INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Vistaar: Diverse Benchmarks and Training Sets for Indian Language ASR

Kaushal Bhogale[1,2], *Sai Sundaresan*[2], *Abhigyan Raman*[2], *Tahir Javed*[1,2], *Mitesh M. Khapra*[1,2], *Pratyush Kumar*[1,2,3]

[1]Indian Institute of Technology Madras, India
[2]AI4Bharat, India
[3]Microsoft, India

cs22d006@cse.iitm.ac.in, {tahirjmakhdoomi,saisundaresan01,ramanabhigyan}@gmail.com, {miteshk,pratyush}@cse.iitm.ac.in

Abstract

Improving ASR systems is necessary to make new LLM-based use-cases accessible to people across the globe. In this paper, we focus on Indian languages, and make the case that diverse benchmarks are required to evaluate and improve ASR systems for Indian languages. To address this, we collate Vistaar as a set of 59 benchmarks across various language and domain combinations, on which we evaluate 3 publicly available ASR systems and 2 commercial systems. We also train IndicWhisper models by fine-tuning the Whisper models on publicly available training datasets across 12 Indian languages totalling to 10.7K hours. We show that IndicWhisper significantly improves on considered ASR systems on the Vistaar benchmark. Indeed, IndicWhisper has the lowest WER in 39 out of the 59 benchmarks, with an average reduction of 4.1 WER. We open-source all datasets, code and models - https://github.com/AI4Bharat/vistaar

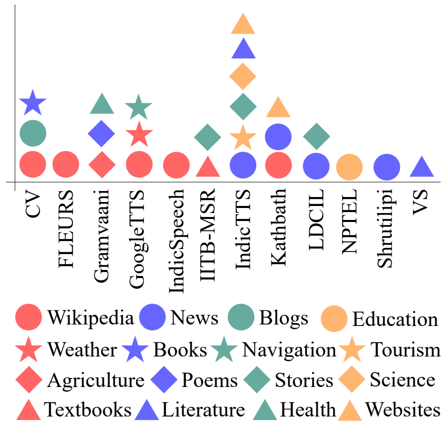

Figure 1: *The diversity domains/types covered by the Vistaar benchmark. (CV: CommonVoice; VS: Vakyasancayah)*

1. Introduction

Large Language Models such as GPT3[1] have demonstrated emergent behaviors which unlock new use-cases for deployment of language AI. For instance, an Indian villager could interact with a LLM-based system to learn about rights and entitlements on existing government schemes. To make such use-cases accessible to a large population, it is essential to make them voice-enabled, which requires support for speech recognition models across languages. Specifically, for Indian languages, given the large linguistic diversity (60 languages with over a million speakers) and the large print illiterate population (over 300 million people [2]), accurate ASR systems have significant societal impact.

Improvement of Machine Learing systems like ASR are dependent on various choices of training architectures, algorithms, datasets, and metrics. Systematic evaluation of these choices is critically dependent on representative benchmarks. Several works, over many years, have contributed publicly available benchmarks of specific languages and specific domains/types of data. For instance, Gramvaani [3], a Hindi dataset, contains data from agriculture domain; Vakyasancayah [4], a Sanskrit dataset, contains data from literature domain. Often these benchmarks are accompanied by leaderboards which earmark performance of different ASR systems. We make the case that such narrow comparison incentivizes model optimization to over-fit for the benchmark's characteristics. For instance, tuning of language models to the dev set of benchmarks have been shown to reduce WER scores by over 6.5 points averaged across 8 languages [5]. This suggests that narrow benchmarking may incentivize such over-fitting and thereby preclude more robust comparison of models.

To address some of these challenges, we propose collation of benchmarks across languages and domains/types of data. We call this Vistaar (meaning broad in Hindi) and it comprises of publicly available benchmarks across 12 languages, leading to 59 computed WER values across benchmarks and languages. We show in Figure 1 the diversity of the languages and domains/types covered by these benchmarks. We will host these datasets in a public resource and release source-code to evaluate an ASR system across these benchmarks.

We then evaluate various ASR systems on Vistaar. Specifically, we consider 3 publicly released ASR systems and 2 commercial systems from Google and Microsoft. We report the results across models and find large variance across benchmarks, languages, and models. Specifically, we show that comparing models based on one benchmark may be highly uncorelated with results on another benchmark. We hope that the Vistaar benchmark, and easy access to it, enables robust evaluation of ASR systems.

Based on our insight that ASR systems' performance varies across domains, we recognize value in training ASR systems with a wide set of training sets. To this end, we curate all publicly available training sets for 12 Indian languages amounting over 10,700 hours of audio. We call this the Vistaar-train dataset. We train the Whisper ASR model [6] with Vistaar-train for each of the 12 languages, to create a family of IndicWhisper models. When evaluated on Vistaar benchmark, we find that IndicWhisper models have the lowest average WER across benchmarks for each of the 12 languages. Indeed, IndicWhipser has the lowest WER in 39 out of the 59 benchmarks with an average reduction of 4.1 WER. We specifically compare IndicWhisper

10.21437/Interspeech.2023-2589

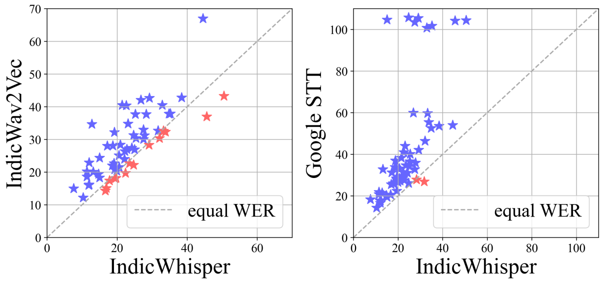

Figure 2: ***WER Comparison on Vistaar:*** *IndicWhisper is better than IndicWav2Vec in 45 out of 59 benchmarks (highlighted by purple) and better than Google STT in 57 out of 59 benchmarks.*

to one publicly available (IndicWav2Vec) and one commercial (Google) ASR system in Figure 2. As shown, the reported WER is lower in IndicWhisper for most of the benchmarks, with large gaps for some of the benchmarks. This validates our proposition that to both evaluate and train ASR systems we need diversity in datasets.

The rest of the paper is organized as follows. We detail all benchmark datasets in Vistaar in Section 2. In Section 3 we detail the training dataset and the procedure used to train IndicWhisper. We compare various public and commercial ASR systems along with IndicWhisper on Vistaar in Section 4. We end with conclusions and our thought-process on further improvements towards building open Indic ASR systems.

2. Vistaar Benchmark Set

In the following we describe various publicly available datasets which we curate as part of the Vistaar benchmark set.

Kathbath [7] This contains *read speech* collected on Android phones with the Karya application [8] by AI4Bharat, where the sentences are sourced from Wikipedia and news articles. This covers all the 12 languages we consider.

Kathbath-Hard [7] We create a hard benchmark for by adding background noise of various types to the audio files of the Kathbath benchmark. We use ESC dataset[9], which consists of 2,000 short clips of background noise from 5 different categories. For each audio, we randomly pick a background clip and add it to the audio signal with a random Signal-to-Noise Ratio (SNR) value between 3 dB and 30 dB.

FLEURS [10] This contains *read speech of translated Wikipedia content* with 3 recordings by different speakers for a sentence and manual validation collected by researchers at Meta and their collaborators. This has a large language coverage with 11 of the 12 languages considered, excluding Sanskrit.

CommonVoice [11] This contains *crowd-sourced read speech* from the popular Commonvoice website of Mozilla Foundation. The exact source of the sentences is not documented but it seems to contain sentences from news, Wiki articles, stories, and literature. This covers 8 languages.

IndicTTS [12] This contains *studio-quality read speech* by professional speakers who are trained to maintain constant pitch and prevent stress phenomenon. Content is sourced from newspapers, websites, blogs, children stories, and tourism and is chosen to consider common day-to-day usage and syllable coverage. This covers 9 languages.

MUCS [13] This contains *read speech* collected by different speakers from high-literacy and semi-literate group, and the text is collected from storybooks. This covers 6 languages.

GramVaani [3] This contains *telephone quality speech data* with specific focus on regional/dialectical variations of Hindi using the Uliza crowdsourcing platform by the farmer-centric NGO GramVaani.

The above benchmarks datasets represent a wide variety. The variation is across several axes - source content, speakers, audio equipment, and collection agencies. And the diversity of the benchmarks is evident from a statistic that we obtained when we compared ASR systems on these benchmarks. The Spearman correlation of the 59 WER values of the 3 ASR systems (IndicWhisper, IndicWav2Vec and Google STT) on the 7 benchmarks were computed pair-wise. Certain pairs of benchmarks showed negative correlation. For instance, CommonVoice and IndicTTS had a negative correlation of -0.26 indicating that the ranking of models on CommonVoice is not informative of ranking of models on IndicTTS. This low correlation is perhaps attributable to the diversity in content and speakers of the two benchmarks - while CommonVoice is crowdsources, IndicTTS is collected from professional speakers in a studio environment.

3. Vistaar-Train Dataset and IndicWhisper Model Training

In this section, we detail the Vistaar-Train dataset, and then architecture and training methodology to train the IndicWhisper family of models.

3.1. Vistaar-Train Set

As discussed in the previous sections, results in correlating the performance of a set of ASR models on a diverse set of benchmarks. Based on this, we hypothesize that ASR models must also be trained on a diverse set of training data. To this end, we first put together Vistaar-Train set which curates 13 publicly released datasets. In the following, we detail those datasets which are included in Vistaar-Train and not already discussed in the earlier section.

Shrutilipi [14] This contains *read speech* collected by mining audio and text pairs from news bulletins aired on All India Radio. This covers all the 12 languages we consider.

NPTEL This contains *conversational speech* from classroom lecture recordings of undergraduate and graduate level engineering courses. The audio files were recorded using a lapel microphone. This covers 8 languages. The recordings are available as videos (most of them are over 30 min long), and the transcripts are available in PDF format. We use the recently proposed document-level alignment technique [14] to get sentence level transcriptions. We will release this dataset in a public resource.

IISc-MILE [15] This contains *read speech* recorded in a *clean, noise-free* environment using USB microphones. Tamil and Kannada data were recorded from 531 and 915 native speakers respectively.

IITB-MSR [8] This contains *read speech* in Marathi recorded from three user groups: (i) low-income rural village, (ii) low-income urban slums, and (iii) university students. It is collected on Android phones with the Karya [8] application, where the sentences are sourced from Marathi textbooks.

Vakyasancayah [4] This contains *read speech* in Sanskrit collected on Android phones with the Audacity platform. The text was sourced from pre-classical, classical, and modern Sanskrit literature.

GoogleTTS [16, 17] This contains *read speech* recorded in a

Table 1: *The Vistaar Benchmark Set - 59 benchmarks over 12 Indian languages*

Datasets	bn	gu	hi	kn	ml	mr	or	pa	sa	ta	te	ur
Kathbath	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Kathbath-Hard	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
FLEURS	☑	☑	☑	☑	☑	☑	☑	☑		☑	☑	☑
CommonVoice	☑		☑		☑	☑	☑	☑		☑		☑
IndicTTS	☑	☑	☑	☑	☑	☑	☑			☑	☑	
MUCS		☑	☑			☑	☑			☑	☑	
Gramvaani			☑									

Table 2: *The Vistaar-Train Dataset With Hours of data across Datasets and Languages*

Datasets	bn	gu	hi	kn	ml	mr	or	pa	sa	ta	te	ur	Total
Shrutilipi	443	460	1620	459	359	1015	601	94	27	794	390	193	6457
Kathbath	103	116	137	153	134	172	99	124	102	172	142	74	1527
CommonVoice	64	-	13	-	1	19	2	2	-	226	-	46	373
NPTEL	44	59	157	80	60	68	-	-	-	206	183	-	857
IISC-mile	-	-	-	347	-	-	-	-	-	150	-	-	497
MUCS	-	40	95	-	-	94	95	-	-	40	40	-	403
IndicTTS	20	21	20	19	18	19	19	-	-	20	36	-	192
IITB-msr	-	-	-	-	-	109	-	-	-	-	-	-	109
Gramvaani	-	-	100	-	-	-	-	-	-	-	-	-	100
FLEURS	11	9	7	8	10	12	3	6	-	9	8	7	90
Vakysancayah	-	-	-	-	-	-	-	-	78	-	-	-	78
Google TTS	5	8	-	8	6	3	-	-	-	7	6	-	43
IIITH-IndicSpeech	2	-	1	2	2	2	-	-	-	1	2	-	12
Total	691	712	2150	1077	590	1513	819	226	207	1625	806	320	10736

quiet room with a fanless laptop by Google employees. Organic sentences were hand-crafted from templates for weather forecasts and navigation. This covers 7 languages.

IIIT-IndicSpeech [18] This contains *studio-quality read speech*. A set of 1000 phonetically balanced sentences was selected from the Wikipedia dump of Indian languages released in 2008. This covers 7 languages.

In combination with the training components of the datasets in Vistaar, the Vistaar-train dataset contains a total of 10,736 hours of data across 12 languages, as detailed in Table 2. The size of data varies across languages with Hindi having the maximum of 2,150 hours of data and Sanskrit with 207 hours having the least. We believe that making this diverse dataset accessible to ASR researchers will speed up research progress.

3.2. IndicWhisper - Model Architecture and Training

To demonstrate improved ASR models on the Vistaar-train dataset, we had to make a choice of the model architecture. Given the improved performance shown by Whisper models [6] released by OpenAI, we chose the pretrained models and fine-tuned them on Vistaar-train. We made this choice based on results for Hindi with parts of the training data where Whisper-based models had substantially lower WER than any of the other model architectures.

We train one model per language and refer to this family as IndicWhisper. These models follow Transformer-based encoder-decoder architecture of the Whisper model. Each model is trained starting from the Whisper-medium model which has 769M parameters and 24 layers in both the encoder and decoder. It has been trained on 680,000 hours of multilingual data [6]. However, the representation of Indian languages in this data is low, and consequently the performance of the model on Indian languages is significantly poor [6].

We fine-tune the Whisper-medium model for each of the 12 languages using Vistaar-train. The audio files are re-sampled to 16KHz, mono-channel format, and an 80-channel log-magnitude Mel spectrogram representation is computed with a window size of 25ms and stride of 10ms. We use the same byte-level BPE multilingual text tokenizer used in Whisper[6], as it supports all the 12 languages that we consider. The multilingual tokenizer makes it possible to fine-tune on languages like Odia, although the Whisper[6] ASR model does not support it.

Whisper uses a multitask format to perform multiple speech processing tasks like multilingual speech recognition, speech translation, spoken language identification and voice activity detection. Specifically, it uses a sequence of input tokens to the decoder for specifying the language (eg. <|hi|>), the task (eg. <|transcription|>, <|translation|>) and whether to predict timestamps or not (<|notimestamps|>). Since the IndicWhisper model is trained specifically for speech recognition, we pass the <|transcription|> and <|notimestamps|> tokens along with the language token to force the decoder to predict the correct language.

4. Evaluation Results

In this section, we evaluate and compare ASR systems on the Vistaar benchmark.

Table 3: *Comparison of publicly-available models on the Hindi subset of the Vistaar benchmark*

Model	Kathbath	Kathbath-Hard	FLEURS	CommonVoice	IndicTTS	MUCS	Gramvaani	Average
Google STT	14.3	16.7	19.4	20.8	18.3	17.8	59.9	23.9
IndicWav2vec	12.2	16.2	18.3	20.2	15.0	22.9	42.1	21.0
Azure STT	13.6	15.1	24.3	14.6	15.2	15.1	42.3	20.0
Nvidia-medium	14.0	15.6	19.4	20.4	12.3	12.4	41.3	19.4
Nvidia-large	12.7	14.2	15.7	21.2	12.2	**11.8**	42.6	18.6
IndicWhisper	**10.3**	**12.0**	**11.4**	**15.0**	**7.6**	12.0	**26.8**	**13.6**

Table 4: *WER(%) of IndicWhisper on the Vistaar Benchmark. Values in bold indicate where IndicWhisper has the lowest WER.*

Datasets	bn	gu	hi	kn	ml	mr	or	pa	sa	ta	te	ur	avg
Kathbath	16.6	17.8	**10.3**	**19.3**	**34.8**	**19.9**	24.7	16.9	45.6	**24.2**	25.0	**11.9**	22.3
Kathbath Hard	19.4	**20.6**	**12.0**	**22.2**	**38.4**	**22.1**	29.1	19.7	50.5	**27.5**	27.8	**14.7**	25.3
CommonVoice	24.7	-	**11.4**	-	**44.5**	22.8	**35.2**	22.4	-	**29.2**	-	31.7	27.8
FLEURS	**20.9**	23.5	**15.0**	**18.6**	**22.6**	**20.5**	**32.9**	**23.1**	-	**25.2**	**25.4**	**19.2**	22.5
IndicTTS	**18.8**	**19.1**	**7.6**	**13.2**	**21.4**	**11.4**	**15.0**	-	-	**17.2**	33.8	-	17.5
MUCS	-	33.2	**12.0**	-	-	**12.8**	**27.5**	-	-	28.3	32.1	-	24.3
Gramvaani	-	-	**26.8**	-	-	-	-	-	-	-	-	-	26.8
Average	20.1	22.8	13.6	18.3	32.3	18.2	27.4	20.5	48.0	25.3	28.8	19.4	24.6

4.1. ASR systems compared

We compare the following 6 ASR systems - 3 are publicly available, 2 are commercial systems, and IndicWhisper that is proposed in this paper. We detail the 5 ASR systems, excepting IndicWhisper, in the following.

IndicWav2Vec [7] - The IndicWav2Vec models are Wav2vec Large models with 317M parameters. The models are trained on the Kathbath [7] dataset and support all 12 languages.

Nvidia-medium [19] It is a Conformer[20] Medium model with 30M parameters, trained on ≈1900 hours of Hindi speech, thus supporting only Hindi.

Nvidia-large [19] It is a Conformer[20] Large model with 120M parameters, trained on about 2900 hours of speech consisting of Hindi and English, thus supporting only Hindi.

Google STT[1] It is a paid API service with a cost of $1.44/hour, supporting all 12 languages.

Azure STT[2] It is a paid API service with a cost of $1/hour, supporting all languages except Odia, Punjabi, Sanskrit and Urdu.

4.2. WER Results

We report the WER results for each of the 6 ASR systems for the different benchmarks in Vistaar for Hindi in Table 3. On 6 of the 7 benchmarks for Hindi, IndicWhisper has the lowest WER. And on average, IndicWhisper has the lowest WER by a significant margin of 5 WER points over the Nvidia-large model. This is a significant improvement on WER values for the two commercial systems - Google (23.9) and Azure (20). Many of the models do not support all 12 languages covered under Vistaar. For instance, Azure model does not support Odia, Punjabi, Sanskrit and Urdu, while Nvidia models are only available for Hindi. We report the the performance of publicly available models on all benchmarks in Table 4. We show a comparison with IndicWav2Vec and Google in Figure 2. The WER across languages shows a large variation going from 13.6 in Hindi to 48 in Sanskrit. This indicates room for improvement by collecting larger datasets for languages with smaller resources such as Sanskrit, Malayalam, and Odia. On this entire set of 59 benchmarks, IndicWhisper has the lowest WER for 39, establishing a highly competitive benchmark for Indian language ASR.

[1] https://cloud.google.com/speech-to-text/

[2] https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text

4.3. Discussion

Training IndicWhisper on diverse training sets clearly improves on all compared ASR systems. The improvement covers a broad range of benchmarks and languages. Indeed, the large gap to commercially available APIs for speech recognition was surprising. This suggests that fine-tuning on Whisper-like encoder-decoder architectures trained on large amounts of weakly supervised data will likely bring large improvements in ASR systems for various languages. However, given the still large WERs for various languages, we see significant work required for improving Indian language ASR, which may include the following:

1. Curation of large (order tens of thousands of hours) corpora of audio for weakly supervised training.
2. Building generic acoustic models trained on 1. which are then combined with domain-specialized language models.
3. Creation of benchmarks with diversity across speakers, content, and collection methodologies, to evaluate models in 2.

5. Conclusion

We made the case that advancing IndicASR requires evaluation of different ASR systems on a diverse set of benchmarks covering languages and types/domains of data. The Vistaar benchmark was presented and used to compare various ASR systems. We also present the IndicWhisper models by finetuning OpenAI's Whisper models on the Vistaar-train set with over 10,000 hours on 12 Indian languages. IndicWhisper achieves significantly lower WER across a large set of benchmarks establishing state-of-the-art performance. However, the obtained WER results indicate further room for improvement, and we outlined potential directions of research.

6. References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] C. Chandramouli and R. General, "Census of india 2011," *Provisional Population Totals. New Delhi: Government of India*, pp. 409–413, 2011.

[3] A. Bhanushali, G. Bridgman, D. G, P. K. Ghosh, P. Kumar, S. Kumar, A. R. Kolladath, N. Ravi, A. Seth, A. Seth, A. Singh, V. N. Sukhadia, S. Umesh, S. Udupa, and L. V. S. V. D. Prasad, "Gram vaani ASR challenge on spontaneous telephone speech recordings in regional variations of hindi," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 3548–3552.

[4] D. Adiga, R. Kumar, A. Krishna, P. Jyothi, G. Ramakrishnan, and P. Goyal, "Automatic speech recognition in sanskrit: A new speech corpus and modelling insights," in *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, ser. Findings of ACL, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., vol. ACL/IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 5039–5050. [Online]. Available: https://doi.org/10.18653/v1/2021.findings-acl.447

[5] T. Javed, S. Doddapaneni, A. Raman, K. S. Bhogale, G. Ramesh, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "Towards building ASR systems for the next billion users," in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 2022, pp. 10 813–10 821. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/21327

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *CoRR*, vol. abs/2212.04356, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2212.04356

[7] T. Javed, K. S. Bhogale, A. Raman, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "Indicsuperb: A speech processing universal performance benchmark for indian languages," *CoRR*, vol. abs/2208.11761, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2208.11761

[8] B. Abraham, D. Goel, D. Siddarth, K. Bali, M. Chopra, M. Choudhury, P. Joshi, P. Jyothi, 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, LREC 2020, Marseille, France, May 11-16, 2020*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. European Language Resources Association, 2020, pp. 2819–2826. [Online]. Available: https://aclanthology.org/2020.lrec-1.343/

[9] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[10] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "FLEURS: few-shot learning evaluation of universal representations of speech," in *IEEE Spoken Language Technology Workshop, SLT 2022, Doha, Qatar, January 9-12, 2023*. IEEE, 2022, pp. 798–805. [Online]. Available: https://doi.org/10.1109/SLT54892.2023.10023141

[11] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. European Language Resources Association, 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520/

[12] A. Baby, A. L. Thomas, N. Nishanthi, T. Consortium *et al.*, "Resources for indian languages," in *Proceedings of Text, Speech and Dialogue*, 2016.

[13] A. Diwan, R. Vaideeswaran, S. Shah, A. Singh, S. R. K. M., S. Khare, V. Unni, S. Vyas, A. Rajpuria, C. Yarra, A. R. 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, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 2446–2450. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-1339

[14] K. S. Bhogale, A. Raman, T. Javed, S. Doddapaneni, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "Effectiveness of mining audio and text pairs from public data for improving ASR systems for low-resource languages," *CoRR*, vol. abs/2208.12666, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2208.12666

[15] M. A, B. Pilar, and R. A. G, "Subword dictionary learning and segmentation techniques for automatic speech recognition in tamil and kannada," 2022. [Online]. Available: https://arxiv.org/abs/2207.13331

[16] F. He, S.-H. C. Chu, O. Kjartansson, C. Rivera, A. Katanova, A. Gutkin, I. Demirsahin, C. Johny, M. Jansche, S. Sarin, and K. Pipatsrisawat, "Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems," in *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 6494–6503. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.800

[17] K. Sodimana, K. Pipatsrisawat, L. Ha, M. Jansche, O. Kjartansson, P. D. Silva, and S. Sarin, "A Step-by-Step Process for Building TTS Voices Using Open Source Data and Framework for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, Gurugram, India, Aug. 2018, pp. 66–70. [Online]. Available: http://dx.doi.org/10.21437/SLTU.2018-14

[18] N. Srivastava, R. Mukhopadhyay, K. R. Prajwal, and C. V. Jawahar, "Indicspeech: Text-to-speech corpus for indian languages," in *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. European Language Resources Association, 2020, pp. 6417–6422. [Online]. Available: https://aclanthology.org/2020.lrec-1.789/

[19] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook, P. Castonguay, M. Popova, J. Huang, and J. M. Cohen, "Nemo: a toolkit for building AI applications using neural modules," *CoRR*, vol. abs/1909.09577, 2019. [Online]. Available: http://arxiv.org/abs/1909.09577

[20] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 5036–5040. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-3015