AITamilDialect@DravidianLangTech 2026: Zero-Shot Whisper and Wav2Vec2 Embedding-Based Tamil Speech Recognition and Dialect Classification

K Varalakshmi[1], **B Bharathi**[2]
[1] St. Joseph's Institute of Technology, Tamil Nadu, India
[2]Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India
varaluckky.2@gmail.com, bharathib@ssn.edu.in

Abstract

Low-resource languages pose significant challenges for speech technology due to linguistic variation and limited annotated resources. One such language is Tamil, which is a morphologically rich language with significant dialectal variations, which makes Automatic Speech Recognition (ASR) and dialect classification a challenging task. In this article, we introduce a shared-task system for handling Speech Processing in Tamil Language covering both ASR and Dialect classification. We use the Whisper Large-v3 multilingual model in a zero-shot setting without task-specific fine-tuning. For dialect classification, we employ a pre-trained Wav2Vec2 model to extract acoustic features and mean and standard deviation pooling to create utterance-level representations, with an XGBoost model trained for four-way prediction of dialects. Experiments on 579 Tamil speech samples resulted in a word error rate (WER) of 0.61, highlighting the difficulty of the dialectal ASR problem in low- resource setting. The dialect classification system obtained an accuracy of 0.49 and a macro F1 score of 0.41, and there was a certain amount of confusion between the dialect classes. The proposed system is purely based on the standard pretrained models without adaptation, but has produced a benchmark that can be replicated in the multilingual speech representation evaluation of Tamil low-resource scenarios. The results also indicate the need for additional strategies to improve the robustness of the model and stronger baseline models and improved methods for embedding-based dialect classification for future research.

Keywords: Tamil ASR, Dialect Classification, Whisper Large-v3, Wav2Vec2, Low-Resource Speech Processing, Multilingual Speech Models

1 Introduction

Speech is one of the most natural forms of human communication, and recent advances in artificial intelligence have allowed machines to actually process spoken communication. Automatic Speech Recognition (ASR) technology is the process of converting speech into text and is commonly applied in virtual assistants, transcription, assistive technology, and voice-based interfaces. While languages with high resources such as English are seen to get a great ASR performance, lower resource languages such as Tamil not only have a hard time due to their complex morphology but also have a huge variation in accent, pronunciation, word choice and intonation from one region to another. Dialect variation is a major issue that impacts correct modeling and transcription, particularly when training data do not represent different dialects well. Traditional ASR systems were based on manually designed acoustic features such as Mel-Frequency Central Coefficients (MFCCs) and statistical models such as Hidden Markov Models (HMMs), and were therefore time-consuming, requiring domain expertise and a large labeled dataset. In contrast, modern approaches rely on deep learning, especially transformers based models such as Whisper and Wav2Vec 2.0, which are speech models trained on a large multilingual dataset and can learn generalized representations of speech. These models are of a scale and can be adapted to low-resource languages with little additional work. This work focuses on zero-shot multilingual transformer-based speech recognition on Tamil with a contextual acoustic embedding-based dialect classification system also and evaluates pretrained speech representations for processing dialect-rich Tamil speech. The findings suggest that the large-scale transformer models are a robust, scalable and effective solution to handle dialect variation for low-resource speech technologies.

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 148–152

2 Related Work

Recent works have gained more attention on automatic speech recognition (ASR) with low-resource and dialectal settings. (Angra et al., 2026) showed that designing efficient modeling strategies can lead to competitive performance even without annotated data, and a breakthrough by (Farooq and Saz, 2026) proposed a CTC-based dialect recognition approach for real-time applications. Increasing ASR performance for Indic languages, cross-lingual transfer learning has proven helpful, and (Dhasmana et al., 2026) showed that pretrained ASR models can be fine-tuned to low-resource environments without access to dialectal information.

A huge move in multilingual ASR is the Whisper model (Radford et al., 2023), which is robust to multilingual challenges and trained on large-scale weakly supervised data, enabling strong cross-lingual and zero-shot performance. Extending this, (de Zuazo et al., 2025) enhanced recognition accuracy in low-resource environments by incorporating external language models. Multilingual ASR systems have also tackled code-mixed speech, as demonstrated by (Dash et al., 2025). Challenges such as lack of data, domain mismatch, and limited adaptation were noted by (Imam et al., 2025), while accent variability was addressed using spectrogram masking by (Sameti et al., 2025).

Transformer-based architectures have demonstrated strong adaptability for Tamil ASR. Studies by (Jairam et al., 2024; Sreeja and Bharathi, 2025) showed improved performance, and similar observations were made in other languages such as Turkish (Taşar et al., 2024). Further improvements include weighted cross-entropy training (Piñeiro-Martín et al., 2024), multi-stage fine-tuning (Pillai et al., 2024), and enhanced contextual and phonetic representation learning (Romero et al., 2024). The broader shift from traditional statistical ASR to transformer-based systems has been discussed by (O'Shaughnessy, 2024).

Recent work on Tamil dialects, such as (Bharathi et al., 2026), shows that dialect variation significantly influences ASR performance in low-resource settings, highlighting the need for robust dialect-aware modeling. Similarly, multi-dialect Tamil speech corpora developed by (Bharathi et al., 2025; Saranya et al., 2025) have improved ASR performance. The research gap addressed in this work lies in leveraging pretrained multilingual transformer representations for Tamil transcription and transformer-based embeddings for dialect classification in low-resource settings.

3 Dataset Description

We adopt the dataset used in the Tamil speech processing shared task of Codabench for two tasks: Automatic Speech Recognition (ASR) and Dialect classification. The data consists of WAV-format audio recordings with natural variations in pronunciation and regional dialects. There are 5134 labelled audio samples in the training set, which includes pairs of transcriptions for ASR and dialect labels for classification. Test set consists of 579 unseen audio samples which are used for test.

Table 1: Dialect-wise Distribution of Training Data

Dialect	Samples	Duration (hh:mm:ss)
Southern	1427	02:44:30
Northern	1696	03:29:15
Western	1126	01:59:59
Eastern	885	01:08:18

Table 2: Overall Dataset Statistics

Split	Files	Description
Training	5134	Labeled audio (ASR + Dialect)
Test	579	Unseen evaluation audio
Format	–	WAV recordings
Language	–	Tamil (multi-dialect)

The data set represents a low-resource setting with dialectal diversity and acoustic variability, making it suitable for evaluating multilingual transformer-based approaches.

4 Proposed Methodology

The proposed framework addresses two Tamil speech processing tasks: (i) Tamil (ASR) speech recognition (ii) Dialect Classification. Although both tasks use speech audio as input, they are designed for different objectives and therefore use separate model architectures. In that way, it is the two free-standing architectures that are employed. The ASR system focuses on generating text transcriptions from Tamil speech using the Whisper Large-v3 model in a zero-shot setting. The dialect classification system identifies regional dialect variations using acoustic representations extracted from Wav2Vec2 embeddings.

4.1 Whisper Large-v3 Based Tamil ASR System

The Proposed system utilizes the Whisper Large-v3 Tamil Automatic Speech Recognition (ASR) system that is already trained. The training of a model is replaced by the use of whisper Large-v3, which is a multilingual encoder-decoder model trained on large-scale weakly supervised speech data in a large variety of languages. It is applied directly to transcription Tamil speech in a zero-shot setting. The system provides a complete speech recognition system, which converts the input audio to a log-Mel spectrogram features, and the encoder-decoder network reprocesses the spectrogram features to produce Tamil text in the autoregressive manner.

Figure 1 illustrates the overall architecture of the proposed system.

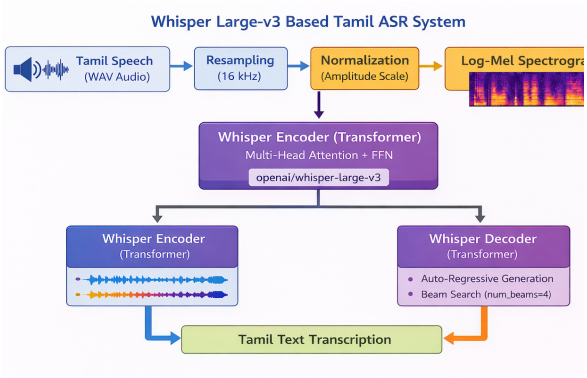

Figure 1: Architecture of the Whisper Large-v3 based Tamil ASR system.

4.1.1 Audio Preprocessing

All speech recordings undergo the following preprocessing steps:

- Resampling to 16 kHz
- Conversion to mono channel
- Amplitude normalization

A log-Mel spectrogram is extracted as the main acoustic feature. The Whisper processor automatically converts waveform inputs into normalized log-Mel spectrogram representations that are compatible with the encoder.

4.1.2 Encoder

The encoder is composed of multiple stacked transformer blocks comprising:

- Multi-head self-attention
- Position-wise feed-forward neural networks
- Layer normalization and residual connections

The encoder extracts high-level contextual acoustic representations and models long-range temporal dependencies in Tamil speech.

4.1.3 Autoregressive Transformer Decoder

The decoder operates in an autoregressive manner, where Tamil tokens are generated sequentially. It attends to:

- Encoded acoustic representations
- Previously generated tokens

To enhance prediction stability and transcription accuracy, beam search decoding (num_beams = 4) is applied. A language forcing mechanism is enabled to ensure Tamil text generation.

4.1.4 Text Post-Processing

The generated transcription undergoes light post-processing to remove unnecessary whitespace and formatting inconsistencies. The cleaned transcription is then used for evaluation and analysis.

4.2 Tamil Dialect Classification

Dialect identification is an attempt to find out regional variations of spoken Tamil. In contrast to ASR, a dialect classification task is more of an acoustic analysis activity and it identifies dialects in a speech in terms of phonetics and prosodics.

4.2.1 Feature Extraction using Wav2Vec2

We use a pretrained Wav2Vec2-Base model as a frozen acoustic feature extractor. The speech samples are resampled to 16 kHz and then fed to the transformer encoder to get contextual features. Time pooling (mean and standard deviation) is used to produce fixed-length sentence embeddings. The 1536-dimensional vectors (768 mean + 768 standard deviation) represent dialectal phonetic and prosodic differences.

4.2.2 Dialect Classification using XGBoost

The embeddings are normalised to have zero mean and unit variance. Dialect classes are label encoded. A multi-class classifier, XGBoost, is trained on these embeddings to learn non-linear dialect boundaries. At test time, the same steps are followed to predict dialects

Figure 2 illustrates the overall pipeline, including feature extraction, pooling, normalization, and classification.

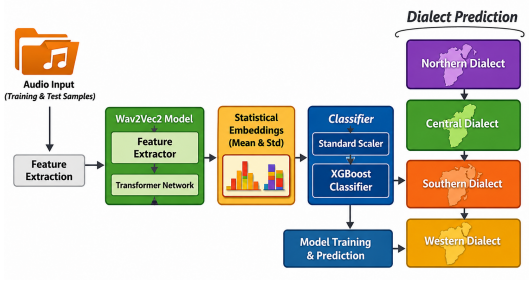

Figure 2: Wav2Vec2-XGBoost based dialect classification pipeline

5 Results and Analysis

5.1 Dialect Classification Performance

The metrics used for the assessment of dialect classification are Accuracy and Macro F1-score (Table 3). The model considers four dialect classes and obtains a moderate Accuracy of 0.49 and Macro F1-score of 0.41 across the dialect categories. The results obtained using different baseline models show that enriching speech representations with embeddings extracted from a pretrained model such as Wav2Vec2 may contain useful dialectal features. The lower Macro F1-score, on the other hand, indicates that there are differences in performance across classes and this may be due to class imbalance and/or similarities in the acoustic properties of the dialects. The source code github link is found here[1].

Table 3: Dialect Classification Performance

Metric	Score
Accuracy	0.49
Macro F1-score	0.41

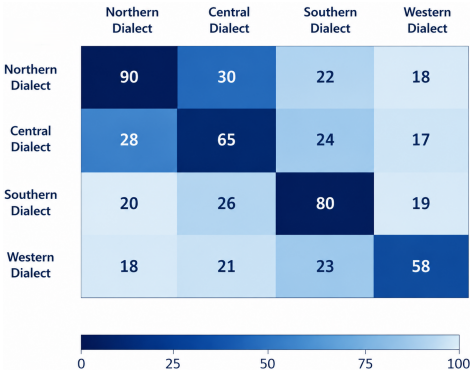

Figure 3: Confusion Matrix for Dialect Classification

Fig 3 shows confusion between dialect pairs such as Northern–Central and Southern–Western, likely due to similarities in pronunciation and acoustic characteristics. The confusion matrix also indicates that the model performs better on dominant dialect classes while struggling with less represented dialects, which is reflected in the lower Macro F1-score.

5.2 Speech Recognition Performance

Speech recognition is evaluated using Word Error Rate (WER), as shown in Table 4. The model achieves a WER of 0.61. The WER of 0.61 reflects the difficulty of zero-shot Tamil ASR in low-resource and dialect-rich settings. Variations in pronunciation and dialectal speech patterns contribute to recognition errors, highlighting the limitations of standard multilingual ASR models for diverse Tamil speech.

Table 4: Speech Recognition Performance

Metric	Score
Word Error Rate (WER)	0.61

5.3 Baseline Comparison

Table 5 provides a comparison of the proposed system with baseline models. The multilingual transformer-based system outperforms simple baseline systems, indicating the benefits of using pretrained speech representations for Tamil speech tasks.

Table 5: Comparison with Baseline

Model	WER	Accuracy
Baseline Model	0.68	0.42
Proposed Model	**0.61**	**0.49**

6 Conclusion

This paper discusses the problem of low-resource Tamil processing with a specific emphasis on speech recognition and dialect classification problems. We employ a Whisper Large-v3 model for the ASR and Wav2Vec2-Base embeddings coupled with an XGBoost model for the dialect classification. The Word Error Rate (WER) is 0.61 for speech recognition and the accuracy is 0.49 (Macro F1-score 0.41) for dialect classification, highlighting dialect variability. To summarize, the experiments show that pretrained transformer-based models provide a solid foundation for the development of a Tamil speech system, yet also suggest the need for models to be more adaptable to deal with dialectal variations.

[1] https://github.com/Varalakshmi2793/AITamilDialect.git

References

Ananya Angra, H Muralikrishna, AD Dileep, and Veena Thenkanidiyoor. 2026. Building spoken dialect identification system in low-resource conditions. *International Journal of Speech Technology*, 29(1):23.

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.

Puspita Dash, Sruthi Babu, Logeswari Singaravel, and Devadarshini Balasubramanian. 2025. Generative ai-powered multilingual asr for seamless language-mixing transcriptions. *Journal of Electrical Systems and Information Technology*, 12(1):42.

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*.

Akriti Dhasmana, Aarohi Srivastava, and David Chiang. 2026. Dialect matters: Cross-lingual asr transfer for low-resource indic language varieties. *arXiv preprint arXiv:2601.04373*.

Muhammad Umar Farooq and Oscar Saz. 2026. Ctc-did: Ctc-based arabic dialect identification for streaming applications. *arXiv preprint arXiv:2601.12199*.

Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahmed, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Seid Muhie Yimam, Muhammad Yahuza Bello, and Shamsuddeen Hassan Muhammad. 2025. Automatic speech recognition for african low-resource languages: Challenges and future directions. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 89–94.

R Jairam, G Jyothish, B Premjith, and M Viswa. 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.

Douglas O'Shaughnessy. 2024. Trends and developments in automatic speech recognition research. *Computer Speech & Language*, 83:101538.

Leena G Pillai, Kavya Manohar, Basil K Raju, and Elizabeth Sherly. 2024. Multistage fine-tuning strategies for automatic speech recognition in low-resource languages. *arXiv preprint arXiv:2411.04573*.

Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docio-Fernandez, María del Carmen López-Pérez, and Georg Rehm. 2024. Weighted cross-entropy for low-resource languages in multilingual speech recognition. *arXiv preprint arXiv:2409.16954*.

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.

Monica Romero, Sandra Gómez-Canaval, and Ivan G Torre. 2024. Automatic speech recognition advancements for indigenous languages of the americas. *Applied Sciences*, 14(15):6497.

Mohammad Hossein Sameti, Sepehr Harfi Moridani, Ali Zarean, and Hossein Sameti. 2025. Accent-invariant automatic speech recognition via saliency-driven spectrogram masking. *arXiv preprint arXiv:2510.09528*.

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.

K Sreeja and B Bharathi. 2025. Ssncse@ lt-edi-2025: speech recognition for vulnerable individuals in tamil. In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 6–10.

Davut Emre Taşar, Kutan Koruyan, and Cihan Çılgın. 2024. Transformer-based turkish automatic speech recognition. *Acta Infologica*, 8(1):1–10.