IndicST: Indian Multilingual Translation Corpus For Evaluating Speech Large Language Models

Sanket Shah*†, Kavya Ranjan Saxena*§‡, Kancharana Manideep Bharadwaj†, Sharath Adavanne†, Nagaraj Adiga†
§IIT Kanpur, †Krutrim AI, Bangalore
§kavyars@iitk.ac.in, †{sanket.shah, kancharana.bharad, sharath.adavanne, nagaraj.adiga}@olakrutrim.com

***Abstract*—The integration of speech modalities into large language models, known as Speech LLMs, is a promising area of research for applications like automatic speech recognition (ASR) and automatic speech translation (AST). While several datasets exist for ASR, there is a critical gap for AST tasks in Indian languages. To fill this gap, we introduce IndicST, a new dataset tailored for training and evaluating Speech LLMs for AST tasks (including ASR and TTS), featuring meticulously curated, automatically and manually verified synthetic data. The dataset offers 10.8k hrs of training data and 1.13k hrs of evaluation data. Additionally, we present a scalable data collection methodology that allows easy development for other speech tasks. Our analysis examines the performance of various Speech LLMs on ASR and AST tasks, accompanied by insightful findings. We also explore the performance of these models across different prompts, highlighting the significant potential of our research in enhancing ASR and AST capabilities for Indian languages.**

***Index Terms*—Speech LLMs, Speech encoders, Automatic Speech Translation, Automatic Speech Recognition**

I. Introduction

Recent research on large language models (LLMs) has focused on integrating various modalities, including audio, speech, images, and their combinations. For inputs beyond text, different types of encoders are used depending on the input type, such as audio events [1], [2] or speech [3] or encoders of multiple input types [4] [5]. This work focuses on integrating the speech modality into large language models (Speech LLMs). Numerous studies have been conducted on the development of Speech LLMs [6]–[9]. One of the recent state-of-the-art models that performs both Automatic Speech Recognition (ASR) and Automatic Speech Translation (AST), along with other tasks, is SALMONN [10]. This model combines the Whisper [11] speech encoder and the BEATs [12] audio encoder. To connect the outputs of these dual encoders to the input space of the Vicuna LLM [13], SALMONN employs a window-level query Transformer (Q-Former) [14]. Additionally, a low-rank adaptation (LoRA) [15] is used to align the input and output spaces of the Vicuna LLM. All of these Speech LLMs are trained on datasets that may or may not include Indian languages for the ASR and AST tasks.

In this study, we focus on Speech LLMs to perform fundamental tasks such as ASR and AST for Indian languages. Our emphasis is currently on English and 8 Indian languages: Hindi (hi), Marathi (mr), Gujarati (gu), Bengali (bn), Tamil (ta), Telugu (te), Malayalam (ml), and Kannada (kn). To train and evaluate these models for ASR tasks, we utilize various datasets in these 9 languages, which contain transcriptions featuring diverse speakers [16] [17] [18]. However, there is a significant gap in the execution of AST tasks for Indian languages, primarily due to the lack of a specific translation dataset that includes audio for training and evaluating models. To address this issue, we release IndicST dataset that contains synthetically generated translations for Indian languages, making it suitable for training and assessing models designed for AST tasks. The details of data curation is explained in the following sections.

Additionally, we conduct experiments with Speech LLMs using various LLMs to identify the best-performing Speech LLM for ASR and AST tasks in Indian languages. Our findings indicate that the performance of Speech LLMs can fluctuate depending on the evaluation prompts used, which is consistent with earlier work [19]. Therefore, we provide a comprehensive analysis of the prompts utilized for ASR and AST tasks, specifically focusing on the best-performing Speech LLM for each Indian language.

The main contributions of this work are as follows:

- We introduce IndicST, a well-curated dataset specifically designed for training and evaluating Speech LLMs on AST tasks in English and 8 Indian languages.
- We present a comprehensive study on the impact of replacing various components of Speech LLMs for ASR and AST tasks in Indian languages.
- We compare single-stage and two-stage training paradigms of Speech LLMs for AST tasks.
- We also conduct an ablation study to assess the performance of the best Speech LLM using different prompts.

The IndicST dataset will be accessible here[1].

II. IndicST Dataset

We outline the steps to create IndicST and provide a detailed procedure for preparing the training and evaluation datasets. This dataset consists of English and 8 Indian languages, specifically tailored for the AST task.

A. Training Dataset at scale

1) ASR dataset: To pre-train the Speech LLMs, we utilize ASR data from 14 open-source datasets, which collectively contain approximately 10.8k hrs of audio in 9 different languages. Each dataset includes input speech audio along

*Equal contribution
‡This work is done during internship at Krutrim AI

[1]https://github.com/KavyaRSaxena/IndicST

TABLE I

SUMMARY OF ASR DATASETS FOR VARIOUS INDIAN LANGUAGES USED FOR TRAINING SPEECH LLM. THE DURATION IS MENTIONED IN K HRS.

Datasets	Languages									Duration
	en	hi	mr	gu	bn	ta	te	ml	kn	
Spring Labs [17]	✓	✓	✗	✗	✗	✗	✗	✗	✗	2.20
Common accent[2]	✓	✗	✗	✗	✗	✗	✗	✗	✗	0.010
MUCS [21]	✗	✓	✓	✗	✓	✗	✗	✗	✗	0.22
CMU [22]	✗	✓	✓	✓	✓	✓	✓	✗	✓	0.06
CommonVoice [23]	✗	✓	✓	✗	✓	✓	✓	✓	✗	1.6
Gramavaani[3]	✗	✓	✗	✗	✗	✗	✗	✗	✗	0.095
Vaani[4]	✗	✓	✓	✓	✓	✓	✓	✗	✓	0.074
Lahaja [24]	✗	✓	✗	✗	✗	✗	✗	✗	✗	0.011
Shrutilipi [25]	✗	✓	✓	✓	✓	✓	✓	✓	✓	5.319
Google Corpus [26]	✗	✗	✓	✓	✗	✓	✓	✓	✓	0.034
Google Fleurs [27]	✗	✗	✓	✓	✓	✓	✓	✓	✓	0.087
Microsoft Speech Corpus[5]	✗	✗	✗	✓	✗	✓	✓	✗	✗	0.12
IISc MILE[6]	✗	✗	✗	✗	✗	✓	✗	✗	✓	0.45
IndicVoices [18]	✗	✓	✓	✓	✓	✓	✓	✓	✓	0.52
Duration	1.4	3.0	1.1	0.5	1.7	1.4	0.5	0.4	0.8	**10.8k hrs**

TABLE II

LANGUAGE-WISE DURATION (HRS) OF THE AUDIOS IN KATHBATH.

Languages							
hi	**mr**	**gu**	**bn**	**ta**	**te**	**ml**	**kn**
137.1	166.5	116.2	104.2	166.3	139.2	132.2	149.2

TABLE III

LANGUAGE-WISE DURATION (MINS) OF THE AUDIOS IN AI4BHARAT.

Languages								
en	**hi**	**mr**	**gu**	**bn**	**ta**	**te**	**ml**	**kn**
28.9	36.1	40.0	36.0	44.3	39.9	45.2	33.1	35.3

with the corresponding transcriptions. The statistics for these datasets are mentioned in Table I. We divide the dataset into three parts: train $D_{ASR}^{(tr)}$, validation, and test, with a split ratio of 8:1:1. The test split is referred to as Generic-ASR.

2) AST dataset: To synthetically generate translations for input speech audio and transcription pairs from the datasets mentioned above, we consider two translation directions: one-to-many, where English (source) transcription is translated to text in 8 Indian languages (target), represented as $en \longrightarrow X$, and many-to-one, where transcription in 8 Indian languages (source) is translated to English (target), represented as $X \longrightarrow en$. For each audio file in the source language, the translation in the target language is obtained by translating the corresponding transcript using IndicTrans2 [20], which is one of the state-of-the-art translation models for Indian languages. Therefore, we create 16 translation pairs from $\sim$10.8k hrs of audio. This constitutes the in-domain data of the IndicST dataset. We use the same train, validation and test split as in ASR dataset. For this task, the train and the in-domain test splits are referred to as $D_{AST}^{(tr)}$ and Generic-AST, respectively.

B. Evaluation Dataset at scale

1) ASR dataset: We evaluate the trained Speech LLM using two datasets. The first dataset, Kathbath [16], contains 1.68k hrs of input speech audio along with transcription pairs in 12 Indian languages, excluding English. For our analysis, we focus on $\sim$1.11k hrs of transcriptions corresponding to the input speech audio in $X = 8$ Indian languages. The language-wise statistics for this dataset is provided in Table II. The second dataset, Svarah [28], includes Indian-accented English speech data. This dataset comprises of 9.6 hrs of transcribed English audio from 117 speakers across 65 districts in 19 states of India, which results in a diverse range of accents.

2) AST dataset: There are two case scenarios from which we can synthetically generate an AST dataset for evaluation, described as:

a) The first scenario is to synthetically generate translations using IndicTrans2 when both input speech audio and transcription pairs are available. This method is already explained in Section II-A, which discusses the creation of training data for AST tasks. For evaluation purposes, we utilize the Kathbath and Svarah datasets. In the case of Kathbath, we consider the translation direction from X to English ($X \longrightarrow en$). For the Svarah dataset, we focus on the translation data obtained by IndicTrans2 in the direction from English to X ($en \longrightarrow X$).

b) The second scenario occurs when no input speech audio is available, and we only have text-to-text translation pairs. In this situation, we utilize StyleTTS2 [29], a multi-lingual text-to-speech synthesis model, to generate synthetic audio that corresponds to the text in the source language. This model is trained on a multilingual dataset and can be adapted for any number of speakers.

One example of a text-to-text translation dataset is AI4Bharat Conv [20] (Ai4B), which comprises of 1503 sentences across 22 Indic languages. We focus on 8 $en \longrightarrow X$ and 8 $X \longrightarrow en$ translation pairs. Thus, we synthesize the audio for the sentences in the source language using StyleTTS2. This results in 16×1503 synthesized pairs for nine languages. Additionally, we employ an automatic verification stage to assess the quality of the synthetic audio generation. For automatic verification, when the source language is English, we utilize synthesized English audio with the NVIDIA STT Conformer-CTC Large model (NC) from the NeMo toolkit [30] to generate transcriptions. This model is chosen because it is one of the best open-source options available. After obtaining the transcriptions from the NC model, we calculate the word error rate (WER) and select the top 500 transcription pairs with the lowest WER. In cases where the source language is Indic, we use the corresponding synthesized audio with the Data2Vec (DV) model [31] to produce transcriptions. Once we have the DV-based transcriptions, we select the top 500 transcription pairs with the lowest WER. For both English and 8 Indic languages, we select translation pairs corresponding to these 500 audio samples with the least WER. In total, we compile 16×500 translation pairs (8 pairs of $en \longrightarrow X$ and 8 pairs of $X \longrightarrow en$). Further, we perform manual verification by selecting 1 annotator per bi-direction translation for each Indian language. The annotators correct the translation text by manually identifying the inaccurate translation pairs. The language-wise statistics for this dataset is provided in Table III. Therefore, we create 16 translation pairs from $\sim$5.6 hrs of audio.

The synthetically generated datasets in both scenarios serve as out-of-domain test data for the IndicST dataset. By using either a speech-transcription pair or a text-to-text translation pair, we can effectively scale these methods to synthesize large volumes

[2]https://huggingface.co/datasets/DTU54DL/common-accent

[3]https://openslr.org/118/

[4]https://vaani.iisc.ac.in/#Data

[5]https://www.microsoft.com/en-us/download

[6]https://www.openslr.org/126/

of data through translation or speech synthesis models. Furthermore, these techniques can produce synthetic data suitable for various downstream speech tasks, such as speech question answering (SQA) and speech entity recognition, among others.

III. Exploring Speech LLMs

Current speech LLMs are developed by integrating speech encoders (SE) with pre-trained text-based LLMs. In this context, we focus on the Whisper family, specifically the *whisper-large-v2* model. This model is a sequence-to-sequence transformer that has been trained on a variety of speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. It has been trained on 680k hrs of labeled data. For ASR tasks, the model is trained on both multilingual and English-only datasets. The AST task, denoted as $X \longrightarrow en$, relies solely on multilingual data that includes X languages. The subset of languages in this multilingual data comprises a limited number of hours of recordings across all 8 Indic languages. For our purposes, we utilize the encoder from the *whisper-large-v2* model (denoted as whisperSE) and do not use the decoder.

We fix the whisperSE and examine two different families of LLaMA models [32]: (1) Vicuna, which contains 13B parameters [13], and (2) the instruction-tuned LLaMA 3.1[7], which contains 8B parameters. These LLMs are pre-trained and are not fine-tuned on our dataset. Both models are based on the LLaMA architecture. The Vicuna-13B model is fine-tuned from LLaMA through supervised instruction fine-tuning using a dataset collected from ShareGPT, while the LLaMA 3.1 model is an instruction-tuned variant. We also conducted experiments using mHuBERT [33] as the speech encoder (SE) alongside the best-performing LLM. Our findings show that the whisperSE outperforms the mHuBERT speech encoder with the LLM combination. Consequently, we have chosen to use only whisperSE as the speech encoder.

IV. Experimental details

To study the ASR and AST performance on the Indian languages, we consider different training paradigms (TPs) mentioned as:

1) **TP1:** In this experiment, we pre-train the Speech LLMs on ASR-only data, i.e., on $D_{ASR}^{(tr)}$ and evaluate the results on in-domain Generic-ASR data and out-of-domain test data, Svarah and Kathbath.

2) **TP2:** In this experiment, we pre-train the Speech LLMs on the AST-only data, i.e., on in-domain IndicST train data $D_{AST}^{(tr)}$ and evaluate the results on in-domain Generic-AST data and out-of-domain IndicST test dataset.

3) **TP3:** In this experiment, we perform Stage 1 pre-training on only ASR data $D_{ASR}^{(tr)}$ followed by a Stage 2 training on only AST data $D_{AST}^{(tr)}$ and evaluate on in-domain Generic-AST data and out-of-domain IndicST test dataset..

We consider the architecture of the Speech LLM to be similar to that of SALMONN [10]. It is important to note that we modify the Speech LLM models by replacing only the LLM, while keeping whisperSE and the BEATs audio encoder unchanged. Thus, model M1 is identical to SALMONN, whereas model M2 utilizes the instruction-tuned LLaMA 3.1 as its LLM, instead of Vicuna-13B. Both the M1 and M2 models are trained using TP1 and TP2. After comparing their results, we proceed to train the best-performing model with TP3. Each model across all training paradigms is trained for one epoch using the Adam optimizer. The learning rate begins with a warmup phase over the first 3k steps, starting from a base learning rate of 1×10^{-6} and gradually increasing to the initial learning rate of 3×10^{-5}. Following this warmup period, the learning rate decays gradually to a minimum of 1×10^{-5}. We compare these multilingual Speech LLM models with *whisper-large-v2* as the baseline. This baseline is chosen as the whisper model is trained with multilingual ASR and AST tasks. It is important to note that we assess the performance of these models on the audio in the test datasets that do not have LID (language identification) information. Therefore, to maintain a fair comparison, we calculate the performance of the baseline model without LID information. The performance metric considered for ASR tasks is WER, and for AST tasks is bilingual evaluation understudy (BLEU) score. The WER and BLEU score is computed by using standard jiwer[8] and nltk [34] library, respectively.

V. Results and Analysis

Table IV compares the WERs on in-domain Generic-ASR and out-of-domain Kathbath and Svarah test sets for different models trained using TP1. For the Generic-ASR dataset, the average WER across all 9 languages is 47.1% for model M1, 40.9% for model M2, while the baseline achieves about 87.8%. Additionally, model M2 outperforms model M1 on both the Kathbath and Svarah datasets by approximately 6% each. M2 also exceeds the baseline on the Kathbath dataset by about 60%, though its performance on the Svarah dataset is comparable to that of the baseline. Therefore, model M2 is identified as the most suitable Speech LLM for Indian languages. This performance analysis of ASR tasks across various languages also allows us to validate the data gathered from different sources for Indian languages.

The benchmarking of IndicST for translation tasks is presented in Table V. This table shows the average BLEU scores calculated for the translation directions *en* $\rightarrow$ *X* and *X* $\rightarrow$ *en*, covering both in-domain (Generic-AST) and out-of-domain (Kathbath, Svarah and AI4B) datasets. We have the following insights-

- The **Baseline** model achieves moderate performance with *X* $\rightarrow$ *en* average BLEU scores across the datasets, specifically scoring 10.9% on Generic-AST, 15.3% on Kathbath, and 18.1% on AI4B. The scores for *en* $\rightarrow$ *X* are not available for this model.
- The **M1 (TP2)** model performs better on Generic-AST, with an *en* $\rightarrow$ *X* average BLEU score of 21.6% and an *X* $\rightarrow$ *en* score of 24.5%, indicating improved performance over the Baseline.

[7]https://ai.meta.com/blog/meta-LLaMA-3-1/

[8]https://github.com/jitsi/jiwer

TABLE IV

PERFORMANCE METRIC (WER↓) WITH TP1 (ASR-ONLY) ACROSS DIFFERENT MODELS ON IN-DOMAIN GENERIC-ASR AND OUT-OF-DOMAIN SVARAH AND KATHBATH TEST SETS. ALL VALUES ARE IN PERCENTAGE.

Languages	**Models**								
	Baseline			**M1 (TP1)**			**M2 (TP1)**		
	Generic-ASR	**Svarah**	**Kathbath**	**Generic-ASR**	**Svarah**	**Kathbath**	**Generic-ASR**	**Svarah**	**Kathbath**
en	23.3	**25.6**	–	17.7	32.0	–	**16.5**	26.4	–
hi	63.7	–	44.5	34.3	–	14.6	**27.3**	–	**9.9**
mr	99.7	–	91.0	29.5	–	31.9	**24.2**	–	**29.7**
gu	109.4	–	109.9	56.3	–	34.2	**41.3**	–	**25.9**
bn	116.6	–	110.9	69.4	–	**26.8**	**63.2**	–	26.9
ta	66.6	–	59.1	**37.1**	–	39.3	38.0	–	**34.6**
te	111.3	–	112.7	75.4	–	51.1	**68.5**	–	**37.1**
ml	111.7	–	117.5	47.6	–	47.2	**47.4**	–	**46.6**
kn	87.7	–	82.4	56.90	–	44.2	**42.1**	–	**30.4**

TABLE V

PERFORMANCE METRIC (BLEU↑) WITH TP2 (AST-ONLY) AND TP3 (ASR + AST) ACROSS DIFFERENT MODELS ON IN-DOMAIN GENERIC-AST AND OUT-OF-DOMAIN SVARAH, KATHBATH, AND AI4BHARAT TEST SETS. ALL VALUES ARE IN PERCENTAGE.

Models	**Datasets**	$en \longrightarrow X$								$X \longrightarrow en$							
		hi	mr	gu	bn	ta	te	ml	kn	hi	mr	gu	bn	ta	te	ml	kn
Baseline	Generic-AST	–	–	–	–	–	–	–	–	16.9	13.1	10.7	7.7	11.0	7.7	11.9	8.1
	Svarah	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Kathbath	–	–	–	–	–	–	–	–	28.1	13.9	16.8	11.8	11.1	12.8	17.6	10.1
	AI4B	–	–	–	–	–	–	–	–	28.8	17.1	19.3	19.7	14.5	17.1	15.7	12.7
M1 (TP2)	Generic-AST	30.2	19.9	25.1	24.4	18.5	19.0	16.7	18.8	29.2	32.4	30.0	13.0	24.2	14.6	29.0	23.8
	Svarah	20.9	10.6	14.9	14.5	7.9	10.2	7.4	11.5	–	–	–	–	–	–	–	–
	Kathbath	–	–	–	–	–	–	–	–	36.6	22.3	25.3	20.8	17.7	19.0	22.0	15.9
	AI4B	8.8	3.8	7.2	5.3	0.9	1.9	0.6	0.8	26.2	18.9	19.5	21.4	14.7	16.3	15.9	12.1
M2 (TP2)	Generic-AST	35.6	22.1	29.0	27.8	21.6	25.0	20.0	**23.9**	**31.0**	32.0	30.3	14.7	**24.6**	15.0	29.6	24.2
	Svarah	**28.9**	**15.1**	17.7	**19.2**	11.0	14.2	**10.6**	11.0	–	–	–	–	–	–	–	–
	Kathbath	–	–	–	–	–	–	–	–	37.2	23.9	25.1	20.6	17.2	19.1	22.4	16.8
	AI4B	13.4	6.9	9.5	6.3	1.6	2.1	1.2	1.2	**26.7**	19.2	**19.4**	**22.1**	**14.7**	**17.4**	16.0	**13.0**
M2 (**TP3**)	Generic-AST	**37.0**	**22.6**	**30.8**	**28.6**	**23.0**	**25.4**	**20.6**	23.7	30.2	**33.0**	**32.3**	**15.4**	24.4	**16.2**	**30.5**	**26.2**
	Svarah	23.9	14.7	**19.3**	18.9	**11.8**	**14.5**	10.1	**15.2**	–	–	–	–	–	–	–	–
	Kathbath	–	–	–	–	–	–	–	–	**38.0**	**24.2**	**25.6**	**22.3**	**18.4**	**20.2**	**22.5**	**17.3**
	AI4B	**14.9**	**7.3**	**11.7**	**8.7**	**1.6**	**2.9**	**1.2**	**1.3**	26.1	**19.6**	18.8	21.2	14.0	17.1	**16.5**	12.9

- The **M2 (TP2)** model shows incremental improvement, with the highest average scores across both translation directions on Generic-AST and competitive results on Kathbath and AI4B.
- Lastly, the **M2 (TP3)** model exhibits the highest scores in both translation directions, demonstrating significant improvements across datasets, with an *en* $\rightarrow$ *X* score of 26.5% on Generic-AST, 16.0% on Svarah, and 6.2% on AI4B, as well as an *X* $\rightarrow$ *en* score of 26.0% on Generic-AST, 23.6% on Kathbath, and 18.3% on AI4B.

For both ASR and AST tasks, the superiority of M2 over M1 can be attributed to their different training backgrounds. M1 uses Vicuna-13B and leverages the pre-trained LLaMA-13B, which was trained on $\sim$1.4 T tokens and subsequently fine-tuned on conversation data sourced from ShareGPT. In contrast, M2, referred to as instruction-tuned LLaMA3.1, utilizes the pre-trained LLaMA3.1-8B, trained on $\sim$15 T tokens and further fine-tuned on publicly available instruction datasets. Also, the two-stage training paradigm (TP3), where pre-training on ASR data precedes training on AST data, achieves the highest BLEU scores across both directions. This strategy benefits from ASR pre-training in Stage 1, boosting both ASR and AST performance when followed by AST fine-tuning in Stage 2. TP3 achieves superior translation accuracy across in-domain and out-of-domain datasets, particularly noted in the consistent improvement across both *en* $\rightarrow$ *X* and *X* $\rightarrow$ *en*.

We also examine the how the model familiarity with prompts effect its performance in ASR and AST tasks. In the ASR task (TP1 with M2 model), we tested Hindi language from Kathbath dataset using the prompt "Recognize the speech and give me the transcription", which is included in the training data, achieved a lower WER of 9.5%. In contrast, the prompt "Give me the transcription", which is not part of the training data, resulted in a WER of 9.9%. Similarly, in the AST task (TP3), we tested for Hindi-to-English translation pair from Kathbath. The prompt "Translate the speech in the given audio to English" which the model encountered during training, produced a slightly higher BLEU score of 38.6. On the other hand, the unseen prompt "Translate to English" achieved a BLEU score of 38.0. These findings suggest that familiar prompts can marginally improve the performance of ASR and AST tasks with Speech LLMs. Mitigating this prompt bias could be a direction for future research.

VI. CONCLUSION

In this work, we present IndicST, a dataset that includes translation pairs between 8 Indian languages and English. This dataset was designed for training and evaluation of Speech LLM for AST task. We also present a scalable data creation methodology for training various speech-related applications. We evaluate the performance of the IndicST dataset utilizing a SALMONN-based model architecture with the LLaMA 3.1 language model instead of the Vicuna LLM to attain the best translation results as evaluated by the BLEU score. Future work will extend the proposed methodology to include SQA tasks.

REFERENCES

[1] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," *arXiv preprint arXiv:2305.10790*, 2023.

[2] C. Lyu, M. Wu, L. Wang, X. Huang, B. Liu, Z. Du, S. Shi, and Z. Tu, "Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration," *arXiv preprint arXiv:2306.09093*, 2023.

[3] F. Chen, M. Han, H. Zhao, Q. Zhang, J. Shi, S. Xu, and B. Xu, "X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages," *arXiv preprint arXiv:2305.04160*, 2023.

[4] Y. Su, T. Lan, H. Li, J. Xu, Y. Wang, and D. Cai, "Pandagpt: One model to instruction-follow them all," *arXiv preprint arXiv:2305.16355*, 2023.

[5] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[6] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. d. C. Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov *et al.*, "Audiopalm: A large language model that can speak and listen," *arXiv preprint arXiv:2306.12925*, 2023.

[7] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," in *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[8] S. Hu, L. Zhou, S. Liu, S. Chen, H. Hao, J. Pan, X. Liu, J. Li, S. Sivasankaran, L. Liu *et al.*, "Wavllm: Towards robust and adaptive speech large language model," *arXiv preprint arXiv:2404.00656*, 2024.

[9] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[10] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," *arXiv preprint arXiv:2310.13289*, 2023.

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[12] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," in *International Conference on Machine Learning*. PMLR, 2023, pp. 5178–5193.

[13] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org*, vol. 2, no. 3, p. 6, 2023.

[14] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[15] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[16] T. Javed, K. Bhogale, A. Raman, P. Kumar, A. Kunchukuttan, and M. M. Khapra, "Indicsuperb: A speech processing universal performance benchmark for indian languages," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 12 942–12 950.

[17] A. Gangwar, S. Umesh, R. Sarab, A. K. Dubey, G. Divakaran, S. V. Gangashetty *et al.*, "Spring-inx: A multilingual indian language speech corpus by spring lab, iit madras," *arXiv preprint arXiv:2310.14654*, 2023.

[18] T. Javed, J. A. Nawale, E. I. George, S. Joshi, K. S. Bhogale, D. Mehendale, I. V. Sethi, A. Ananthanarayanan, H. Faquih, P. Palit *et al.*, "Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages," *arXiv preprint arXiv:2403.01926*, 2024.

[19] B. Wang, X. Zou, G. Lin, S. Sun, Z. Liu, W. Zhang, Z. Liu, A. Aw, and N. F. Chen, "Audiobench: A universal benchmark for audio large language models," *arXiv preprint arXiv:2406.16020*, 2024.

[20] J. Gala, P. A. Chitale, R. AK, V. Gumma, S. Doddapaneni, A. Kumar, J. Nawale, A. Sujatha, R. Puduppully, V. Raghavan *et al.*, "Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages," *arXiv preprint arXiv:2305.16307*, 2023.

[21] A. Diwan, R. Vaideeswaran, S. Shah, A. Singh, S. Raghavan, S. Khare, V. Unni, S. Vyas, A. Rajpuria, C. Yarra *et al.*, "Multilingual and code-switching asr challenges for low resource indian languages," *arXiv preprint arXiv:2104.00235*, 2021.

[22] K. Prahallad, N. K. Elluru, V. Keri, S. Rajendran, and A. W. Black, "The iiit-h indic speech databases." in *Interspeech*, 2012, pp. 2546–2549.

[23] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[24] T. Javed, J. Nawale, S. Joshi, E. George, K. Bhogale, D. Mehendale, and M. M. Khapra, "Lahaja: A robust multi-accent benchmark for evaluating hindi asr systems," *arXiv preprint arXiv:2408.11440*, 2024.

[25] K. 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," in *Proc. ICASSP)*. IEEE, 2023, pp. 1–5.

[26] A. Butryna, S.-H. C. Chu, I. Demirsahin, A. Gutkin, L. Ha, F. He, M. Jansche, C. Johny, A. Katanova, O. Kjartansson *et al.*, "Google crowdsourced speech corpora and related open-source resources for low-resource languages and dialects: an overview," *arXiv preprint arXiv:2010.06778*, 2020.

[27] 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*, 2023, pp. 798–805.

[28] T. Javed, S. Joshi, V. Nagarajan, S. Sundaresan, J. Nawale, A. Raman, K. Bhogale, P. Kumar, and M. M. Khapra, "Svarah: Evaluating english asr systems on indian accents," *arXiv preprint arXiv:2305.15760*, 2023.

[29] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[30] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook *et al.*, "Nemo: A toolkit for building ai applications using neural modules.(2019)," *arXiv preprint arXiv:1909.09577*, 1909.

[31] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[32] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[33] M. Z. Boito, V. Iyer, N. Lagos, L. Besacier, and I. Calapodescu, "mhubert-147: A compact multilingual hubert model," *arXiv preprint arXiv:2406.06371*, 2024.

[34] S. Bird, "Nltk: the natural language toolkit," in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 2006, pp. 69–72.