Annotated Speech Corpus for Low Resource Indian Languages: Awadhi, Bhojpuri, Braj and Magahi

Ritesh Kumar[1], *Siddharth Singh*[1], *Shyam Ratan*[1], *Mohit Raj*[1], *Sonal Sinha*[1], *Sumitra Mishra*[1], *Bornini Lahiri*[2], *Vivek Seshadri*[3], *Kalika Bali*[3], *Atul Kr. Ojha*[4,5]

[1]Dr. Bhimrao Ambedkar University, Agra
[2]Indian Institute of Technology-Kharagpur
[3]Microsoft Research India, Bangalore
[4]Panlingua Language Processing LLP, New Delhi
[5]National University of Ireland, Galway

riteshkr.kmi@gmail.com

Abstract

In this paper we discuss an in-progress work on the development of a speech corpus for four low-resource Indo-Aryan languages - Awadhi, Bhojpuri, Braj and Magahi - using the field methods of linguistic data collection. The total size of the corpus currently stands at approximately 18 hours (approx. 4-5 hours each language) and it is transcribed and annotated with grammatical information such as part-of-speech tags, morphological features and Universal dependency relationships. We discuss our methodology for data collection in these languages, most of which was done in the middle of the COVID - 19 pandemic, with one of the aims being to generate some additional income for low-income groups speaking these languages. In the paper, we also discuss the results of the baseline experiments for automatic speech recognition system in these languages.

Index Terms: Speech Dataset, Low-Resource Language, Low-income groups, Awadhi, Magahi, Bhojpuri, Braj, Linguistic Fieldwork, ASR

1. Introduction

Development of reliable speech technology for several low-resource languages of India has always been a challenge. Although many automatic speech recognition systems (ASR) [1] have been built for some major languages of India, still there is lack of appropriately transcribed speech corpus and models for several other languages [2, 3]. To overcome the lack of these low resources of data, many initiatives have been taken by several projects and teams in recent times. For example, Interspeech 2018 Low Resources ASR challenge [4] released speech data of phrasal and conversational speech, with transliteration of 50 hours each in Gujarati, Tamil and Telugu.

More recently [2] discusses the development of a speech corpus for 16 low-resource languages of Eastern and North-Eastern India including those for languages like Adi, Angami, Ao, Hrangkhawl, Khasi, Lotha, Mizo, Nagamese, Sumi and others. The speech data, collected from 240 native speakers, totals around 2.17 to 6.67 hrs for each language. The data was used for developing speaker and language identification systems for these languages.

Table 1 shows some other speech corpora prepared for Indian languages by group of researchers and some organisations under different projects.

As is evident, speech corpora of various kinds for "official" and relatively richer languages are now available. This has resulted in remarkable progress in the development of Automatic Speech Recognition (ASR) systems for these languages in the last few years. However, it is still an extremely challenging tasks for non-scheduled, extremely low-resourced languages, generally spoken by people with low income primarily because of the absence of any datasets for these languages. This situation is further exacerbated by the fact that there is a rather minimal government or industrial funding or support for most of these languages because of various reasons. This has led to a situation where millions of people do not have access to speech and language technologies in their own language and may not have access in the next several years.

In this paper, we discuss an in-progress work of the development of a speech dataset in four Indo-Aryan languages - Awadhi, Bhojpuri, Braj and Magahi - using field methods of linguistic data collection, remodelled as limited crowdsourcing-like micro-tasks. This method proved to be advantageous in three ways (see, however, Section 4 for challenges and issues with this) -

- It allowed us to collect the data remotely during the COVID-19 pandemic.
- It provided the (mostly low-income) speakers an opportunity to make an earning during the extremely difficult period of the pandemic.
- It led to the development of speech datasets for hitherto largely ignored languages.

In the paper, we also discuss some of our baseline experiments with the dataset that we have completed till now and demonstrate how even a small dataset like ours could give significant improvement in ASR for these languages over a zero-shot system based on a high-resource closely-related language like Hindi.

2. Our Dataset

2.1. Speaker Selection

We have collected speech data for four languages - Awadhi, Bhojpuri, Braj, and Magahi - given in Table 5. We have recorded speech data via Karya app - a mobile-based crowdsourcing tool, especially aimed at generating income for low-income groups by providing them opportunities to work on crowd-based tasks [16]. The process of sampling the speakers took into consideration the following criteria.

- Age – The age of the language expert should be more than eighteen years.

Corpus Name	Languages	Speakers	Data Size
Hindi Speech Database[5]	Hindi	50	500 Sentences
Standard Speech Corpora[6]	Hindi, Indian English and Bengali	1,500 female and male	-
Bangali Speech Corpus[7]	Bengali	40 female and 70 male	26 hours
PPRT (Phonetic and Prosodically Rich Transcibed) Speech Corpus[8]	Bengali and Odiya	-	20 hours
IITKGP-MLILSC (Indian Institute of Technology Kharagpur - Multilingual Indian Language Speech Corpus) Database[9]	27 Indian languages: Arunachali, Assamese, Bengali etc.	-	1 hour speech data for each language
Speech data for Bangla[10]	Bangla	-	-
North-East Speech Corpus[11]	Assamese, Bengali and Nepali	27	1,000 sentences
Mizo Tones Database[12]	Mizo	5	4,384 syllables
ALS-DB (Arunachali Language Speech Database)[13]	Apatani, Adi, Galo and Nyishi	100 female and 100 male	4-5 minutes in each language by each speaker
Marathi Speech Database[14]	Marathi	1,500	-
Hindi and Marathi Corpus[15]	Hindi and Marathi	-	9 hours

Table 1: *Prior Works*

Details	Awadhi	Bhojpuri	Braj	Magahi
Region	Pratapgarh, Uttar Pradesh East	Patna, Sasaram, Banaras, Ballia	Agra, Uttar Pradesh West	Patna District
Gender	5 Female, 5 Male	7 Female, 3 Male	5 Female, 5 Male	5 Female, 5 Male
Age (yrs)	18 - 35	24 - 75	18 - 30	18 - 35
Lingual	Multi	Mono-Multi	Multi	Multi

Table 2: *Speaker Details*

- Gender – The number of male and female language experts should be roughly equal for each language.
- Region – They should have spent most of their time in the same region from where the data was being collected - this was needed to avoid excessive influence of Hindi on these languages.
- Language – The speakers use their native language in the home domain - again this was needed since all of the languages under study have witnessed a shift to Hindi in the major urban centres, especially among the educated population .

We have selected a total of 40 language experts, ten from each language, by considering the above-mentioned criteria (see Table 2). In order to collect clean and noise-free data the speakers were asked to record speech in the absence of any noise like the sound of the fan, birds chirping, dog barking, family chattering etc.

2.2. Data Collection via Field Methods

We recorded the speech using the methods used by field linguists for linguistic data collection. The rationale for using the field methods for data collection instead of the usual method of recording read or narrated speech using some random sets of texts from web sources or other texts is explained here.

The questionnaires designed by field linguists for linguistic data collection are generally prepared (and perfected over years of fieldwork in hundreds of languages across the globe) with lot of care and attention such that the data collected from using these questionnaires could be utilised for grammatical description of specific phenomena and, possibly, of the language as a whole. This is not generally possible with any random text. Since the languages under the current study are not only under-resourced but are underrepresented, minoritised and devalued (by being mistakenly referred to as "dialects" of Hindi despite very robust linguistic studies demonstrating that it is both historically and structurally erroneous to make this claim), the speakers themselves hesitate in using the language in public, especially when they are being recorded. In such a situation, it could be assumed that it would be extremely difficult to collect data from these languages again - it could also be gauged from the fact that there is hardly any resource available for these languages prior to our work. As such we wanted our data to be useful not only for just one task - Automatic Speech Recognition - but to be more generally useful for the larger community of NLP practitioners, linguists, speakers and other stakeholders.

As a common practice in field-based data elicitation methods, our initial attempt at data collection involved two independent phases - translation phase and narration phase. In the translation phase, we provided 369 sentences in Hindi and asked the speakers to record the translation of these sentences into the respective languages viz. Awadhi, Bhojpuri, Braj, and Magahi. These sentences are Hindi translation of the questionnaire prepared by [17] for collecting data for language documentation and description in Indian languages. These sentences belong to the domain of daily routine life situations like domestic work, food, cooking, etc. More importantly, these sentences have been designed to elicit patterns of a large number of grammatical phenomena such as case, classifiers, reflexives and reciprocals,

tense, mood and aspect, ECV, reduplication and others. Thus it allowed us to not only get data for describing the different phenomena in the language but we also get representation of various kinds of sentences in the dataset.

In the narration phase, we prepared 39 questions related to the three main lifecycle events in our culture - birth, marriage, and death. These questions were prompts asking the speakers to talk about their rituals and tradition related to these events. This yielded a more naturalistic narrative data (in comparison to the translated sentences elicited in the first phase) and also allowed us to collect those kinds of sentences that were not possible to elicit via the translations.

Both the questionnaires (and all the future questionnaires) used for data collection in the project and the dataset has been made publicly available on the project's GitHub repo[1].

2.3. Data Transcription and Annotation

After the completion of speech recording, we have sliced the speech signal based on sentence completion and then transcribed it into Devanagari script and exported the data in TextGrid format. The transcribed data is then further exported to the CONLL-U format and annotated with part-of-speech labels, morphological features and Universal Dependencies relation. The statistics of the transcribed and annotated dataset is give in Table 3.

3. How does this dataset help?

As mentioned earlier, since these languages are often mistaken to be the "varieties" of Hindi, it is generally assumed that the systems built for Hindi should also work fine for these languages. This is expected to be true even if these languages are considered closely-related to Hindi but not its variety. As such we conducted some experiments to understand if this is indeed the case and how well Hindi ASR models perform for these languages. We then also compared the performance of these Hindi models with some baselines that have been developed for these languages - either from scratch or using transfer learning. We used the two commercially available Hindi ASR systems for transcribing the dataset in all of the 4 languages and calculating their WER on our test set -

- Speech-to-text API by Google Cloud[2]
- Speech-to-text API by Azure Cognitive Services (Microsoft)[3]

In the second step, we augmented these models with the language-specific vocabulary and language model to see if this helps or not. Finally in the third step, we trained models for these languages using two approaches -

- Training from scratch: We used Kaldi recipes [18] provided by [19][4] to train the ASR models for these languages from scratch. We experimented with four different models - monophone model (mono), triphone model (tri1), triphones with Delta feature augmentation (tri2b) and triphones with both delta feature augmentation and speaker normalisation (tri3b).
- Transfer learning: We fine-tuned wav2vec-large-xlsr-53 [20] model using the complete, multilingual dataset and evaluated the performance of the model for the whole test set.

For both the approaches. we used two kinds of setups for training and evaluation. In the first setup, we trained monolingual models of each of the languages and calculated average WER of these. In the second setup, we trained a multilingual model by taking a combined dataset from all the languages for training and evaluated all the languages together. This was also meant to test if a single multilingual model gives a performance improvement over multiple monolingual models even with this small dataset or not. We report the WER for each of these models for each language in Table 4.

As is evident, WER obtained on at least one of the Hindi models for all the languages is better than the models trained for these languages from scratch. However, the wav2vec2.0 model (transfer learning) for each of the language is almost half of those for the models trained from scratch or the Hindi models. Moreover, in all cases, in both the models trained from scratch as well as those trained using transfer learning, multilingual models outperform the multiple monolingual models. These results are on expected lines.

4. Challenges

The human speech and its variations due to different dialects and accents, social factors like gender, age and speed of utterance create challenges in building ASR systems. Therefore it is generally recommended to keep the datasets as balanced as possible in these different ways.

In our case, in addition to finding this balance among these factors, use of the 'Karya' app, especially during the pandemic, proved to be tricky and challenging. It was not easy for those speakers who were not acquainted with smartphones to understand the workings of even a relatively simple app like Karya. This resulted in occasional incomplete and corrupted data. Moreover, the recording was done by the users themselves (without any significant guidance or training), thus there were some lack of expertise as how to use the mobile speaker adequately for recording tasks like these. Some speakers' voice were too low or murmuring and breaking and not usable; in others silences were too long before and after the actual recordings. Some of the recordings also contained noises of birds, vehicles, fans, or voices of some other people momentarily. These challenges had become more acute because of the inability of the researchers to actually visit the field, train the speakers in recording and then ask them to do the recordings. Unlike the earlier data collection efforts involving Karya, that was done via actual field visit, all the instructions had to be passed through the mobile phone in our case because of the restrictions imposed by the COVID-19 pandemic; thus there was no possibility of demonstrating the usage of the app. This has affected both the quality and the quantity of the dataset that have been collected till now. However, despite this, the app did enable us to collect and transfer the data even in the middle of the pandemic, when there was no other means of collecting data. On the other hand, the possibility of extra income that the task provided for the low-income groups in a very difficult time, proved to be an additional incentive for the completion of the task.

[1] https://github.com/kmi-linguistics/SpeeD-IA

[2] https://cloud.google.com/apis/docs/overview

[3] https://azure.microsoft.com/en-in/

[4] the script is available here: https://github.com/navana-tech/baseline_recipe_is21s_indic_asr_challenge/blob/master/is21-subtask1-kaldi/s5/run.sh

Language	Translation Sentences	Translation Tokens	Narration Sentences	Narration Tokens
Awadhi	2,320	15,692	620	16,601
Bhojpuri	2,466	16,228	482	7,705
Braj	1,057	6,961	1,298	21,216
Magahi	2,311	15,797	630	10,563
Total	**8,154**	**54,678**	**3,030**	**56,085**

Table 3: *The Speech Dataset Counts*

Models	Awadhi	Bhojpuri	Braj	Magahi	Avg. WER	Multilingual
Azure Hindi	83.2	76.8	91.0	83.3	83.6	-
Google Hindi	**79.9**	67.0	82.8	77.5	76.8	-
mono	86.3	82.7	93.2	88.4	87.7	87.5
tri1	80.9	77.9	90.1	84.2	83.3	81.2
tri2b	82.1	79.7	90.3	82.5	83.7	81.1
tri3b	82.5	79.4	89.2	82.5	83.4	82.3
wav2vec 2.0	-	**37.6**	**56.7**	**38.0**	**44.1**	**40.4**

Table 4: *WER of the Baseline Models*

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
Total	**08:30:27**	**09:30:24**	**18:00:51**

Table 5: *The Speech Dataset*

5. Ethical and Societal Implications

Since the data for this research is collected directly from the speakers, it involved the usual ethical considerations for working with speech communities for linguistic data collection. It included explaining the purpose of the data collection, how the data will be used, speakers' intellectual property rights over the data contributed by them and all its derivatives and finally getting their informed consent for using the dataset for research. In addition to this, more specific societal implications of the research is discussed below -

Short-term good of the speakers' community: As we mentioned earlier the data was collected in the middle of the pandemic and it provided the speakers, who were all selected from the low-income groups [5], an opportunity for some additional income in very trying and difficult times. Although this was nowhere substantial, we would like to think that it did help the speakers in some minimal way.

Possible enhancement of Language Prestige: Since all of the languages under study are relegated to the status of 'illiterate', 'uncouth' or 'rural' varieties of Hindi, the speakers do not generally feel very comfortable owning up to the language or speaking it for "official" purposes like ours. However, our interest and insistence on this very "version" of the language rubbed a little on the speakers as well and led to a possible enhancement of language prestige, with many speakers admitting that they never thought that it could be of any use for highly educated people like University students and professors. Moreover, it was also surprising and encouraging that they could earn some money by knowing and speaking their own language (as opposed to English or Hindi).

The Dataset and its application: Finally, of course, as the dataset is further augmented, it could lead to the development of usable speech technologies for millions of people such that they could access the technologies and content in their own language, without the need to switch to Hindi. Moreover, since a large population of speakers of these languages may not be able to access the writing-based content and technologies, working speech-based technologies might prove to be an essential tool for them.

[5] Since Karya is designed to provide income to low-income groups, it was a conscious decision on our part. However, along with this, since these languages are now generally not preferred to be spoken by the "elite", urban, educated population, it also proved to be a pre-requisite to get the data that we were aiming to collect.

6. Summing Up

In this paper, we have discussed an in-progress work on the development of an annotated speech corpus and baseline ASR systems for 4 extremely low-resource Indian languages, spoken largely by low-income groups. We have demonstrated how even a small dataset like ours could prove to be effective in building a reasonably better system for these languages. We have also argued that our research has the potential to have a positive social impact at multiple levels. The complete dataset with transcriptions and annotations has been made available publicly on our GitHub repo - `https://github.com/kmi-linguistics/SpeeD-IA` - and as further annotations are being finalised, they will be incrementally made available.

7. Acknowledgments

We would like to thank Karya Inc., Microsoft Research India and Panlingua Language Processing LLP for providing funds and financial assistance for this project.

8. References

[1] M. Malik, M. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimedia Tools and Applications*, vol. 80, pp. 1–47, 03 2021.

[2] 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.

[3] H. Yadav and S. Sitaram, "A survey of multilingual models for automatic speech recognition," 2022. [Online]. Available: https://arxiv.org/abs/2202.12576

[4] B. Srivastava, S. Sitaram, R. Mehta, K. Mohan, P. Matani, S. Satpal, K. Bali, R. Srikanth, and N. Nayak, "Interspeech 2018 low resource automatic speech recognition challenge for indian languages," 08 2018, pp. 11–14.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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

[11] 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.

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

[13] 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

[14] 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.

[15] 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

[16] B. Abraham, D. Goel, D. Siddarth, K. Bali, M. Chopra, M. Choudhury, P. Joshi, P. Jyoti, S. Sitaram, and V. Seshadri, "Crowdsourcing speech data for low-resource languages from low-income workers," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 2819–2826.

[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] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Vesel, "The kaldi speech recognition toolkit," *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 01 2011.

[19] A. Diwan, R. Vaideeswaran, S. Shah, A. Singh, S. Raghavan, S. Khare, V. Unni, S. Vyas, A. Rajpuria, C. Yarra, A. Mittal, P. K. Ghosh, P. Jyothi, K. Bali, V. Seshadri, S. Sitaram, S. Bharadwaj, J. Nanavati, R. Nanavati, and K. Sankaranarayanan, "MUCS 2021: Multilingual and code-switching ASR challenges for low resource indian languages," in *Interspeech 2021*. ISCA, aug 2021. [Online]. Available: https://doi.org/10.21437\%2Finterspeech.2021-1339

[20] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," 2020. [Online]. Available: https://arxiv.org/abs/2006.13979