Exploring some general issues in Digital representation of Low-Resource Languages from North Eastern India

Sushanta Kabir Dutta1

North Eastern Hill University, India

**Abstract.** Low-resource languages can be defined as having a very minimal amount of dataset, less computerized, not having sufficient of linguistic information about them etc. They have relatively less amount of data available for training conventional artificially intelligent (AI) systems. This leads to difficulties in their digital representation and consequently, non-availability for speech and language technologies. In this paper, we try to find out some general issues attached to such representations and suggest language technologies to low-resource language communities, particularly belonging to North Eastern India. In that, we focus our attention to various success and failure stories of some past works in the area and relevant challenges faced by researchers. In addition, we attempt to provide some general ideas on how one can tackle these challenges while working for the low-resource languages in North Eastern India.

**Keywords:** Low-Resource Languages, Language technologies, issues, challenges, North Eastern India

# Introduction

Low-resource (LR) languages may be loosely defined as less studied, having scarcity of resources, less computerized with low density amidst other languages. Thus, we may say that LR languages often lack a unique orthographic system, very rare representation over the internet, do not have a linguistic expertise and also suffer from lack of electronic resources for speech and language processing etc. Thus, these languages differ from the rest and form a separate group for studies and research.

## 1.1 Need for the representation in digital domain

Well-resourced languages are represented in digital domain due to availability of data-sets with pronunciation dictionaries and stable orthographies. But in LR languages it is very difficult to get these and also at times we don’t have any knowledge of them. However, it is needed to represent such LR languages in digital domain so as to avail the advantages of speech technology. This can also help the researchers to communicate with the people with these language groups.

One of the needs for developing digital representations is mainly to meet with the necessities of modern language users across all communities equally. It can help in contributing to language revitalization. Further, it can be useful for language learners, students with dyslexia, people unable to read, vision impaired individuals and many more. As many well-resourced languages are already there in the digital domain, so representing LR languages in this would be very useful. For example, one application of effective multilingual language technology implementation can be cited from Africa [9] ‘**Viamo**’, also known as VOTO Mobile, is being started in 2012 which implemented Interactive Voice Response systems (IVRS) in various African languages. This has been done to air public health information.

## Languages of North East India

India is a country with considerable linguistic diversity. Many a times it so hap- pens that very little knowledge is available for many languages [9] spoken across the country. As per the official language reports, in India languages are divided into several categories, such as scheduled, non-scheduled and mother tongues etc, which offers difficulties in uniform development in language technology. Governments usually provide funds for technology development particularly to the scheduled languages and thus non-scheduled languages and mother tongues are overlooked. The 8-th schedule of the constitution of India lists a total of 22 languages as the officiating languages which are also called ‘Scheduled Lan- gauges’. The constitution, however, does not prevent a state in India from choosing any other language apart for the scheduled ones as the official language of it. Apart from the census, India has considered 99 languages as ‘non-scheduled languages’ that have at least, more than 10,000 speakers each. The Northeastern part of India comprises the ‘seven sister’-states of Assam, Arunachal Pradesh, Meghalaya, Manipur, Mizoram, Nagaland and Tripura. The state of Sikkim has also been included recently into North Eastern (NE) India. According to the last census there are about 220 languages spoken across the NE states, which belong to mainly Indo-European, Sino-Tibetan, Kra-Dai, Austric and some other Creole language groups[9]. Indo-European is represented by Asomiya (naive Assamese langauge), Sino-Tibetan is represented by the Tibeto-Burman languages of Bodo, Karbi, Garo, Mising, Rabha, Dimasa, Kachari, Tiwa, Deuri etc and Tai repre- sented by a few dialects of Tai-Ahom, Tai-Phake, Tai-Khamyang, Tai-Turung, Tai-Aiton and Tai-Khamti. The sole representative of Austric family is Khasi language. It is one of the major languages spoken in the state of Meghalaya. Tibeto-Burman language Meitei is the official language of Manipur, where also spoken Tangkhul-Naga of the same group. Different Tibeto-Burman languages like Ao, Angami, Sema, Lotha, Konyak, Dzemietc are spoken in Nagaland. Tibeto- Burman kokborak, sometimes also called Tripuri and Bangla are the main languages of Tripura. Mizo and Hmar of the Tibeto-Burman group are the major languages spoken in Mizoram. In Arunachal Pradesh all the major languages belong to the Tibeto-Burman group namely, Hrusso, Tanee, Nisi, Adi, Abor, Nocte, Apatani, Misimi, Galong, etc. Nepali, an Indo-Aryan language, is the dominant one in Sikkim, besides the Sino-Tibetan languages Limbu, Bhutia, Lepcha, Rai, Tamang, Sherpa, etc

The table-1 below shows the official language which are spoken across NE India. The majority of languages spoken in the region are thus Tibeto-Burman

**Table 1.** Official languages of NE states

|  |  |  |
| --- | --- | --- |
| Sl No | Sate | Official language |
| 1 | Assam | Assamese, Bengali, Bodo |
| 2 | Arunachal Pradesh | English |
| 3 | Manipur | Meitelion |
| 4 | Meghalaya | Khasi, Garo, English |
| 6 | Mizoram | Mizo, English |
| 7 | Nagaland | English |
| 8 | Sikkim | Nepali, English |
| 9 | Tripura | Kokborak, Bengali, English |

languages. The Tibeto-Burman languages have their own phonological features which are completely different from the mainland Indian languages. Many of these features are not commonly found and hence, they emerge as linguistic challenges to deal with while developing speech and language technology. We explore a few of such features here.

## Some distinctive features of NE languages

The NE India is the part of India constituted of eight states having a population of which is around 45 million. This region is linguistically diverse and can be categorized into three major language families. Each family is represented by native speakers of more than 200 languages. Indo-European and Austro-Asiatic languages are mostly spoken in this area. There is a large number of Tibeto- Burman languages which are spoken across this region, although only two major Indio-European languages are used as native languages: Assamese and Bengali. However, the total number of speakers of these languages is around 27 million. The remaining 18 million speakers share the remaining 200 odd languages. This makes the linguistic situation of the area extremely complicated, and several languages of this region are considered minority languages. Thus, there is a huge scarcity of linguistic resources among these languages.

Tibeto-Burman languages are known to be mostly tonal in nature, except a few. As the majority of languages spoken in NE India belong to this family, most of them have lexical tones. Lexical tones are generally classified into two groups, viz, (i) Register tone and (ii) Contour tone [10]. The tonal contrast varies in NE Indian languages ranging from one tone to five tones (or sometimes even more!), which include both Register and Contour tones. While, Bodo language is reported to be a two-tone language [11], Mizo has four lexical tones in its inventory [12]. Acoustic studies have been conducted for several languages and dialects but there still remains a large number of tonal languages in the region with little knowledge available. For language technology development incorporation of tone information is of utmost necessity, because it affects the recognition rate (RR) accuracy by disambiguating the words. Several works have shown the advantages of incorporating this in the development of Automatic Speech Recognition (ASR) systems in tonal languages [13]. Moreover, tone modeling needs to be exhaustive since some studies showed that tones and segments interact with each other in a predictable manner [14]. Although nasal sounds are phonemically voiced, several Tibeto-Burman languages exhibit phonemic contrast between voiced and voiceless nasals. Tibeto-Burman languages spoken in NE India, viz, Mizo and Angami, also exhibit evidences for voicing distinctions in nasals [9]. Mizo language 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 [15]. Such phonetically similar segments may pose challenges in speech technology development. Languages in NE India also show the aspiration of the fricative sounds. Bodo and Rabha have reported the existence of aspiration associated with voiceless, alveolar fricatives [16]. It is reported that phone recognition in Rabha langauge is better when aspiration in fricatives is taken into account [17].

# Progress in Speech technology for LR language around the globe and within India

We report here some recent progress in development of speech technology over world-wide and then in Indian context, before highlighting the issues related to NE languages.

* A significant study on LR Language Identification (LID) is done by [1] in 2019. Here two approaches are there that deal with the language identification (LID) tasks. The first approach uses data augmentation by incorporating various distortions in the original dataset. The second approach relates to a multi-lingual bottleneck feature extraction (BNF) using on speech recognition systems for different languages. Experiments conducted by various research groups on both the i-vector and x-vector models demonstrated that both approaches are effective and can obtain promising results on in-domain data and out-of-domain data.
* In a separate study, in the year 2019, Kexin Feng et.al [2] worked on LR LID from Speech using ‘Transfer Learning’. Here, they explored transfer learning systems that employ various neural network (NN) architectures. Here, they created the large data-sets for language identification models using feed- forward neural networks. These are further fine-tuned on the LR data from a target domain to improve the system performance. They applied the pro- posed approach to the automatic identification of some African languages, which comprises a challenging task due to being LR languages themselves.

They conducted the experiments using two publicly available data-sets: the ‘VoxForge’ corpus which contains 7 Indo-European languages as source data, and the ‘Lwazi’ corpus which includes 11 African languages as target data.

* Recently, in the year 2021,Tharindu Ranasinghe and Marcos Zampieri

[3] had worked on Multilingual offensive Language Identification for LR Languages. Here, they used the available English data sets by applying cross-lingual contextual word embedding and transfer learning to make predictions in such languages. They projected predictions on comparable data in Arabic, Bengali, Danish, Greek, Hindi, Spanish, and Turkish. They reported competitive performance on Arabic, and Turkish using the training and development sets of ‘OffensEval 2020’ shared task.

* In the year 2021, yet another study by R. Bedyankin and N. Mikhaylovskiy

[4] showed that a convolutional neural network (CNN) with a Self-Attentive Pooling layer produced promising results in LR language identification task. They set up a system for the LR data ASR challenge. The confusion matrix for the LID system bears the language similarity measures as well.

With these recent advances across the world, we now focus our attention to the Indian context and report a few progress in speech technology developments for Indian languages

* In the year 2021, Kusampudi et.al [5] introduced two Telugu-English manually annotated data-sets (Twitter data-set and Blog data-set). The Twitter data-set has a lot of Romanization variability and also misspelled words as compared to the blog data-set. Authors then compared them across various classification models. Additionally, they performed extensive bench-marking using both Classical and Deep Learning Models for LID as compared to existing models. They proposed two models for language classification in the data:
  + Word Level Classification
  + Sentence Level word-by-word Classification
* In another study, S. Gaikwad et.al [6] worked on Cross-lingual Offensive Language Identification for LR Languages: the case of Marathi, where they introduced the Marathi Offensive Language Data-set (MOLD). MOLD is the first data-set of its kind compiled for Marathi, thus opening a new domain for research in LR Indo-Aryan category of languages.
* In the year 2021, D.N. Krishna [7] worked on Multilingual Speech Recognition for LR Indian Languages using Multi-Task transformer. He proposed a multi-task learning-based transformer model for LR multilingual speech recognition for Indian languages. His proposed model consists of a transformer encoder and two parallel transformer decoders. He used a phoneme decoder (PHN-DEC) for the phoneme recognition task and a grapheme decoder (GRP-DEC) to predict grapheme sequences. He considered the phoneme recognition task as an auxiliary task for our multi-task learning framework.

He jointly optimized the network for both phoneme and grapheme recognition tasks using Joint CTC-Attention training. He used a conditional de- coding scheme to inject the language information into the model before predicting the grapheme sequence. His experiments showed that his proposed approach could obtain significant improvement over previous approaches.

* In the year 2021, Joyanta Basu et.al [8] worked on a Multilingual Speech Corpus in LR Eastern and NE Indian Languages for Speaker and Language Identification. Here, they illustrated the creation process of such an LRL corpus comprising of sixteen rarely studied Eastern and Northeastern (E & NE) Indian languages and presented the data variability with different statistics. Furthermore, several experiments were carried out using the collected LR corpus to build baseline speaker identification (SID) and LID systems for acceptance evaluation. Vector quantization (VQ), Gaussian mixture models (GMMs), support vector machine (SVM), and multilayer perception (MLP)-based models were developed to represent the speaker and language-specific information captured through the spectral features. Apart from this, i-vectors, time delay neural networks (TDNN), and recurrent neural network with long short-term memory (LSTM-RNN) method-based SID and LID models were being experimented with to comply with the recent approaches. Performances of the developed systems were analyzed with LRL corpus in terms of SID and LID accuracy

# Issues related to low resource languages in North East (NE) India

With the recent progress in development of speech and language technologies a few concerns may be highlighted here for the LR languages from NE India. The major issue in NE India languages is that many languages are now endangered. They do not have proper

orthography, phonological structure, dictionary etc. So with the limited amount of knowledge available, it is very difficult to build the speech technology. There are many more challenges one gets prompted to face while dealing with LR languages in the region. These may be categorized as under-

Ethical Challenges: Often faced while collecting the data. One cannot be forced provide data. So, it is needed to know how to collect datasets properly.

1. Recording of sensitive information is often a difficulty.
2. One cannot loose any data due to mismanagement or handling.
3. Amount of Compensation which may vary for collecting data.

. Procedural Challenges:

1. One has to take care for collecting the data.
2. Finding right person/resource is very important.
3. Selection of the right variety of data.
4. Finding suitable environments of data Recording.

Technical Challenges:

1. Manual annotation of data sometimes requires native level knowledge of languages.
2. Spectrogram Analysis often requires good quality hardware/software
3. Deep learning based methods often require costly hardware

Acoustic-phonetic challenges:

1. Estimating exact phoneme Boundaries
2. Classifying articulatory features
3. Identifying phonemes from spectrogram
4. Requires mastery of softwares and sound understanding
5. With a large data-set, manual annotation is often time consuming

Theoretical Challenges:

1. Minimal pairs: a common test to decide whether two phones represent different phonemes or are allophones of the same phoneme. However, absence of minimal pairs for a given pair of phones does not always mean that they belong to the same phoneme.
2. Minimal pairs are not always easy to find.
3. Near-minimal pair: A pair of words differing by a few (but more than one) phonetic segments.
4. Virtually impossible to find a minimal pair to distinguish among English sounds
5. Often sounds are affected by the environments in which they occur.
6. Vowels are longer before the voiceless consonants.
7. Words are in isolation, in framed contexts, in spontaneous speech.

## Possible ways to overcome the challenges

With these challenges in mind, we outline the most relevant techniques that may be adopted for effective results with LR languages from NE India.

* + - Transfer Learning: Transfer learning is a way of solving new tasks by lever- aging prior knowledge in combination with new information. It is a common phenomenon observed in humans. For example, a random athlete is much more likely to beat a random individual with no athletic background in a physical sport new to both. More importantly, the athlete will likely take fewer resources (time) to learn the new sport.

So here, we base the foundations of our models on language models that are like general athletes who can adapt to a new sport even in low-resource setting. Base language models themselves do not require “annotated” data

and learn generic language capabilities by self-learning in an unsupervised fashion. Nonetheless, they are not very useful for specific tasks like classifying user intents off-the-shelf. So, we must fine-tune these base language models to accurately solve user-specific tasks that normally have very small amounts of annotated data. Consequently, our models learn to solve the tasks accurately despite low resources for annotated data, via transfer learning.

* + - Multilingual Learning: Multilingual Learning is a technique where a single model is trained on multiple languages. The assumption is that the model learns representations that are very similar for similar words and sentences of different languages. Thus, this can also assist cross-lingual transfer learning as knowledge from data for high-resourced languages like Hindi can transfer to the model’s representations for LR languages like Bodo. This way, base models can perform better on LR languages despite lack of enough text corpora.
    - Data Augmentation: Data augmentation is a data pre-processing strategy that automatically creates new data without collecting it explicitly. For in- stance, in a sentiment classification task, “Today is wonderful” can be altered to “today is a great day”. This alteration increases and possibly diversifies training data in an automatic way. Importantly, augmentation should be such that the ground-truth of any new instance does not change, in this case, “positive sentiment”. Unlike other data collection strategies, data augmentation is very cheap, fast, and usually does not require human involvement.

# Conclusion

So in this paper, we have tried to explore some issues related to LR LID tasks for NE-India and provide some general solutions to it. There are some practical difficulties in actual implementation because collection of raw data from all LR languages of NE would require enough efforts. Here, we proposed some general idea about what is an LR language and how to identify one. Further, we discuss how such languages are different from Main land Indian languages. Besides, we talked about what general challenges are faced by researchers while working on such LR languages and finally some broad directions to overcome such problems. Thus the extension of this work may be considered for digital representation of at least a few LR languages chosen from NE India and use technology that are possible.

# References

1. Qi, Z., Ma, Y., & Gu, M. (2019, November). A Study on Low-resource Language Identification. In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 1897-1902). IEEE.
2. Feng, K., & Chaspari, T. (2019, October). Low-resource language identification from speech using transfer learning. In 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1-6). IEEE.
3. Ranasinghe, T., & Zampieri, M. (2021). Multilingual offensive language identifi- cation for low-resource languages. Transactions on Asian and Low-Resource Lan- guage Information Processing, 21(1), 1-13.
4. Bedyakin, R., & Mikhaylovskiy, N. (2021). Low-resource spoken language identi- fication using self-attentive pooling and deep 1D time-channel separable convolu- tions. arXiv preprint arXiv:2106.00052.
5. Kusampudi, S. S. V., Chaluvadi, A., & Mamidi, R. (2021, September). Corpus Cre- ation and Language Identification in Low-Resource Code-Mixed Telugu-English Text. In Proceedings of the International Conference on Recent Advances in Nat- ural Language Processing (RANLP 2021) (pp. 744-752).
6. Gaikwad, S., Ranasinghe, T., Zampieri, M., & Homan, C. M. (2021). Cross-lingual offensive language identification for low resource languages: The case of Marathi. arXiv preprint arXiv:2109.03552.
7. Krishna, D. N. (2021). Multilingual Speech Recognition for Low-Resource Indian Languages using Multi-Task conformer.
8. Basu, J., Khan, S., Roy, R., Basu, T. K., & Majumder, S. (2021). Multilin- gual speech corpus in low-resource eastern and northeastern Indian languages for speaker and language identification. Circuits, Systems, and Signal Processing, 40(10), 4986-5013.
9. Hyslop, G., Konnerth, L., Morey, S., & Sarmah, P. (2014). North East Indian Linguistics 6. Asia-Pacific Linguistics, School of Culture, History and Language, College of Asia and the Pacific, The Australian National University.
10. Yip, M. (2002). Tone. Cambridge University Press.
11. Sarmah, P. (2004). Some aspects of the tonal phonology of Bodo (Doctoral disser- tation, English and Foreign Languages University, Hyderabad).
12. Sarmah, P., & Wiltshire, C. R. (2010). A preliminary acoustic study of Mizo vowels and tones. J. Acoust. Soc. Ind, 37(3), 121-129.
13. Hu, W., Qian, Y., & Soong, F. K. (2014, May). A DNN-based acoustic model- ing of tonal language and its application to Mandarin pronunciation training. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3206-3210). IEEE.
14. Lalhminghlui, W., Terhiija, V., & Sarmah, P. (2019). Vowel-Tone Interaction in Two Tibeto-Burman Languages. In INTERSPEECH (pp. 3970-3974).
15. Bhaskararao, P., & Ladefoged, P. (1991). Two types of voiceless nasals. Journal of the International Phonetic Association, 21(2), 80-88.
16. Sarmah, P., & Mazumdar, P. (2015). Aspiration in alveolar fricatives in Bodo.
17. Rabha, S., Sarmah, P., & 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.