2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)
18-20 August 2023, Dublin, Ireland

Collecting Speech Data for Endangered and Under-resourced Indian Languages

Ritesh Kumar[1] [2], *Meiraba Takhellambam*[3], *Bornini Lahiri*[4], *Amalesh Gope*[5], *Shyam Ratan*[1], *Neerav Mathur*[1], *Siddharth Singh*[1]

[1]UnReaL-TecE LLP, India; [2]Council for Strategic and Defense Research, India; [3]Manipur University, India; [4]Indian Institute of Technology-Kharagpur, India; [5]Tezpur University, India

riteshkr.kmi@gmail.com, unreal.tece@gmail.com

Abstract

The preparation of speech corpora for languages un(der)represented on the web largely depends on the manual methods of data collection and processing from different sources. The methods used in field linguistics and documentary linguistics for collecting data from the speech communities provide a valuable set of resources and methodologies for such data collection but these methods were not developed and optimised for large-scale data collection. However, this limitation could be overcome by combining linguistic field methods with crowdsourcing for data collection. In this paper, we discuss two such ongoing projects - SpeeD-TB and SpeeD-IA - in which we are experimenting with different methods and developing software and other infrastructure to rapidly collect speech data in six Tibeto-Burman - Toto, Chokri, Nyishi, Kok Borok, Bodo and Meitei - and four Indo-Aryan - Awadhi, Bhojpuri, Braj and Magahi - languages in India. Till now we have collected over 40 hours of speech data in these languages and over the period of the next year, we plan to collect a total of approximately 1,200 hours of speech data.

Index Terms: Speech data, Tibeto-Burman, Indo-Aryan, Indian languages, Field methods, Documentary linguistics, Endangered languages, Minoritised languages

1. Introduction

Over the last decade or so, research in speech technologies has seen a rapid and successful shift towards exclusively data-driven techniques such as machine learning and deep learning methods. Over the years, experiments with well-resourced languages such as English have demonstrated the success of these systems given sufficient data for training the systems. However, barring a handful of languages, this technological revolution has escaped most of the languages (including the officially supported, scheduled languages) spoken in India. This could be gauged from the commercial support for very few Indian languages across different speech-based products - Amazon Alexa supports Hindi among seven other international languages; Google Home supports 13 languages, including Hindi, as the only Indian language; Microsoft supports Indian English, Hindi, Tamil, Telugu, Gujarati, and Marathi for its ASR systems - there is no support whatsoever for most of the other Indian languages, especially languages belonging to the Tibeto-Burman and Austro-Asiatic language families.

One of the primary reasons behind this could be the non-availability of sufficient speech datasets for most Indian languages. This is even more so for the non-scheduled Indo-Aryan and Dravidian languages and even the scheduled languages from the Tibeto-Burman and Austro-Asiatic language families, largely spoken in Eastern and North-Eastern parts of India. As per our survey of the resources and corpora available for building speech technologies in Indian languages, the publicly accessible resources available for the three scheduled languages from these language families are listed below -

- Approximately 177 hours of speech data collected from 456 speakers are available in Bodo - this dataset is provided by the LDC-IL Speech Corpus.
- Slightly over 156 hours of speech data collected from 620 speakers is available in Meetei through the LDC-IL Speech Corpus.
- As far as we are aware, no publicly available dataset is available for Santhali. [1] mention IIITH-ILSC Speech Database which contains 4.5 hours of speech data collected from 50 speakers but we could not find a way to access the dataset.

Most of the other major and state official languages do not have even these minimal resources.

In order to alleviate this situation, we have started the 'SpeeD-IL' (Speech Datasets and Models for Indian Languages) project for developing speech corpora and other resources and models for un(der)represented languages in India. The stated aims and objectives of the project are listed below -

- To build a transcribed speech dataset of approximately 1000 hours each in at least 10 un(der)represented languages across each of the four major language families of India - Tibeto-Burman, Austro-Asiatic, Dravidian and Indo-Aryan - and the other language families with fewer languages viz. Tai-Kadai and Great Andamanese. The transcriptions will be in the native script of the respective languages and IPA as well.
- To develop a phone set for each of the languages under study.
- To build baseline wav2vec 2.0 (or other state-of-the-art techniques) pre-trained models based on the data collected in the project for each language family under study and use that for developing a baseline ASR system for each of the languages.
- To build a language model for the languages under consideration.
- To make the dataset and pre-trained and fine-tuned models publicly available through appropriate platforms under CC-By-NC-SA 4.0 license (for the dataset) and AGPL v3 (for the model).

In the first phase, the main objective of the project is to build a speech dataset of at least 2,000 hours consisting of around 200 hours in 6 Tibeto-Burman languages - Toto, Chokri, Nyishi, Kok Borok, Bodo and Meitei - and 4 Indo-Aryan languages - Awadhi, Bhojpuri, Braj and Magahi. In the next stages/phases of the project, we plan to expand to more languages including those from the Austro-Asiatic and Dravidian language families.

In the following sections of the paper, we discuss the methodology that we have adopted for collecting the data and a

10.21437/SIGUL.2023-4

summary of the data collected till now in the project.

2. Related work

In the last decade or so, there have been continuous and consistent efforts at developing speech datasets in some of the major languages in both the Indo-Aryan and Tibeto-Burman language families. Some of the prominent efforts to build speech datasets for Indo-Aryan languages include a speech database of 500 spoken sentences by 50 speakers in Hindi [2]; a read corpus of Bangla prepared by recording the readings of a popular Bangla newspaper, Anandabazar, by 40 female and 70 male speakers [3]; a dataset in Bangla and Odia that consisted of a mix of five hours each of conversational speech and extempore and 10 hours of read speech [4]; IVR-based speech data for Bangla, representing different varieties of the language [5], and Marathi speech database consisting of varieties collected from 34 districts of Maharashtra [6]; a Hindi and Marathi corpus consisting of data collected in both noisy and quiet environments [7]; a corpus of Hindi, Bangla and Indian English [8] and another corpus of Assamese, Bangla and Nepali [9]; low-resource Indo-Aryan languages speech corpus, which is a speech dataset of approximatey 4-5 hours of speech recordings in four extremely low-resourced Indo-Aryan languages - Awadhi, Braj, Bhojpuri and Magahi is developed through field methods of linguistic data collection [10].

Besides these and other efforts in the development of speech databases for scheduled languages of India, there were some notable efforts at developing speech databases for non-scheduled languages as well, including Tibeto-Burman languages. Most notable of these include ALS-DB (Arunachali Language Speech Database), which is a multilingual and multi-channel read speech database of four languages from Arunachal Pradesh viz. Apatani, Adi, Galo and Nyishi, along with English and Hindi as secondary languages[11], collected from 100 females and 100 males of 20-50 years age group; a database of Mizo tones consisting of a total of 4,384 syllables were collected from 338 three-syllable/word sentences from five Mizo speakers [12]; IITKGP-MLILSC (Indian Institute of Technology Kharagpur - Multilingual Indian Language Speech Corpus) speech database consisting of data from 27 Indian languages including four languages - Arunachali, Manipuri, Mizo and Nagamese - of Tibeto-Burman language family [13]; low-Resource Eastern and Northeastern speech corpus, which is a speech corpus with recordings of 16 low-resource languages from Eastern and North-Eastern India including languages like Adi, Angami, Ao, Hrangkhawl, Khasi, Lotha, Mizo, Nagamese, Sumi and others [14].

3. Data collection

3.1. Languages

The first part of this project involves the collection of data in the following languages across the two language families -

1. Tibeto-burman languages:
   - **Meetei [mni] -** It is one of the scheduled languages, spoken across Manipur by around 1.8 million speakers (Census 2011). It uses both the Bangla and Mayek scripts for writing - in the current project, we will be using Mayek for transcription. Meetei is a tonal language and maintains two-way tonal contrasts.
   - **Bodo [brx] -** It is also one of the scheduled languages, spoken across Bodoland in Assam and neighbouring states by around 1.5 million people (Census 2011). The official script for Bodo is Devanagari (which we will use for transcription in the current project) but the Bangla script has been traditionally used as well.
   - **Kok borok [trp / xtr] -** Kok Borok falls within the Boro-Garo group of the Tibeto-Burman branch of languages, spoken by 84% of the tribal population of Tripura (over 1 million speakers approximately as per 2011 Census) and some of the tribal population of the Chittagong hill tracts of Bangladesh. This language exhibits two-way tonal contrasts. Bangla script will be used for transcribing the language in the project.
   - **Nyishi [njz] -** Nyishi is one of the largest languages of Arunachal Pradesh, spoken by approx. 300k speakers (Census 2011). It belongs to the Tani branch of the Tibeto-Burman language family. Phonologically, it exhibits a 3-way tonal contrast. We will use Roman script for transcription in the language.
   - **Chokri [nri] -** Chokri belongs to the Angami sub-group of Naga languages, spoken in Nagaland (primarily in the Phek district) and parts of Manipur by over 100k speakers (Census 2011). Chokri is a highly tonal language that exhibits 5-way tonal contrasts. This community has adopted the Roman writing system and added a few diacritics to mark the tonal contrasts, which will be used for transcription in the project.
   - **Toto [txo] -** Toto is affiliated to a Himalayan subgroup of the Tibeto-Burman Language family. It is spoken by around 1,500 speakers in an area of Totopara. It is situated in Madarihat, which is part of the Alipurduar Subdivision of the Jalpaiguri district in West Bengal. It has six small sub-divisons namely, Mondalgaon, Dumsigaon, Pujagaon, Subagaon, Ponchayetgaon, and Mitrangaon [15]. Toto uses the Bangla script (which will be used for transcription) and follows two-way tonal contrast.
2. Indo-aryan languages:
   - **Awadhi [awa] -** Awadhi is an East Central language of the Indo-Aryan language family and mainly spoken in the Awadh region (Kanpur, Unnao, Hardoi, Barabanki, Lucknow, Amethi, Sitapur, etc.) of Uttar Pradesh and some parts of Nepal by 3.9 million speakers (Census 2011).
   - **Bhojpuri [bho] -** Bhojpuri is a widely spoken language in the Hindi belt with a vast number of speakers. It is an Eastern Indo-Aryan language and is primarily spoken in western Bihar, parts of eastern Uttar Pradesh, and northwestern Jharkhand by 50 million people (Census 2011).
   - **Braj [bra] -** Braj is a western Indo-Aryan language, it is widely spoken in the states of Western Uttar Pradesh and some parts of Rajasthan and Madhya Pradesh. According to Census 2011, it is spoken by 1.6 million people in these states.
   - **Magahi [mag] -** It is a member of the Eastern Indo-Aryan language family, Magahi is primarily spoken in the eastern Indian states of Bihar and Jharkhand as well as in some areas of West Bengal and Odisha spoken by 13 million speakers (census 2011).

   All these four languages use Devanagari for writing and we will use it for transcription across all languages.

3.2. Preparation of questionnaires

Although a large number of speech data collection efforts in Indian languages have relied on read speech, because of their relatively poor generalisability to real-life speech recognition situations, we have relied minimally on such methods for data collection and are collecting the data mainly using the methods discussed below. These methods are commonly employed in Linguistic field methods (where diversity of structures is focussed but the 'quantity' of data collected is generally low) but have been rarely utilised for building large-scale datasets that may be used for building speech technologies. While preparing these questionnaires, our primary focus was on adapting the existing questionnaires for collecting large-scale datasets.

1. **Translation and stimuli method:** In the translation method, we provide sentences to the speakers in a contact language (such as Hindi and English) either as speech or as text and they will be required to translate and record those in their language. We have prepared over 1,500 such sentences across various domains that are being used for data collection. Some of the most common and standardised resources that we have used for preparing our questionnaire include the following -
   (a) Abbi's questionnaire for Indian languages [16].
   (b) Lahiri's questionnaire for Indian languages [17] and for eliciting data on case system [18]
   (c) Takhellambam's Questionnaire for Language Documentation of Tibeto-Burman languages (unpublished till now).
   (d) Max-Planck Institute's Typological tools for field linguistics [1], L&C Field Manuals and Stimulus Materials [2] and TulQuest Archive of Questionnaires [3].
2. **Stimuli-guided narrations:** Narrations of various kinds including those based on speech/text prompts/stimuli, picture stories and video prompts are also common and standard methods of data collection in Linguistics. We have prepared a questionnaire of 635 question sentences for the collection from eight broad domains as mentioned in Section 3.3.
3. **Role-play method:** Role-play is a fun way of data collection whereby multiple participants are given a situation and a role and they are supposed to act out as real persons in that situation. We are currently in the process of designing situations for role-play, specific to different linguistic communities.
4. **Spontaneous conversation:** Finally, spontaneous conversation in natural settings is considered to be the holy grail of speech data. While it is tricky at multiple levels including ethical and privacy considerations, given appropriate permission and setup, we plan to collect data using this method as well by visiting the field.

These questionnaires are developed in different phases, through the processes of adapting the relevant questions from the above-mentioned resources, translating those into the contact language of the community, adapting those for specific cultures and communities, recording those to enable data collection using audio prompts and adding new questions relevant to specific communities. The complete questionnaire being used in the project is being made publicly available for researchers [4].

[1] https://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaires.php

[2] http://fieldmanuals.mpi.nl/

[3] http://tulquest.huma-num.fr/en

[4] All data and models will be made available through the project's GitHub repository - https://github.com/unrealtecellp/SpeeD-IL

3.3. Domains

In the first phase of the project, we focused on collecting most of the data from three domains - education, agriculture and science and technology. However, since there are no other large-scale, general datasets available for these languages, in order to ensure good coverage of the generated dataset, we are also collecting at least one-third of the data from other domains as well. The details of the questionnaires are given below:

Agriculture: This domain has 257 question sentences regarding the most common vegetables, crops of different seasons, insecticides and pesticides, fruits, cereals, fertilizers, animal husbandry, seeds, and spices, etc. It also includes questions about the agricultural diversity of a particular region.

Culture: There are 31 narration questions in this domain related to the people's occupations, indigenous administration system, lifestyle, tradition, customs, ornaments, festivals, art, marriage, folk songs and stories, traditional tools and weapons, hunting, and food of a particular community/society.

Education: This domain has 115 narration questions and is framed to elicit the importance, infrastructure, standards, benefits and challenges, improvements, co-curricular activities, sports, and hygiene level of/in education for a community/society.

General-oral-history: Here we have 58 questions to elicit the oral history of the community, cultural background, language relationship, urban and rural lifestyle, religions, family planning, economic background, traditional music, etc.

Healthcare: It has 20 questions to elicit details about the modern and indigenous healthcare system, the use of indigenous remedies as medicine for day-to-day issues, etc.

Lifecycle: Here, 42 narration questions cover the three major lifecycle events that are significant in our culture: birth, marriage, and death.

Sports: This domain includes 16 narration questions about indigenous and modern sports and games.

Science-technology: It has 96 narration questions related to natural, physical and social sciences as well as some general questions related to household tools and technology.

3.4. Data collection in the field

In the first phase of the project, we plan to collect approximately 10,000 hours of data in 10 languages from at least 2,000 speakers in each language. While collecting the dataset, we have ensured balance in terms of 4 social parameters - gender, age (our primary target age group is 20 - 50 years), socio-economic status, education and linguistic varieties. Moreover, the dataset will represent data from most if not all varieties of each language by collecting data from both urban and rural populations in each of the districts where the language is spoken. The process of data collection ensures that all this information about each speaker is maintained as part of the metadata.

3.5. Data management and processing

We are primarily using the LiFE app[5] for questionnaire management, data management, processing and sharing of the dataset. It is an open-source web-based field linguistic data management system that allows for the storage of field data and its metadata in standard formats. It also allows for exporting the data to multiple formats, sharing it with other people for working on a project together and also an automation component that could

[5] http://life.unreal-tece.co.in/

be used for accelerating the process of data transcription, glossing, etc. This app is being developed in-house to support the use of field methods for large-scale data collection, wherein it provides a complete pipeline from questionnaire development to model training.

The app is integrated with the Karya mobile app [6] which we are using to collect speech data using the translation and narration methods. Karya is a mobile-based crowdsourcing application that is especially aimed at providing additional income to low-income groups in India. It runs on any Android-based smartphone and provides an easy interface for completing tasks. The platform has already been successfully deployed to collect speech data in Marathi (109 hours), Hindi (500 hours), and Odia (1700 hours) and is currently used by different organizations to collect different types of language data in many Indian languages. As a result of its crowdsourcing architecture, it allows presenting the fieldwork questions as microtasks to several speakers in parallel in the field, thereby, eliminating the need for the researchers to be present with the speakers to collect data - thus it allows for quick and large-scale data collection from the field for specific purposes. All recordings by the app are in standard 16-bit PCM encoding sampled at 16Khz

4. The data collection pilot

We have currently completed the pilot phase of the project, which involved data collection of four Tibeto-Burman languages - Bodo, Chokri, Meitei and Toto and all four Indo-Aryan languages. For the pilot, we only used the translation and narration questionnaires. Since we needed extensive feedback from the participants, in this phase, we mostly tried to hire students and people directly in contact with the researchers. The major insights gained from this pilot survey are given below -

4.1. Toto

As Toto is spoken in an isolated area and the area Totopara is still not affected by the modern lifestyle, the questionnaire for Toto is to be further customised according to their worldview, otherwise translating the sentences and the words becomes an impossible task for the Toto speakers. Based on the data collected from the pilot survey a list of words was prepared which are alien to the Toto community. These words can not be translated into Toto as these concepts are not present in the community and hence there are no equivalent words in Toto. The list at present has 36 words of common use which can be easily found in major Indian languages, e.g. thief, dacoit, blessing, problem, lock, celebration, mirror, guest, solve, everyday, law etc. The list has words from different parts of speech though most of such words are nouns. This list will help us to further adapt the questionnaire suited for data collection from the community. In addition to this, it was noticed that some sentence structures are difficult to construct in Toto like sentences with anaphora. Short sentences are preferred. Complex and long sentences are difficult to be spoken in Toto. This was communicated by the Toto speakers. We accordingly further modified the questionnaire based on this feedback as well.

4.2. Meitei

A total of 81 such suggestions and feedback were received for Meitei. Specific feedback on Causatives and Double Causatives, Reciprocals and Reflexives were considered and were further adapted and modified for future elicitation sessions to suit Meitei. Apart from the structural and grammaticality judgment issues, some instances of the translatability of some lexical items were also highlighted, which were incorporated into the questionnaire. Another important aspect highlighted was the adaptability and localisation of the names used in the questionnaire.

The main motivation behind the pilot survey was to reduce the colonial design of the questionnaires and adapt that to suit the specific characteristics and nuances of each language, culture and community. We believe it is an essential component of data collection from the communities directly - using the same set of generic texts or questionnaires, which might be common for the majority community but not so much for the communities and languages that one is working with, is neither ethical (since it might lead to the further development of negative attitude towards the language as the speakers start noticing what is *not* possible in their language and take that as a sign of inferiority) nor appropriate from a practical perspective (since it will yield a biased dataset which is more representative of the kind of data that one encounters in the majority, more powerful, well-resourced languages and may not include samples of language as it is actually used in the community). As such adaptation of elicitation tools specifically to the community and the language that we are collecting data from is an essential prerequisite for collecting a balanced and representative sample of the language.

5. The dataset till now

The statistics of the data collected till now is given in Table 1

Language	Translation	Narration	Total
Awadhi	01:52:29	02:54:01	04:46:30
Bhojpuri	01:55:32	02:56:06	04:51:38
Braj	02:14:59	02:20:01	04:35:00
Magahi	02:27:27	01:20:16	03:47:43
Toto	02:31:02	03:24:21	05:55:23
Chokri	02:32:22	03:58:19	06:30:41
Bodo	02:04:47	03:32:03	05:36:50
Meitei	02:05:41	02:42:58	04:48:39
Total	**17:44:19**	**23:08:05**	**40:52:24**

Table 1: *The Speech Dataset*

For each language, the data is collected from 20 speakers and mainly from the urban speakers till now.

6. Summary

We have completed the pilot survey in our project and within a period of one month, we have managed to collect approximately 50 hours of data using a combination of crowdsourcing and linguistic field methods of data collection wherein we have presented the elicitation questions as microtasks to the speakers. This has led to rapid collection of a larger dataset. Methodologically, we are adapting the questionnaires and elicitation methods developed by the field linguists for data collection from the field for collecting large datasets, which may be later reused and replicated for other un(der)represented languages. Using these methods, we plan to collect a larger dataset for a larger pool of under-resourced and un(der)represented languages in India.

[6] https://karya.in/

7. Acknowledgements

We would like to thank the Ministry of Electronics and Information Technology, Government of India for the grant to the SpeeD-TB project under Mission BHASHINI. We would also like to thank Karya Inc for supporting the SpeeD-IA project.

8. References

[1] R. Kumar Vuddagiri, K. Gurugubelli, P. Jain, H. K. Vydana, and A. Kumar Vuppala, "IIITH-ILSC Speech Database for Indain Language Identification," in *Proc. 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, 2018, pp. 56–60.

[2] K. Samudravijaya, P. V. S. Rao, and S. S. Agrawal, "Hindi speech database," in *Proc. 6th International Conference on Spoken Language Processing (ICSLP 2000)*, 2000, pp. vol. 4, 456–459.

[3] B. Das, S. Mandal, and P. Mitra, "Bengali speech corpus for continuous auutomatic speech recognition system," in *2011 International conference on speech database and assessments (Oriental COCOSDA)*. IEEE, 2011, pp. 51–55.

[4] S. B. Sunil Kumar, K. S. Rao, and D. Pati, "Phonetic and prosodically rich transcribed speech corpus in indian languages: Bengali and odia," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013, pp. 1–5.

[5] J. Basu, S. Khan, R. Roy, and M. S. Bepari, "Commodity price retrieval system in bangla: An ivr based application," in *Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction*, ser. APCHI '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 406–415. [Online]. Available: https://doi.org/10.1145/2525194.2525310

[6] T. Godambe, N. Bondale, K. Samudravijaya, and P. Rao, "Multi-speaker, narrowband, continuous marathi speech database," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013, pp. 1–6.

[7] A. Mohan, R. Rose, S. H. Ghalehjegh, and S. Umesh, "Acoustic modelling for speech recognition in indian languages in an agricultural commodities task domain," *Speech Communication*, vol. 56, pp. 167–180, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639313000952

[8] J. Basu, S. Khan, R. Roy, B. Saxena, D. Ganguly, S. Arora, K. K. Arora, S. Bansal, and S. S. Agrawal, "Indian languages corpus for speech recognition," in *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2019, pp. 1–6.

[9] B. Deka, J. Chakraborty, A. Dey, S. Nath, P. Sarmah, S. Nirmala, and S. Vijaya, "Speech corpora of under resourced languages of north-east india," in *2018 Oriental COCOSDA - International Conference on Speech Database and Assessments*, 2018, pp. 72–77.

[10] R. Kumar, S. Singh, S. Ratan, M. Raj, S. Sinha, S. Mishra, B. Lahiri, V. Seshadri, K. Bali, and A. K. Ojha, "Annotated Speech Corpus for Low Resource Indian Langauges: Awadhi, Bhojpuri, Braj and Magahi," in *Proc. 1st Workshop on Speech for Social Good (S4SG)*, 2022, pp. 1–5.

[11] K. Sarmah and U. Bhattacharjee, "Gmm based language identification using mfcc and sdc features," *International Journal of Computer Applications*, vol. 85, 12 2013. [Online]. Available: https://doi.org/10.5120/14840-3103

[12] B. D. Sarma, P. Sarmah, W. Lalhminghlui, and S. R. M. Prasanna, "Detection of mizo tones," in *INTERSPEECH*, 2015.

[13] S. Maity, A. Vuppala, K. Rao, and D. Nandi, "Iitkgp-mlilsc speech database for language identification," *2012 National Conference on Communications, NCC 2012*, 02 2012.

[14] J. Basu, S. Khan, R. Roy, T. Basu, and S. Majumder, "Multilingual speech corpus in low-resource eastern and northeastern indian languages for speaker and language identification," *Circuits, Systems, and Signal Processing*, vol. 40, 10 2021.

[15] C. Basumatary, "The phonological study of toto language," *Language In India*, vol. 14, 2014. [Online]. Available: www.languageinindia.com

[16] A. Abbi, *A Manual of Linguistic Field Work and Structures of Indian Languages*, ser. LINCOM handbooks in linguistics. Lincom GmbH, München, 2001. [Online]. Available: https://books.google.co.in/books?id=0HJiAAAAMAAJ

[17] B. Lahiri and A. Saha, "Words and sentences," *Jadavpur Journal of Languages and Linguistics A Questionnaire Developed for Conducting Fieldwork on Endangered and Indigenous Languages*, vol. 2, no. 3, pp. 11–42, 2018.

[18] B. Lahiri, *The Case System of Eastern Indo-Aryan Languages: A Typological Overview*. Routledge, 03 2021.