Int. j. inf. tecnol. (October 2023) 15(7):3513–3519
https://doi.org/10.1007/s41870-023-01451-z

ORIGINAL RESEARCH

A zero-resourced indigenous language phones occurrence and durations analysis for an automatic speech recognition system

Sajal Sasmal[1] · **Yang Saring**[1]

Received: 16 March 2023 / Accepted: 26 August 2023 / Published online: 8 September 2023

Abstract This research illustrates phone occurrence analysis for an automatic speech recognition (ASR) model of 'Adi.' 'Adi' is a low-resourced endangered tribal language of Arunachal Pradesh, a northeastern state of India. Phones alignment analysis is crucial to know the actual silence and nonsilence phone occurrences of an ASR design of a specific language. This investigation was executed on a continuous speech corpus of 84 native Adi speakers with five distinct ASR models, monophone, triphone (tri-1, tri-2, tri-3), and Sub Space Gaussian Mixture Model (SGMM). Phones duration is also exhibited in frames using 'median, mean, 95-percentile' for all models. In monophone, tri-1, tri-2, tri-3, and SGMM models, the silence 'sil' is glimpsed at 97.6%, 97.4%, 97.6%, 97.1%, and 97.3%, respectively, at the time of utterances begin and 49.42%, 74.68%, 73.48%, 71.23%, and 71.57% respectively at the time of utterance end. Overall occurrences for a:_I, ɔ_I, ɔ_E, k_I, ŋ_E, s_B, ɔə_I, y_B phones in the SGMM model are 6.3%, 5.3%, 3.1%, 2.7%, 2.0%, 1.2%, 0.7%, and 0.5%.

Keywords Adi · Zero-resource endangered tribal language · Automatic speech recognition · Phones occurrence · Phones alignment analysis · Arunachal Pradesh

✉ Sajal Sasmal
sajal.sasmal@gmail.com

Yang Saring
ysaring@nitap.ac.in

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

1 Introduction

Language is a fascinating aspect of human nature that has evolved primarily as a speech form, a series of symbols that convey messages from speaker to listener. It develops along with the modern man, Homo sapiens. Therefore, there is no best language as there are no best species [1]. Language is the most fragile thing. All languages are equally complex, if not simple, similarly powerful, and equipped to perform any task, store, process, and transfer information in exchange [2]. They grow as much as one wishes to use. They are flexible, as much as speakers want to bend them. However, the range of that flexibility is reached by native speakers.

India has 480 tribal languages, most of which are endangered. There are 197 endangered languages in India, 33 found only in Arunachal Pradesh, as reported by the UNESCO Atlas in 2017. Adi is one of them. When a language disappears, the culture, history, and mythology associated with it also die. It is excruciating to see at least two languages die with their last speaker every month in this high-tech technology-enabled 'Anno Domini era.' There is no script, dictionary, or writing system for the vast majority of tribal languages.

Although the development of the ASR [3, 4] system for some commercially beneficial and well-resourced languages started a few decades ago, very infrequent initiatives have been taken for low-resourced tribal languages. Cross-lingual acoustic modeling using SGMM for recognition of low-resource speech implemented in [5]. An experimental analysis of six vowels in Nagaland's Ao, Lotha, and Nagamese languages was accomplished with nucleus vowel duration, formants, and intensity for speech recognition, language identification, and synthesis research [6]. The acoustic–phonetic investigation and categorization of the Sora vowels of the Munda language were carried out

in [7]. An ASR system of Sora Language was developed using 20 speakers and 146 min of audio corpus [8]. Tanwar et al. performed an in-depth machine translation investigation of morphologically rich Indo-Aryan and Dravidian languages under zero-resource conditions [9]. Tzudir et al. [10] worked on automatic dialect identification of Ao, a Tibeto-Burman under-resourced language of Nagaland. An experimental study was performed on two low-resource tribal languages, Santali and Hrangkhawl, for automatic language identification [11]. Kumar et al. [12] developed 18 h of speech corpus (4–5 h for each language) for four under-resourced Indo-Aryan languages, Awadhi, Braj, Bhojpuri, and Magahi, for ASR system modeling. Pre-trained self-supervised models with only 10 h of labeled data for 15 low-resource languages have been developed to boost the performance of ASR for these languages [13]. The generative adversarial network used for phone identification from unpaired speech utterances and phone sequencing was proposed in [14]. Both frame-wise and segment-wise generator categorization techniques were used for better performance. On any enormous spontaneous speech corpus, Akita et al. suggested a statistical pronunciation modeling method for pronunciation changes between base form and surface [15]. This model was trained for phone context-dependent dissimilarity patterns, variable length, and incidence probabilities. Cao et al. worked on the reliability of confidence measures for local phone mismatch of noisy and unclear Chinese speech [16]. This approach predicted the mismatch phone from the recognition phone series by analyzing individual phone occurrence frequency in speech frames. A statistical phone duration model was developed by Lo et al. to identify entire utterances in a corpus of 100 Chinese learners' English speech used in a computer-controlled pronunciation instruction system [17]. In [18], feature-based pronunciation modeling was implemented to improve ASR system performance for pronunciation variability considering phonetic insertion, substitution, and deletion rules. Stănescu et al. created phonetically balanced speech data for the ASR system for under-resourced Romanian languages [19]. Stolcke et al. [20] focused on precise phonetic segmentation utilizing boundary correction models and system fusion considering phonetic context and duration features. Acoustic phonetic analysis of the ASR system using the statistical and landmark-based approach for phone and syllable recognition was documented in [21]. Phuong et al. created a phonetically balanced high-quality Vietnamese speech corpus of 5400 utterances from 12 speakers for analyzing speech characteristics and building speech synthesis models [22].

In [23], Lalrempuii worked on the Adi language morphology considering 29 phones. The spectral properties of Adi consonants were analyzed using formants frequencies [24]. In [25], researchers created a continuous ASR system for the Adi language.

In this current research, overall phone occurrence analysis of all 50 Adi phone are determined for five different ASR models, monophone, triphone (tri-1, tri-2, tri-3), and SGMM. Phone durations are also shown in frames using 'median, mean, 95-percentile' for all models.

The significant contributions of this research are:

- A detailed phone alignment and occurrence analysis was conducted on various speech recognition models.
- A corpus was built using 7528 continuous Adi sentences having 54,252-word utterances from 84 native speakers.
- The 7384 Adi words have been phonetically transcribed.
- This type of analysis can make the ASR system more robust and efficient.

2 Phones of Adi language

The Adi language is classified as Tibeto-Burman, often associated with the language family of Sino-Tibetan. The majority of Arunachal Pradesh's 2.49 lakh native Adi people dwell in the state's upper, east, and west Siang districts [26].

In order to develop the ASR system, the authors used a total of 50 distinct Adi phonemes. Seven long and seven short monophthongs (vowels), 19 diphthongs, 16 consonants, and one triphthong make up the language's phonetic inventory. In Adi, seven short vowels are /a/ [a], /é/ [ə], /í/ [ɨ], /i/ [i], /e/ [e], /o/ [ɔ], /u/ [u], and long vowels are /ii/ [iː], /íí/ [ɨː], /uu/ [uː], /ee/ [eː], /éé/ [əː], /aa/ [aː], /oo/ [ɔː]. Among them, /u/, /uː/, /ɔ/, /ɔː/ are rounded, and reaming ten are unrounded. The vowel's duration (long or short) in the Adi can change the sense of the same word. Consonants /d/, /b/, /j/, /g/, /m/, /n/, /l/, /r/, /ñ/ [ny], /ŋ/ [ng], and /y/ are all voiced /h/, /k/, /s/, /p/, and /t/ are all unvoiced. The Adi consonant phonetic properties are alveolar (/j/, /s/, /l/, /n/, /t/, /d/), bilabial (/b/, /m/, /p/), glottal (/h/), palatal (/ñ/, /y/), and velar (/g/, /k/, /r/, /ŋ/). Affricates (/j/), fricatives (/h/, /s/), approximants (/y/), nasals (/n/, /m/, /ñ/, /ŋ/), liquids (/l/, /r/), and stops (/b/, /d/, /g/, /k/, /p/, /t/) are all examples of articulation styles. In this language, some instances of diphthongs are /aé/ [aə], /au/ [au], /éi/ [əi], /ia/ [ia], /ié/ [iə], /íа/ [ɨa], /íé/ [ɨə], /oa/ [ɔa], and /ui/ [ui]. The /uai/ is the only triphthong in Adi.

The noteworthy challenges of this research are

- Adi has adopted a modified Roman script for writing but is still in the developing stage; therefore, it is very difficult to represent Adi utterances in proper phonetic transcripts.
- Data recording is more difficult since most native Adi speakers cannot understand Adi sentences written in modified Roman scripts.

- Aboriginal speakers have inhabited the various mountainous areas of Arunachal Pradesh.

3 Speech dataset

The continuous speech utterances of 36 male and 48 female native Adi speakers (age 17–45 years) of Arunachal Pradesh were recorded using digital voice recording tools by the authors. The recording samples were stored in WAVE file format (.wav) in the corpus. The bit rate of the speech data was 256 k, and the sampling rate was 16 kHz using signed PCM encoding. The corpus consists of 7384 unique Adi words. In this dataset 54,252-word utterances are present in 7528 sentences.

4 Model construction

The architecture of the phone alignment analysis for the Adi Language ASR model is shown in Fig. 1. 84 native Adi speakers' continuous speech data have been recorded. Then Mel frequency cepstral coefficients (MFCC) features of the speech utterances were extracted. Finally, overall silence, nonsilence phone occurrences, and phone duration of the continuous ASR system for five different models are predicted using the Kaldi decoder with the support of a lexicon, an acoustic model, and a language model.

A continuous ASR system has been developed using five models: monophone, triphone (tri-1, tri-2, tri-3), and SGMM. In the acoustic data of the ASR system, each speech file location keeps in 'wav.scp'. The 'utt2spk' and 'spk2utt' files map speakers and utterances. The complete utterance transcripts are put in 'corpus.txt.'

In the language data of the ASR system, 'lexicon.txt' holds phonetic transcriptions of all Adi words present in the corpus using 50 Adi phones considered for this research. The 'nonsilence_phones.txt' includes all non-silence phones available in the models. All non-silence and silence phone records are put in 'phone.txt.'

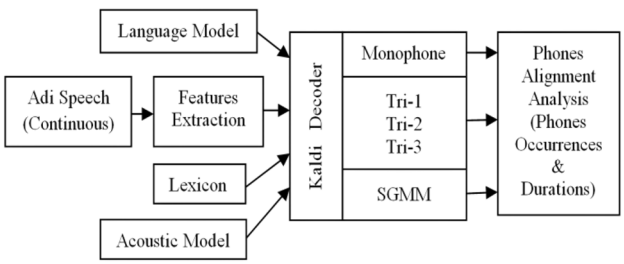

Fig. 1 The architecture of phones alignment analysis

The file 'word.txt' contains the information on every word used by this speech recognition system. The 'utt_spk_trans_train' includes a mapping between utterances, speakers, and related transcripts. Table 1 reveals the phonetic translations of some Adi utterances.

In the align_lexicon file, each word available in the lexicon is split into phoneme sequences using 50 nonsilence phones of the Adi language. Every nonsilence phone has three different categories: begin (B), end (E), and internal (I). For example, 'm' nonsilence phone categorization are m_B, m_E, and m_I. Table 2 shows the phonetic alignment of some Adi words. For example 'dooying' word consists of five nonsilence phones /d/, /ɔ:/, /y/, /i/, and /ŋ/. The starting phone /d/ is categorized as 'begin' (d_B). The phones /ɔ:/, /y/, and /i/are categorized as 'internal' (ɔ:_I, y_I, and i_I). The final phone is labeled as 'end' (ŋ_E).

The Kaldi is a FST (finite state transducer) toolkit. L.fst (phonetic dictionary) is prepared to give the lexicon with phonetic characters in.fst form. G.fst was constructed for language model or grammar, L_disambig.fst contains a list of disambiguation symbols useful in debugging. Statistical language models for the ASR system may be built with the help of the SRILM toolkit. All phones in this recognition system have their probability calculated using an ARPA trigram model file.

The ASR system's acoustic model converts the voice samples to phonetic sequences. The goal of the models here is accurate phoneme identification, which is reported as posteriorgrams (posterior probability for each phone in a speech frame). A decoding model indicates the occurrence of phonemes in continuous speech sentences based on the speaker's utterances, and its output is a decoding graph.

Table 1 Phonetic transcription of few Adi utterances in Lexicon

Adi utterance	Phonetic transcription
yagope duk mangkom	y a: g ɔ p e - d u k - m a: ŋ k ɔ m
no ngom tomduneya	n ɔ - ŋ ɔ m - t ɔ m d u n a y a:
idola bunyi annyi mato	i d ɔ l a: - b u ñ I - a: n ñ i - m a: t ɔ
sinying ke duitak si aido	s i ñ i ŋ - k a - d eu t a: k - s i - a: i d ɔ
irea inyo la imapeka	i r e: - i ñ ɔ - l a - i m a: p a k a:

Table 2 Phonetic alignment of few Adi words

Adi word	Phonetic alignment
buiroem	b_B eu_I r_I ɔ_I a_I m_E
dooying	d_B ɔ:_I y_I i_I ŋ_E
eluing	a_B l_I eu_I ŋ_E
gooralo	g_B ɔ:_I r_I a:_I l_I ɔ_E
menyok	m_B a_I ñ_I ɔ_I k_E
yuutgoo	y_B u:_I t_I g_I ɔ:_E

Decoding is accomplished using a lexicon, language model, and acoustic model.

5 Results and discussion

The monophone and triphone (tri1, tri2, tri3) approaches are founded on Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM). The acoustic model size recedes in the SGMM approach. In this technique, all phonetic states convey a common GMM structure where mixture and means weights differ in a subspace of the entire parameter space. Each silence, nonsilence phone duration, and occurrence are investigated for all five models.

Table 3 exhibits that in the monophone model, 'sil' (silence) is noticed for 97.6% of phone occurrences at the time of utterance initiates, and optional silence 'sil' is caught merely 49.42% when utterance stop. In the tri1, tri2, tri3, and SGMM approach, the silence was observed at 97.4%, 97.6%, 97.1%, and 97.3%, respectively, at utterances begin and 74.68%, 73.48%, 71.23%, and 71.57% respectively at utterance end.

5.1 Phones alignment analyzation

A phone alignment assessment is essential to understand the actual nonsilence and silence phone occurrences included in an ASR system of any language. Table 4 demonstrates overall phone occurrences for five different ASR models of Adi. The phone's duration is exhibited in frames by 'median, mean, 95-percentile'.

5.1.1 Monophone model

In the monophone model, silence (sil) accounts for 97.6% of phone occurrences at utterance beginning with duration (median, mean, 95-percentile) of (56, 61.2, 118) frames and optimal silence 49.42% at the time of utterance end with duration (144, 171.4, 463) frames. So, at utterance begin, nonsilence accounts for 2.4% of phone occurrences, with a duration of (5, 11.9, 19) frames, and 50.58% at utterance ending with a duration of (37, 103.7, 321) frames. At utterance end, nonsilence phone occurrences of ŋ_E, e:_E, u_E, a:_E, and ɔ_E are 22.7%, 16.1%,8.2%,0.9%, and 0.7% with a duration of (10, 41.1, 170), (136, 126.9, 280), (252, 259.8, 491), (3, 8.0, 14) and (3, 6.0, 7) frames. The overall phone alignment and occurrences analysis of the monophone model shows that the nonsilence phones account for 92.7% of phone occurrences, with a duration of (9, 16.2, 33) frames. Table 4 shows different nonsilence phone occurrences for the monophone model. The phone a:_I has the highest overall occurrences of 6.3% with a duration of (11, 11.3, 21) frames, whereas ɔə_I and n_E show minor occurrences of 0.6% for each phone with a duration of (12, 12.8, 24) and (6, 7.5, 16) frames, respectively. The optional-silence phone 'sil' occupies 28.2% of frames overall.

5.1.2 Tri-1 model

In the tri-1 model, silence accounts for 97.4% of phone occurrences at the beginning of utterance with a duration of (59, 63.4, 118) frames, and optimal silence 74.68% at the time of utterance ending with a duration of (208, 216.1, 516) frames. So, at utterance begin, nonsilence accounts for 2.6% of phone occurrences, with a duration of (4, 13.1, 20) frames where n_B occurrence is 0.6% with (4, 5.2, 7) frames duration and 25.32% at utterance end with the duration of (4, 59.2, 290) frames. At utterance end, non-silence phone occurrences of ŋ_E, e:_E, u_E, a:_E, ɔ_E and e_E are 11.5%, 4.8%, 2.0%, 1.8%, 1.8% and 1.6% with duration of (9, 47.2, 226), (3, 92.0, 268), (107, 190.4, 474), (3, 3.1, 3), (3, 5.1, 11) and (3, 3.6, 3) frames. The overall phone alignment and occurrences analysis of the tri-1 model shows that the nonsilence phones account for 92.3% of phone occurrences, with a duration of (8, 13.2, 30) frames. Table 4 shows individual nonsilence phone occurrences where a:_I has the highest overall occurrences of 6.3% with a duration of (10, 11.0, 24) frames, and u_E, y_B, and b_I show the minor occurrences of 0.5% for each phone with the duration of (4, 39.6, 205), (8, 21.4, 46) and (5, 14.4, 21) frames respectively. The optional-silence phone 'sil' occupies 39.9% of frames overall.

Table 3 Silence and non-silence phones occurrences for all models

Model	Silence at utterance begin (%)	Optional-silence at utterance end (%)	Overall non-silence (%)	Overall silence occupancy (%)
Mono	97.6	49.42	92.7	28.2
Tri-1	97.4	74.68	92.3	39.9
Tri-2	97.6	73.48	92.3	39.5
Tri-3	97.1	71.23	92.4	38.8
SGMM	97.3	71.57	92.4	38.6

Table 4 Overall phones occurrences for different ASR model

Phone	Phone occurrences (%)				
	Mono	Tri-1	Tri-2	Tri-3	SGMM
Nonsilence	92.7	92.3	92.3	92.4	92.5
Sil	7.2	7.6	7.6	7.7	7.5
a:_I	6.3	6.3	6.3	6.2	6.3
ɔ_I	5.3	5.3	5.3	5.3	5.3
a_I	4.4	4.1	4.1	4.1	4.1
u_I	3.9	3.8	3.8	3.8	3.7
l_I	3.4	3.5	3.5	3.5	3.5
m_I	3.3	3.2	3.3	3.3	3.3
ɔ_E	2.9	3.1	3.0	3.1	3.1
m_E	2.7	3.0	3.0	3.0	3.0
e_E	2.9	3.0	3.0	3.0	3.0
a:_B	2.5	2.9	2.8	3.0	3.0
n_I	2.6	2.9	2.8	2.8	2.8
d_I	3.0	2.7	2.8	2.7	2.8
k_I	3.1	2.7	2.7	2.8	2.7
eu_I	3.0	2.7	2.7	2.7	2.7
i_I	2.3	2.4	2.4	2.4	2.4
p_I	2.6	2.2	2.2	2.1	2.2
k_B	2.4	2.1	2.1	2.1	2.0
ŋ_E	2.4	2.0	2.0	1.9	2.0
t_I	1.8	1.8	1.8	1.8	1.8
a:_E	1.7	1.7	1.6	1.6	1.6
d_B	2.0	1.5	1.5	1.6	1.5
e:_I	1.2	1.4	1.4	1.4	1.4
e:_E	0.9	X	X	X	X
ŋ_I	1.4	1.5	1.5	1.5	1.5
b_B	1.3	1.4	1.4	1.4	1.4
a_B	1.3	1.4	1.4	1.3	1.3
y_I	1.2	1.3	1.3	1.3	1.3
r_I	1.1	1.3	1.2	1.2	1.3
s_B	1.4	1.2	1.2	1.2	1.2
a_E	1.0	1.2	1.2	1.2	1.2
m_B	1.2	1.2	1.2	1.2	1.2
eu_E	1.3	1.2	1.2	1.2	1.2
g_I	0.9	1.1	1.1	1.1	1.1
i_B	1.0	1.0	1.0	1.0	1.0
s_I	1.1	1.0	1.0	1.1	1.1
n_B	1.1	1.0	1.0	1.0	1.0
l_B	0.8	0.9	0.9	0.9	0.9
ŋ_B	0.8	0.9	0.9	0.9	0.9
t_B	0.9	0.9	0.9	0.9	0.9
k_E	0.8	0.8	0.8	0.9	0.8
i_E	0.9	0.8	0.8	0.8	0.8
p_B	0.7	0.8	0.8	0.8	0.8
ɔə_I	0.6	0.7	0.7	0.7	0.7
n_E	0.6	0.6	0.6	0.6	0.7
u_E	0.7	0.5	0.5	0.5	0.5
b_I	X	0.5	X	X	X
y_B	0.8	0.5	X	X	0.5

5.1.3 Tri-2 model

In the tri-2 model, silence accounts for 97.6% of phone occurrences at the beginning of utterance with a duration of (59, 63.4, 120) frames, and optimal silence 73.48% at the time of utterance ending with a duration of (207, 217.0, 513) frames. So, at utterance begin, nonsilence accounts for 2.4% of phone occurrences, with a duration of (4, 12.9, 18) frames where n_B and s_B both phones occurrences are 0.6% with frames duration of (3, 5.2, 7) and (3, 5.0, 4). At utterance end, nonsilence accounts for 26.52% with a duration of (3, 58.6, 269) frames where phone occurrences of ŋ_E, e:_E, a:_E, e_E, ɔ_E, and u_E are 12.0%, 4.7%, 2.2%, 2.0%, 2.0%, and 1.9% with a duration of (5, 49.7, 220), (5, 94.8, 269), (3, 4.7, 5), (3, 4.2, 10), (3, 5.3, 11) and (111, 201.0, 474) frames. The overall phone alignment and occurrences analysis of the tri-2 model shows that the nonsilence phones account for 92.3% of phone occurrences, with a duration of (8, 13.2, 30) frames. Table 4 shows individual nonsilence phone occurrences where a:_I has the highest overall occurrences of 6.3% with a duration of (10, 11.0, 24) frames, and u_E shows the minor occurrences of 0.5% for each phone with a duration of (5, 37.5, 207) frames. The optional-silence phone 'sil' occupies 39.5% of frames overall.

5.1.4 Tri-3 model

In the tri-3 model, silence accounts for 97.1% of phone occurrences at the beginning of utterance with a duration of (59, 63.4, 121) frames, and optimal silence, 71.23% at the time of utterance ending with a duration of (207, 218.5, 517) frames. So, at utterance begin, nonsilence accounts for 2.9% of phone occurrences, with a duration of (4, 12.7, 18) frames where n_B and s_B phone occurrences are 0.7% and 0.6% with frames duration of (4, 6.0, 7) and (3, 5.0, 4). At utterance end, nonsilence accounts for 28.77% with a duration of (3, 59.4, 269) frames where phone occurrences of ŋ_E, e:_E, a:_E, ɔ_E, u_E, e_E, and i_E are 11.7%, 5.0%, 2.3%, 2.2%, 1.9% 1.8% and 0.7% with a duration of (5, 54.3, 227), (3, 89.7, 269), (3, 5.0, 8), (3, 6.8, 13), (111, 200.7, 474), (3, 5.0, 10) and (3, 3.0, 3) frames. The overall phone alignment and occurrences analysis of the tri-3 model shows that the nonsilence phones account for 92.4% of phone occurrences, with a duration of (8, 13.3, 30) frames. Table 4 shows individual nonsilence phone occurrences where a:_I has the highest overall occurrences of 6.3% with a duration of (9, 10.7, 23)

frames, and u_E show the minor occurrences of 0.5% with the duration of (5, 36.4, 214) frames. The optional-silence phone 'sil' occupies 38.8% of frames overall.

5.1.5 SGMM model

In the SGMM model, silence accounts for 97.3% of phone occurrences at the beginning of utterance with a duration of (57, 62.1, 121) frames, and optimal silence 71.57% at the time of utterance ending with a duration of (207, 217.0, 506) frames. So, at utterance beginning, nonsilence accounts for 2.7% of phone occurrences, with a duration of (4, 12.4, 19) frames where a:_B and n_B phone occurrences are 0.7% and 0.6% with frames duration of (3, 6.2, 3) and (4, 6.8, 7). At utterance end, nonsilence accounts for 28.43% with a duration of (4, 58.8, 270) frames where phone occurrences of 'ŋ_E,' 'e:_E,' 'ɔ_E,' 'a:_E,' 'u_E,' 'e_E,' 'i_E' and 'a_E' are 11.8%, 5.0%, 2.9%, 2.6%, 2.0%, 2.0%, 0.6%, and 0.6% with a duration of (7, 59.2, 226), (5, 89.9, 270), (3, 4.8, 11), (3, 3.4, 5), (3, 4.7, 9), (111, 189.7, 473), (3, 3.0, 3) and (3, 3.0, 3) frames. The overall phone alignment and occurrences analysis of the SGMM model shows that the nonsilence phones account for 92.5% of phone occurrences, with a duration of (8, 13.2, 30) frames. Table 4 shows individual nonsilence phone occurrences where a:_I has the highest overall occurrences of 6.3% with a duration of (9, 10.5, 23) frames, whereas phones u_E and y_B both show minor occurrences of 0.5% with a duration of (5, 37.4, 213) and (9, 22.8, 61) frames. The optional-silence phone 'sil' occupies 38.6% of frames overall.

Occurrences of phones at utterance begin, and utterance end has been displayed in Figs. 2 and 3, respectively.

6 Conclusion

In this work, authors investigated phone occurrence and alignment analysis for the ASR system of a low-resourced, critically endangered tribal language of Arunachal Pradesh.

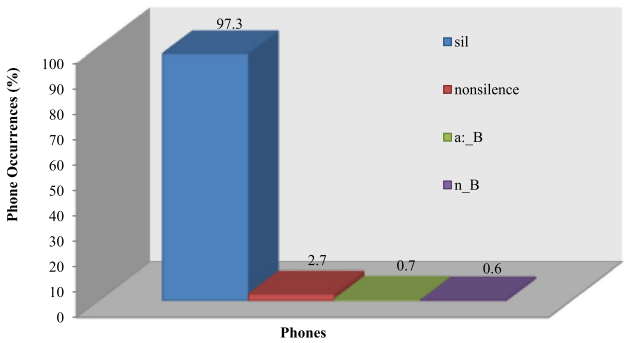

Fig. 2 Occurrences of phones at utterance begin for SGMM

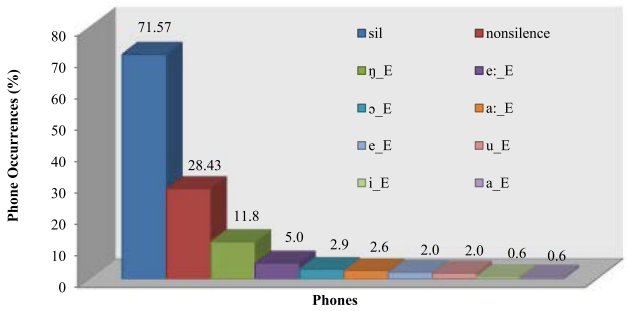

Fig. 3 Occurrences of phones at utterance end for SGMM

Actual silence and nonsilence phone occurrences are calculated for different models. The duration of each phone is measured in frames using 'median, mean, 95-percentile'. This type of analysis with oversized speech corpus can predict individual phone occurrences of a specific language.

- Phone alignment and occurrence analysis are executed employing the Kaldi toolkit for the Adi ASR system.
- The authors build a corpus of 84 native Adi speakers' utterances comprising 54,252 words.
- Five distinct ASR models, monophone, triphone (tri-1, tri-2, tri-3), and SGMM consider for analysis.

7 Future scope

The future scope of the present research is testing models with a larger speech corpus of different dialects and age groups. In-depth phone occurrences analysis will make the Adi ASR system more efficient and robust. Future research will be beneficial to preserve this low-recourse endangered tribal language in this digitalized era.

Acknowledgements The authors are incredibly grateful to all the 84 Adi inhabitants of Arunachal Pradesh for sharing their audio samples to make the speech corpus.

Funding No funding has been obtained for this research work.

Code and dataset availability The custom code dataset are accessible from the corresponding author on demand.

Declarations

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

References

1. Bickerton D (1990) Language and species. University of Chicago Press, Chicago

2. Locke JL, Bogin B (2006) Language and life history: a new perspective on the development and evolution of human language. Behav Brain Sci 29(3):259–280. https://doi.org/10.1017/S0140525X0600906X
3. Pillai LG, Mubarak DMN (2021) A stacked auto-encoder with scaled conjugate gradient algorithm for Malayalam ASR. Int J Inf Technol 13:1473–1479. https://doi.org/10.1007/s41870-020-00573-y
4. Kumar A, Mittal V (2021) Hindi speech recognition in noisy environment using hybrid technique. Int J Inf Technol 13:483–492. https://doi.org/10.1007/s41870-020-00586-7
5. Lu L, Ghoshal A, Renals S (2013) Cross-lingual subspace Gaussian mixture models for low-resource speech recognition. IEEE/ACM Trans Audio Speech Lang Process 22(1):17–27. https://doi.org/10.1109/TASL.2013.2281575
6. Basu J, Basu T, Khan S, Pal M, Roy R, Basu TK (2016) Experimental study of vowels in Nagamese, Ao and Lotha: Languages of Nagaland. Proc. of the 13th Intl. Conference on Natural Language Processing. pp 315–323, Varanasi, India. NLP Association of India (NLPAI)
7. Horo L, Sarmah P, Anderson GD (2020) Acoustic phonetic study of the Sora vowel system. J Acoust Soc Am 147(4):3000–3011. https://doi.org/10.1121/10.0001011
8. Chakraborty K, Horo L, Sarmah P (2018) Building an automatic speech recognition system in Sora language using data collected for acoustic phonetic studies. In: SLTU, pp 239–242. Gurugram, India. https://doi.org/10.21437/SLTU.2018-49
9. Tanwar A, Majumder P (2020) Translating morphologically rich indian languages under zero-resource conditions. ACM Trans Asian Low Resour Lang Inf Process 19(6):1–15. https://doi.org/10.1145/3407912
10. Tzudir M, Sarmah P, Prasanna SM (2021) Analysis and modeling of dialect information in Ao, a low resource language. J Acoust Soc Am 149(5):2976–2987. https://doi.org/10.1121/10.0004822
11. Basu J, Hrangkhawl TR, Basu TK, Majumder S (2021) Identification of two tribal languages of India: An experimental study. In: Dev A, Sharma A, Agrawal SS (eds) Artificial Intelligence and Speech Technology, CRC Press, pp 221–229. https://doi.org/10.1201/9781003150664-25
12. Kumar R, Singh S, Ratan S, Raj M, Sinha S, Seshadri V, Bali K, Ojha AK (2022) Annotated speech corpus for low resource Indian languages: Awadhi, Bhojpuri, Braj and Magahi. arXiv preprint arXiv:2206.12931
13. Zhao J, Zhang WQ (2022) Improving automatic speech recognition performance for low-resource languages with self-supervised models. IEEE J Sel Top Signal Process 16(6):1227–1241. https://doi.org/10.1109/JSTSP.2022.3184480
14. Liu DR, Hsu PC, Chen YC, Huang SF, Chuang SP, Wu DY, Lee HY (2021) Learning phone recognition from unpaired audio and phone sequences based on generative adversarial network. IEEE/ACM Trans Audio Speech Lang Process 30: 230–243. arXiv.2207.14568
15. Akita Y, Kawahara T (2005) Generalized statistical modeling of pronunciation variations using variable-length phone context. In: IEEE international conference on acoustics, speech, and signal processing. IEEE, pp I-689. https://doi.org/10.1109/ICASSP.2005.1415207
16. Cao W, Liu Y, Zheng TF (2008) Local mismatch phone for confidence measure in standard and accented Chinese speech recognition. In: 6th international symposium on Chinese spoken language processing. IEEE, pp 1–4. https://doi.org/10.1109/CHINSL.2008.ECP.64
17. Lo WK, Harrison AM, Meng H (2010) Statistical phone duration modeling to filter for intact utterances in a computer-assisted pronunciation training system. In: IEEE international conference on acoustics, speech and signal processing. IEEE, pp 5238–5241. https://doi.org/10.1109/ICASSP.2010.5494988
18. Livescu K (2005) Feature-based pronunciation modeling for automatic speech recognition. Doctoral dissertation, Massachusetts Institute of Technology, Cambridge
19. Stănescu M, Cucu H, Buzo A, Burileanu C (2012) ASR for low-resourced languages: building a phonetically balanced Romanian speech corpus. In: Proceedings of the 20th European signal processing conference. IEEE, pp 2060–2064
20. Stolcke A, Ryant N, Mitra V, Yuan J, Wang W, Liberman M (2014) Highly accurate phonetic segmentation using boundary correction models and system fusion. In: IEEE international conference on acoustics, speech and signal processing. IEEE, pp 5552–5556. https://doi.org/10.1109/ICASSP.2014.6854665
21. Sarma BD, Prasanna SM (2018) Acoustic–phonetic analysis for speech recognition: a review. IETE Tech Rev 35(3):305–327. https://doi.org/10.1080/02564602.2017.1293570
22. Phuong PN, Do QT, Mai LC (2019) A high quality and phonetic balanced speech corpus for Vietnamese. arXiv:1904.05569
23. Lalrempuii C (2005) Morphology of the Adi language of Arunachal Pradesh. Doctoral dissertation, NEHU, Shillong
24. Sasmal S, Saring Y (2020) Spectral analysis of consonants in Arunachali Native language-Adi. In: Mallick PK, Meher P, Majumder A, Das SK (eds) Electronic Systems and Intelligent Computing Springer, Singapore, pp 783–790. https://doi.org/10.1007/978-981-15-7031-5_74
25. 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
26. Office of the Registrar General, India (2018) Language-India, States and Union Territories. https://censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf. Accessed 14 Jan 2022