NSR@LT-EDI-2025: Automatic speech recognition in Tamil

Nishanth S, Shruthi Rengarajan, Burugu Rahul, G. Jyothish Lal
Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
{cb.en.u4aie22149, cb.en.u4aie22154,
cb.en.u4aie22161}@cb.students.amrita.edu
g_jyothishlal@cb.amrita.edu

Abstract

Automatic Speech Recognition (ASR) technology can potentially make marginalized communities more accessible. However, older adults and transgender speakers are usually highly disadvantaged in accessing valuable services due to low digital literacy and social biases. In Tamil-speaking regions, these are further compounded by the inability of ASR models to address their unique speech types, accents, and spontaneous speaking styles. To bridge this gap, the LT-EDI-2025 shared task is designed to develop robust ASR systems for Tamil speech from vulnerable populations. Using whisper-based models, this task is designed to improve recognition rates in speech data collected from older adults and transgender speakers in naturalistic settings such as banks, hospitals and public offices. By bridging the linguistic heterogeneity and acoustic variability among this underrepresented population, the shared task is designed to develop inclusive AI solutions that break communication barriers and empower vulnerable populations in Tamil Nadu.

Keywords: Speech Recognition, Indian languages, Tamil, Whisper model

1 Introduction

Speech is the most natural form of human communication. As Speech Technologies and AI advance rapidly, Automatic Speech Recognition (ASR) systems have become essential for human-computer interactions, offering convenience especially in multilingual countries like India. However, a significant gap exists in how these technologies reach vulnerable communities.But there is a big disconnect in how these technologies get to communities that are at risk.

Elderly people and transgender communities are marginalized groups that struggle to access essential services in Tamil-speaking areas. While transgender communities, who are frequently the targets of discrimination in society, encounter obstacles in education and digital literacy, older adults struggle with age-related disabilities that make interacting with digital interfaces more difficult. The most dependable method of expressing needs for both groups is still face-to-face interaction; however, current ASR systems that were trained on mainstream data are unable to accurately transcribe their distinct speech patterns.

Developing reliable ASR systems for these groups involves multiple challenges. Elderly speakers show acoustic variability due to physiological changes like decreased vocal cord elasticity and altered speech rhythm. Transgender speakers may present varied vocal characteristics influenced by hormonal treatments or voice therapy. Tamil itself poses linguistic challenges as a Dravidian language with complex morphology, multiple dialects, and code-mixing tendencies.

Compounding these issues is the severe shortage of high-quality speech corpora from older and transgender Tamil speakers. Available resources typically focus on standard varieties with minimal coverage of marginalized voices.

To address these gaps, the LT-EDI 2025 Shared Task on Speech Recognition for Vulnerable Speakers in Tamil(B. Bharathi, 2025) aims to stimulate research in this underexplored area. The task focuses on developing ASR systems to transcribe spontaneous Tamil speech from older and transgender speakers, providing curated datasets and leveraging models like Whisper to tackle the unique challenges these populations present.

This collaborative initiative primarily seeks to bridge the technological divide and enhance digital inclusion for underserved communities, empowering vulnerable Tamil speakers with improved access to essential services through accurate speech recognition.

More details about the shared task can be found

at[1].

2 Related works

Recent work in ASR has gained interest in calibrating systems to function proficiently with underrepresented and non-standard speech categories. (S and B, 2022) introduced a Transformer-based Tamil conversational ASR for the speech of older and trans women. Training their system with actual audio captured from public spaces, they confronted natural speech variations and obtained a WER of 39.65%, divulging the sophistication in creating accessible ASR. In a similar vein, (R et al., 2024) developed ASR for vulnerable Tamil speakers based on fine-tuned Whisper and XLS-R models. Fine-tuned on LT-EDI@EACL2024 data, the Whisper model performed with improved robustness with a WER of 24.452, proving its ability to deal with age, gender, and social background-caused variability.

(Radford et al., 2022) presented a large-scale method by training ASR models on 680,000 hours of weakly labeled web audio. Their zero-shot models saw success across several benchmarks and demonstrated nearly human-level performance and extreme generalizability over languages and tasks. (Shraddha et al., 2022), in contrast, investigated ASR performance on child speech, a field usually overlooked because of the scarcity of data and pitch/articulation contrasts. Benchmarking six end-to-end models, they highlighted the limitation of adult-trained ASR systems when applied to child speech.

(Biswas et al., 2022) suggested the application of Weighted Finite-State Transducers (WFSTs) to integrate acoustic, lexical, and language models within a probabilistic and efficient decoding pipeline. As opposed to end-to-end neural systems, WFSTs are modular and flexible and thus effective for structured ASR frameworks. Overall, these studies highlight the requirement of ASR systems that are robust, flexible, and inclusive, capable of accommodating linguistic, acoustic, and demographic variation.

3 Dataset

The dataset was distributed by the shared task organisers of Speech Recognition for Vulnerable Individuals in Tamil - LT-EDI@LDK 2025 (B et al., 2022)(B et al., 2024). The audio samples are collected from individuals whose mother tongue is Tamil and was presented in a .wav format. The total audio length is 7 hours and 30 minutes and is divided into 5.5 hours(approx.)

4 Methodology

A speech recognition model (Whisper-V3-large) (Graham and Roll, 2024) is trained using a specially created audio-text dataset. The suggested methodology makes use of two folders, one of which contains audio recordings and the other of which contains the text transcripts that go with them. Every audio file has a corresponding text file with the same name but a different extension. First, each transcript's encoding is automatically identified to guarantee accurate text file reading. Because text files can be saved in a variety of formats and incorrect reading can result in errors or misinterpreted characters, this step is crucial. After being identified, the transcript is read and cleared of any extraneous words before being paired with the matching audio file in a structured list.

Following the collection of all legitimate audio-transcript pairs, the information is transformed into a specific format that facilitates effective audio data handling. The way the audio column is handled enables the system to directly load and process waveform data, which makes it appropriate for speech recognition model training. The dataset is divided into two sections, usually in a 90-10 ratio, one for training and one for testing, in order to accurately assess model performance.

Other components are initialized to get this data ready for model training. These comprise tools for converting text into numerical form, extracting significant features from unprocessed audio, and combining the two to expedite input processing. In addition, a pre-trained model architecture is loaded, which is intended to convert spoken language into written formis loaded. All of these elements are set up to function exclusively with a selected language, in this example Tamil, guaranteeing that the input and output match the features of that language. When combined, the dataset and these tools form a comprehensive pipeline that is prepared for training a model to comprehend and record spoken Tamil.

[1] https://codalab.lisn.upsaclay.fr/competitions/21879

4.1 Data Preprocessing

To be able to train a model that translates spoken language into written text, the suggested methodology first entails the preparation and compilation of audio-text data. Each data sample is first processed, with audio signals being converted into model-appropriate numerical features and the corresponding transcriptions being changed into token identifier sequences that the model can comprehend. To make sure that the input and output are in a format that the model can use for learning, this transformation is carried out for the entire dataset. A distinctive component is defined to appropriately arrange and align the inputs and outputs in order to manage training in batches. Since text labels and audio inputs need different handling, they are separated. To enable effective computation, the audio features are gathered and padded in a batch so that they are all the same size. In a similar manner, the text labels are padded to guarantee size alignment, but with special handling: portions of the padding are designated to be disregarded during training to prevent the model from treating them as actual words. Additionally, since the training process will add the starting tokens separately, any extraneous ones that are already in the labels are eliminated to prevent duplication. By ensuring that the model receives clean, consistent, and well-aligned data at every stage, this meticulous preparation raises the training process's accuracy and efficiency.

4.2 Training and Model Evaluation

A training framework for a sequence-to-sequence model by defining key components like training parameters, datasets for training and evaluation, and a data preparation function is set up. It contains a preprocessor to format model input data and a metric calculation function to track performance. By automating the training and evaluation process, this setup makes it possible to fine-tune the model effectively.

The parameters for training a sequence-to-sequence model are established by the given configuration. It outlines specifics such as the model's storage location, the training and evaluation batch sizes, the number of training epochs, and the frequency of evaluations and checkpoint saves. To handle large models, the training process is optimized using techniques like gradient accumulation, mixed-precision training, and memory-efficient methods. The model that performs the best is saved for later use after its performance is assessed on a regular basis using a particular metric. Furthermore, the generated sequences are constrained to a specific length during prediction, and the learning rate is initially increased gradually.

The two main metrics derived for evaluation are the Word Error Rate (WER) and the Character Error Rate (CER), both of which are used in common practice to measure the precision of automatic speech recognition (ASR) systems (Hamed et al., 2023). WER calculates the number of total errors in the transcription of the model against the ground truth set. This metric is calculated by comparing the predicted transcription with the reference text and penalizing insertions, deletions, and word substitutions. The lower the WER, the better the performance. CER is similar in operation but at the character level rather than the word level. CER calculates the number of errors at the character level, which can be useful when the model transcribes text with various spelling or formatting errors. Both WER and CER have significance since both provide a precise understanding of the accuracy of the model at varying levels of granularity (word vs. character).

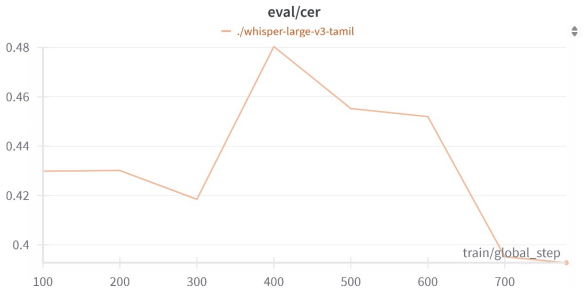

Figure 1: CER evaluation graph

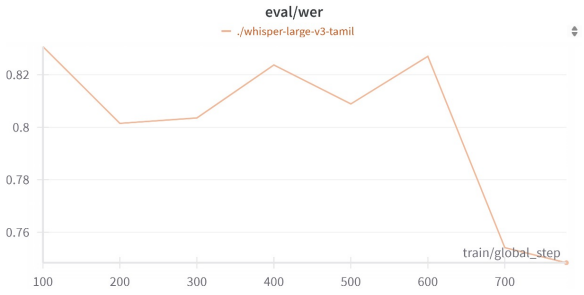

Figure 2: WER evaluation graph

Table 1: Step-wise Performance of the Whisper Model on the Tamil Dataset

Step	Training Loss	Validation Loss	WER ()	CER ()
100	0.3993	0.6537	0.8307	0.4299
200	0.1820	0.8540	0.8015	0.4301
300	0.1460	0.9717	0.8035	0.4185
400	0.1226	0.9849	0.8237	0.4803
500	0.1006	1.0915	0.8089	0.4552
600	0.0272	1.1300	0.8270	0.4519
700	0.0035	1.1713	0.7541	0.3952

Parameter	Value
output_dir	./whisper-large-v3-tamil
per_device_train_batch_size	16
gradient_accumulation_steps	2
learning_rate	3e-4
warmup_steps	500
num_train_epochs	30
gradient_checkpointing	True
fp16	True
evaluation_strategy	steps
per_device_eval_batch_size	8
predict_with_generate	True
generation_max_length	225
save_steps	500
eval_steps	100
logging_steps	50
report_to	wandb
load_best_model_at_end	True
metric_for_best_model	wer
greater_is_better	False
save_total_limit	2
push_to_hub	False
dataloader_num_workers	7
dataloader_prefetch_factor	2
dataloader_pin_memory	True

Table 2: Updated Training Parameters for Seq2Seq Model

Evaluation Type	Score
Word Error Rate (WER)	0.7484
Character Error Rate (CER)	0.3927

Table 3: Model Evaluation

5 Experimental Inference

In this proposed methodology, a pre-trained Whisper model from Hugging Face is leveraged to perform automatic speech recognition (ASR) (Amorese et al., 2023).

The script randomly selects a sample from the test set of a preloaded dataset. The sample contains an audio file and its corresponding transcribed text. The audio file is extracted from the sample, along with its sampling rate, which is important for the subsequent feature extraction process. The audio data, in the form of a raw waveform, is passed to the Whisper Processor, which is responsible for converting the raw audio into features that the Whisper model can understand. Then, the waveform is transformed into a spectrogram, a 2D representation of the audio, which is the input format the model expects.

Once the audio has been transformed into input features, it is sent to the model which takes the processed audio features as input and outputs a sequence of predicted token IDs that represent the transcription of the speech. These token IDs are converted from the numerical token IDs back into readable text, skipping any special tokens that may have been used during training (such as padding or end-of-sequence markers).

Finally, the original and machine-predicted scripts are compared, which allows an evaluation of the models performance, offering insight into how well the Whisper model has learned to transcribe speech. By printing both the original and predicted text, users can directly observe how accurately the model has transcribed the audio sample, which is the primary goal of the ASR process.

With the same method used to obtain inference on the test data, We secured **2nd** rank in this task.

The code files for this project can be accessed from[2]

6 Conclusion

This paper presented the results of the shared task addresses a challenging area in Automatic Speech Recognition: vulnerable old-aged and transgender people in Tamil. Many elderly people do not know how to use the equipment available to them. Speech is the only medium that could help transgender people meet their needs because they are denied access to primary education due to societal prejudice. Data on spontaneous speech are

[2] https://github.com/BURUGURAHUL/NSR-LT-EDI-2025-Automatic-speech-recognition-in-Tamil

collected from elderly and transgender individuals who cannot take advantage of these resources.

7 Limitations

While working on this topic, the major limitation we faced was the use of Whisper V3 large, which significantly increased computational requirements. Due to the models large size, standard GPUs were insufficient, and an NVIDIA A6000 was required to handle the memory load. This made the approach less accessible in environments with limited hardware resources.

References

Terry Amorese, Claudia Greco, Marialucia Cuciniello, Rosa Milo, Olga Sheveleva, and Neil Glackin. 2023. Automatic speech recognition (asr) with whisper: Testing performances in different languages. In *S3C@ CHItaly*, pages 1–8.

Bharathi B, Bharathi Raja Chakravarthi, Subalalitha Cn, Sripriya N, Arunaggiri Pandian, and Swetha Valli. 2022. Findings of the shared task on speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 339–345, Dublin, Ireland. Association for Computational Linguistics.

Bharathi B, Bharathi Raja Chakravarthi, Sripirya N, Rajeswari Natarajan, Suhasini S, and Swetha Valli. 2024. Overview of the third shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

N. Sripriya Rajeswari Natarajan Rajalakshmi R S. Suhasini B. Bharathi, Bharathi Raja Chakravarthi. 2025. Overview of the Fifth Shared Task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Dipshikha Biswas, Suneel Nadipalli, B. Sneha, and M. Supriya. 2022. Speech recognition using weighted finite-state transducers. In *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, pages 1–5.

Calbert Graham and Nathan Roll. 2024. Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits. *The Journal of the Acoustical Society of America*, 4.

Injy Hamed, Amir Hussein, Oumnia Chellah, Shammur Chowdhury, Hamdy Mubarak, Sunayana Sitaram, Nizar Habash, and Ahmed Ali. 2023. Benchmarking evaluation metrics for code-switching automatic speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 999–1005.

Jairam R, Jyothish G, Premjith B, and Viswa M. 2024. CEN_Amrita@LT-EDI 2024: A transformer based speech recognition system for vulnerable individuals in Tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 190–195, St. Julian's, Malta. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Suhasini S and Bharathi B. 2022. SUH_ASR@LT-EDI-ACL2022: Transformer based approach for speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 177–182, Dublin, Ireland. Association for Computational Linguistics.

S Shraddha, Jyothish Lal G, and Sachin Kumar S. 2022. Child speech recognition on end-to-end neural asr models. In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pages 1–6.