



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

Ritesh Kumar¹, Siddharth Singh¹, Shyam Ratan¹, Mohit Raj¹, Sonal Sinha¹, Sumitra Mishra¹, Bornini Lahiri², Vivek Seshadri³, Kalika Bali³, Atul Kr. Ojha^{4,5}

¹Dr. Bhimrao Ambedkar University, Agra

²Indian Institute of Technology, Kharagpur

³Microsoft Research India, Bangalore

⁴Panlingua Language Processing LLP, New Delhi

⁵National University of Ireland, Galway



Overview

- Introduction
- Language Demography
- Dataset Preparation
- Experimental Setup
- Ethical and Societal Implications
- Summary



Introduction

- Development of reliable speech technology for several low-resource languages of India: always a challenge.
- Many automatic speech recognition systems (ASR) built, still lack of appropriately transcribed speech corpus and models for several minor ones.
- To overcome the lack these low resources of data, many initiatives have been taken by several projects and teams including ours in recent times.



Objectives

- Transcribed speech corpus and ASR for four low-resource Indo-Aryan languages
 - **Awadhi, Bhojpuri, Braj and Magahi.**



Language Demography

- Language Family: Indo-Aryan
 - Awadhi: Native Speaker - 3.85 million (2011 census), Central Indo-Aryan language, spoken in the Awadh region of Uttar Pradesh.
 - Bhojpuri: Native Speaker - 51 million (2011 census), Eastern Indo-Aryan language, spoken in western Bihar, eastern Uttar Pradesh, western Jharkhand, northeastern Madhya Pradesh, northeastern Chhattisgarh and in the Nepal.



Cont...

- Braj: Native Speaker - 1.6 million (2011 census), Western Indo-Aryan language, spoken in the states of Western Uttar Pradesh and parts of Rajasthan.
- Magahi: Native Speaker: 20.7 million (2011 census), Eastern Indo-Aryan language, spoken mainly in Eastern Indian states including Bihar and Jharkhand, along with some parts of West Bengal and Odisha.




Dataset Collection

- 'KARYA' App: a mobile-based crowdsourcing tool.
- Field methods of linguistic data collection, remodelled as limited crowdsourcing creating micro-tasks.

Karya User Input Data





1:57 0 KB/s HD 4G 74%

नमस्ते 


रहुआ एकरा भोजपुरी में कइसे बोलब / बोली ला?


"भूख से मेरा अतरि
ककुरा रहा है।"

000

9:51 PM 0.4KB/s

नमस्ते 

 नए कामन ए लएँ जां दबाऔ - करे भए कामनि ए जमा करिबे - जंचौ भऔ अपडेट करौ - कमाई अपडेट करौ


बृज भासा वाक्य अनुवाद
बोले भए डेटा को संग्रै


362 काम हतो ऐ

7 काम पूरौ है गऔ

7 काम जमा है गऔ

9:39 AM 0.2KB/s 30%

नमस्ते 


 नए कामन ए लएँ जां दबाऔ - करे भए कामनि ए जमा करिबे - जंचौ भऔ अपडेट करौ - कमाई अपडेट करौ

**जीवनचक्र ति जुरे भए
सबाल**
बोले भए डेटा को संग्रै

39 काम पूरौ है गऔ

39 काम जमा है गऔ

39 जमा भऔ ई जांचौ गऔ ऐ

 234.00



Speaker Details

Details	Awadhi	Bhojpuri	Braj	Magahi
Region	Pratapgarh, Uttar Pradesh East	Patna, Sasaram - Bihar Varanasi, Ballia - Uttar Pradesh	Agra, Uttar Pradesh West	Patna District
Gender	5 F/M	7 F & 3 M	5 F/M	5 F/M
Age	18-35	24-75	18-30	18-35
Lingual	Multi	Mono-Multi	Multi	Multi



Questionnaire Method

- Utilised for grammatical description of specific phenomena and, possibly, of the language as a whole.
- “Words and sentences,” Jadavpur Journal of Languages and Linguistics A Questionnaire Developed for Conducting Fieldwork on Endangered and Indigenous Languages.
- Data collection involved two phases - Translation phase and Narration phase.
- https://docs.google.com/spreadsheets/d/1n0pRHrzrGByQAFU37II_9TuLbi-XtpPVJUXSid9Glf0/edit#gid=0



Translation Phase

- Translation of 369 Hindi sentences into Awadhi, Bhojpuri, Braj, and Magahi.
- Domain: domestic work, travelling, cooking, etc.
- Produced good amount and various aspects of sentences.
- Elicited patterns of grammatical phenomenon: case, classifiers, reflexives, reciprocals, tense, mood, aspect, ECV, reduplication and others.



Narration Phase

- Total 39 questions
- Lifecycle events:

Birth, Marriage, Death.

- Yielded naturalistic narrative data (in comparison to the translated sentences elicited in the first phase). Prompted asking the speakers to talk about their rituals and tradition related to those events.



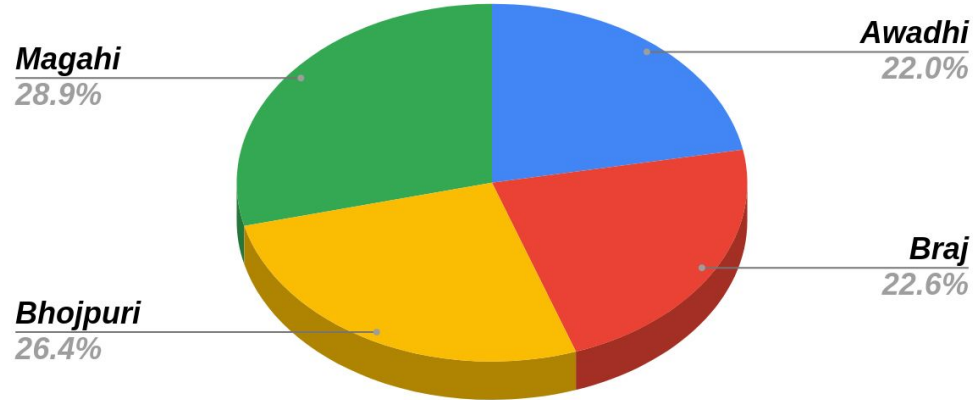
The Speech Data Set

Language	Translation (HH:MM:SS)	Narration (HH:MM:SS)	Total (HH:MM:SS)
Awadhi	01:52:29	02:54:01	04:46:30
Braj	01:55:32	02:56:06	04:51:38
Bhojpuri	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



Translation Data Set

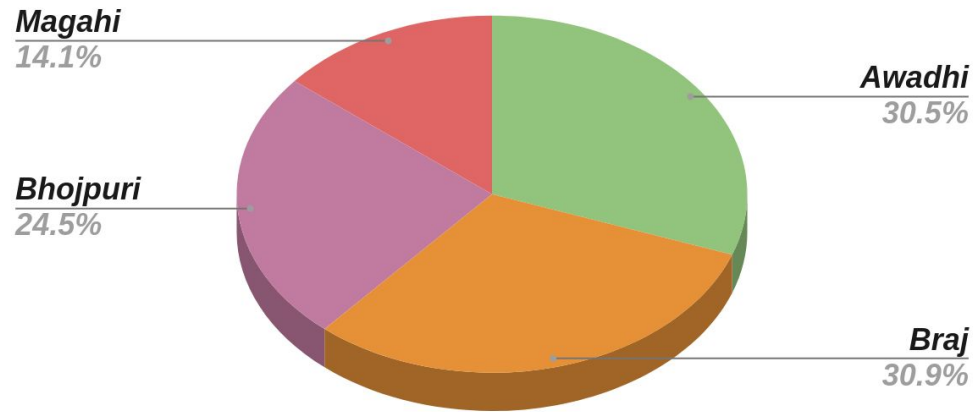
Translation





Narration Data Set

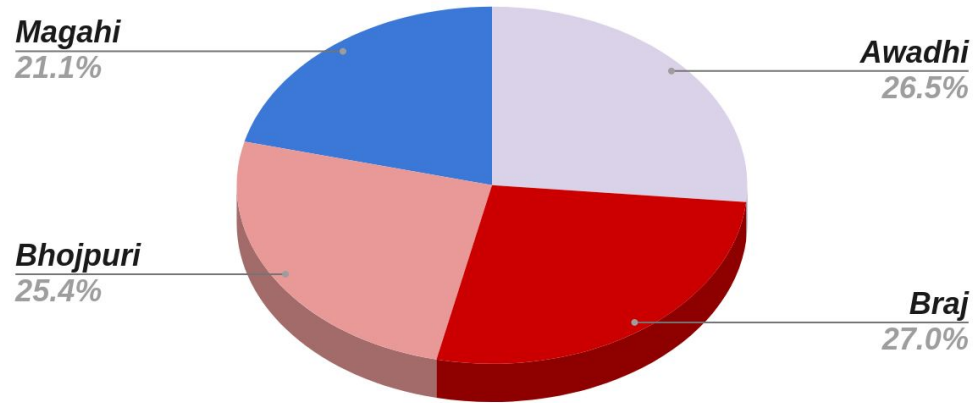
Narration



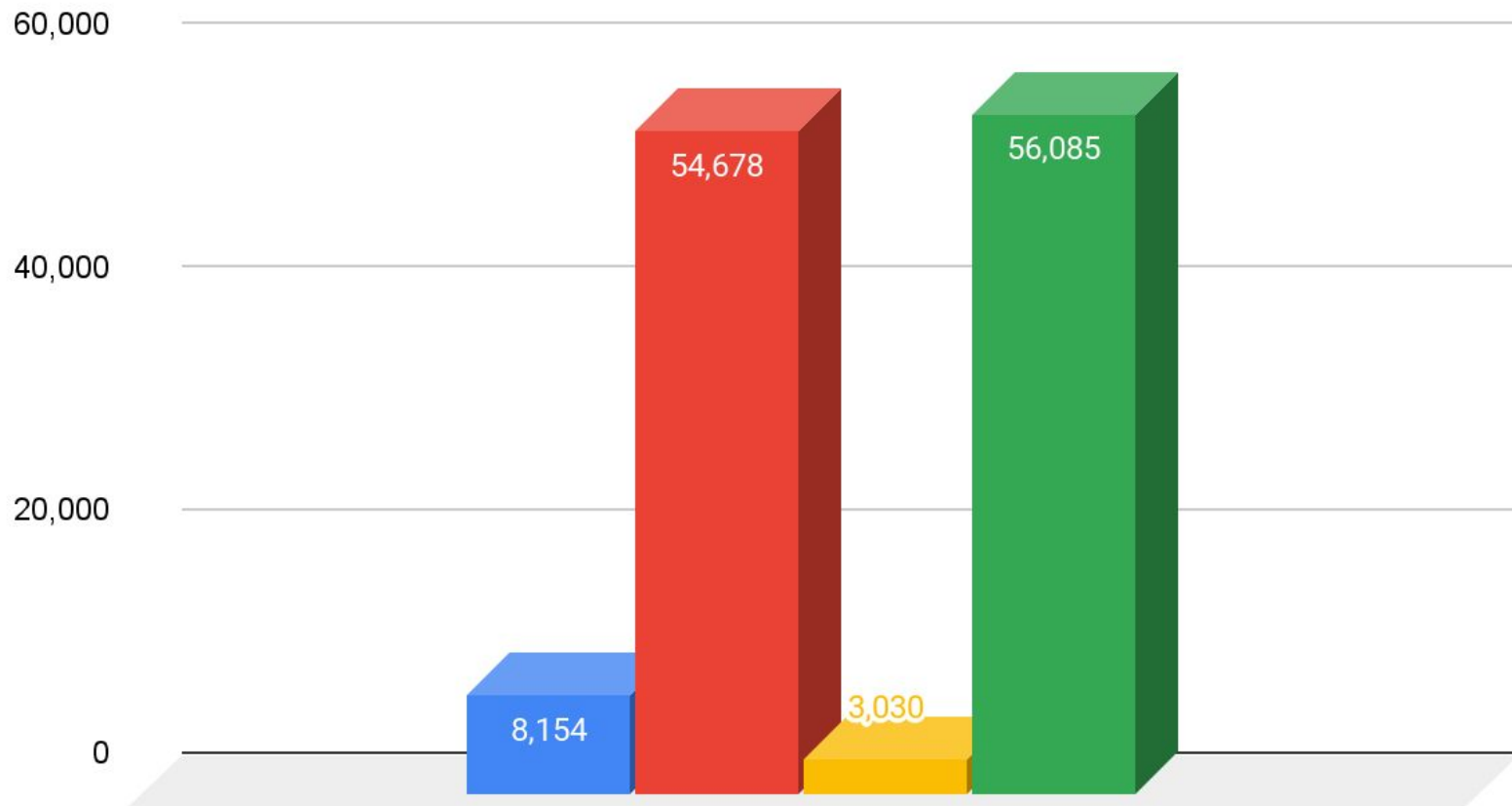


Total Data Set

Total



Translation Sentences Translation Tokens Narration Sentences Narration Tokens





Experimental setup

- 80:20 split of the dataset into Train:Test
- We used the two commercially available Hindi ASR systems for transcribing the dataset and calculated WER on our test set -
 - Speech-to-text API by Google Cloud
 - Speech-to-text API by Azure Cognitive Services (Microsoft)



Cont...

- **Training from scratch:** We used Kaldi recipes 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 model using the complete, multilingual dataset and evaluated the performance of the model for the whole test set.



Cont...

- 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 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.



Cont...

- 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.
- 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.



WER of the Baseline Models

Models	Awadhi	Bhojpuri	Braj	Magahi	Avg. WER	Multilingual
Azure-Hi	83.28	76.84	91.08	83.38	83.64	-
Google-Hi	79.93	67.06	82.89	77.53	76.85	-
mono	86.37	82.76	93.25	88.45	87.70	87.54
tri1	80.97	77.93	90.14	84.23	83.31	81.21
tri2b	82.16	79.76	90.38	82.53	83.70	81.17
tri3b	82.56	79.40	89.25	82.55	83.44	82.34
Wav2vec 2.0	-	37.65	56.78	38.03	44.15	40.44



Ethical and Societal Implications

- **Short-term good of the speakers' community:** an opportunity for some additional income in very tiering and difficult times during COVID 19 pandemic. Although this was nowhere substantial, we would like to think that it did help the speakers in some minimal way !



Cont...

- **Possible enhancement of Language Prestige:** 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, by giving a sense of importance of the culture and language through the questionnaires.



The Dataset Application

- 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 others.
- Suitable working speech-based technologies might prove to be an essential tool for them rather the writing-based content and technologies in these times.



THANK YOU ALL

Contact for more details:

ritesh78_llh@jnu.ac.in

<https://github.com/kmi-linguistics/Speed-IA>



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



Cont...

- [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 automatic 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.



Cont...

- [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 COCODA - International Conference on Speech Database and Assessments, 2018, pp. 72– 77.
- [12] B. D. Sarma, P. Sarmah, W. Lalhminghlu, and S. R. M. Prasanna, "Detection of mizo tones," in INTERSPEECH, 2015.



Cont...

- [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 COCODA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCODA/CASLRE), 2013, pp. 1–6.
- [15] A. Mohan, R. Rose, S. H. Ghalehjeh, 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>



Cont...

- [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. Motl’ı̇cek, Y. Qian, P. Schwarz, J. Silovsk’ı̇y, G. Stemmer, and K. Vesel, “The kaldi speech recognition toolkit,” *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 01 2011.



Cont...

[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>