

# Spoken Language Technology for North-East Indian Languages

Viyazonuo Terhija<sup>1</sup>, Priyankoo Sarmah<sup>1,2</sup>, Samudra Vijaya<sup>2</sup>

<sup>1</sup>Department of Humanities and Social Sciences

<sup>2</sup>Centre for Linguistic Science and Technology

Indian Institute of Technology Guwahati

Guwahati 781039, Assam, INDIA

{viyazonuo, priyankoo, samudravijaya}@iitg.ac.in

## Abstract

The North-East India hosts a myriad of languages that belong to three different language families and that have salient phonological inventories. Officially, of the 99 non-scheduled languages, 60 are spoken in North-East India. Among these 60, 34 languages have more than 50,000 speakers. However, these languages do not have enough linguistic resources for language technology development. Many of them do not even have detailed linguistic descriptions that would be helpful in understanding the challenges that lay ahead in building language technology based tools in the languages. In this paper, we provide an overview of the languages of North-East India and outline a few phonetic features distinct to some of the languages. We also provide an overview of the attempts to build spoken language technology for some of these languages and finally we conclude by outlining the challenges in building speech technology in these languages and suggest some approaches to overcome the challenges.

**Keywords:** North-East India, Speech technology, Languages, Phonetic features

## Cayie

North-East India nu die kekreikecii kekra baya. Die siiko die kikru se nu ba mu siikoe puo pfhephra kemeyie kekra baya. Diezho nu ba kemo die hiethepfiiethepfii donu die hiesorou-e North-East India nu puya mu siiko donu die seredia-e mia nyie hiepengou mese pu tuoya. Derei die hako chii kehielie kevi cha ba mo. Die hako puo dze puo se puo nyi di u khruohipie u mhodziitsatie u chiekezhiiko die chii kehielie kevi sikelie cha tuo mo. Leshii hau nu, North East India nu die pete donu rei die huo se par puo kepu pfhephra kekreikecii se kezashii. Siisie die-e kikemhie di chii kehie morosuo shi ikecii rei kezashii. Thelanu, die u chiekezhiu mu kikemhie di u chiekezhiu die siiko mho kuolietuo shi ikecii rei kezashii.

## 1. Introduction

The North-East India is part of India constituted of eight provinces with a population of 45 million. In spite of representing only 4% of the population of India, this area is linguistically diverse area with three major language families represented by native speakers of over 200 languages. While Indo-European and Austro-Asiatic languages are spoken in the area, there is a large number of Tibeto-Burman languages spoken in this area. There are only two major Indo-European languages, spoken as native languages in this area, namely Assamese and Bengali. However, the total number of speakers of these two languages is 27 million. In other words, the remaining 18 million speakers share the remaining 200 odd languages in the region. This makes the linguistic situation of the area extremely complicated, as several languages in the area are considered minority languages, and as a result of that there are not much linguistic resources available in these languages. The lack of such resources stand out as the first block in technology development in the languages of North-East India. Additionally, as the languages with smaller number of speakers belong to Tibeto-Burman or Austro-Asiatic languages, the linguistic features are quite distinct from other Indian languages. Hence, the approaches in developing speech technology in major Indian languages, such as Hindi or Bengali, may not be entirely applicable in the North-East Indian languages.

This paper is organized as follows. Section 2. provides an overview of the languages of North-East India. Section 3. provides examples of some of the phonetic fea-

tures of North-East Indian languages that require special attention in building language technology, Section 4. provides a summary of the language technology developments in the North-East Indian languages and finally Section 5. discusses the way forward and concludes the paper.

## 2. Languages of North-East India

The eighth schedule of the constitution of India lists 22 languages as official languages, also called ‘scheduled languages’. While English is not one of them, it is considered as a subsidiary official language (CoI, 1950). The constitution, however, does not prevent the states or provinces in India from choosing another languages, apart from the ‘scheduled languages’, as official language of the state. Apart from that the census of India has considered 99 languages as ‘non-scheduled languages’ that have more than 10,000 speakers each.

Table 1 shows the number of scheduled and non-scheduled languages spoken in North-East India, according to the census of India. The rightmost column lists the number of languages, subsumed under the scheduled and non-scheduled categories, that have more than 10,000 speakers. Of the total 90 mother tongues, 6 are Indo-European, 4 are Austrosiatic and 80 are Tibeto-Burman. As seen in the Table, the North-East Indian region has several Tibeto-Burman languages (Van Driem, 2018). Figure 1, shows the distribution of Tibeto-Burman languages in the world with North-East India hosting a sizeable number of them. The official languages of the eight states of North-East India are provided in Table 2 (Nag, 1963; Meg, 2005; Sik, 1977; Man, 1979;

Tri, 1964; Sik, 1977; Ass, 1960).

Table 1: Distribution of languages across North-East India

Languages	Number	Incl. mother tongues
Scheduled	4	10
Non-scheduled	60	80



Figure 1: Geographical distribution of the major Tibeto-Burman languages, argued to be Trans-Himalayan languages, (Van Driem, 2018). Each dot represents the putative historical geographical centre of each of 41 major linguistic subgroups. Source: (Van Driem, 2018)

Table 2: Official state languages across North-East India

States	Official languages
Arunachal Pradesh	English
Assam	Assamese, Bengali, Bodo and English
Meghalaya	English, Khasi and Garo
Manipur	Meiteilon, English
Nagaland	English
Tripura	Bengali, English and Kokborok
Mizoram	Mizo and English
Sikkim	Nepali, Bhutia, and English

### 3. Features of North-East Indian Languages

As seen in Table 1, the majority of languages spoken in the region are Tibeto-Burman languages. The Tibeto-Burman languages have their own distinct phonological features, some of which will be discussed in the sections below. Many of these features are not commonly found and hence, they emerge as linguistic challenges to deal with in speech technology development.

#### 3.1. Lexical tones in North-East India

Tibeto-Burman languages are known to be tonal. There are only a few Tibeto-Burman languages that do not have lexical tones. As the majority of languages spoken in North-East India belong to the Tibeto-Burman family, almost all

of them, barring a few, have lexical tones. Lexical tones are generally classified into two major groups namely, register tones and the contour tones (Yip, 2002). The range of tonal contrast varies in North-East India ranging from atonal to five or more and the inventory includes both register and contour tones. While, Bodo is reported to be a two-tone system (Sarmah, 2004), Mizo has four lexical tones in its inventory (Sarmah and Wiltshire, 2010). Acoustic studies have been conducted on Ao, Angami, Bodo, Dimasa, Mizo, Paite, Poula, Rabha, Tiwa etc. and the types and acoustic properties of lexical tones in these languages are fairly well understood. However, this still leaves out quite a number of tone languages in the region of which not much is known. In case of language technology development for tone languages, incorporation of tone information is of utmost importance as it improves recognition by disambiguating words. Several works have shown the advantage of incorporating tonal information in the development of Automatic Speech Recognition (ASR) systems in tone languages (Hu et al., 2014; Metze et al., 2013). Moreover, tone modeling needs to be exhaustive as it is also noticed that tones and segments interact with each other in a predictable manner (Lahminghlu et al., 2019; Coupe, 1998; Sarmah, 2009).

#### 3.2. Voiceless nasals

While nasals are known to be phonemically voiced, several Tibeto-Burman languages, such as Burmese, are known to phonemically contrast between voiced and voiceless nasals. Tibeto-Burman languages spoken in North-East India, namely, Mizo and Angami, also show evidence for voicing distinction in nasals. Mizo voiceless nasals are primarily voiceless with a bit of voicing towards the end of the nasal segment. On the other hand, Angami voiceless nasals are entirely voiceless with aspiration at the end (Bhaskararao and Ladefoged, 1991). Hence, Angami contrasts between /m/ and /m<sup>h</sup>/, /n/ and /n<sup>h</sup>/, /ŋ/ and /ŋ<sup>h</sup>/. In Mizo, the contrastive nasals are: /m/ and /m̥/, /n/ and /n̥/, /ŋ/ and /ŋ̥/. Such phonetically similar segments may pose challenges in speech technology development.

#### 3.3. Fricative Aspiration

Another recently reported phenomenon in Tibeto-Burman languages in North-East India is the aspiration of the fricative sounds. Bodo and Rabha have reported the existence of aspiration associated with voiceless, alveolar fricatives, /s/ (Sarmah and Mazumdar, 2015; Rabha et al., 2019). It is reported that phone recognition in Rabha is better when aspiration in fricatives is taken into account using acoustic features such as strength of excitation (SoE) and variance of successive epoch intervals (VSEI) (Rabha et al., 2019).

### 4. Spoken Language Technologies

A brief summary of the spoken language technologies developed for the languages of North-East India is presented in this section. The 3 ‘scheduled’ languages of North-East India (Assamese, Bodo, Manipuri) can be considered as under-resourced languages. Most of the speech technologies were developed for these languages, thanks to the support from the Government of India. Most other languages

Table 3: Speech technologies developed for the languages of North-East India

Language	ISO 639-3 code	No. of speakers	Speech Systems
Angami	njm	152,796	ASR
Ao	njo	260,008	DID
Assamese	asm	15,311,351	ASR,TTS, PE
Bodo	brx	482,929	ASR,TTS
Khasi	kha	1,431,344	DID
Manipuri	mni	1,761,079	ASR,TTS, KWS
Mizo	lus	830,846	ASR,PE
Sora	srb	5,900	ASR

of North-East India can be considered as zero-resource languages. The spoken language technologies developed for the languages of North East India are listed in Table 3. The ISO 639-3 code as well as the number of persons speaking the language (cen, 2011) are also given in the Table. A brief account of various speech systems implemented for these languages is given below.

#### 4.1. Angami

Angami (also Tenyidie) language is spoken in the state of Nagaland. A preliminary Automatic Speech Recognition (ASR) system was developed using speech data of sentences read by 11 native speakers (Terhijja et al., 2019). The Word Error Rate (WER) of the ASR system on the training data was under 5%. In a ‘leave one speaker out’ cross validation experiment, the average WER of the ASR system using context independent phone Hidden Markov Model (HMM) was 17.3%.

#### 4.2. Ao

Ao is spoken in Nagaland. A Gaussian Mixture Model (GMM) based Dialect Identification (DID) system was implemented to identify two Ao dialects, namely, Changki and Mongsen (Tzudir et al., 2018). Augmentation of spectral features with tonal features resulted in better DID accuracy.

#### 4.3. Assamese

Several speech systems were implemented for Assamese, a ‘scheduled’ language, spoken in the state of Assam. An Assamese ASR system was implemented using speech data from 209 speakers. The ASR system that employed a Deep Neural Network (DNN) along with HMM yielded with the lowest WER of 12.4% (Deka et al., 2019b). A Phonetic Engine (PE) was implemented with a phone recognition accuracy of 47.31%, 45.30%, 36.13% in reading, lecture and conversation modes respectively (Sarma et al., 2013). A text-to-speech (TTS) system using a DNN was developed. The quality of the synthesized speech was distinctively better than that of the speech synthesized by the GMM-HMM based TTS system (Deka et al., 2019a).

#### 4.4. Bodo

Bodo, a ‘scheduled’ language, is spoken in the state of Assam. A GMM-HMM based ASR system was implemented in 2014 (Laba Kr. Thakuria, 2014). A TTS system for Bodo was built using the concatenative approach (IITG TTS group, 2013).

#### 4.5. Khasi

Khasi is a language spoken in the state of Meghalaya. A GMM based Dialect Identification system was implemented that recognises the dialect of the input Khasi speech as one of the two dialects: Khyntiem or Bhoi-Jirang with an accuracy of 97% (Arjunasor Syiem, 2016).

#### 4.6. Manipuri

Manipuri (also Meiteilon) is spoken in the state of Manipur. Development of speech technology for Manipuri language in form of ASR and Keyword Search (KWS) system is reported. The authors collected and transcribed telephonic read speech data of A speech database of over 90 hours telephonic speech from more than 300 speakers was created for implementation of an KeyWord Spotting (KWS) system as well an ASR system (Patel et al., 2018b). The WER of the DNN-HMM based ASR system was 13.5%. The equal error rate of the KWS system was 7.64% (Patel et al., 2018a). A HMM based TTS system for Manipuri language was implemented (IITG speech group, 2013). A toolkit to build TTS enabled one to build TTS system in many languages including Manipuri (Ghone et al., 2017).

#### 4.7. Mizo

Mizo is spoken in Mizoram state. Phonetic Engine (Dey et al., 2017) as well as ASR systems (Kothapalli et al., ), (Dey et al., 2018) were implemented for Mizo language. The phone recognition rate of the PE Mizo phonetic engine was 13.9% when DNN-HMM acoustic model was used in conjunction with language model (Dey et al., 2017). The WER of DNN-HMM based ASR system for clean speech was 13% (Dey et al., 2018).

#### 4.8. Sora

Sora is mainly spoken in the states of Orissa and Andhra Pradesh. In the state of Assam, there are 5,900 Sora speakers whose ancestors migrated to Assam in the 19th century. A DNN-HMM based ASR system was implemented, and had a WER of 13.9% (Chakraborty et al., 2018).

## 5. Conclusion

This paper presented salient features of languages spoken in North-East India, and gave an account of speech systems developed for some of these languages. The zero-resource status of the most of these languages is a major barrier in enabling people of all strata to reap the benefit of language technology. A good study of the linguistic properties of these languages would set a foundation for building spoken language systems for these languages via transfer of knowledge from related languages.

## References

- Arjunasor Syiem, Gaurab Krishnan Deka, T. I. L. J. S. (2016). Khasi dialects identification based on gaussian mixture model. *Int. J. Engg. Sci. Computing*, 6(4):3882–3885.
- (1960). *The Assam Official Language Act, 1960*.
- Bhaskararao, P. and Ladefoged, P. (1991). Two types of voiceless nasals. *Journal of the International Phonetic Association*, 21(2):80–88.
- (2011). *Office of the Registrar General & Census Commissioner India, 'Statement-1 Part-B Languages not specified in the eighth schedule (non-scheduled languages)*.
- Chakraborty, K., Horo, L., and Sarmah, P. (2018). Building an automatic speech recognition system in sora language using data collected for acoustic phonetic studies. In *SLTU*, pages 239–242.
- (1950). The Constitution of India: Part XVI: Official language.
- Coupe, A. R. (1998). The acoustic and perceptual features of tone in the tibeto-burman language ao naga. In *Fifth International Conference on Spoken Language Processing*.
- Deka, A., Sarmah, P., Samudravijaya, K., and Prasanna, S. (2019a). Development of assamese text-to-speech system using deep neural network. In *2019 National Conference on Communications (NCC)*, pages 1–5. IEEE.
- Deka, B., Sarmah, P., and Vijaya, S. (2019b). Assamese database and speech recognition. *22nd Oriental-COCOSDA, Cebu, Philippines*.
- Dey, A., Lalhminghlu, W., Sarmah, P., Samudravijaya, K., Prasanna, S. M., Sinha, R., and Nirrnala, S. (2017). Mizo phone recognition system. In *2017 14th IEEE India Council International Conference (INDICON)*, pages 1–5. IEEE.
- Dey, A., Sarma, B. D., Lalhminghlu, W., Ngente, L., Gogoi, P., Sarmah, P., Prasanna, S. R. M., Sinha, R., and S.R., N. (2018). Robust mizo continuous speech recognition. In *Proc. Interspeech 2018*, pages 1036–1040.
- Ghone, A., Nerpagar, R., Kumar, P., Baby, A., Shanmugam, A., Mukundan, S., and Murthy, H. (2017). Tbt(toolkit to build tts): A high performance framework to build multiple language hts voice. 08.
- Hu, W., Qian, Y., and Soong, F. K. (2014). A dnn-based acoustic modeling of tonal language and its application to mandarin pronunciation training. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3206–3210, May.
- IITG speech group, (2013). *IIT Guwahati Text-to-Speech Synthesis for Manipuri Language*.
- IITG TTS group. (2013). Text-to-speech synthesis for bodo language. <http://www.iitg.ac.in/cseweb/tts/tts/Bodo/>.
- Kothapalli, V., Sarma, B. D., Dey, A., Gogoi, P., Lalhminghlu, W., Sarmah, P., Prasanna, S. M., Nirmala, S., and Sinha, R. ). Robust recognition of tone specified mizo digits using cnn-lstm and nonlinear spectral resolution.
- Laba Kr. Thakuria, Purnendu Acharjee, A. D. P. T. (2014). Bodo speech recognition based on hidden markov model toolkit(htk). *Int. J. Sci. Engg. Res.*, 5(1):339–343.
- Lalhminghlu, W., Terhijja, V., and Sarmah, P. (2019). Vowel-tone interaction in two tibeto-burman languages. *Proc. Interspeech 2019*, pages 3970–3974.
- (1979). *The Manipuri Official Language Act, 1979*.
- (2005). *The Meghalaya Official Language Act, 2005*.
- Metze, F., Sheikh, Z. A., Waibel, A., Gehring, J., Kilgour, K., Nguyen, Q. B., and Nguyen, V. H. (2013). Models of tone for tonal and non-tonal languages. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 261–266. IEEE.
- (1963). *The Nagaland Official Language Act, 1963*.
- Patel, T., Krishna, D., Fathima, N., Shah, N., Mahima, C., Kumar, D., and Iyengar, A. (2018a). An automatic speech transcription system for manipuri language. In *Interspeech*, pages 2388–2389.
- Patel, T., Krishna, D., Fathima, N., Shah, N., Mahima, C., Kumar, D., and Iyengar, A. (2018b). Development of large vocabulary speech recognition system with keyword search for manipuri. In *Interspeech*, pages 1031–1035.
- Rabha, S., Sarmah, P., and Prasanna, S. M. (2019). Aspiration in fricative and nasal consonants: Properties and detection. *The Journal of the Acoustical Society of America*, 146(1):614–625.
- Sarma, B. D., Sarma, M., Sarma, M., and Prasanna, S. M. (2013). Development of assamese phonetic engine: Some issues. In *2013 Annual IEEE India Conference (INDICON)*, pages 1–6. IEEE.
- Sarmah, P. and Mazumdar, P. (2015). Aspiration in alveolar fricatives in bodo. In *ICPhS*.
- Sarmah, P. and Wiltshire, C. R. (2010). A preliminary acoustic study of mizo vowels and tones. *J. Acoust. Soc. Ind*, 37(3):121–129.
- Sarmah, P. (2004). *Some aspects of the tonal phonology of Bodo*. Ph.D. thesis, English and Foreign Languages University, Hyderabad.
- Sarmah, P. (2009). *Tone systems of Dimasa and Rabha: a phonetic and phonological study*. University of Florida.
- (1977). *The Sikkim Official Language Act, 1977*.
- Terhijja, V., Sarmah, P., and Vijaya, S. (2019). Development of speech corpus and automatic speech recognition of angami. *22nd Oriental-COCOSDA, Cebu, Philippines*.
- (1964). *The Tripura Official Language Act, 1964*.
- Tzudir, M., Sarmah, P., and Prasanna, S. M. (2018). Dialect identification using tonal and spectral features in two dialects of ao. In *SLTU*, pages 137–141.
- Van Driem, G. (2018). The East Asian linguistic phylum: A reconstruction based on language and genes. *The Asiatic Society*, 60(4):1.
- Yip, M. (2002). *Tone*. Cambridge University Press.