TamilVoiceLab@DravidianLangTech 2026: Investigating Whisper Tamil Large-v2 for Dialectal Tamil Speech Recognition

S.B.PRIYA
St. Joseph's Institute of Technology
Chennai, India
priyait0843@gmail.com

B.BHARATHI
Sri Sivasubramaniya Nadar College of Engineering
Chennai, India
bharathib@ssn.edu.in

Abstract

Automatic Speech Recognition (ASR) for languages rich in dialects and those with limited resources presents significant challenges due to the variations in pronunciation and vocabulary across different regions. This study offers a baseline evaluation of the Whisper Tamil Large-v2 model without fine-tuning for the shared task of Tamil Dialect Speech Recognition. The focus is on the ASR subtask, utilising dialectal Tamil speech recordings gathered from various regional dialects within Tamil Nadu. The pretrained Whisper Tamil Large-v2 model was assessed directly, without any supplementary fine-tuning or domain adaptation. A total of 579 dialect speech samples were used for experimentation, with performance evaluated based on Word Error Rate (WER). The model recorded a WER of 0.71, indicating that even robust multilingual pretrained models encounter challenges in dialect-rich and low-resource environments. These findings underscore the necessity for dialect-aware adaptation and the importance of balanced dialect training data to develop effective Tamil ASR systems.

Keywords: Tamil ASR, Dialect Speech Recognition, Whisper, Low-Resource Speech Processing, Zero Fine-Tuning

1 Introduction

ASR, or automatic speech recognition, uses computer models to translate speech into text. ASR performance has significantly improved for high-resource languages like Mandarin and English thanks to deep learning and transformer-based models. For low-resource languages and dialect-rich environments with little training data, ASR is still difficult.More than 75 million individuals speak Tamil, a significant Dravidian language. Northern, Southern, Western, and Central Tamil Nadu all have strong regional dialects. Pronunciation, rhythm, vocabulary, and sentence structure vary among these dialects. Speech from one area may sound significantly different from speech from another due to this diversity. Such dialect discrepancies are a common problem for ASR systems that were mostly trained on standard Tamil. Strong cross-lingual abilities have recently been demonstrated by multilingual models such as Whisper. Whisper can do recognition without task-specific fine-tuning because it was trained on vast amounts of multilingual speech samples. Its performance on unadapted dialectal Tamil speech, however, has not been thoroughly investigated. Whisper Tamil Large v2 is assessed in this study using dialectal Tamil ASR with zero fine-tuning. 580 dialect voice samples that were captured in natural settings are used to directly evaluate the model. Word Error Rate, which contrasts predicted text with ground truth transcription, is used to gauge performance. With a Word Error Rate of 0.71, the model demonstrated a notable decline in performance under dialect-rich circumstances. The model struggles with intra-linguistic variances present in Tamil dialect speech, despite having extensive multilingual training. These results emphasise the necessity of fine-tuning low-resource models with balanced training data, dialect-aware fine-tuning, and parameter-efficient adaptation strategies. In addition to supporting future research toward more inclusive and regionally resilient speech systems, this work establishes a baseline for enhancing Tamil dialect ASR.

2 Related Work

Recent research indicates strong growth in Tamil speech recognition, particularly in dialectal and low-resource settings. Many studies focus on improving data quality, adapting large models like Whisper, and handling real-world challenges such as code switching and speaker variation.

Low Rank Adaptation has recently been applied

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 397–402

to Whisper for inclusive Tamil speech recognition. This method fine-tuned Whisper efficiently while reducing computational cost and improving recognition for diverse and vulnerable speaker groups (Acharya et al., 2025). Code switching remains a major challenge in end-to-end ASR, especially for Indian languages such as Tamil, where mixed language speech is common. A systematic review explained the difficulties in handling code-mixed speech and highlighted model adaptation strategies for better performance (Agro et al., 2025). Large multilingual speech models have been shown to improve low-resource recognition through cross-lingual transfer. XLS R demonstrated that training on many languages enhances recognition accuracy for dialect-rich languages (Babu et al., 2021). Self-supervised learning has played a major role in advancing ASR. The wav2vec 2.0 framework showed that learning speech representations from raw audio before fine-tuning significantly improves performance for low-resource languages (Baevski et al., 2020). Recent work has also explored multilingual speech-to-speech translation using unified end-to-end architectures. These systems combine recognition and translation within a single framework (Beltrán Lobato, 2025). Dialect-rich corpora are important for robust Tamil ASR. A multi-dialect Tamil speech corpus demonstrated that including diverse regional accents in training data improves recognition robustness (Bharathi et al., 2025). Integrating external language models with Whisper has been shown to enhance low-resource recognition. Whisper LM reduced recognition errors by strengthening language modelling alongside acoustic modelling (de Zuazo et al., 2025). Fine-tuning Whisper for low-resource languages has produced strong improvements. Adaptation experiments on Amharic confirmed that even limited domain data can significantly enhance performance (Gete et al., 2025). Benchmarking is essential for the fair evaluation of Indian language speech systems. IndicSuperb introduced a standardised evaluation framework for speech processing tasks across Indian languages (Javed et al., 2023). Different fine-tuning strategies for Whisper have been compared in low-resource ASR settings. Targeted adaptation methods,, such as parameter-efficient tuning,, significantly improved recognition accuracy (Liu et al., 2024). Multilingual adaptation techniques have also been studied for RNN-based ASR systems. Knowledge transfer among languages was shown to improve recognition accuracy (Miiller et al., 2018). Whisper has been adapted for Indian language speech tasks beyond recognition, including sentiment analysis and speaker diarization. These studies confirm that proper fine-tuning enables effective adaptation to Indian languages (Papala et al., 2023). Transformer-based ASR models have been successfully developed for other Dravidian languages such as Kannada. These findings support the effectiveness of transformer encoders for speech recognition in related languages like Tamil (Prasad et al., 2026). Whisper, trained with large-scale weak supervision, demonstrated strong robustness to noise, accents, and spontaneous speech across multiple languages (Radford et al., 2023). Cross-lingual generalisation and neuron sharing have been studied in multilingual models. Combining shared and language-specific representations improves performance in low-resource settings (Riemenschneider and Frank, 2025). Real-time Tamil dialect speech recognition has also been explored. A system combining recognition and summarisation showed the importance of dialect-aware modelling for practical applications (Saranya et al., 2025). Multilingual Indian speech translation corpora support cross-task evaluation of speech systems. Such resources highlight the growing need for multilingual benchmarking in Indian languages (Shah et al., 2025). Neural models have been used for dialect recognition using convolutional networks and language embeddings. These models effectively distinguish closely related dialects (Shon et al., 2018). Sequence-to-sequence ASR systems have been combined with Indic large language models for speech translation. Integrating stronger language models improves overall system performance (Wei et al., 2025). Large language models and multilingual training strategies have also been explored for low-resource speech recognition. Transfer learning was shown to be effective in improving performance under limited data conditions (Zhang and Huang, 2025). Overall, recent literature shows three major trends. First, dialect-specific and diverse datasets improve ASR quality. Second, fine-tuning large pretrained models like Whisper reduces error rates in low-resource conditions. Third, multilingual and cross-lingual learning helps transfer knowledge across related languages. These studies strongly support the need for adapting Whisper-based models for Tamil dialect recognition, especially when training data is limited.

Table 1: Tamil Dialect Speech Dataset Overview

Category	Details
Language	Tamil
Number of Dialects	4
Dialects Included	Northern, Southern, Western, Central
Dialect Regions	Regions across Tamil Nadu
Speech Type	Spontaneous and Read Speech
Sampling Rate	16 kHz
Recording Environment	Natural acoustic conditions
Training Duration	9.22 hours
Test Duration	2.05 hours
Test Samples Used	580 audio files
Evaluation Metric	Word Error Rate

3 Dataset Description

A shared job on Tamil voice processing provided the data for this investigation. The Tamil Dialect Speech dataset used in this work was released through the DravidianLangTech shared task (Bharathi et al., 2026). The dataset includes speech recordings from four dialect regions of Tamil Nadu and supports both dialect identification and speech recognition research. Speech recordings from four dialects—Northern, Southern, Western, and Central—are included in the corpus. There is a corresponding Tamil transcription for every training sample. 5,134 recordings totalling roughly 9.22 hours make up the training set. Although the distribution among dialects is not entirely balanced, the data shows regional variance and natural pronunciation. There are 579 unlabeled audio samples in a different test set. The review is objective because the transcripts are not accessible. Table 1 provides a fair comparison across systems and enumerates the dataset's key features.

4 Proposed Methodology

The proposed system uses Whisper Tamil Large v2 for dialectal Tamil speech recognition. The main objective is to convert Tamil speech audio into text using a pretrained multilingual Automatic Speech Recognition model. No additional training or fine tuning was performed in this work. Instead, the study evaluates how well a large pretrained model performs directly on dialect rich Tamil speech.

The speech recordings were obtained from the shared task dataset and maintained at a sampling rate of 16 kHz, which matches the requirement of the Whisper model. Each audio waveform was processed using the Whisper processor to extract log Mel spectrogram features. These features capture important acoustic information such as frequency patterns and energy variations present in the speech signal.

The extracted features were then passed into the Whisper encoder. The encoder learned hidden speech representations related to pronunciation, rhythm, accent, and speaking style. These learned representations were used by the decoder to generate Tamil text in an end to end manner. Beam search decoding was applied to improve prediction accuracy by selecting the most probable sequence of words.

All test samples were processed using the same inference pipeline. The generated transcriptions were stored for analysis and reproducibility.

System performance was evaluated using Word Error Rate. The model achieved a WER of 0.71, showing that dialectal Tamil speech remains challenging in a zero fine tuning setting. Most recognition errors occurred in region specific vocabulary, colloquial expressions, and pronunciation variations across dialects.

Even though the error rate was high, the model was able to generate partially understandable Tamil text and correctly identify many common words. This indicates that large pretrained multilingual speech models possess reasonable generalization capability even without dialect specific adaptation. However, the results also highlight the importance of dialect-aware fine-tuning for robust Tamil speech recognition. Figure 1 illustrates the

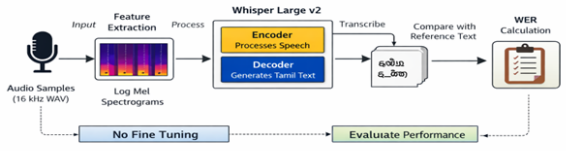

Figure 1: Proposed Tamil Dialect ASR System Architecture

overall architecture of the proposed Tamil dialect speech recognition system. The system accepts Tamil speech audio in 16 kHz WAV format as input. The audio signal is first converted into log Mel spectrogram features using the Whisper processor. These features capture important acoustic patterns from the speech signal and are then passed to the pretrained Whisper Tamil Large v2 encoder-decoder model for Tamil text generation.

The generated transcriptions are finally evaluated using Word Error Rate.

5 Evaluation and Results

Standard metrics for voice recognition are used to assess this system. Measuring the degree to which the predicted text differs from the original reference text is the primary objective. **Word Error Rate** Word Error Rate, or WER, is the most common metric in ASR. It measures errors at the word level. WER is calculated as:

$$WER = \frac{S + D + I}{N} \quad (1)$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference transcription. Substitution means one word is replaced by another word. Deletion means a word is missing. Insertion means an extra word is added. A lower WER means better performance. In this system, the obtained WER is: [WER = 0.71] This means about 71 per cent of the words are incorrect when compared to the reference transcript. **Sample Model Outputs** Below are sample transcriptions generated by the system:

File number	Predicted output	Target Output
test_0006	செய்து என் பையனுக்கு பிடிச்சது பிரியாணி தான் சிக்கன் பிரியாணி அதனால வரத்து ஒருநாள் எப்படியாவது செஞ்சு கொடுப்போம்	அதிகமா செய்றது, அவனுக்கு ஏன் பையனுக்கு பிடிச்சது பிரியாணி தான் சிக்கன் பிரியாணி தான். அதனால வாரத்துல ஒரு நாள் எப்படியாவது அவனுக்கு செஞ்சு கொடுப்போம்.
test_0015	நான் வந்து ஜீ தமிழ் தான் ரொம்ப பாப்பேன் எனக்கு தான் ரொம்ப பிடிக்கும்	நான் வந்து ஜீ தமிழ் தான் ரொம்ப பாப்பேன். எனக்கு அதுதான் ரொம்ப புடிக்கும்.
test_0113	பொழுது அனைக்கும் கேம் தான் விளையாடு நீ என்ன வந்து பேசுற	பொழுதன்னைக்கும் கேம் தான வெளையாடு நீ என்ன வந்து பேசுற

Figure 2: Comparison of Predicted and Target Transcriptions

Figure 2 shows the sample output generated by the proposed Tamil ASR system along with the target transcription. The highlighted words indicate recognition errors. Most errors occur in colloquial expressions and regional speech patterns. Table 2 compares the proposed system with recent Tamil speech recognition systems reported in the literature. The comparison is based on Word Error Rate, where lower values indicate better recognition performance. The proposed system achieved a WER of 0.71 using Whisper Tamil Large v2 without any fine tuning. Compared with other systems that use larger datasets and adaptation techniques, the results highlight the challenges of dialectal Tamil ASR in low resource settings. This comparison provides a baseline for future improvement in Tamil speech recognition systems.

Table 2: Comparison of Tamil ASR Systems

Paper / System	Model Used	Fine Tuning	Dataset Type	Reported WER	Notes
Real-time Continuous Tamil Dialect Speech Recognition and Summarisationn (2025)	Custom Dialect ASR Model	Yes	Multi-dialect Tamil speech	0.35–0.45	Trained specifically on dialect speech
Multi-Dialect Speech Corpus Creation for Enhancing Tamil ASR (2025)	Dialect-specific ASR	Yes	Tamil dialect corpus	0.40	Focused on corpus development
Junlp@LT-EDI 2025 Efficient Low-Rank Adaptation of Whisper	Whisper with LoRA	Yes	Inclusive Tamil speech	0.38	Used parameter-efficient fine tuning
Exploration of Whisper Fine-Tuning Strategies for Low-Resource ASR (2024)	Whisper	Yes	Low-resource languages, including Tamil	0.30–0.45	Fine-tuning improves results clearly
This System: Whisper Tamil Large-v2	Whisper Tamil Large-v2	No	Tamil dialect speech (580 samples)	0.71	Zero fine-tuningng baseline

Result The system was tested on 579 dialect voice samples using Whisper Tamil Large v2 with zero fine-tuning. The Word Error Rate was high at 0.71 because no domain adaptation was used. Nevertheless, the model generated comprehensible Tamil text, demonstrating that huge pretrained models may function fairly well in environments with limited resources. Better training data and dialect-specific tuning can lead to further advances. Code for the proposed system is available in the link ([1])

6 Conclusion

This study evaluated Whisper Tamil Large v2 for Tamil dialect speech recognition using 579 speech samples in a zero fine-tuning setting. The model achieved a Word Error Rate of 0.71, showing that dialect-rich Tamil speech remains challenging for pretrained multilingual ASR systems. Most recognition errors occurred in region-specific vocabulary, colloquial expressions, and pronunciation variations. Even so, the model generated partially understandable Tamil text and correctly identified many common words, showing reasonable generalisation capability without dialect-specific adaptation. This work provides a baseline for dialectal Tamil ASR and highlights the importance of dialect-aware fine-tuning and balanced training data for improving speech recognition performance in low-resourced dialect settings.

[1] https://github.com/Prisur2013/TAMILVOICELABS_1991/tree/main

References

Priyobroto Acharya, Soham Chaudhuri, Sayan Das, Dipanjan Saha, and Dipankar Das. 2025. Junlp@ lt-edi-2025: Efficient low-rank adaptation of whisper for inclusive tamil speech recognition targeting vulnerable populations. In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 17–25.

Maha Tufail Agro, Atharva Kulkarni, Karima Kadaoui, Zeerak Talat, and Hanan Aldarmaki. 2025. Code-switching in end-to-end automatic speech recognition: A systematic literature review. *arXiv preprint arXiv:2507.07741*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, and 1 others. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Érik Beltrán Lobato. 2025. Multilingual speech-to-speech machine translation.

B. Bharathi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, S. Saranya, and S. Suhasini. 2026. Findings in Tamil Dialect Speech Recognition and Classification. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

B Bharathi, S Saranya, P Vijayalakshmi, and T Nagarajan. 2025. Multi-dialect speech corpus creation for enhancing tamil automatic speech recognition. *Circuits, Systems, and Signal Processing*, pages 1–19.

Xabier de Zuazo, Eva Navas, Ibon Saratxaga, and Inma Hernáez Rioja. 2025. Whisper-lm: Improving asr models with language models for low-resource languages. *arXiv preprint arXiv:2503.23542*.

Dawit Ketema Gete, Bedru Yimam Ahmed, Tadesse Destaw Belay, Yohannes Ayana Ejigu, Sukairaj Hafiz Imam, Alemu Belay Tessema, Mohammed Oumer Adem, Tadesse Amare Belay, Robert Geislinger, Umma Aliyu Musa, and 1 others. 2025. Whispering in amharic: Fine-tuning whisper for low-resource language. *arXiv preprint arXiv:2503.18485*.

Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M Khapra. 2023. Indicsuperb: A speech processing universal performance benchmark for indian languages. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 12942–12950.

Yunpeng Liu, Xukui Yang, and Dan Qu. 2024. Exploration of whisper fine-tuning strategies for low-resource asr. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):29.

Markus Miiller, Sebastian Stiiker, and Alex Waibel. 2018. Multilingual adaptation of rnn based asr systems. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5219–5223. IEEE.

Gowtham Papala, Aniket Ransing, and Pooja Jain. 2023. Sentiment analysis and speaker diarization in hindi and marathi using using finetuned whisper: sentiment analysis in hindi and marathi. *Scalable Computing: Practice and Experience*, 24(4):835–846.

Chandrika Prasad, Veenga Gode Swamy Rao, R China Appala Naidu, and 1 others. 2026. An asr transformer-based model for kannada speech-to-text transcription. *Journal of Artificial Intelligence and Technology*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Frederick Riemenschneider and Anette Frank. 2025. Cross-lingual generalization and compression: from language-specific to shared neurons. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13470–13491.

S Saranya, B Bharathi, S Gomathy Dhanya, and Aishwarya Krishnakumar. 2025. Real-time continuous tamil dialect speech recognition and summarization. *Circuits, Systems, and Signal Processing*, 44(4):2855–2881.

Sanket Shah, Kavya Ranjan Saxena, Kancharana Manideep Bharadwaj, Sharath Adavanne, and Nagaraj Adiga. 2025. Indicst: Indian multilingual translation corpus for evaluating speech large language models. In *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE.

Suwon Shon, Ahmed Ali, and James Glass. 2018. Convolutional neural networks and language embeddings for end-to-end dialect recognition. *arXiv preprint arXiv:1803.04567*.

Xuchen Wei, Yangxin Wu, Yaoyin Zhang, Henglyu Liu, Kehai Chen, Xuefeng Bai, and Min Zhang. 2025. Hitsz's end-to-end speech translation systems combining sequence-to-sequence auto speech recognition model and indic large language model for iwslt 2025 in indic track. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 405–411.

Jiqiao Zhang and Degen Huang. 2025. Speech recognition for low-resource languages using large language models and related-language data. In *International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2025)*, volume 13664, pages 1264–1269. SPIE.