Sādhanā (2022)47:271
https://doi.org/10.1007/s12046-022-02051-6

Robust automatic continuous speech recognition for 'Adi', a zero-resource indigenous language of Arunachal Pradesh

SAJAL SASMAL* and YANG SARING

Department of Electronics and Communication Engineering, National Institute of Technology Arunachal Pradesh, Jote, Arunachal Pradesh 791113, India
e-mail: sajal.sasmal@gmail.com

MS received 29 July 2022; revised 31 October 2022; accepted 4 November 2022

Abstract. This article depicts an automatic speech recognition system (ASR) for the continuous speech of 'Adi,' a zero-resourced endangered indigenous language of Arunachal Pradesh, India. This ASR system uses a speech corpus of 40 native Adi speakers of Arunachal Pradesh. Mel frequency cepstral coefficients (MFCC) features extracted from recorded Adi speech samples. Different speech recognition models, such as Sub Space Gaussian Mixture Model (SGMM), Monophone, and Triphone (tri1, tri2, tri3), were applied in this ASR system. The monophone model's word recognition accuracy (WRA) was 58.5%. In triphone models, the recognition efficiency of tri1, tri2, and tri3 was enhanced at 78.18%, 83.08%, and 88.62%, correspondingly. SGMM was the most proficient model in this ASR system, with a minimum word error rate (WER) of 10.12%. This proffered ASR model may be beneficial in the physical world to set up various ASR applications of man-machine interfaces in the Adi language.

Keywords. Automatic speech recognition; Adi; zero-resource indigenous language; Arunachal Pradesh.

1. Introduction

Many indigenous languages nowadays are struggling to survive, and those languages are at risk of vanishing [1], despite making up 6.2% of the world's population. Approximately 7,151 languages are spoken today in the world. In this fragile time, around 40% of languages are endangered, often with less than 1,000 speakers remaining. A language dies with its last speaker every two weeks; within the next century, 50–90% of them are expected to be extinct [2]. Just 23 languages are spoken by more than half of the world's population. Asia has the most indigenous language after Africa. Those indigenous languages have the best libraries from which one can acquire the most pertinent information, history, mythology, and perception of the entire populace [1, 2]. Their diversity is being lost at an alarming rate because of external forces, social pressures, and demographic change. This methodical disappearance of the aboriginal languages intimidates thousands of families, including children and native communities, and the survival of their languages globally. It is tough to analyze and process indigenous languages using computational models as most of this does not have any writing system. It is imperative to preserve these languages as they are enriched in oral tradition and amazingly consistent and reliable over time.

To enjoy the reimbursement of contemporary technology, it needs to be language-independent. Merely 10–12 % of Indian citizens are comfy in English, according to a report on open voice data in Indian languages in the year 2020 [3]. The rest of the Indian inhabitants are just cozy in their mother tongue. Prominent companies like Google Assistance, Amazon Alexa, Apple Siri, Microsoft Cortana digital assistant, etc., are interested in developing ASR [4] systems only on commercially beneficial languages. Google Home and Microsoft Cortana currently support only 12 and 8 languages worldwide, including only Hindi from India. Apple's Siri voice assistant supports 21 languages worldwide but has not included any Indian language.

Already some researchers have taken the initiative to build speech recognition systems in some significant and well-resourced Indian languages. In [5], a Monophone-based Hindi ASR using a corpus of 15 speakers of 240 unique Hindi words with a WRA of 80.28% is designed. Guchhait *et al* [6] implemented an ASR of Bengali using the Kaldi toolkit. Paulose *et al* [7] developed a Marathi ASR system with a recognition accuracy of 78.8% with SGMM. An ASR system was designed to transcribe Telugu TV news automatically [8]. An End-to-End Tamil speech recognition system was developed by Changrampadi *et al* [9] gave a WER of 24.7%. According to the 2011 census, the number of speakers of Hindi, Bengali, Marathi, Telugu, and Tamil are 52.83 crore, 9.72 crore, 8.30 crore, 8.11 crore, and 6.90 crore, respectively [10]. Besides, India

*For correspondence
Published online: 12 December 2022

has 116 languages with more than 10000 speakers and many low-resource endangered indigenous languages. Very few research initiative has been taken for these low-resource endangered languages. The vowel's acoustic analysis of five under-resourced languages, Nagamese, Ao, Lotha, Sumi, and Angami of Nagaland done in [11]. Dey *et al* [12] built 670 words speech corpus of 81 native Mizo speakers and implemented an ASR with a WRA of 87.37%. Nyodu *et al* [13] constructed an ASR of the Galo language of Arunachal Pradesh in Kaldi with an accuracy of 80%. Kaldi-based Santhali language digits recognition model executed in [14].

In this current research, an initiative has been made to build an ASR system for an endangered tribal language, 'Adi.' Monophone, Triphones (tri1, tri2, tri3), and SGMM [15], a total of five models, are employed in this ASR system to recognize continuous Adi speech.

Thus, the novelties/contributions of the present research work are:

- Here, authors include 21 new Adi phonemes.
- A corpus created with 25358-word utterances from 3498 sentences of 40 native Adi speakers.
- The phonetic transcriptions of 3654 Adi words.
- This proposed ASR system of Adi may become an opening step to preserve this zero-resource indigenous language from the risk of disappearing.
- This article shows excellent recognition efficiency on monophone, triphones, and SGMM models.

2. Adi language

As per the 'UNESCO Atlas of the World's languages in Danger 2017,' India has 197 endangered languages, whereas only Arunachal Pradesh counts 33, including Adi. The United Nations announced 2019 as the 'Year of Indigenous Languages (IY2019)' to lift their consciousness for benefiting the native speakers of indigenous languages and also to be grateful for their significant contribution to our world's rich cultural diversity. 2,48,834 native speakers of Adi mostly live in the east, west, and upper Siang districts [10] of Arunachal Pradesh. The Adi language came from the Tibeto-Burman family, typically linked with the Sino-Tibetan family.

The significant challenges of working with Adi are

- Adi does not have any proper script, dictionary, or writing system. So it is too challenging to represent phonetic transcripts of utterances.
- Adi is a zero-Resource Indigenous Language. So no audio recordings are available on the internet or any other digital medium.
- Generally, native speakers encounter difficulties in reading the English transcripts of Adi, which leads to data collection being more complicated.
- The Adi tribes inhabited the diverse mountainous regions of Arunachal Pradesh, making the job more complicated.

Lalrempuii studied the morphology of the Adi language [16]. Lalrempuii enlisted 29 phonemes, including 15 consonants, eight vowels, and six diphthongs, in the linguistic research. In 2021 Sasmal *et al* examined the spectral characteristics of 16 Adi consonants [17].

In this current research, the authors include 21 new Adi phonemes, i.e., 50 unique phonemes with seven short and seven long vowels, 16 consonants, 19 diphthongs, and a single triphthong. The Adi vowels are shown in table 1.

The span of vowels (short or long) in the Adi language can alter the same word's meaning. So, for a short vowel, there is a corresponding long vowel that provides a different sense to the word. Table 2 shows the list of Adi consonants. The phonetic properties of Adi consonants are alveolar, bilabial, glottal, velar, and palatal. This language's known manners of articulation are affricates, approximants, fricatives, liquids, nasals, and stops. Some examples of diphthongs in this language are /aé/ [aə], /ao/ [aɔ], /ai/ [ai], /ié/ [iə], /io/ [iɔ], /oé/ [ɔə], /íé/ [iə], /ué/ [uə] and single triphthong is 'uai.' The retroflex, aspirated, dental, and labio-dental fricative sounds are missing in the Adi language.

3. Model construction

Figure 1 illustrates the architecture of the ASR model of Adi language. Continuous speech samples have been collected from 40 native Adi speakers. The next step is to extract the MFCC feature from the recorded Adi speech. Finally, continuous speeches are decoded and recognized using an acoustic model, language model, and phonetic transcript.

Figure 2 depicts the steps of the ASR system. First continuous audio samples are collected and stored in '.wav format' (waveform audio file). The language data was

Table 1. List of Adi vowels.

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

Table 2. List of Adi consonants.

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

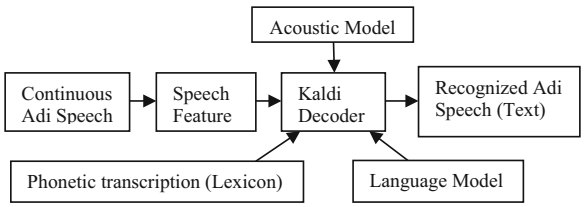

Figure 1. The architecture of the ASR model of Adi Language.

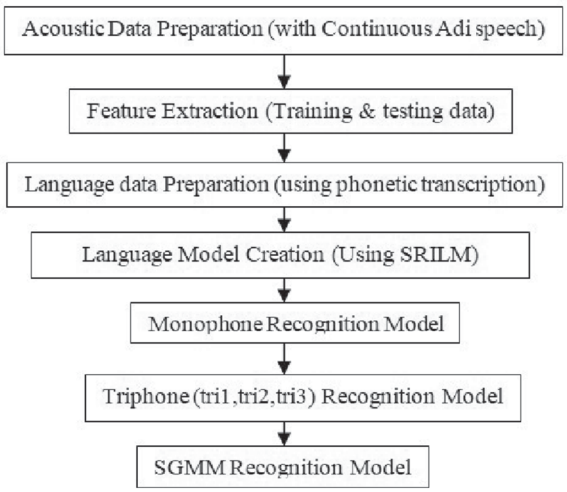

Figure 2. Step by step procedure of ASR model.

prepared using the total number of phonemes available in the Adi language. All recorded sentences in the speech corpus are phonetically represented at the word level. A language model has been constructed with the SRILM toolkit. The continuous Adi speech is recognized using monophone, triphones, and SGMM models.

3.1 *Speech data collection*

The continuous audio samples of native Adi speakers were recorded in WAVE file format. The sampling rate is 16 kHz with a 256k bit rate of signed PCM encoding and a 16-bit mono channel. The speech corpus comprises data from 40 Adi speakers (16 male and 24 female). This dataset incorporates 3498 continuous sentences, 25358-word utterances, and 3654 unique Adi words.

3.2 *Acoustic data*

The acoustic data contains important files like 'wav.scp', 'text,' 'utt2spk', 'spk2utt', 'corpus.txt,' and 'spk2gender'. The 'wav.scp' includes the location of every acoustic file to link all sentence utterances with an associated acoustic file used in this model in .wav format. The 'utt2spk' file connects each utterance with the correlated speaker. The 'spk2utt' holds the speakers to utterance mapping. The 'corpus.txt' includes every single utterance transcription that may take place in this recognition system.

3.3 *Language data*

The language data consists of a lexicon, non-silence, silence, and optional silence phones. The 'lexicon.txt' contains phone transcriptions of each Adi word from the dictionary with the help of a phoneme set developed for this research. The 'nonsilence_phones.txt' encloses a list of each non-silence phone present in this model. The 'phone.txt' has all silence and non-silence phones. The 'word.txt' file keeps the documentation of all words in this recognition system. The 'ovv.txt' holds the record of out-of-vocabulary (OOV) words. The 'utt_spk_trans_train' file maps utterance ID, speakers ID, and corresponding transcripts. The phonetic representations of a few Adi sentences are revealed in table 3.

4. Language model creation

Kaldi utilizes a configuration that relies on finite state transducers (FST). The phonetic dictionary FST (L.fst) is designed to present the lexicon in .fst format with phonetic symbols at the input and word symbols at the output. G.fst was created for grammar, whereas L_disambig.fst has

Table 3. Phoneme sequences of some Adi sentences in Lexicon.

Adi sentence	Phoneme Sequence
bui berokpe aayea	b eu-b a r ɔ k p e-a: y e:
nonyi akone aatoka	n ɔ ñ i-a: k ɔ n e-a: t ɔ k a
ngo nyok kai	ŋ ɔ -ñ ɔ k- k a: i
uirgape agear ibosulangka	eu r g a: p e-a: g e: r-i b ɔ s u l a: ŋ k a:
ngo aguike supe ekum moto	ŋ ɔ-a: g eu k e-s u p e-a k u m-m ɔ t ɔ

disambiguation symbols for debugging. The SRILM (SRI Language Modeling) toolkit makes statistical language models for speech recognition. An ARPA tri-gram model file is created to compute the probabilities of all available phones in this ASR system.

5. Training and decoding

A total of 50 non-silence phones were used to represent 25358-word utterances from 3498 continuous sentences of the corpus. Every word of the Adi speech is alienated into equal alignments to mark the phone time, and each section is mapped to a specific phoneme symbol in the sequence. In the monophone model, each phoneme was compared disjointedly, ignoring all neighboring phones during the training. A speech or word utterance not only depends on its individual phone's progression but also has a sturdy influence on the adjacent phonemes. So, overall ASR model efficiency may improve with adjoining phonemes. The triphone model adopts three consecutive phonemes to determine the emission probability of a particular phoneme. At the time of model training, the decision trees were created using the collection of triphones. The triphone model comprises three training models named tri1, tri2, and tri3. The tri1 model utilizes MFCC with its delta and delta-delta features, whereas the tri2 model uses maximum likelihood linear transform and linear discriminant analysis (MLLT+ LDA). The tri3 model uses speaker adaptive training (SAT) in addition to LDA and MLLT [18]. SGMM is analogous to a Gaussian Mixture Model (GMM) system that asks all Hidden Markov model (HMM) states to share the same GMM configuration with an equal number of Gaussians in each state. SGMM does not include SAT. Subspaces are introduced instead of larger models to reduce the number of parameter estimation concerns by decreasing system dimensions. SGMM can give superior recognition efficiency to monophone and triphone models of an ASR system with inadequate training data. The acoustic model of an ASR system converts the speech samples into a phonetic attribute. Here the models seek to identify the phonemes accurately and output them as a posteriorgram which indicates the posterior probability for each phone in an individual speech frame. A decoding model predicts the continuous speech sentences from the most probable words of the speaker's utterance as an output with the help of a decoding graph. The decoding process is implemented with the support of decoding graphs with the assistance of a lexicon, grammar, and acoustic model.

6. Experimental results and discussion

The Authors used 28 Adi speakers (11 male and 17 female) speech samples to train and 12 speakers (5 male and 7 female) to test the ASR system. The effectiveness of an ASR system may be assessed using the WER and WRA.

$$\mathrm{WER} = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{N}} \tag{1}$$

$$\mathrm{WER}\,(\%) = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{S} + \mathrm{D} + \mathrm{C}} \times 100 \tag{2}$$

Here N is the total number of words present in an ASR system. S, D, and I are substitutions, deletion, and insertion errors, and C is the correct word. So, WRA may be determined as

$$\mathrm{WRA} = (1 - \mathrm{WER}) \tag{3}$$

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

Figure 3 illustrates the WERs (%) for different ASR models by considering all 33 recognition outputs of each model. The monophone model showed a maximum WER of 46.57% and a minimum of 41.5%. In the triphone model,

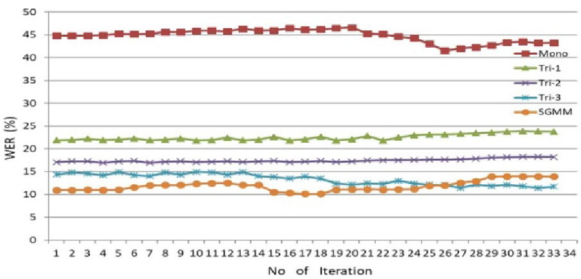

Figure 3. Recognition accuracy of different ASR models.

Table 4. Performance of ASR models.

ASR Model	Sentences in testing data	Words in testing data	Wrongly recognize words	Insertion Error	Deletion Error	Substitution Error	WER (%)	WAR (%)
Mono	1036	7452	3093	234	842	2017	41.5	58.5
Tri-1			1628	158	661	809	21.82	78.18
Tri-2			1261	107	457	697	16.92	83.08
Tri-3			848	93	243	512	11.38	88.62
SGMM			754	65	368	321	10.12	89.88

tri1, tri2, and tri3 produced the highest WER of 23.85%, 18.26%, 14.94%, and the lowest WER of 21.82%, 16.92%, 11.38%, respectively. The SGMM has a peak WER of 13.97% and a bottom of 10.12%. Although SGMM enlisted the lowest WER, in some recognition outputs, the tri3 model beats the recognition efficiencies of the SGMM approach.

In table 4, the best performance of the five ASR models has been included. After analyzing the system outputs, the authors find that SGMM is the superior model, with the highest recognition accuracy of 89.88%, whereas the monophone model offered the lowest accuracy of 58.5%. In this research, the authors applied 17906-word utterances (70.61%) for training and 7452 words utterances (29.39%) for testing.

7. Conclusion

This research is the first initiative to make a continuous ASR system of zero-resource Arunachali endangered indigenous language, 'Adi.' The voice utterances of native Adi speakers are recorded and used to create a speech recognition model. The WER and word WRA are the key parameters to compute the efficiency of the proposed model.

- The ASR system of the Adi language was implemented using the Kaldi toolkit.
- The authors created a corpus of 3498 continuous sentences having 25358-word utterances from 40 Adi speakers (age group 17–55).
- Monophone, triphones, and SGMM analyze the system's performance.
- The WER is calculated for different models.

Funding No funding has been received for this work.

Declarations

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

References

[1] Billings D 2022 The loss of indigenous languages. *Вестник антропологии* 100–112

[2] Angarova G 2021 Our future is tied to indigenous languages. *Cult. Surv.* 45: 1

[3] 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

[4] Yu D and Deng L 2016 Automatic Speech Recognition (Vol. 1). Springer, Berlin

[5] Bhatt S, Jain A and Dev A 2021 Monophone-based connected word Hindi speech recognition improvement. *Sādhanā* 46: 1–17

[6] Guchhait S, Hans A S A and Augustine J 2022 Automatic Speech Recognition of Bengali Using Kaldi. In: *Proceedings of 2nd International Conference on Sustainable Expert Systems*, Springer, Singapore, pp. 153–166

[7] Paulose S, Nath S and Samudravijaya K 2018 Marathi Speech Recognition. In: *SLTU*, pp. 235–238

[8] Reddy M R, Laxminarayana P, Ramana A V, Markandeya J L, Bhaskar J I, Harish B, Jagadheesh S and Sumalatha E 2015 Transcription of Telugu TV news using ASR. *International Conference on Advances in Computing, Communications and Informatics, IEEE,* pp. 1542–1545

[9] Changrampadi M H, Shahina A, Narayanan M B and Khan A N 2022 End-to-end speech recognition of Tamil language. *Intell. Autom. Soft Comput.* 32: 1309–1323

[10] Office of the Registrar General, India 2018 LANGUAGE – INDIA, STATES AND UNION TERRITORIES. https://censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf

[11] Basu J, Basu T, Khan S, Pal M, Roy R, Bepari M S, Nandi S, Basu T K, Majumder S and Chatterjee S 2017 Acoustic analysis of vowels in five low resource north East Indian languages of Nagaland. *O-COCOSDA, IEEE*, pp. 1–6

[12] Dey A, Sarma B D, Lalhminghlui W, Ngente L, Gogoi P, Sarmah P, Prasanna S M, Sinha R and Nirmala S R 2018 Robust Mizo Continuous Speech Recognition. In: *Interspeech*, pp. 1036–1040

[13] Nyodu K and Vijaya S 2020 Automatic speech recognition of Galo. In: *Electronic Systems and Intelligent Computing* Springer, Singapore, pp 663–671

[14] Kumar A, Kumar R and Kishore K 2020 Performance analysis of ASR Model for Santhali language on Kaldi and Matlab Toolkit. *RTEICT, IEEE*, pp. 88–92

[15] Povey D, Burget L, Agarwal M, Akyazi P, Feng K, Ghoshal A, Glembek O, Goel N K, Karafiát M, Rastrow A and Rose R C 2010 Subspace Gaussian mixture models for speech recognition. *International Conference on Acoustics, Speech and Signal Processing, IEEE*, pp. 4330–4333

[16] Lalrempuii C 2005 Morphology of the Adi language of Arunachal Pradesh. Doctoral dissertation, NEHU, Shillong

[17] Sasmal S and Saring Y 2020 Spectral analysis of consonants in Arunachali Native language-Adi. In: *Electronic Systems and Intelligent Computing* Springer, Singapore, pp 783–790

[18] Miao Y, Zhang H and 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