Received 24 August 2023; revised 16 February 2024; accepted 7 March 2024. Date of publication 25 March 2024; date of current version 18 June 2024. The review of this article was arranged by Associate Editor E. Variani.

*Digital Object Identifier 10.1109/OJSP.2024.3379092*

# Lightweight, Multi-Speaker, Multi-Lingual Indic Text-to-Speech

**ABHAYJEET SINGH [1], AMALA NAGIREDDI [1], ANJALI JAYAKUMAR [1], DEEKSHITHA G [1], JESURAJA BANDEKAR [1], ROOPA R [1], SANDHYA BADIGER [1], SATHVIK UDUPA [1], SAURABH KUMAR [1], PRASANTA KUMAR GHOSH [1], HEMA A MURTHY [2], HEIGA ZEN [3], PRANAW KUMAR [4], KAMAL KANT [4], AMOL BOLE [4], BIRA CHANDRA SINGH [4], KEIICHI TOKUDA [5], MARK HASEGAWA-JOHNSON [6], AND PHILIPP OLBRICH [7]**

[1]Electrical Engineering Department, Indian Institute of Science (IISc), Bangalore 560012, India
[2]Department of Computer Science and Engineering, Indian Institute of Technology, Madras 600036, India
[3]Google, Tokyo 150-0002, Japan
[4]CDAC, Pune University Campus, Pune, Maharashtra 411007, India
[5]Department of Computer Science, Nagoya Institute of Technology, Nagoya 466-8555, Japan
[6]Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801 USA
[7]Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, 53113 Bonn, Germany

CORRESPONDING AUTHOR: DEEKSHITHA G (email: deekshu50@gmail.com; deekshitha@ieee.org).

This work was supported by Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) on behalf of the German Ministry for Economic Cooperation and Development for the creation of the dataset.

**ABSTRACT** The Lightweight, Multi-speaker, Multi-lingual Indic Text-to-Speech (LIMMITS'23) challenge is organized as part of the ICASSP 2023 Signal Processing Grand Challenge. LIMMITS'23 aims at the development of a lightweight, multi-speaker, multi-lingual Text to Speech (TTS) model using datasets in Marathi, Hindi, and Telugu, with at least 40 hours of data released for each of the male and female voice artists in each language. The challenge encourages the advancement of TTS in Indian Languages as well as the development of techniques involved in TTS data selection and model compression. The 3 tracks of LIMMITS'23 have provided an opportunity for various researchers and practitioners around the world to explore the state-of-the-art techniques in TTS research.

**INDEX TERMS** End-to-end model, data-constrained multi-speaker, model compression, multi-lingual TTS, speech synthesis, text-to-speech (TTS).

## I. INTRODUCTION

HIGH-quality speech data is necessary to build good speech synthesis systems. In a country such as India, which has 121 major languages [1], and many more variations and dialects, collecting high-quality data in large volume will require monumental effort and resources. There is a need for research in low-resource speech synthesis models that can cater to such language diversities.

A practical approach towards simulating a large volume high-quality of data is to combine low-resource data from multiple languages. This is enabled by the similarity between different languages due to their language groups. For example, languages such as Hindi, Marathi, Bengali, Gujarati, and Maithili belong to the Indo-Aryan group while languages such as Kannada, Tamil, Telugu etc belong to the Dravidian group. Additionally, many Indo-Aryan languages share the same written script - Devanagiri. Also, through multi-lingual training, a text-to-speech (TTS) system may benefit from speaker and accent transfer across languages.

State-of-the-art TTS models achieve good quality synthesis but at a cost of high complexity, and large size of models [2]. For example, the non-autoregressive TTS models can speed up the inference by leveraging parallel computation. However, the number of model parameters and the computation cost are high, which, in turn, makes the deployment of these large, high-performing models in mobile or embedded devices sub-optimal, as these devices don't have enough computation power. This demands a need for lightweight models that consume less computation resources and, in turn, speed up the inference. Survey [3] on neural synthesis model, reveals several studies implementing lightweight TTS models using knowledge distillation, neural architecture search,

quantization, pruning, compression using tensor decomposition, etc.

These have been works in the past on TTS in low-resource Indian languages. However, very little work has been done to develop a lightweight Indic TTS model, trained in multi-lingual scenarios. For example, the development of an end-to-end TTS model has been done in some works for selected Indian languages [4], [5]. Transfer learning techniques have been used to improve performance with low-resource data [6], [7]. Experiments have also been done with language adaptation across different languages and across language families [8], [9].

Lack of research in multi-speaker, multi-lingual TTS models in Indian languages originates the LIMMITS'23 challenge.[1] It calls for research on lightweight and data-constrained multi-speaker, multi-lingual TTS systems in Indian languages in a controlled environment with a large amount of high-quality data. These topics have a lot of practical utility and have not been addressed in Indian languages, potentially generalizable to various other languages around the world as well [10]. While there are other high-quality corpora in some of the LIMMITS'23 challenge languages including IndicTTS,[2] IndicSpeech,[3] they are not as large as the corpus being shared in this challenge. The corpus from this challenge is from the SYSPIN project, which is an initiative by the Indian Institute of Science (IISc), Bangalore to create open-sourced corpora in nine Indian languages including low-resourced ones.

The corpora shared as part of LIMMITS'23 allows research on multilingual TTS models in three Indian languages, namely Hindi, Marathi, and Telugu. We have selected these three languages based on the amount of data collected in the SYSPIN project at the time of the challenge. Our goal also was to take into consideration the geographical distribution of the languages across India and select the languages that can cover major parts of the country. This challenge enables the development of large-scale, multi-speaker, multi-lingual TTS models with 3 different problem statements through 3 challenge tracks enabling research on data-constrained and lightweight TTS systems. This challenge also helps to understand and compare various approaches to build TTS models and simultaneously share a platform to identify competitive TTS research groups across the world.

## II. CHALLENGE DATASET DESCRIPTION

The challenge corpora[4] include TTS datasets in Marathi, Hindi, and Telugu (1 male, and 1 female voice artist from each language) that are being collected as a part of the SYSPIN project.[5] Currently, speech data available for these three Indian languages is scarce, specifically for multi-speaker models as it is necessary to have data balanced across multiple speakers. Table 1 summarizes some of the publically available TTS corpora for Hindi, Marathi, and Telugu. As can be seen in the table most of them do not have speaker-related metadata available. Also, the gender ratio whenever reported, is very imbalanced.

**TABLE 1. Summary of TTS Data Available Across Different Languages**

| Corpus | Speaker details | Duration (hour) | | |
|---|---|---|---|---|
| | | Hindi | Marathi | Telugu |
| IIIT-H Voices [11] | NA | 1 | 2 | 1.5 |
| CMU wilderness [12] | Mostly males | 20 | 20 | 20 |
| IITM Indic TTS [13] | NA | 41.86 | 22.36 | 36.71 |
| OpenSLR multi speaker [14] | MR: 9 F<br>TE: 47 (24F+23M) | NA | 3 | 7 |

*NA: Not Available; MR: Marathi; TE: Telugu; F: Female; M: Male*

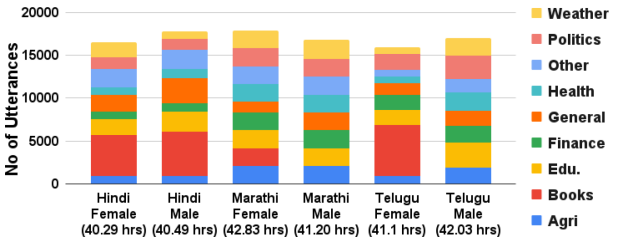


**FIGURE 1. Histogram showing the domain-wise distribution of sentences for each voice artist. The total duration for each voice artist is given in brackets.**

Unlike the existing TTS corpora in these languages, LIMMITS'23 datasets are unique in terms of balanced duration per voice artist and the variety of domains covered in the process of preparing the text. Sentences were mined from online sources as well as printed textbooks and included 9 different domains like agriculture, books, education, finance, general, health, politics, weather, and miscellaneous as shown in Fig. 1. The challenge corpus includes studio-quality audio with 48 kHz, 24 bits per sample from every voice artist. As it is clear from Fig. 1, the number of sentences varies across languages. It also varies between two voice artists within a language. This is because, at the time of recording an artist, the available mined sentences are used. For example, the Telugu male artist's recording was done before the Telugu Female. At that time, the sentences in the books category were not available. However, it became available during the recording of Telugu Male. The number of common sentences between male and female artists is 5425, 2587, and 8481 in the case of Hindi, Marathi, and Telugu, respectively. It is important to note that the scripts in these three languages are not identical. Fig. 2 shows the domain-wise distribution of the duration of all utterances.

The audio recordings in this dataset, while primarily matched to their respective text, might have some errors too. The types of errors present in the corpus are listed below -

[1] [Online]. Available: https://sites.google.com/view/syspinttschallenge2023/

[2] [Online]. Available: https://www.iitm.ac.in/donlab/tts/database.php

[3] [Online]. Available: http://cvit.iiit.ac.in/research/projects/cvit-projects/text-to-speech-dataset-for-indian-languages

[4] [Online]. Available: https://ee.iisc.ac.in/limmits23dataset/

[5] [Online]. Available: https://syspin.iisc.ac.in/

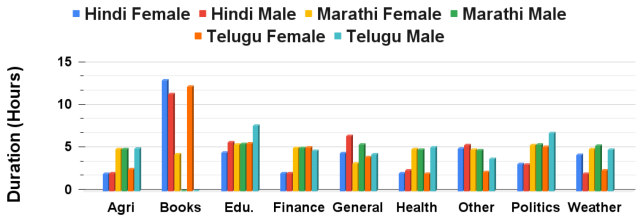


**FIGURE 2.** Histogram showing the domain-wise distribution of total duration of utterances in the dataset.

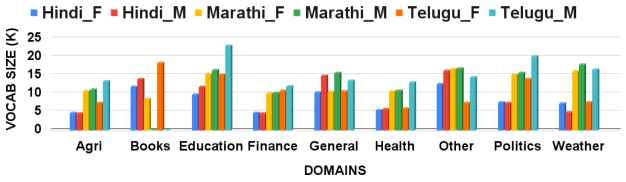


**FIGURE 3.** Histogram showing the domain-wise distribution of vocabulary size in the dataset.

**TABLE 2.** Summary of Challenge Dataset

| Language and Gender | Total No. of utterances | Total Duration (hrs) | Manual validation completed? | % Duration of matched Audio-Text pairs |
|---|---|---|---|---|
| Hindi F | 16512 | 40.29 | Yes | 86.61 % |
| Hindi M | 17800 | 40.49 | No | - |
| Marathi F | 17876 | 42.83 | No | - |
| Marathi M | 16748 | 41.20 | Yes | 65.38 % |
| Telugu F | 15950 | 41.09 | Yes | 91.05 % |
| Telugu M | 16939 | 42.03 | Yes | 100.00 % |

**TABLE 3.** Vocabulary Size and Utterance Duration and Sentence Length

| Voice Artist | Vocab size | Utterance duration in sec Avg (SD) | Utterance word count Avg (SD) |
|---|---|---|---|
| Hindi_F | 40069 | 8.78 (2.77) | 19.13 (5.75) |
| Hindi_M | 46384 | 8.19 (1.99) | 17.94 (4.08) |
| Marathi_F | 65722 | 8.63 (2.29) | 15.63 (4.23) |
| Marathi_M | 68668 | 8.86 (2.27) | 16.5 (3.89) |
| Telugu_F | 60355 | 9.28 (3.16) | 13.84 (4.76) |
| Telugu_M | 77070 | 8.93 (2.37) | 15.72 (3.71) |

- The audio is fine, but the text contains typographical errors.
- Audio and text do match, but there are problems in audio like the wrong pronunciation of words, the speaker speaking too fast, difficulty in understanding the speech, long pauses in between words, etc.
- Audio and text may match, but the sentence might be incomplete, or meaningless.
- The audio matches the text exactly but has distortion during the speech.
- The audio matches the text exactly but has distortion in the beginning and/or end silence only.
- The sentence is from a language, but not written in the script of that language (Hindi sentence written in Telugu script).

Table 2 summarises the details of the corpus. As the text was collected from various domains, data has a rich vocabulary set for all subjects. Table 3 shows the vocab size for all artists. It is interesting to note that, for all subjects, the number of utterances is almost the same. But, Telugu Male, Marathi Male, and Female have larger vocab sizes than the rest of the artists. Fig. 1, shows that the domain composition for these 3 speakers was the most evenly balanced, indicating that balancing of domains results in rich vocabulary. Fig. 3 presents the vocabulary size for each of the 9 domains. There is no domain-specific consistent pattern in vocabulary size. Table 3 also contains the average utterance duration (in seconds) per subject. It is clear that the utterance duration does not change significantly across voice artists. The average length of the sentence together with the standard deviation indicates that both short and long sentences are present in the corpus.

## III. CHALLENGE TRACKS

Three different tracks were designed as a part of the challenge. Each of the tracks demanded the participating teams develop one multi-speaker, multi-lingual TTS model. The evaluation procedure inspected the model's control over the speaker, over input text from different languages. The models developed were evaluated on mono-lingual pairs and cross-lingual combinations.

### A. TRACK 1: DATA SELECTION

Recent literature has shown that a limited amount of speaker-specific data may be sufficient to build high-quality speech synthesis models for multiple speakers and languages [10]. In this sub-track, at least 40 hours of data have been shared from each of the six voice artists, from which participants can use a maximum of 5 hours of data, from each voice artist, to train one multi-speaker, multi-lingual model. The goal is to identify subsets of data from the entire corpus, which can be used for multi-speaker, multi-lingual TTS training. The presence of parallel sentences across speakers in a language can also be exploited in this formulation.

With the advancements in Text-to-Speech techniques, there are several TTS models that perform well and can produce natural, human-like speech when trained on large volumes of data. However, these data-hungry approaches can't be used for low-resource languages. One way to resolve this is by employing data selection strategies on the available data. These strategies involve utterance level selection [15], speaker-based selection [16], utterance selection on ASR corpora [17], data selection with a loop of training and evaluation based on predicted quality of synthetic speech [18]. Some studies have also improved upon the naturalness of synthesized text-to-speech voices using data selection techniques [19], [20], [21]. Track 1 focuses on this aspect of data selection from a given dataset and will aid the TTS corpus creation and selection for low-resource languages. It might improve the naturalness of synthesized speech.

Several studies on multi-lingual speech synthesizers for Indian languages have taken 5 hours of speech per language for training [5], [8], [22]. In these studies, 5 h data is taken from a highly curated TTS corpus which enables TTS models

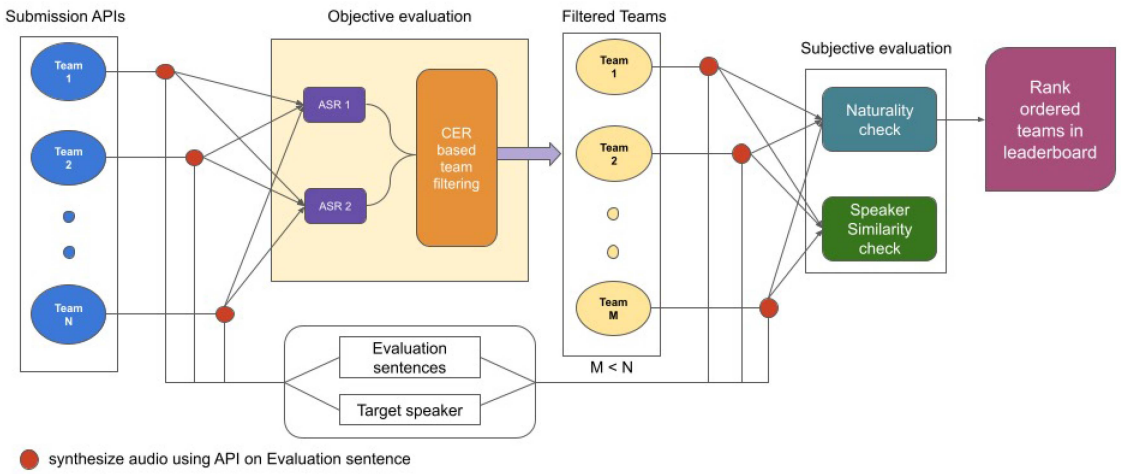


**FIGURE 4.** Flowchart showing the steps involved in submission followed by evaluation for leaderboard preparation.

to achieve decent MOS scores. Keeping 5 hours as a constraint, motivates effective data-selection strategies, especially in a multi-speaker, multi-lingual setup. In this setup, data-selection methods can reveal the best combinations of languages as well as the most conducive speaker attributes to achieve the best performance.

### B. TRACK 2: LIGHTWEIGHT TTS

The size of the TTS model is an important aspect to be considered while deploying such models for practical use. It would be ideal to incorporate a multitude of speakers and languages in a single model to lower the cost of hosting models. Towards this, track 2 has been proposed to build lightweight multi-speaker, multi-lingual TTS models, which can employ techniques such as model distillation, compression, lighter model architectures, etc. This track limits the TTS model (text to Mel spectrogram) to have 5 M usable parameters while a fixed vocoder is to be used as provided by the organizers.

There have been many efforts in the recent past towards building lightweight TTS systems. Light-TTS [10] used a novel model architecture consisting of dynamic convolution, low-rank approximation, and knowledge distillation to reduce model parameters to 1.4 M. SpeedySpeech [23] used a student-teacher network to achieve high-quality synthesis with an attention-free network of 4.3 M parameters. LightSpeech [24] uses neural architecture search to achieve 15X model compression, resulting in the final model with 1.8 M parameters. Nix-TTS [25] builds an end-to-end TTS system with 5.23 M using knowledge distillation. These previous works built models on a single-speaker dataset of LJSpeech. We fix the 5 M parameter limit for track 2 based on the progress made in the literature and accounting for multiple languages in the challenge dataset.

### C. TRACK 3: LIGHTWEIGHT MODEL DEVELOPMENT FROM BEST DATA

This track is a combination of Track 1 and Track 2, in which the participants have to build one multi-speaker, multi-lingual, lightweight speech synthesis model by utilizing at max 5 hours of data from each voice artist, and overall model (text to Mel spectrogram) size had to be less than 5 M. The participants must use the vocoder provided by the organizers.

## IV. EVALUATION CRITERIA AND METRICS

In the following section, we discuss the details of objective and subjective evaluation of the synthesized files from the participating teams, as summarised in Fig. 4.

We first finalize the evaluation sentences for each speaker, for mono-lingual, and cross-lingual synthesis setup. We use these evaluation sentences to synthesize audio using the submitted APIs from different teams. The teams are first filtered by an objective evaluation of the synthesized files. The subjective evaluation is performed on the submissions of the filter team. Finally, the leaderboard is rank-ordered according to the naturalness score obtained through subjective evaluation. These steps are followed for all three tracks.

### A. EVALUATION SENTENCES PREPARATION FOR AUDIO SAMPLE SYNTHESIS

The evaluation set contains 400 sentences (200- conversational, 100- grammatically incorrect, and 100- out-of-domain sentences), along with sentences from different input domains. These sentences are used to create monolingual and cross-lingual combinations as shown in Fig. 5. We give equal weightage to mono-lingual, as well as cross-lingual synthesis for the evaluation of teams. The cross-lingual pairs (Marathi to Hindi, Telugu to Marathi, and Hindi to Telugu) are predefined, and the same is maintained for all teams, and all tracks. 144 sentences are randomly picked and are used to synthesize audio samples for each submission. From Fig. 5, for each of the six blocks, 24 samples were allocated, 12 from each speaker. The text domains for each of the 12 sentences are distributed as follows - 6 from conversational, 3 from grammatically incorrect, and 3 from out-of-domain. With this formulation, 72 samples were used for mono-lingual evaluation and 72 samples were used for cross-lingual evaluation.

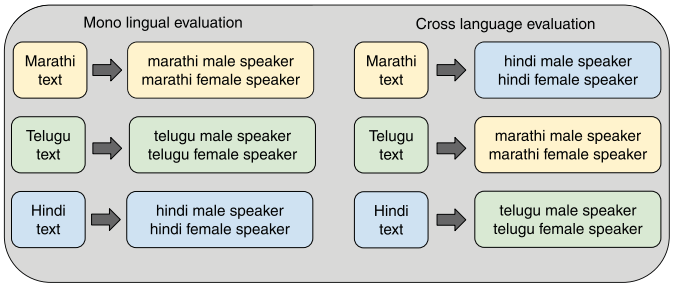


**FIGURE 5.** **Mono-lingual and cross-lingual sets for evaluation. For cross-lingual evaluation, we ensure pairs within and across language groups. For example, Marathi text - Hindi target speaker (within Indo-Aryan), Telugu text—Marathi target speaker (Dravidian to Indo-Aryan).**

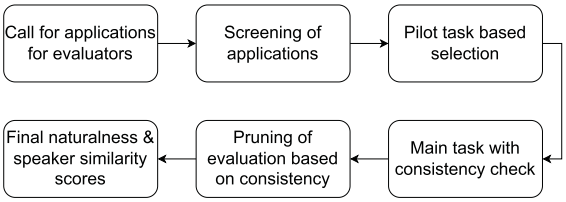


**FIGURE 6.** **Figure shows the steps for selection of validators for subjective and objective evaluation.**

### B. ASR-BASED OBJECTIVE EVALUATION

We have defined Automatic Speech Recognition (ASR) based objective metrics to filter the participating team submissions. For each language, we use two different high-performant ASRs to compute utterance-level character error rates (CER). For Hindi, we use two wav2vec2 fine-tuned models [26], [27]. For Telugu, we use an espnet model trained on CSTD corpus [28] and a wav2vec2 fine-tuned model [26]. For Marathi, we use two wav2vec2 models fine-tuned on [14]. We compute the final global CER for submission as follows:

1) Compute utterance level CER from both ASRs for the corresponding language.
2) Store the Selection, Insertion and Deletion errors for the lower CER (best performing)
3) Perform steps 1 and 2 for all utterances to be considered for evaluation
4) Use the stored utterance level Selections, Insertions, and Deletions to compute a global CER which represents a submission's objective score

We computed the global CER for all teams, in each track and all the submitted teams ended up being selected for the next phase of subjective evaluation.

### C. SUBJECTIVE EVALUATION

For performing the subjective evaluation on the synthesized files in a short span, we required a good panel of experts in each language. To reach out to the experts, we advertised on different platforms and shortlisted the best evaluators with the help of 3 rounds of pilot tasks as shown in Fig. 6.

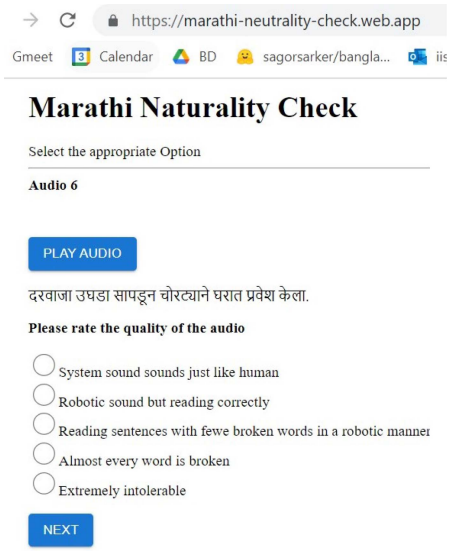


**FIGURE 7.** **Screenshot of web app used for subjective evaluation- *Naturality check* The synthesized audio is shown, along with the reference text.**

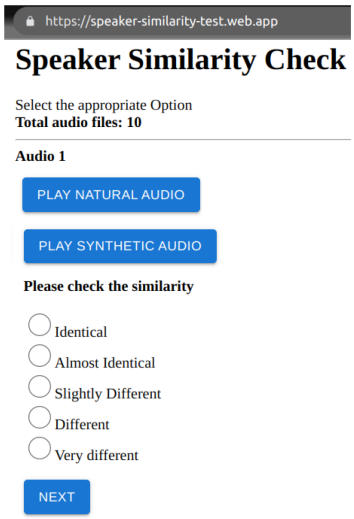


**FIGURE 8.** **Screenshot of webapp used for subjective evaluation- *Speaker similarity check.* The synthesized audio is shared, along with the reference audio for the target speaker.**

Separate subjective evaluations have been done for naturality and speaker similarity checking using the evaluation sentences. These evaluations have been conducted on web applications built with *react* and deployed on *Firebase*. Figs. 7 and 8 show the screenshots of the web apps created for naturality and speaker similarity checks respectively. The final listeners for evaluation were selected based on multiple rounds of pilot listening tests, based on the scores on the synthesized audios of varying quality. Separate pilot tests have been conducted for naturality and speaker similarity-checking tasks. All sentences from all teams were randomly shuffled

**TABLE 4. Challenge Timeline**

| | |
|---|---|
| Nov 26, 2022: | Registration for the challenge opened |
| Dec 04, 2022: | Release of training data |
| Dec 15, 2022: | Release of baseline recipe |
| | Webapp designing for subjective evaluation |
| Jan 7, 14, 21, 2023: | Pilots for evaluators selection |
| Jan 15, 2023: | API submission begins |
| Jan 20, 2023: | Last date for submission of a working API |
| Jan 30, 2023: | Final API submission deadline |
| Feb 10, 2023: | Announcement of winners in all tracks |
| Feb 20, 2023: | Grand Challenge 2-page papers due |

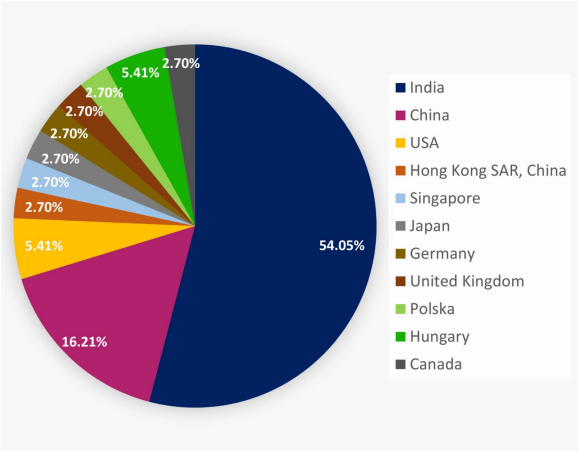


**FIGURE 9. Distribution of registrants across the globe who downloaded the data.**

and provided to evaluators to maximize the number of evaluators for a participating team. A predefined number of natural audio was also added. Along with this, a few repetitions were also included to check the consistency of evaluators. Evaluator scores were rejected if the consistency score was low. Selected evaluators verified all the final evaluation files as a paid task in a span of 7 days.

The naturalness MOS and speaker similarity score is obtained after taking the average over the scores from 144 ratings, for each submission. Equal weightage is provided to mono-lingual and cross-lingual synthesis scores.

## V. CHALLENGE SUMMARY

As mentioned in Table 4, we started the challenge registration on November 26th 2022, and in 2 months we completed the challenge and announced the winners on February 10th 2023. The challenge submissions have been handled through REST API hosted by the participants. We were able to complete the subjective evaluations of naturality and speaker similarity, for all submissions, within 10 days. Table 5 summarises the leaderboard for all three tracks. Based on the performance on the leaderboard, the top teams were invited to submit papers to ICASSP, on their submitted TTS systems.

Table 6 presents a few key statistics of the LIMMITS'23 challenge. We started the outreach for the challenge by first assessing the interest in the community to participate in our challenge, towards which we obtained 79 interested individuals. Then, we proceeded with the official challenge registration and a data download portal. We had 37 teams register, and 43 different downloads of the corpus. Finally, we had 8 teams who submitted APIs for their TTS systems during the challenge timeline. Additionally, for the subjective evaluation, we had over 100 interested evaluators for each language, and we ended up having 29, 20, and 26 evaluators for Hindi, Marathi, and Telugu, respectively.

LIMMITS'23 datasets have been downloaded by researchers from different countries worldwide, as shown in Fig. 9. While over 50% of downloads are from India, there is a lot of participation in other countries such as China, USA, Singapore, Japan, Hungary, Canada etc.

### A. BASELINE

A fastspeech-based [29] multilingual TTS baseline has been released as a GitHub repository,[6] as part of the challenge. The repository includes codes for preprocessing, training, and inference of the multi-lingual TTS model. The text was modelled on the character level for all languages, through a combined embedding space for all symbols. The fastspeech model was trained to predict mel-spectrograms, generated for 22 kHz audio. The alignment between the input text and the output mel spectrogram was learned using a duration predictor. The ground truth durations were generated using a CTC-based forced aligner.[7] The monolingual and cross-lingual predictions are controlled by varying the speaker embeddings and input text. Speaker-specific waveglow [30] vocoders were trained, which were used to synthesize the speech waveform from mel-spectrograms. The force-aligned durations were shared with the participants on Dropbox and speaker-specific waveglow [30] vocoders were shared on huggingface.[8]

### B. SUBMITTED SYSTEMS

The participating teams submitted their synthesized audio from APIs. A tutorial for creating the API from a hosted cloud instance was shared with the participants, along with the necessary metadata inputs required such as text, speaker ID, language ID, and file name to save. Additionally, a script to test the API functionality was shared and a few days of API testing were also observed. After the challenge submission was officially opened, the files were synthesized for evaluation.

Even though we had 8 unique final submissions as illustrated in Table 6, only 7 teams occupied the leaderboard and among them, only 4 teams got selected for the ICASSP 2023

[6][Online]. Available: https://github.com/bloodraven66/ICASSP_LIMMITS23
[7][Online]. Available: https://github.com/as-ideas/DeepForcedAligner
[8][Online]. Available: https://huggingface.co/SYSPIN

**TABLE 5.** Table Summarizing the Top Five Teams on the LIMMITS'23 Overall Challenge Leaderboard

| Rank | TRACK 1 | | | | TRACK 2 | | | | TRACK 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NATURAL | | SPKR. SIMI. | | NATURAL | | SPKR. SIMI. | | NATURAL | | SPKR. SIMI. | |
| | Team | Av. Sc. | Team | Av. Sc. | Team | Av. Sc. | Team | Av. Sc. | Team | Av. Sc. | Team | Av. Sc. |
| 1 | TEAM 1 | 4.77 | TEAM 2 | 3.98 | TEAM 6 | 4.56 | TEAM 6 | 3.98 | TEAM 1 | 4.44 | TEAM 3 | 3.8 |
| 2 | TEAM 2 | 4.71 | TEAM 3 | 3.96 | TEAM 3 | 4.47 | TEAM 3 | 3.95 | TEAM 3 | 4.4 | TEAM 1 | 3.54 |
| 3 | TEAM 3 | 4.46 | TEAM 1 | 3.86 | TEAM 1 | 4.4 | TEAM 1 | 3.6 | TEAM 2 | 4.04 | TEAM 4 | 2.87 |
| 4 | TEAM 4 | 4.4 | TEAM 4 | 3.72 | TEAM 2 | 4.12 | TEAM 2 | 3.02 | TEAM 4 | 3.67 | TEAM 2 | 2.76 |
| 5 | TEAM 5 | 4.27 | TEAM 5 | 3.28 | TEAM 4 | 3.67 | TEAM 4 | 2.74 | TEAM 7 | 1.61 | TEAM 7 | 1.4 |

[*Naturality: Score 5 = Human-like sound .. Score 1 = Extremely intolerable] [* Speaker similarity: Score 5 = Identical .. Score 1 = Very different]
NATURAL- Naturality, SPKR. SIMI.- Speaker similarity, Av. Sc.-Average Score

**TABLE 6.** Challenge Summary

| Item | Count |
|---|---|
| # Interested participants | 79 (12) |
| # Registered participants | 37 (11) |
| # Track 1:2:3 registrants | 27:35:30 |
| # Data downloads | 43 (10) |
| # Unique final submissions | 8 |
| # Submissions per track 1:2:3 | 7:7:6 |
| # Registered evaluators for Hindi:Marathi:Telugu | 196:121:141 |
| # Final evaluators for Hindi:Marathi:Telugu | 29:20:26 |

Number of the unique countries shown in brackets.

presentation. A brief summary of techniques from some of these teams (as available to us) is given below.

*Team 1:* Team 1 [31] used a Vector quantized (VQ) acoustic feature target [32] based TTS for track 1 and GradTTS [33] for tracks 2 and 3. The text is processed as raw characters, with numbers being mapped to their pronunciation. A $< sil >$ token is also added from forced alignment with Kaldi, and character level durations are extracted. For all tracks, speaker and language embeddings were used for speaker and language control. X-vector embeddings are computed for all utterances, and the average representation for each speaker is used as the speaker embedding. For cross-lingual synthesis, a native speaker's embedding was provided to the acoustic model, and the target speaker's embedding was shared with the vocoder. For track 1, features extracted from vq-wav2vec [34] are used as the TTS training target. Additionally, pitch, energy, and POV features are also used. As the VQ features are quantized, an additional layer is used at the output to smoothen the features before giving it to a HifiGAN vocoder. For tracks 2 and 3, GradTTS is used since it is a diffusion model, which leverages an iterative denoising process for inference. The same U-net architecture is used for infinitesimal steps which keeps the model parameters low. The GradTTS parameter size from 15 M is reduced to 5 M using fewer layers in the text encoder and lower channels in U-net.

*Team 2:* Team 2 [35] builds upon the RADMMM [36] to disentangle speaker characteristics for multilingual TTS. The data selection is performed by transcribing the audio through a speech recognizer and rank-ordering the files in terms of character error rate (CER). RADMMM is used for track 1, which supports explicit control of accent, language, speaker, and fine-grained F0, and energy features for speech synthesis. Format scaling is used to disentangle speaker and accent attributes. The augmented samples are treated as new speakers, which helps reduce the correlation between speaker, text, and accent. For tracks 2 and 3, a lightweight version of the model is built using an autoregressive decoder, conditioned on text, accent, speaker, F0, and energy. The architecture is similar to Flowtron [37], with 2 steps of flow, using LSTMs. A HifiGAN vocoder is used for track 1 and Waveglow vocoders are used for tracks 2 and 3.

*Team 3:* Team 3 [38] used a prosody TTS [39] system with F0 estimation and VAD prediction to obtain high-quality synthesis. The text is encoded using a bidirectional gated recurrent unit (GRU) encoder, and language, speaker embeddings are incorporated into it. The alignment is performed using a duration predictor, which is learned through a few dense layers. The durations are then used to upsample the features to the required mel-spectrogram frame length. These upsampled features are used for F0 estimation and VAD flag prediction. The estimated F0 is conditioned on the upsampled features, which act as the input to the decoder. The decoder is built using GRU and transformer neural network layers. The waveglow vocoder provided by challenge organizers is used to synthesize the waveform. Additionally, acoustic parameters are extracted through a pretrained model and used as an auxiliary loss function for training the TTS model. The text from all languages is mapped to IPA, which resulted in 90 total symbols. The data selection and ground truth durations were obtained using a 3-state HMM.

*Team 6:* Team 6 [40] used knowledge distillation, with DelightfulTTS [41] as a teacher model to obtain high-quality synthesis for a parameter-efficient model. The dataset is processed by converting text to phones using a G2P model, and the audio is resampled to 22 kHz. An alignment model is trained to extract ground truth durations. The Teacher model is trained as a multispeaker, multi-lingual model, which also models prosody features such as pitch. The teacher model is then used to predict durations and mel spectrogram for all the data, which is further used to train the teacher model. Additionally, bottleneck prosody variance [42] is used for cross-lingual inference. The encoder, decoder, and duration predictor in the student model consists of 1D-convolution and

LSTM units, due to which the overall student network ends up with 4.98 M parameters. The student model is trained for loss of Mean Squared Error for mel spectrogram and duration values. The model is trained for 450 k steps, for 20 hours with 8 GPUs.

## VI. RESULTS AND DISCUSSIONS

The top teams have observed high scores on both naturalness as well as speaker similarity measures as shown in Table 5. The track-specific analysis of the top results is given below.

### A. TRACK 1

For Track 1, Team 1 [31] obtains the best naturalness subjective score with MOS of 4.77, though it has a speaker similarity score of 3.86. On the other hand, Team 2 [35] obtains the best speaker similarity score of 3.98 while placing second in naturalness. The overall naturalness score in this track is in the range of 4.27-4.77, with the speaker similarly range of 3.98-3.28. The high naturalness scores across teams indicate that 5 hours of data (30 hours across 6 speakers) is sufficient to build high-quality multi-lingual TTS systems.

### B. TRACK 2

In Track 2, Team 6 [40] obtains the best naturalness as well as speaker similarity scores. Further, the best naturalness score of 4.56 is impressive considering the track is for parameter-efficient TTS. Additionally, the Team 6 speaker similarity score matches the best similarity score of Track 1, indicating high retention of speaker characteristics with parameter-efficient TTS. For track 2, the naturalness scores observe a range of 3.67-4.56. There is a degradation in performance when compared to Track 1, which indicates training a larger model with limited data (Track 1) can achieve higher performance than training a smaller model with an abundance of data (Track 2).

### C. TRACK 3

In Track 3, we find that Team 1 [31] obtains the best naturalness score of 4.44, with speaker similarity of 3.54. Team 2 [35] obtains the best naturalness score of 3.8 while placing second in naturalness. The top 3 submissions have naturalness scores over 4, and the top 2 teams have scores over 3. This indicates strong synthesis capability in resource and model parameter-constrained TTS systems. The naturalness scores for this track are in the range of 1.61-4.44. On comparing the teams' performance between track 2 and track 3, we find that teams 1-4 observe similar naturalness scores, while Team 7 has low scores. This indicates that a limited amount of training data is sufficient to train a good-quality low-parameter TTS model.

### D. OVERALL SUMMARY

We can observe that Team 1 [31] has obtained rank 1 on naturalness for two different tracks - this indicates the promise of VQ features as targets (rather than mel spectrograms) for TTS. On the other hand, Team 6 has shown improvements in naturalness for low parameter TTS in track 2, which shows the benefits of distilling a strong teacher model such as DelightfulTTS. Meanwhile, there is no single approach that has topped the leaderboard across tracks for speaker similarity. We can observe that different methods such as speaker disentanglement (Team 2), prosody features (Team 3, 6) can be important in different data, and model parameter settings. The presence of cross-lingual samples could be responsible for the lower speaker similarity scores, indicating the need for more research in this area.

## VII. CONCLUSION

LIMMITS'23 challenge provided an opportunity for the speech community to work on data-constrained and lightweight multilingual TTS in Indian languages. The challenge received a large number of registrations and interest from varying demographics. Challenging tracks and evaluation constraints such as cross-lingual synthesis were presented in this challenge. Many participants brought in novel ideas and, as a result, the top challenge submissions have high subjective scores in naturalness and speaker similarity, contributing to the state-of-the-art TTS research on these topics. We plan to provide more such opportunities in the future in terms of challenging tracks and datasets.

The dataset, open-sourced as part of the LIMMITS'23 challenge, could play an important role beyond the challenge itself. It could spur innovation in the research and development of TTS in the Indian languages in both commercial and academic settings. This is similar to the initiatives on creating an open source, multi-language dataset of voices by Common Voice.

## ACKNOWLEDGMENT

The authors would like to thank Deutsche Gesellschaft fur Internationale Zusammenarbeit (GIZ) on behalf of the German Ministry for Economic Cooperation and Development for supporting data collection.

## REFERENCES

[1] C. Chandramouli and G. Registrar, "Census of India 2011," Provisional Population Totals. New Delhi: Government of India, 2011, pp. 409–413.
[2] Y. Ren et al., "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. 9th Int. Conf. Learn. Representations*, 2021, pp. 1–15. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA
[3] X. Tan, T. Qin, F. K. Soong, and T. Liu, "A survey on neural speech synthesis," 2021, *arXiv:2106.15561*.
[4] S. Srivastava and H. A. Murthy, "USS directed E2E speech synthesis for Indian languages," in *Proc. IEEE Int. Conf. Signal Process. Commun.*, 2022, pp. 1–5.
[5] A. Prakash, A. L. Thomas, S. Umesh, and H. A Murthy, "Building multilingual end-to-end speech synthesisers for Indian languages," in *Proc. 10th ISCA Workshop Speech Synth.*, 2019, pp. 194–199.
[6] A. Debnath, S. S. Patil, G. Nadiger, and R. A. Ganesan, "Low-resource end-to-end Sanskrit TTS using Tacotron2, WaveGlow and transfer learning," in *Proc. IEEE 17th India Council Int. Conf.*, 2020, pp. 1–5.
[7] K. K. A. Kumar, H. R. S. Kumar, R. A. Ganesan, and K. P. Jnanesh, "Efficient human-quality Kannada TTS using transfer learning on NVIDIA's Tacotron2," in *Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol.*, 2021, pp. 1–6.

[8] A. Prakash and H. Murthy, "Generic Indic text-to-speech synthesisers with rapid adaptation in an end-to-end framework," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 2962–2966.
[9] A. Prakash and H. A. Murthy, "Exploring the role of language families for building Indic speech synthesisers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 734–747, 2023.
[10] S. Li, B. Ouyang, L. Li, and Q. Hong, "Light-TTS: Lightweight multi-speaker multi-lingual text-to-speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 8383–8387.
[11] K. Prahallad, E. N. Kumar, V. Keri, S. Rajendran, and A. W. Black, "The IIIT-H Indic speech databases," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 2546–2549.
[12] A. W. Black, "CMU wilderness multilingual speech dataset," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 5971–5975.
[13] A. Baby et al., "Resources for Indian languages," in *Proc. Community-Based Building Lang. Resour.*, 2016, pp. 37–43.
[14] F. He et al., "Open-source multi-speaker speech corpora for building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu speech synthesis systems," in *Proc. 12th Lang. Resour. Eval. Conf. Eur. Lang. Resour. Assoc.*, 2020, pp. 6494–6503. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.800
[15] P. Baljekar and A. W. Black, "Utterance selection techniques for TTS systems using found speech," in *Proc. 9th ISCA Workshop Speech Synth. Workshop*, 2016, pp. 184–189.
[16] K.-Z. Lee, E. Cooper, and J. Hirschberg, "A comparison of speaker-based and utterance-based data selection for text-to-speech synthesis," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 2873–2877.
[17] E. Cooper, X. Wang, A. Chang, Y. Levitan, and J. Hirschberg, "Utterance selection for optimizing intelligibility of TTS voices trained on ASR data," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 3971–3975.
[18] K. Seki, S. Takamichi, T. Saeki, and H. Saruwatari, "Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection," 2022, *arXiv:2210.14850*.
[19] F.-Y. Kuo, I. C. Ouyang, S. Aryal, and P. Lanchantin, "Selection and training schemes for improving TTS voice built on found data," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 1516–1520.
[20] E. Cooper, A. Chang, Y. Levitan, and J. Hirschberg, "Data selection and adaptation for naturalness in HMM-based speech synthesis," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 357–361.
[21] E. Cooper, Y. Levitan, and J. Hirschberg, "Data selection for naturalness in HMM-based speech synthesis," in *Proc. Speech Prosody*, 2016, pp. 791–795.
[22] A. L. Thomas, A. Prakash, A. Baby, and H. A. Murthy, "Code-switching in Indic speech synthesisers," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 1948–1952.
[23] J. Vainer and O. Dušek, "SpeedySpeech: Efficient neural speech synthesis," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3575–3579.
[24] R. Luo et al., "Lightspeech: Lightweight and fast text to speech with neural architecture search," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 5699–5703.
[25] R. Chevi, R. E. Prasojo, and A. F. Aji, "NIX-TTS: An incredibly lightweight end-to-end text-to-speech model via non end-to-end distillation," 2022, *arXiv:2203.15643*.
[26] H. S. Chadha et al., "Vakyansh: ASR toolkit for low resource Indic languages," 2022, *arXiv:2203.16512*.
[27] T. Javed et al., "Towards building ASR systems for the next billion users," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 10813–10821.
[28] G. S. Mirishkar, V. V. R. V, M. D. Naroju, S. Maity, P. Yalla, and A. K. Vuppala, "CSTD-Telugu corpus: Crowd-sourced approach for large-scale speech data collection," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2021, pp. 511–517.
[29] Y. Ren et al., "FastSpeech: Fast, robust and controllable text to speech," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 3171–3180.
[30] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 3617–3621.
[31] C. Du, Y. Guo, F. Shen, and K. Yu, "Multi-speaker multi-lingual VQTTS system for LIMMITS 2023 Challenge," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–2.
[32] C. Du, Y. Guo, X. Chen, and K. Yu, "VQTTS: High-fidelity text-to-speech synthesis with self-supervised VQ acoustic feature," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2022, pp. 1596–1600.
[33] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. A. Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," in *Int. Conf. Mach. Learn.*, 2021, pp. 8599–8608. [Online]. Available: https://api.semanticscholar.org/CorpusID:234483016
[34] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *Proc. 8th Int. Conf. Learn. Representations*, 2020, pp. 1–12. [Online]. Available: https://openreview.net/forum?id=rylwJxrYDS
[35] R. Badlani et al., "VANI: Very-lightweight accent-controllable TTS for native and non-native speakers with identity preservation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–2.
[36] R. Badlani, R. Valle, K. J. Shih, J. F. Santos, S. Gururani, and B. Catanzaro, "Multilingual multiaccented multispeaker TTS with RADTTS," 2023, *arXiv:2301.10335*.
[37] R. Valle, K. J. Shih, R. Prenger, and B. Catanzaro, "Flowtron: An autoregressive flow-based generative network for text-to-speech synthesis," in *Proc. 9th Int. Conf. Learn. Representations*, 2021, pp. 1–10. [Online]. Available: https://openreview.net/forum?id=Ig53hpHxS4
[38] G. Pamisetty, S. C. Varun, and K. S. R. Murty, "Lightweight prosody-TTS for multi-lingual multi-speaker scenario," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–2.
[39] G. Pamisetty and K. Sri Rama Murty, "Prosody-TTS: An end-to-end speech synthesis system with prosody control," *Circuits, Syst., Signal Process.*, vol. 42, no. 1, pp. 361–384, 2023.
[40] C. Zhang et al., "LeanSpeech: The Microsoft lightweight speech synthesis system for LIMMITS Challenge 2023," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–2.
[41] Y. Liu et al., "DelightfulTTS: The Microsoft speech synthesis system for Blizzard Challenge 2021," 2021, *arXiv:2110.12612*.
[42] S. Pan and L. He, "Cross-speaker style transfer with prosody bottleneck in neural speech synthesis," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 4678–4682.