Speech corpus for native language of Hindi: Rajasthani, Bhojpuri, Chhattisgarhi, Magahi, Haryanvi

Jogender Singh
Research Scholar (CSE), UIET, Maharshi Dayanand University, Rohtak

Dr. Yudhvir Singh
Professor in CSE, UIET, Maharshi Dayanand University, Rohtak

Abstract - This paper focuses on the collection and creation of a speech corpus for five Hindi dialects: Rajasthani, Bhojpuri, Chhattisgarhi, Magahi, and Haryanvi, using field methods for gathering language data from speakers. While various speech databases exist for different languages, there is a noticeable lack of such resources for Indian native languages. To address this gap, a total of 18 hours of speech data was recorded from 500 native speakers, with 100 speakers representing each dialect. The data was recorded in real-time across diverse environments such as colleges, offices, roadside shops, and homes, ensuring its authenticity. This speech database aims to aid the development of a native language recognition system in India. The paper also outlines the methodology used for collecting the speech corpus, aiming to provide a foundation for future research in areas such as natural language processing, forensic linguistics, language classification, and language identification. In language classification speech corpus gives 80% accuracy on convolutional neural network model.

Keywords: Hindi native speech corpus, natural language processing, language recognition system, language classification, language identification.

I INTRODUCTION

Speech is the most crucial and essential form of communication within human society. It played a key role in the cognitive revolution that shaped human development. Unlike other species, which have limited communication abilities and cannot think like humans or form large groups, humans are inherently social beings who require connections and verbal or communicative understanding. Across the world, numerous languages are spoken, and India, being a multilingual country, highlights the critical importance of communication in human society. In today's era of globalization, the need for speech interfaces has become more vital than ever.

As reported in 2011 census and linguistic report, most of the native languages are on the edge of their extinction[1]. In the 8th Schedule of Constitution of India[2] 22 languages have been classified as scheduled languages and 99 languages have been put in non-scheduled or non- specified languages. Scheduled languages are 1) Assamese 2) Bengali 3) Bodo 4) Dogri 5) Gujarati 6) Hindi 7) Kannada 8) Kashmiri 9) Konkani 10) Maithili 11) Malayalam 12) Meitei (Manipuri) 13) Marathi 14) Nepali 15) Odia 16) Punjabi 17) Sanskrit 18) Santhali 19) Sindhi 20) Tamil 21) Telugu and 22) Urdu. Scheduled languages are recognized for the official purposes at the national level. But non-scheduled languages are spoken by various communities in India and are not recognized at national level. However, they may still be recognized at the state or provincial level for certain cultural and linguistic purposes. This linguistic diversity includes dialects, tribal slangs/ languages and small diverse linguistic communities. There is a demand of 38 languages to add in the 8th schedule for their enrichment and promotion at national level.

The linguistic recognition is important to procure diverse linguistic heritage. As per the report [3], after the census of 2011, 99.82% of total population of India acknowledges their mother tongue as their first language among these 122 languages. For further addition, 0.17% of total population communicates in other languages and dialects. These languages and dialects are non-specified because they are spoken by less than 10,000 people as their mother tongue. They are almost 768 languages which are spoken by only 0.17% of Indian population.

According to PLSI statistics, in every decade, almost 32 native languages are on the edge of extinction without procurement and preservation[4]. All the statistics and reports about native languages are quite alarming for the diverse multilingual country like India. Extremely alarming facts about the endangered states of languages are giving threatening shift to the vitality of language diversity. As it was published in TOI [1], August 9, 2013 that presently India speaks more than 880 languages. But up till 1961, approx. 1100 languages would be in vogue and spoken by native speakers in India. Therefore, we have already faced large extinction of 220 languages in last few years. Vulnerability of native languages is not easy to grasp [4].

For considerable progress of India, Speech Corpus Development will enable generations about the variations of native languages. From various researchers, it is evident that various dataset and speech corpora are prepared but there are gaps. Still, it is challenging task for non-scheduled and less spoken languages to survive in the era of competition.

For low income and less resourced people, it becomes quite difficult to dominate and compete with the resourced languages[5]. Absence of dataset fueled by the non-funding of government to these minimal income groups of people. Absence of dataset leads the absence of these native languages from history. Precarious state of these groups and non-accessibility of datasets are major things to leads the gradual extinction of native languages. Therefore, the development of effective speech corpora can be a major step for the procurement of languages for cultural heritage of linguistically diverse India. As India is a multilingual country and has multilingual society which has about 1652 dialects/Native languages[6], there is the dire need of speech corpus. Most of the work in different native languages is not recorded at critical level.

Language audio Corpus can play an important part in the preservation of native languages which are on the verge of extinction in following ways:

a) Help to secure the endangered native languages for the upcoming generations;
b) It facilitates the role of primary materials, audio recordings; and thereby assists native languages preservation, maintenance and revitalization.
c) It will help the researcher for anthropological, typological, historical and comparative studies.
d) It will assist the further researcher for keeping records and documentation of ancestral history of culturally, lingual diverse India.
e) It will provide platform to the natives of India in the resurgence of native languages in future.
f) Technology like Artificial Intelligence is in vogue and in coming years it will expand more, speech recording of native languages will assist for creating speech interfaces for natural language processing.

Why choose Hindi native language

Hindi has more than 50 native languages which are used by 43.65% of Indian population according to censor 2011 data[7]. Some of Hindi native language shown in table 2 have more speaker strength than half of scheduled language which is shown in table 1. Due to some reasons, they are not recognized as scheduled language in 8th Schedule of Constitution of India but have more speaker strength. So, we have to decide to target these languages to find the diversity of these native languages of Hindi and make speaker data corpus that target speaker strength of these languages.

Language	*Population According Censor 2011*
Assamese	153113551
Bodo	1482929
Dogri	2596767
Kashmari	6797587
Konkoni	2256502
Maithili	13583464
Manpuri	1761079
Nepali	2926168
sanskrit	24821
santali	7368192
sindhi	277264

Table 1 Scheduled Language other than Hindi

Language	*Population According Censor 2011*
Bhojpuri	50579447
Rajasthani	25806344
Chattisgarhi	16245190
Maghi	12706825
Haryanvi	9806519

Table 2 Native language of Hindi

Indian Speech Corpus Review

The present section revealed the relevant research study carried out by the different researchers for understanding and development of speech database in language technology. The Linguistic Data Consortium for Indian languages (LDC-IL) is hosted and managed by CIIL-Mysore in India collect data in various Indian Languages that can be used by researcher for speech application. LDC-IL collect 163:25:47 hours Hindi language data[8] like other language speaker data. Many institutes work in natural language processing create their speaker corpus for different languages with different type of environment and different recording method. Indian Languages Speech Resources Development for Speech Applications create a speech database of Hindi, English and Bangali languages funded by MeitY, Govt. of India[9]. A speech database with large vocabulary is developed at IIIT Hyderabad in collaboration with HP labs Bengaluru collect speaker database for Tamil, Marathi and Telugu languages. The acoustic data is recorded by the use of mobile phone and microphone. The speech data is collected from 559 speakers of different age group with native speaking skill of three different languages. While recording the background music and surround sound are also recorded with speech corpus[10]. Dept. of computer science, Panjab university, Patiala has developed a text to speech synthesis system for Panjabi language. For the development of the above said speech database from text to speech synthesis system, syllables were considered as these are selected as basic unit of concatenate. This speech database for Panjabi languages comprise of 3312 syllables which attributes above 99% of commutative percentage frequency in the preferred corpus[11]. In this series, Department of computer science and application, Utkal University, Bhubaneswar, has developed text to speech synthesis system for Bengali, Hindi, Odiya and Telugu. The speech of native speaker of these four languages was recoded in laboratory environment and the native speakers of languages were directed to read the text[12]. A team of Researcher from TIFR Mumbai and CDAC Noida has developed a multi-speaker general purpose continuous speech database for the Hindi language. This database is detailed enough to recognize acoustic, phonetic inter and intra speaker variability in Hindi language. This speech database comprising of a set of 10 Hindi sentences which are spoken by group of 100 native speakers of Hindi. Two microphones were engaged in the recording procedure and the speech data was recorded in digital format in

a noise less environment. Each of the speakers read 10 sentences containing two parts. The first part consists of two 'Dialect' sentences which rather covers the maximum phonemes of Hindi languages. The second part consist of 8 sentences which covered maximum possible phonetic content. This continuous speech database has been designed and developed in such a manner that it can be used for speaker recognition, study and acoustic-phonetic correlational of the languages[13].

TIFR Mumbai also create a Marathi speaker database with collaboration with IIT Bombay [14]. IIITH-ILSC Speech Database created by International Institute of Information Technology Hyderabad using 23 Indian languages for speech recognition [15]. IITKGP-MLILSC speech database for language identification using 27 Indian Languages[16]. Many institutes create and develop their speaker corpus with Hindi but there is no database created for native language of Hindi. So, we start take voice sample from different geographical area in Hindi native language and create a voice corpus for native Hindi language.

Data-Collection

Data was recorded in offices, colleges, universities, roadsides and in villages in noisy environment. Recorded data is real-time random voice recorded data which also have background sound of speaker. Database was collected and during collection a transition from native languages to Hindi language was observed. The native languages are trying to survive only in their domestic home domain.

Initially, we relied on smartphone for recording conversations. While this approach had the advantage of being discreet and readily available, it quickly became apparent that the audio quality was insufficient for the detailed analysis required in computational linguistics. Subtle phonetic nuances and regional accents were sometimes lost in the background noise and limited recording fidelity. Recognizing the need for better equipment, so we invested in a high-quality voice recorder microphone. This upgrade significantly improved the clarity of the recordings, capturing the full spectrum of sounds and making subsequent transcription and analysis more accurate.

The data collection process involved interviewing over 1,000 individuals, with 200 speakers representing each of the five languages. Each interview was designed to be a natural conversation rather than a formal questionnaire, allowing speakers to express themselves freely and authentically. This method helped in capturing a wide range of linguistic features, including variations in pronunciation, intonation, and colloquial expressions. The involvement of translators was instrumental not only in breaking the ice but also in ensuring that the conversations remained relevant and engaging for the speakers.

Challenges during data collection

Dataset was directly collected from linguistic communities after communicating them the purpose of data collection. Close variations of languages were recorded and many challenges were being faced: -

1. Data privacy and consent from the participants was challenging. To ensure the native language speaker of how their speech would be used for speech interfaces was quite challenging.
2. Different accents and dialects in native languages and their subtle variations was challenging because it encompasses wide range of variations. Collecting representative data of variations was difficult to gauge.
3. Ensuring high quality of data at the time of audio recording was being assured. It was crucial for effective speech processing. Mixed noise of animals and the quality of recording equipment was also being assured time to time.
4. Speaker's variability at the time of recording was quite challenging. The variations even on a single place in single native language were challenging.
5. Volume and scalability of the speakers were kept in check and sometimes, the murmuring of speakers at the time of recording was challenging for data collection.
6. Process of data collection from native speakers-First and foremost thing is to define objectives of data collection to participants so that they can be aware about their data use.
7. Standardized recording equipment was used and data was stored in a proper way for evaluating the variations.

Collecting speech data from native speakers of Rajasthani, Maghi, Bhojpuri, Chhattisgarhi, and Haryanvi presented a unique and enriching challenge that provided profound insights into the intricacies of these languages. The endeavor, aimed at creating robust speech interfaces, was underpinned by the understanding that while Hindi often serves as the lingua franca in many parts of India, regional languages hold the key to preserving cultural identity and enabling more inclusive technological solutions. This not only aimed to gather data but also to contribute to the relatively under explored field of computational linguistics for these languages.

The process began with the realization that while many native speakers are bilingual or multilingual, they tend to default to their native language in informal and familiar settings, particularly when speaking with neighbors or within their community[4]. This bilingualism often masks the rich linguistic diversity that exists and poses a challenge for data collection. To overcome this, I employed a translator familiar with each language to facilitate more natural and comfortable interactions. This step was crucial in making the speakers feel at ease and encouraging them to use their native language instead of switching to Hindi, which they might consider more appropriate for outsiders or formal interactions.

One of the challenges faced during the data collection was the initial hesitation of the speakers. Many were reluctant to participate, either due to a lack of understanding of the purpose or due to privacy concerns. To address this, I conducted preliminary sessions to explain the goals of data collection and how the data would be used to develop speech interfaces that could benefit their communities. Emphasizing the potential impact of their participation in preserving and promoting their native languages through technology helped in gaining their trust and cooperation.

The collected data holds immense potential for the development of speech interfaces, which can significantly enhance accessibility for speakers of these languages. Currently, there is a dearth of technological solutions that cater specifically to these linguistic groups, partly due to the limited availability of comprehensive linguistic data. By creating a substantial database of natural speech samples, this project lays the groundwork for developing more inclusive and user-friendly voice-activated systems, automated transcription services, and language learning tools.

The significance of this project extends beyond technological development. Linguistically, it contributes to the preservation and documentation of these languages, many of which are at risk of decline as younger generations increasingly adopt Hindi or English. Each recorded conversation is a valuable record of contemporary usage, capturing evolving trends and regional dialects. This data can serve as a resource for linguists and researchers studying language change and preservation. Moreover, the project's methodological approach offers a template for future linguistic research in multilingual settings. The combination of using translators and high-quality recording equipment proved effective in ensuring the authenticity and quality of the data. This approach can be replicated or adapted for similar projects involving other under-represented languages, thereby expanding the scope of computational linguistics research.

The next steps involve the meticulous process of transcribing the recordings, annotating them for various linguistic features, and developing algorithms capable of processing and understanding these languages. This will likely involve collaboration with linguists, computer scientists, and native speakers to ensure accuracy and cultural sensitivity. The ultimate goal is to create speech recognition and synthesis models that can handle the unique phonetic and syntactic characteristics of each language.

Speech Corpus

Speech corpus created using voice sample from 500 speakers from different geographical area of native language. More than 100 raw audio sample in each language is taken from different speaker with different length according to real time situation. Sample also have environmental noise, nearby sound and conversation of people in specific language.

Language Classification Experiment

Language corpus will be helpful in many natural languages processing like language identification, language classification and speaker identification etc. So, we start with language classification which help to identify the language of a user and classifies one language from other.

We use a language classification framework[17]with Kapre Layer[18] using convolutional neural network. Voice sample are clipped into one second sample with noise cancelation for input into convolutional neural network model. After clipped voice sample in one second sample the sample base makes a support of 29500 samples for training purpose. With

Convolutional Neural network model get a validation accuracy about 80% at training time which is shown in figure 1. At the time of predication using convolutional neural network model get 88% accuracy for five languages that is shown in table 3 which show the classification report of the model. Overall performance of data corpus with real time noisy environment is good for classification with a support of 29503 sample data from 500 speakers which are used in training and validation process also.

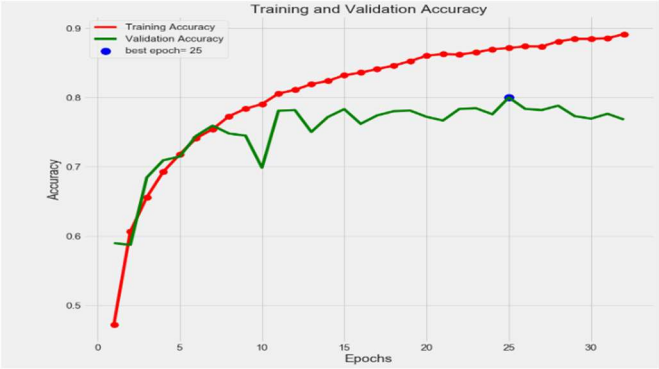

Fig.1 Convolutional neural network Training and validation Graph

Language	Precision	Recall	F1-score	Support
Bhojpuri	0.89	0.97	0.93	6395
Chhattisgarhi	0.88	0.86	0.87	7058
Haryanvi	0.86	0.74	0.80	2402
Maghi	0.85	0.80	0.82	5318
Rajasthani	0.90	0.93	0.92	8330
Accuracy			**0.88**	**29503**

Table 3 Classification Report with Convolutional Neural network

Conclusion

In conclusion, the collection of speech data from native speakers of Rajasthani, Maghi, Bhojpuri, Chhattisgarhi, and Haryanvi represents a significant stride towards more inclusive and representative speech technology. By overcoming initial challenges and leveraging improved recording techniques, samples are gathered to make a rich dataset that will not only aid in technological advancements but also contribute to the preservation and appreciation of linguistic diversity. As technology continues to evolve, initiatives like this ensure that the voices of smaller linguistic communities are heard and integrated into the digital landscape, fostering a more inclusive and accessible future.

References:

[1] T. T. O. I. India, “India-lost-220-languages-in-last-50-years-survey-finds,” 2013.
[2] G. India, “The 8th Schedule of Constitution of India.”
[3] Census of India, “Census of India 2011: Language,” *Census of India*, pp. 1–52, 2011, [Online]. Available:

http://www.censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf
[4] Bhasha Research, "People's Linguistic Survey of India," 2011. [Online]. Available: https://www.bhasharesearch.org/plsi
[5] B. Abraham *et al.*, "Crowdsourcing speech data for low-resource languages from low-income workers," *Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc.*, no. May, pp. 2819–2826, 2020.
[6] P. and others Padmanabha, *Indian Census and Anthropological Investigations*. Registrar General and Census Commissioner, India, 1978.
[7] G. of India, "Census 2011," 2011. [Online]. Available: https://censusindia.gov.in/census.website/
[8] A. K. T. & S. K. A. Ramamoorthy, L., Narayan Choudhary, Jitendra Kumar Singh, Richa, Anjali Sinha, Dheeraj Kumar Mishra, "Hindi Raw Speech Corpus," Languages, Central Institute of Indian. [Online]. Available: https://data.ldcil.org/speech/speech-raw-corpus/hindi-raw-speech-corpus
[9] J. Basu *et al.*, "Indian Languages Corpus for Speech Recognition," *2019 22nd Conf. Orient. COCOSDA Int. Comm. Co-ord. Stand. Speech Databases Assess. Tech. O-COCOSDA 2019*, pp. 1–6, 2019, doi: 10.1109/O-COCOSDA46868.2019.9041171.
[10] R. Kumar *et al.*, "Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems by Vocabulary Speech Recognition Systems," no. July, 2007.
[11] P. Singh and G. S. Lehal, "Text-to-Speech Synthesis system for Punjabi language," in *Proceedings of international conference on multidisciplinary information sciences and technologies, Merida, Spain*, 2006.
[12] S. Ystem, "S Yllable Based I Ndian L Anguage T Ext T O S Peech," vol. 1, no. 2, pp. 138–143, 2011.
[13] P. V. S. R. and S. S. A. Sumudravijaya .K, "Hindi Speech Database," *6th Int. Conf. Spok. Lang. Process. 2000) Beijing,China*, no. 6th, 2000.
[14] S. Paulose, S. Nath, and K. Samudravijaya, "Computer Engineering , Mukesh Patel School of Technology Management and Engineering , Centre for Linguistic Science and Technology ," *6th Intl. Work. Spok. Lang. Technol. Under-Resourced Lang. 29-31 August 2018, Gurugram, India*, no. August, pp. 235–238, 2018.
[15] R. K. Vuddagiri, K. Gurugubelli, P. Jain, H. K. Vydana, and A. K. Vuppala, "IIITH-ILSC Speech Database for Indian Language Identification," *6th Work. Spok. Lang. Technol. Under-Resourced Lang. SLTU 2018*, no. August, pp. 56–60, 2018, doi: 10.21437/SLTU.2018-12.
[16] S. Maity, A. K. Vuppala, K. S. Rao, and D. Nandi, "IITKGP-MLILSC Speech Database for Language Identificatio," *2012 Natl. Conf. Commun. NCC 2012*, 2012.
[17] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, 2017, doi: 10.1109/LSP.2017.2657381.
[18] K. Choi, D. Joo, and J. Kim, "Kapre: On-GPU Audio Preprocessing Layers for a Quick Implementation of Deep Neural Network Models with Keras," 2017, [Online]. Available: http://arxiv.org/abs/1706.05781