Interspeech 2024
1-5 September 2024, Kos, Greece

LAHAJA: A Robust Multi-accent Benchmark for Evaluating Hindi ASR Systems

Tahir Javed, Janki Nawale, Sakshi Joshi, Eldho George, Kaushal Bhogale, Deovrat Mehendale, Mitesh M. Khapra

AI4Bharat, Indian Institute of Technology Madras, India
tahirjmakhdoomi@gmail.com, miteshk@cse.iitm.ac.in

Abstract

Hindi, one of the most spoken language of India, exhibits a diverse array of accents due to its usage among individuals from diverse linguistic origins. To enable a robust evaluation of Hindi ASR systems on multiple accents, we create a benchmark, LAHAJA, which contains read and extempore speech on a diverse set of topics and use cases, with a total of 12.5 hours of Hindi audio, sourced from 132 speakers spanning 83 districts of India. We evaluate existing open-source and commercial models on LAHAJA and find their performance to be poor. We then train models using different datasets and find that our model trained on multilingual data with good speaker diversity outperforms existing models by a significant margin. We also present a fine-grained analysis which shows that the performance declines for speakers from North-East and South India, especially with content heavy in named entities and specialized terminology.
Index Terms: non-native speech recognition, Indian accents.

1. Introduction

Hindi is one of the most widely spoken languages of India with 528M speakers identifying it as their first language and another 163M identifying it as their second or third language. People across the country learn and speak Hindi for personal, political and/or employment reasons, and it serves as an unofficial lingua franca for day-to-day activities in several parts of the country. As a result there is significant variation in the accents of people speaking Hindi across the country with regional influences as well as influences from the primary language. These regional influences stem from the rich linguistic diversity of India which has 22 scheduled languages, 122 major languages, and 1599 other languages, as per the Census of 2011. Speakers of languages from the Dravidian family, like Tamil and Malayalam, showcase unique speech rhythms and ways of articulating words that stand in contrast to those from the Indo-Aryan group, including languages like Hindi, Marathi, and Gujarati. Accentual differences are also prominent within the Indo-Aryan languages, reflecting the diverse linguistic landscapes of India's northern, western, and eastern regions. Given the widespread usage and diversity, it is imperative to develop automatic speech recognition systems for Hindi which cater to multiple accents.

While there are efforts to collect voice samples from *native* speakers of Hindi [1, 2, 3, 4, 5] for training ASR systems, there is no benchmark which has Hindi speakers from *diverse backgrounds*, speaking with *different accents*. In this work, we address this gap by releasing LAHAJA, an ASR benchmark containing multi-accent Hindi data. We follow the same collection methodology as used in SVARAH [6], which is an ASR benchmark containing Indian-accent English data and INDICVOICES [1] which is an effort to collect data from native speakers only (as opposed to non-native speakers in our case). LAHAJA contains a total of 12.5 hours of Hindi data collected from 132 speakers of which 122 are non-native speakers. These speakers were spread across 82 districts spanning 18 states in India as shown in Figure 1. The set of native languages spoken by these speakers encompasses 19 of the 22 constitutionally recognised languages of India, spread across 4 distinct language families.

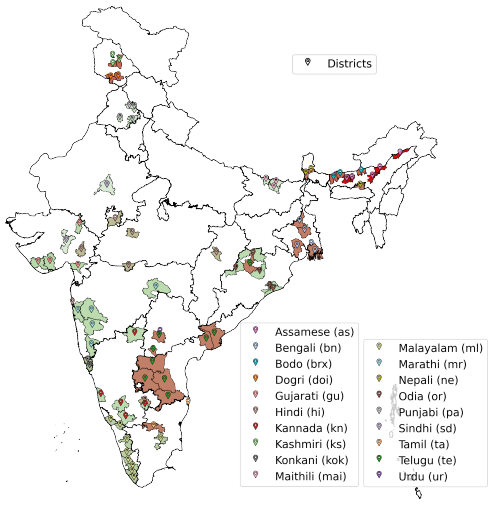

Figure 1: *Different districts of India from which data was collected. The colors show how the WER of our best model varies across different regions of India (shades of green being relatively better and shades of red and brown being relatively poor).*

We evaluate existing open source and commercial models on LAHAJA to understand the current state of ASR for Hindi. In addition, we train multiple model variants based on the Conformer architecture by using different data sources. We observe that (i) our model outperforms all existing models on LAHAJA (ii) our model trained on multilingual data performs better perhaps due to better speaker diversity and (iii) in low resource monolingual settings adding synthetic code-mixed data helps. We also present a fine-grained analysis across different accents and content categories and observe that the performance is poor on speakers from North-East India and South India, with a sharp drop on content rich in named entities and terminology from specific domains. All the code, datasets, models and scripts have been made publicly available[1] and we hope that they will enable further research on multi-accent Hindi ASR systems.

[1] https://github.com/AI4Bharat/Lahaja

10.21437/Interspeech.2024-2376

2. LAHAJA

We now describe the process of creating LAHAJA. As mentioned earlier, we largely follow the same methodology as used in [6, 1] but focus on non-native speakers of Hindi.

2.1. Recruitment of Speakers

We selected 132 participants from 18 out of the 28 states of India, of which 122 were non-native speakers and identified Hindi as their second, third or fourth language. The primary languages of these non-native speakers covered 19 of the 22 scheduled languages of India belonging to Indo-Aryan, Dravidian, and Tibeto-Burman language families. For each of the 19 languages, we recruited 3–5 participants who could speak Hindi. This included 65 males and 67 females, with 6.5 hours of male speech and 6.0 hours of female speech. We included participants from diverse age groups: 18–30, 30–45, 45–60, and 60+, with roughly equal representation in each age group. Participants came from various segments (unemployed, students, blue-collar, and white-collar) with varying education levels (upto 12th grade, undergraduates, graduates, and postgraduates). Participants were briefed about the task, and were clearly informed that their voice samples will be used to develop and evaluate speech recognition models. Their voice samples were recorded only after they willingly agreed and signed a consent form. The participants were appropriately compensated for their work according to daily wages in their region. The entire process was reviewed and approved by our Institute Ethics Committee.

2.2. Data collection

For recording voice samples, we used Microsoft's open-source Karya platform [7]. Once a participant is identified, we onboard them by asking them to fill a web-form which collects participant's meta-data such as age, gender, district, primary language and topics/domains of interest. Once registered, the participants are asked to download and install the Karya application. The participant then performs the following tasks on the Karya.

Read speech: To ensure good vocabulary coverage we use 1K sentences from Wikipedia articles covering 13 domains, as released by [1]. We ask non-native speakers of Hindi to read out these sentences as it is.

Digital interactions with voice assistants: Following [1], we ask speakers to record utterances typically found in digital transactions with voice assistants. These digital transactions cover interactions with (i) in-home assistants for everyday tasks such as *setting an alarm*, *switching on the light, playing music*, etc. (ii) digital payment services covering multiple intents such as *checking account balance, transferring money, paying electricity bill*, etc. (iii) online grocery shopping apps covering multiple intents such as *placing an order, seeking a refund, changing delivery address, etc.* and (iv) online government services covering multiple intents such as *applying for a service*, *checking the status of application*, *renewing a service*, etc. The diversity in the applications covered ensures that the benchmark has a good representation of number sequences, alphanumeric codes, brand names, product names, bank names, government scheme names, application specific terminology and code mixed content (English-Hindi) typically found in such interactions.

Extempore conversations: We use a carefully curated list [1] of 2.5K questions from 21 domains such as tourism, government etc., and 28 topics of interest such as reading, painting etc. Next, we request each participant to select two topics they are interested in and two domains with which they are familiar and capable of answering questions about. Some sample questions include "Technology: How have smartphones made life better?", "Government: Given a chance, what policies will you

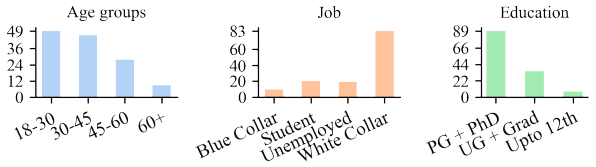

Figure 2: *Demographic distribution of participants in* LAHAJA *across age group, job segment and educational background.*

Table 1: *Statistics of* LAHAJA *across different native speakers. (# Mins = # Minutes, # Sp = # Speakers, # Exclusives = # words which are unique to that splice of data)*

Native speakers	# Mins	# Sp	# Words	# Exclusives
Assamese (as)	32.6	6	1158	225
Bengali (bn)	52.0	10	1786	471
Bodo (brx)	43.6	6	1462	316
Dogri (doi)	30.4	6	1377	325
Gujarati (gu)	37.0	4	1463	339
Hindi (hi)	61.1	10	2053	521
Kannada (kn)	37.0	8	1367	294
Kashmiri (ks)	39.5	7	1505	416
Konkani (kok)	50.3	7	1432	280
Maithili (mai)	34.6	6	1576	381
Malayalam (ml)	38.8	11	1566	319
Marathi (mr)	52.5	6	1911	498
Nepali (ne)	42.3	5	1475	329
Odia (or)	44.7	8	1553	333
Punjabi (pa)	30.8	7	1357	262
Sindhi (sd)	17.3	5	851	119
Tamil (ta)	24.7	5	1030	201
Telugu (te)	55.1	12	1786	387
Urdu (ur)	14.8	3	727	114

Table 2: *Statistics about different content categories in* LAHAJA. *(# Mins = # Minutes, # Utt = # Utterances, # Exclusives = # words which are unique to that splice of data)*

Content categories	# Mins	# Utt	# Exclusives
Read speech			
Wikipedia sentences	41.1	268	754
Digital interactions with voice assistants			
Digital payment services	38.2	263	188
Everyday tasks	13.5	257	144
Online government services	52.9	281	138
Online grocery shopping	32.3	264	558
Extempore conversations			
Agriculture and fisheries	6.5	47	70
Business and finance	16.1	133	131
Humanities and culture	92.6	697	723
Icebreakers	216.6	1647	1898
Leisure activities	65.5	511	633
Mass communication	22.3	160	169
Product reviews	28.6	203	248
Public resources	35.9	275	284
Science and technology	18.9	153	181
Sports and travel	31.8	264	260
Named entities			
Task of five	22.7	476	349

introduce to aid farmers in your area" , "Reading: Do you have a favorite book? If so, what is it and why do you like it?" and so on. While the examples shown here are in English, the questions are translated to Hindi and shown to the participants. In addition to the above we also use some icebreaker questions to warm up the participants. These included questions about their mother tongue, their everyday life, their state/district and so on.

Named entities: To get a good representation of named entities typically encountered in downstream applications, we ask users to speak any 5 numbers , any 5 dates, any 5 person names, names of any 5 Indian cities, any 5 Indian states, any 5 Indian districts, any 5 countries, and any 5 international cities.

Each participant thus reads 20 sentences across both read speech and digital interactions, and answers 8 questions on selected domains and topics of interest.

2.3. Transcription

We adhere to the guidelines as outlined in [1] for transcribing the collected audio samples. We use an open-source platform, Shoonya [8], for transcription which supports multiple Indian languages and a maker-checker workflow. The workflow ensures that the initial transcript (maker) is verified by a senior transcriber (checker). All transcripts are generated in the native Devanagari script of Hindi. Our transcribers are language experts with several years of experience in transcription and translation tasks. We first split the larger audio files into segments using Silero Voice Activity Detection [9] and then provide these segments to the transcribers for transcription. Finally, we downsample the chunked audios to 16kHz, resulting in 16kHz mono 16-bit PCM wav audios.

2.4. Statistics

Table 1 shows statistics of LAHAJA split across native speakers belonging to different languages. Table 2 shows the statistics grouped across different content categories which allows for a fine-grained evaluation of downstream models on LAHAJA. These categories were created by grouping roughly related domains and topics of interest. For example domains like education, govt., health, legal are grouped into 'public resources'.

3. Experimental Setup

Baselines: To establish a baseline, we evaluate the performance of the following existing models on LAHAJA.

- *MMS [10]:* This is Meta's open-source 300M wav2-vec2 [11] model, supporting 1107 languages, including Hindi.
- *WhisperV3:* This refers to the latest open-source Whisper [12] model, trained on 680k hours of data, having 1550M parameters and supporting 100+ languages, including Hindi
- *Azure:* This refers to the Hindi speech to text systems, commercially made available by Microsoft through their SDKs.
- *Google Chirp:* This refers to Google USM [13] model which is made commercially available through Google Cloud APIs.

IndicASR model: We train a Conformer-L [14] with a hybrid RNNT-CTC [15] decoder. We trained 4 different variants of the model by starting with the pretrained checkpoint of an English ASR model, Nvidia-En-SSL [16], and fine-tuning it on different datasets as described below. We found that starting with english checkpoint helped in faster convergence.

- **M1:** This model was trained on the Hindi subset of the INDICVOICES dataset which contains 65 hours.
- **M2:** This is a multilingual model trained on the entire INDICVOICES dataset which contains 1509 hours summed up across 22 Indian languages.
- **M3:** This is a monolingual model trained on 2285 hours by combining the Hindi subsets of VISTAAR [17], SPRING-INX [2] and INDICVOICES.
- **M4:** A lot of extempore content in Indian languages is code-mixed with English, especially for non-native speakers. Hence, we do an interesting experiment where we train a model using 65 hours of Hindi data from INDICVOICES plus an additional 65 hours of synthetic data. This synthetic data is obtained by taking 65 hours of English ULCA ASR data [18] which contains English content spoken by Indian users. We transliterate the English transcripts to Devanagari script using an open source transliteration model, IndicXlit [19]. Thus we created a English-Hindi *mixed* dataset which contains original Hindi audios as well as English audios which are transcribed using Devanagari script (the content is in English but written in Devanagari script, as is the case in code-mixing).

We trained all the models for a maximum of 130k steps and employed early stopping with a patience of 5k steps. We set the max sequence length to 30 secs, used batch size of 16 audios per GPU on 8 GPUs with gradient accumulation of 4, resulting in an effective batch size of 512 audios. We used AdamW [20] as the optimizer with *lr* of 2.0 and Noam [21] as the LR scheduler.

Evaluation metric: We used Word Error Rate (WER) as the metric to compare performance across models.

4. Results and Discussion

Performance across models: Referring to Table 3, we observe that our base model **M1** outperforms all existing models with a minimum and maximum improvement of 2.9% WER and 13% WER, respectively. Among the baseline models, the massively multilingual open source models perform poorly as compared to the closed source commercial models from Azure and Google.

Next, we compare our monolingual model, **M1**, with our multilingual model, **M2**, both trained on IndicVoices. It is observed that **M2** outperforms its monolingual counterpart, by a margin of 2.7% WER. There could be two reasons for the better performance of the multilingual model (i) on aggregate it uses much more training data than the monolingual model although the amount of Hindi data is the same (ii) it sees training data in the native language of the accents studied in this work (although from a different set of speakers). We hypothesise that the second reason is more likely as otherwise the massively multilingual MMS and Whisper V3 models which have arguably trained on much larger data would also have performed better.

Lastly, again referring to Table 3, we compare our multilingual model (**M2**) and our monolingual model (**M3**), trained using two additional sources of Hindi data: VISTAAR [17], SPRING-INX. Here, we clearly see the effect of adding more Hindi data and observe a further reduction of 2.4% WER while moving from **M2** to **M3**. It would have been interesting to see the effect of training a multilingual model by combining all multilingual subsets of Vistaar, Spring and IndicVoices but due to computational constraints, we leave this as future work.

Effect of adding English-Hindi *mixed* data: Referring to Table 4, we compare our monolingual model **M1** with **M4**, which is trained with English-Hindi mixed data. Interestingly, **M4** performs better than **M1** by $\approx$1% WER. We hypothesize that since LAHAJA contains a significant code-mixed data, adding synthetically created code-mixed content helps when the training

Table 3: *WERs (%) of different models on* LAHAJA, *(ML = Multilingual models, **?** = Information Unavailable)*

Model	ML	WER%
MMS	✓	34.4
WhisperV3	✓	32.4
Azure	?	28.6
Google Chirp	?	22.3
M1: IndicASR *(Trained on IndicVoices)*	✗	19.4
M2: IndicASR *(Trained on IndicVoices-Multilingual)*	✓	16.7
M3: IndicASR *(Trained on Vistaar, Spring-Inx and IndicVoices)*	✗	**14.3**

Table 4: ***(Left)**: Effect of adding 65 hours of English-Hindi mixed data to IndicVoices. Both models are trained in a monolingual setting. **(Right)** Comparison of performance of **M3** on native and non-native accents*

Model	WER%
M1: IndicASR *(Trained on IndicVoices)*	19.4
M4: IndicASR *(Trained on IndicVoices, English-Hindi mixed)*	18.6

M3	WER%
LAHAJA	14.3
INDICVOICES	13.1

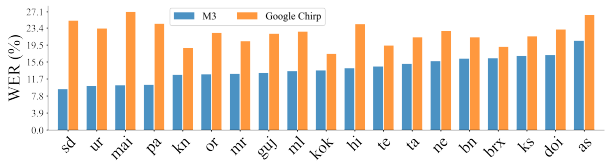

Figure 3: *Performance breakdown of **M3** & **Google Chirp** across non-native speakers.*

data is less (**M1** uses only INDICVOICES). In a separate experiment, we found that adding synthetic code-mixed on top of resource rich settings as in **M2** and **M3** does not help.

Performance across accents: Figure 3 contains the spliced WER of our best model **M3** on different accents. It is evident that performance of **M3** decreases as we move from regions where Hindi is more popularly spoken as the 2^{nd} language or is closely related to the region's native language to regions where this is not the case. We observe the model performs best for languages like Urdu and Sindhi with 10.5% WER and worst for Assamese with 20.5% WER. More generally, from Figure 1 we understand that moving from Central and West India (where languages related to Hindi like Maithili, Urdu are spoken) to North East India (where languages like Assamese, Nepali are spoken) and South India (where Dravidian languages like Tamil, Telugu are spoken), we see a clear decline of performance. We hypothesise that this is due to strong influences of the primary language of the speaker which is increasingly different from Hindi. We do see surprises (e.g., we expected WER on South Indian languages, 'kn' and 'ml' to also be poor).

Comparison with native accents: We now compare the the performance of our model on LAHAJA and the Hindi subset of the INDICVOICES which only contains native speakers (see the right half of Table 4). We observe that the performance of **M3** on IndicVoices, which consist of only native speakers of Hindi, is better than that on LAHAJA. This combined with the fine-grained results in Figure 3 implies that LAHAJA is a good benchmark for evaluating performance across different accents.

Table 5: *WERs (%) of **M3** & **Google Chirp** across content categories of* LAHAJA

Content categories	Codename	Chirp	M3
Read speech			
Wikipedia sentences	R1	18.6	12.9
Digital interactions with voice assistants			
Digital payment services	D1	40.8	12.7
Everyday tasks	D2	15.0	13.2
Online government services	D3	52.3	13.0
Online grocery shopping	D4	26.1	21.5
Extempore conversations			
Business and finance	E1	17.7	14.0
Humanities and culture	E2	16.4	13.5
Icebreakers	E3	17.3	13.3
Sports and Travel	E4	19.3	13.5
Leisure activities	E5	16.9	14.4
Public resources	E6	18.4	14.6
Product Reviews	E7	19.7	15.1
Mass communication	E8	17.1	16.2
Science and Technology	E9	25.5	18.4
Agriculture and fisheries	E10	21.3	20.5
Named entities			
Task of five	N1	63.2	24.4

Table 6: *Examples of errors. AC = Accent, CN = Codename*

Reference	Predicted	AC	CN	Comment
Meghalaya	Mai kha li ya	ml	N1	Unnecessary splitting
Shonitput	Chaunitpur	ne	N1	Confuses 'sh' and 'ch'
Pathological	{}	as	E6	No output
Glucon D	Blookandee	as	D4	Confuses 'glu' and 'blu'
Bade vyapaari	Body paper	or	E10	Predicts En word instead
Ansh	Aaj	ta	E9	Predicts frequent Hi word

Performance across different content categories: In Table 5, we present a fine-grained evaluation of the model across different content categories. The model performs well on read speech with standard vocabulary from Wikipedia, as well as everyday tasks, icebreaker questions and some domains like business and culture. The model particularly struggles in utterances which are rich in named entities (task of fives, product reviews, online grocery shopping) and in certain domains (science and technology, agriculture and fisheries) which may have very domain-specific vocabulary. We list examples of errors in Table 6.

5. Conclusion

We present LAHAJA, a comprehensive benchmark featuring 12.5 hours of Hindi audio from 132 speakers across 83 districts, allowing evaluation of Hindi ASR systems on multiple accents. Our evaluations reveal that existing open-source and commercial models fall short in accurately recognizing multi-accent Hindi speech, underscoring the challenge of accent diversity. However, by training models on multilingual data that encompass a broad range of speakers, we have achieved notable improvements, surpassing existing models by a significant margin. Our fine-grained analysis further emphasizes the performance gaps for speakers from North-East and South India, particularly with content laden with named entities and specialized terminology. By making our code, datasets, and models publicly available, we aim to spur further research and development of ASR systems supporting multiple accents.

6. Acknowledgements

We would like to thank Digital India Bhashini, the Ministry of Electronics and Information Technology (MeitY[2]) of the Government of India and the Centre for Development of Advanced Computing (C-DAC[3]), Pune for generously supporting this work and providing us access to multiple GPU nodes on the Param Siddhi Supercomputer. We would like to thank the EkStep Foundation and Nilekani Philanthropies for their generous grant which went into hiring human resources as well as cloud resources needed for this work. We would like to thank the team of AI4Bharat for helping us to collect data from native speakers of different languages across the country.

7. References

[1] T. Javed, J. A. Nawale, E. I. George, S. Joshi, K. S. Bhogale, D. Mehendale, I. V. Sethi, A. Ananthanarayanan, H. Faquih, P. Palit, S. Ravishankar, S. Sukumaran, T. Panchagnula, S. Murali, K. S. Gandhi, A. R, M. K. M, C. V. Vaijayanthi, K. S. R. Karunganni, P. Kumar, and M. M. Khapra, "Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages," 2024.

[2] N. R, M. S, J. F, A. Gangwar, M. N. J, S. Umesh, R. Sarab, A. K. Dubey, G. Divakaran, S. V. K, and S. V. Gangashetty, "SPRING-INX: A multilingual indian language speech corpus by SPRING lab, IIT madras," *CoRR*, vol. abs/2310.14654, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.14654

[3] A. Bhanushali, G. Bridgman, D. G, P. Ghosh, P. Kumar, S. Kumar, A. Raj Kolladath, N. Ravi, A. Seth, A. Seth, A. Singh, V. Sukhadia, U. S, S. Udupa, and L. V. S. V. D. Prasad, "Gram Vaani ASR Challenge on spontaneous telephone speech recordings in regional variations of Hindi," in *Proc. Interspeech 2022*, 2022, pp. 3548–3552.

[4] 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, K. Sankaranarayanan, T. Seeram, and B. Abraham, "Multilingual and code-switching asr challenges for low resource indian languages," *Proceedings of Interspeech*, 2021.

[5] T. Javed, K. S. Bhogale, A. Raman, P. Kumar, A. Kunchukuttan, and M. M. Khapra, "Indicsuperb: A speech processing universal performance benchmark for indian languages," in *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, B. Williams, Y. Chen, and J. Neville, Eds. AAAI Press, 2023, pp. 12 942–12 950. [Online]. Available: https://doi.org/10.1609/aaai.v37i11.26521

[6] T. Javed, S. Joshi, V. Nagarajan, S. Sundaresan, J. Nawale, A. Raman, K. S. Bhogale, P. Kumar, and M. M. Khapra, "Svarah: Evaluating english ASR systems on indian accents," *CoRR*, vol. abs/2305.15760, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.15760

[7] M. Chopra, I. Medhi Thies, J. Pal, C. Scott, W. Thies, and V. Seshadri, "Exploring crowdsourced work in low-resource settings," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–13. [Online]. Available: https://doi.org/10.1145/3290605.3300611

[8] AI4Bharat. (2023) Shoonya: An open source platform to annotate and label data at scale. [Online]. Available: https://github.com/AI4Bharat/Shoonya

[9] S. Team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," https://github.com/snakers4/silero-vad, 2021.

[10] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W. Hsu, A. Conneau, and M. Auli, "Scaling speech technology to 1, 000+ languages," *CoRR*, vol. abs/2305.13516, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.13516

[11] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[13] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang, Z. Meng, K. Hu, A. Rosenberg, R. Prabhavalkar, D. S. Park, P. Haghani, J. Riesa, G. Perng, H. Soltau, T. Strohman, B. Ramabhadran, T. N. Sainath, P. J. Moreno, C. Chiu, J. Schalkwyk, F. Beaufays, and Y. Wu, "Google USM: scaling automatic speech recognition beyond 100 languages," *CoRR*, vol. abs/2303.01037, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.01037

[14] Z. Peng, Z. Guo, W. Huang, Y. Wang, L. Xie, J. Jiao, Q. Tian, and Q. Ye, "Conformer: Local features coupling global representations for recognition and detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9454–9468, 2023. [Online]. Available: https://doi.org/10.1109/TPAMI.2023.3243048

[15] F. Boyer, Y. Shinohara, T. Ishii, H. Inaguma, and S. Watanabe, "A study of transducer based end-to-end ASR with espnet: Architecture, auxiliary loss and decoding strategies," *CoRR*, vol. abs/2201.05420, 2022. [Online]. Available: https://arxiv.org/abs/2201.05420

[16] Nvidia. (2023) Nvidia-en-ssl pretrained checkpoint. [Online]. Available: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/ssl_en_conformer_large

[17] K. S. Bhogale, S. Sundaresan, A. Raman, T. Javed, M. M. Khapra, and P. Kumar, "Vistaar: Diverse benchmarks and training sets for indian language ASR," *CoRR*, vol. abs/2305.15386, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.15386

[18] ULCA. (2022) English asr data. [Online]. Available: https://github.com/Open-Speech-EkStep/ULCA-asr-dataset-corpus

[19] Y. Madhani, S. Parthan, P. Bedekar, G. NC, R. Khapra, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "Aksharantar: Open indic-language transliteration datasets and models for the next billion users," in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, 2023, pp. 40–57. [Online]. Available: https://doi.org/10.18653/v1/2023.findings-emnlp.4

[20] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[2]https://www.meity.gov.in/

[3]https://www.cdac.in/index.aspx?id=pune