Systems Interaction. Advances in Intelligent Systems and Computing, vol 592. Springer, Cham, pp 241–250. https://doi.org/10.1007/978-3-319-60366-7_23
6. Levis J, Suvorov R (2020) Automatic speech recognition. In: Concise encyclopedia of applied linguistics. Wiley, New York https://doi.org/10.1002/9781405198431.wbeal0066.pub2
7. Kumar A, Mittal V (2021) Hindi speech recognition in noisy environment using hybrid technique. Int J Inf Tecnol 13:483–492. https://doi.org/10.1007/s41870-020-00586-7
8. Babhulgaonkar AR, Sonavane SP (2022) Experimenting with factored language model and generalized back-off for Hindi. Int J Inf Tecnol 14:2105–2118. https://doi.org/10.1007/s41870-020-00503-y
9. Guchhait S, Hans ASA, Augustine J (2022) Automatic speech recognition of Bengali using Kaldi. In: Proceedings of second international conference on sustainable expert systems, Springer, Singapore, pp 153–166. https://doi.org/10.1007/978-981-16-7657-4_14
10. Supriya S, Handore SM (2017) Speech recognition using HTK toolkit for Marathi language. In: International conference on power, control, signals and instrumentation engineering (ICPCSI), IEEE, pp 1591–1597. https://doi.org/10.1109/ICPCSI.2017.8391979
11. Reddy MR, Laxminarayana P, Ramana AV, Markandeya JL, Bhaskar JI, Harish B, Jagadheesh S, Sumalatha E (2015) Transcription of Telugu TV news using ASR. ICACCI, IEEE, pp 1542–1545. https://doi.org/10.1109/ICACCI.2015.7275832
12. Lokesh S, Malarvizhi KP, Ramya DM, Parthasarathy P, Gokulnath C (2019) An automatic tamil speech recognition system by using bidirectional recurrent neural network with self-organizing map. Neural Comput and Appl 31(5):1521–1531. https://doi.org/10.1007/s00521-018-3466-5
13. Ananthi S, Dhanalakshmi P (2013) Speech recognition system and isolated word recognition based on hidden Markov model, HMM for hearing impaired. Int J Comput Appl 73:30–34. https://doi.org/10.5120/13012-0241
14. Bharali SS, Kalita SK (2015) A comparative study of different features for isolated spoken word recognition using HMM with reference to assamese language. Int J Speech Technol 18:673–684. https://doi.org/10.1007/s10772-015-9311-7
15. Sawant S, Deshpande M(2018) Isolated spoken Marathi words recognition using HMM. In: 2018 4th International conference on computing communication, control, automation, IEEE, pp 1–4. https://doi.org/10.1109/ICCUBEA.2018.8697457
16. Slívová M, Partila P, Továrek J, Vozňák M (2020) Isolated word automatic speech recognition system. In: Multimedia communications, services and security. communications in computer and information science, 1284, Springer, Cham, pp 252–264. https://doi.org/10.1007/978-3-030-59000-0_19
17. Sri KVL, Srinivasan M, Nair RR, Priya KJ, Gupta D (2020) Kaldi recipe in Hindi for word level recognition and phoneme level transcription. Procedia Comput Sci 171:2476–2485. https://doi.org/10.1016/j.procs.2020.04.268
18. Paul B, Bera S, Paul R, Phadikar S (2021) Bengali spoken numerals recognition by MFCC and GMM technique. Advances in Electronics, Comm and Comput. Springer, Singapore, pp 85–96. https://doi.org/10.1007/978-981-15-8752-8_9
19. Boumehdi A, Yousfi A (2022) Arabic speech recognition independent of vocabulary for isolated words. In: Proceedings of sixth international congress on information and communication

technology, Springer, Singapore, pp 585–595. https://doi.org/10.1007/978-981-16-1781-2_52

20. Nimachow G, Taga T, Tag H, Dai O (2010) Linkages between bio-resources and human livelihood: a case study of Adi Tribes of Mirem Village, Arunachal Pradesh, India. The Initiation 2:183–198. https://doi.org/10.3126/init.v2i1.2542
21. Lalrempuii C (2005) Morphology of the Adi language of Arunachal Pradesh. Doctoral dissertation, NEHU, Shillong
22. Sasmal S, Saring Y (2020) Spectral analysis of consonants in Arunachali native language-adi. Electronic Systems and Intelligent Computing. Springer, Singapore, pp 783–790. https://doi.org/10.1007/978-981-15-7031-5_74.
23. Sasmal S, Saring Y (2022) Robust automatic continuous speech recognition for 'Adi', a zero-resource indigenous language of Arunachal Pradesh. Sādhanā 47(4):1–5. https://doi.org/10.1007/s12046-022-02051-6
24. Hamidi M, Zealouk O, Satori H et al (2023) COVID-19 assessment using HMM cough recognition system. Int J Inf Tecnol 15:193–201. https://doi.org/10.1007/s41870-022-01120-7
25. Lad NR, Nirmal JH, Naikare KD (2019) Total variability factor analysis for dysphonia detection. Int J Inf Tecnol 11:67–74. https://doi.org/10.1007/s41870-018-0112-5
26. Povey D, Burget L et al (2010) Subspace Gaussian mixture models for speech recognition. In: International conference on acoustics, speech and signal processing, IEEE, pp 4330–4333. https://doi.org/10.1109/ICASSP.2010.5495662
27. Mizera P, Pollak P (2013) Accuracy of HMM-based phonetic segmentation using monophone or triphone acoustic model. In: International conference on applied electronics, IEEE, pp 1–4
28. Povey D, Ghoshal A et al (2011) The Kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding, IEEE signal processing society
29. Ahmed N, Natarajan T, Rao K (1974) Discrete cosine transform. IEEE Trans Computers C 23(1):90–93. https://doi.org/10.1109/T-C.1974.223784
30. Miao Y, Zhang H, Metze F (2015) Speaker adaptive training of deep neural network acoustic models using i-vectors. IEEE/ACM Trans Audio Speech Lang Process 23:1938–1949. https://doi.org/10.1109/TASLP.2015.2457612

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.