Wave2Word@DravidianLangTech 2026: WhisTam: A unified framework for dialect based Tamil speech recognition and classification

Ruwad Naswan[1], Shadab Tanjeed Ahmad[2],
[1]Bangladesh University of Engineering and Technology, [2]Islamic University of Technology,
Correspondence: ruwad45678@gmail.com

Abstract

While Automatic Speech Recognition (ASR) systems have shown impressive performance in languages having sufficient annotated speech data like English, their performance is still limited for low-resource, dialect rich languages like Tamil. Tamil poses further challenges because of its extremely high regional variation in dialects that manifest in varying vocabulary, pronunciations, and even syntactic structures. To address these challenges, we present a unified framework WhisTam based on the Whisper medium model, which performs speech transcription and dialect classification jointly within a single system. Our method is evaluated against speech samples from four regional dialects and achieves a macro F1-score of 0.53 and a Word Error Rate (WER) of 0.55 for dialect classification and transcription respectively, ranking 2nd in the dialect classification task and 3rd in the transcription task in the DravidianLangTech@ACL 2026 shared task on Dialect-based Speech Recognition and Classification in Tamil. These findings emphasize the challenges in dialectal Tamil ASR as well as the promise of multi-task learning for low-resource languages. Our implementation is publicly available at: https://github.com/rwd51/DravidianLangTech-Wave2Word.

1 Introduction

Recent advances in deep learning and transformer-based architectures have greatly improved ASR for well-resourced languages (Radford et al., 2022; Baevski et al., 2020; Hinton et al., 2012). However, progress remains uneven across languages, particularly for low-resource, dialect-rich languages where annotated speech data is scarce (Besacier et al., 2014; Pratap et al., 2020). Insufficient transcribed data combined with phonetic and acoustic variability leads to substantially higher error rates for such languages (Besacier et al., 2014). Multilingual training, transfer learning, and parameter-efficient adaptation of pretrained models have emerged as promising directions to bridge this gap (Conneau et al., 2020).

Tamil, a major Dravidian language spoken by over 80 million people (Nanmalar et al., 2024a), remains relatively underexplored in ASR research (Pratap et al., 2020). Its high regional dialectal variation in pronunciation, vocabulary, and syntax poses a significant challenge, especially when ASR systems are trained on limited or standardized corpora.

To address this, we propose WhisTam, a unified framework that jointly performs speech transcription and dialect classification for Tamil. Built on a pre-trained Whisper model (Radford et al., 2022), WhisTam introduces a lightweight classifier adapter on the encoder to simultaneously identify the dialect and generate transcriptions across four regional groups: northern, southern, western, and central. Our experiments yield a macro F1-score of 0.53 for dialect classification and a WER of 0.55 for transcription, highlighting both the difficulty of dialectal Tamil ASR and the promise of multi-task learning for low-resource settings.

2 Related Works

Early Dravidian speech processing relied on classical features like MFCC combined with GMMs, HMMs, and SVMs for language and dialect classification (Ismail and Singh, 2017; Koolagudi et al., 2012), achieving around 88% accuracy across Tamil, Telugu, Kannada, and Malayalam (Abdul and Al-Talabani, 2022; Koolagudi et al., 2012). Deep learning approaches have since pushed these results further: CNN-based spectrogram representations reached over 98% on Indic speech datasets (Godbole et al., 2020), and MFCC with handcrafted feature combinations achieved F1 scores above 0.99 for Tamil dialect identification (Nan-

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 442–446

malar et al., 2019, 2022, 2024b). Multilingual ASR training and multitask phone mapping have also shown gains in low-resource Dravidian settings by leveraging cross-language acoustic similarities (Madhavaraj and Ganesan, 2022).

Despite this progress, dialectal variation, speaker diversity, and limited labeled data continue to hinder robust ASR for most Dravidian languages (Das and Bhattacharjee, 2024; Ahmad Dar and Pushparaj, 2025), motivating frameworks that can generalize across dialects and low-resource conditions.

3 Methodology

3.1 Problem Statement

This work is conducted as part of the DravidianLangTech 2026 (Bharathi et al., 2025, 2026) shared task at ACL, which focuses on advancing speech technologies for dialectal Tamil.

The shared task consists of two subtasks designed to evaluate models on dialect-aware speech understanding.

3.1.1 Subtask 1: Speech-Based Dialect Classification

The goal of this subtask is to determine the dialect of a provided Tamil speech recording and categorize each given audio sample into one of four dialect categories: *Northern*, *Southern*, *Western*, or *Central*.

3.1.2 Subtask 2: Dialectal Tamil Automatic Speech Recognition

The second subtask focuses on automatic speech recognition for dialectal Tamil. Participants are required to generate Tamil text transcriptions from speech recordings containing dialectal variation.

3.2 Architecture Overview

Our framework builds upon a Tamil-adapted version of the Whisper model (Radford et al., 2022). Specifically, we initialize from vasista22/whisper-tamil-medium, a publicly available model fine-tuned on multiple Tamil ASR corpora. Whisper is a model that transforms input speech into log-Mel spectrogram features. These features are fed through a transformer encoder to create contextual acoustic representations, which are then used by a transformer decoder to produce autoregressive text. Our approach retains the core encoder and decoder architecture but introduces lightweight extensions that enables dialect-aware learning.

Regional Adapter. To incorporate dialect awareness, we introduce a *Regional Adapter* layered on top of the encoder. A small embedding layer maps each dialect label (Northern, Southern, Western, Central) into a low-dimensional vector. This vector is projected to the encoder hidden dimension using a linear layer and normalized via LayerNorm. During training, the projected embedding is broadcast-added to every timestep of the encoder hidden states:

$$\mathbf{h}'_t = \mathbf{h}_t + \mathbf{e}_{region}$$

This allows the model to condition its representations on dialect-specific characteristics such as phonetic cues and prosody. During inference, the dialect embedding is omitted, and the model operates as a standard Whisper ASR system.

Dialect Classification Head. To explicitly learn dialect-discriminative features, we attach an auxiliary classification head. Encoder hidden states are aggregated using an attention-based pooling mechanism, which assigns higher weights to frames informative for dialect recognition. The pooled representation is passed through dropout and a linear classifier to predict the dialect category. Attention pooling had been used over simple mean or max pooling as dialect markers are not uniformly distributed across time.

Multi-task Learning. The model is trained with a multi-task objective combining ASR and dialect classification. Let $\mathcal{L}_{ASR}$ denote the standard Whisper cross-entropy loss for next-token prediction, and $\mathcal{L}_{dialect}$ denote the cross-entropy loss for dialect classification. The total training loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{ASR} + \alpha \cdot \mathcal{L}_{dialect} \quad (1)$$

where α balances the contribution of the dialect classification loss. This multi-task setup encourages the shared encoder to organize its representations around dialect-specific patterns while maintaining strong transcription performance.

Overall, the Regional Adapter introduces a small number of additional parameters relative to the base Whisper model, enabling dialect-aware ASR with minimal architectural modification.

4 Experiments

4.1 Dataset and Preprocessing

Dataset. The Tamil Dialect Speech Dataset (Bharathi et al., 2025, 2026) covers four major

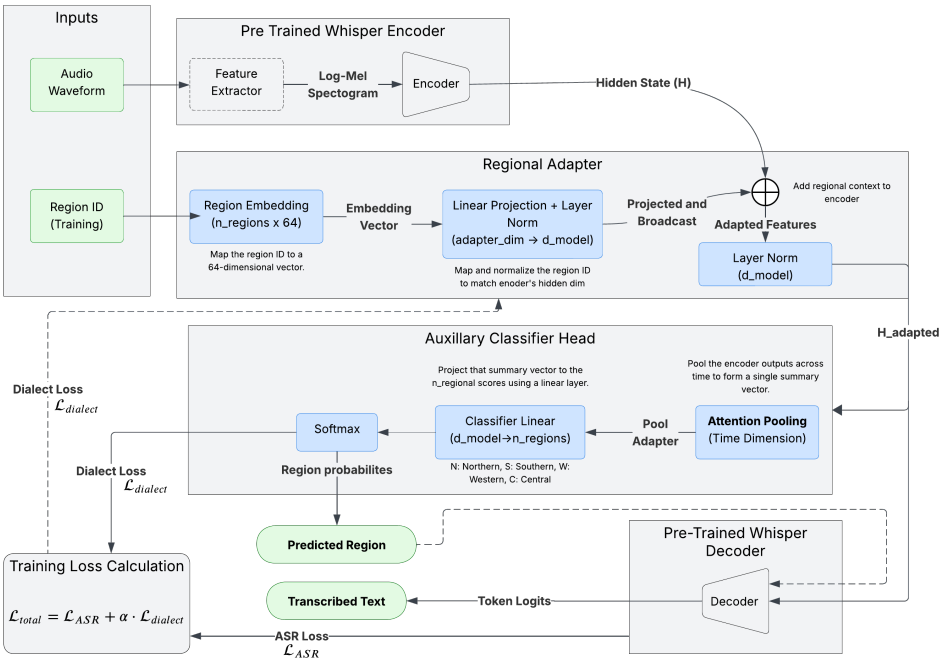

Figure 1: Classifier adapter architecture used in the regional Whisper model.

dialect regions of Tamil Nadu: Northern, Southern, Western, and Central. It contains both spontaneous and read speech recorded by male and female speakers across diverse age groups in natural acoustic environments at 16 kHz. The training set consists of 9.22 hours of manually transcribed speech, and 2.05 hours are released as test set for evaluation. The 9.22-hour training corpus is partitioned into a 90/10 train/validation split using stratified sampling.

Preprocessing. Text normalization is applied to transcriptions, including Unicode NFC normalization, removal of zero-width characters, Tamil visarga normalization, punctuation removal, English lowercasing, and whitespace normalization. Audio is converted to log-mel spectrograms using the Whisper feature extractor, zero-padded or truncated to a fixed length.

During training, audio-level augmentations are applied via curriculum learning, including time-stretching, pitch shifting, additive Gaussian or pink noise, volume shifts, time shifts, and random cropping. SpecAugment is further applied to spectrograms, masking frequency and time bands to improve robustness to acoustic variability.

4.2 Evaluation

For **dialect classification**, system performance is measured using macro-averaged F1 score across the four dialect classes.

For **dialectal ASR**, performance is measured using Word Error Rate (WER) between predicted transcriptions and reference human transcriptions:

$$WER = \frac{S + D + I}{N}$$

where S, D, and I are the numbers of substitutions, deletions, and insertions, respectively, and N is the total number of words in the reference.

4.3 Results

We report results for both tasks on the stratified validation split from the training set, as well as on the hidden test set used in the shared task competition.

4.3.1 Automatic Speech Recognition (ASR)

Table 1 shows per-dialect WER on the validation set and the overall WER (mean across dialects) on validation as well as the overall WER on the hidden test set. Lower WER indicates better transcription quality.

Figure 2: t-SNE visualization of encoder embeddings on the validation set. Left: pretrained Whisper encoder (mean-pooled). Right: trained regional adapter (attention-pooled). The trained model produces clearly separated dialect clusters.

Dialect / Set	**Validation WER (%)**	**Hidden Test WER (%)**
Central Dialect	30.23	–
Northern Dialect	36.70	–
Southern Dialect	40.08	–
Western Dialect	41.68	–
Mean / Overall	37.17	55.0

Table 1: ASR performance (WER %) for each Tamil dialect on the validation set and overall on the hidden test set.

4.3.2 Dialect Classification

Table 2 shows precision, recall, and F1 for each dialect on the validation set, and the overall macro F1 on the hidden test set. Higher values indicate better classification performance.

Table 2: Dialect classification performance on validation and hidden test set.

Dialect	**Precision**	**Recall**	**F1 Score**	**Hidden Macro F1**
Northern	0.97	1.00	0.98	–
Southern	0.98	1.00	0.99	–
Western	1.00	0.90	0.95	–
Central	0.94	1.00	0.97	–
Overall	0.97	0.98	0.97	0.53

4.3.3 Embedding Space Analysis

To qualitatively assess whether the regional adapter encourages dialect-discriminative representations, we visualize the attention-pooled encoder outputs using t-SNE on the held-out validation set. As shown in Figure 2, the pretrained Whisper encoder produces embeddings with no visible dialect structure, while the trained regional adapter yields clearly separated clusters corresponding to the four dialect regions. This confirms that the multi-task objective guides the shared encoder to organize its representations around dialect-specific acoustic patterns.

Conclusion

We introduce WhisTam, a framework developed on top of a Tamil-adapted Whisper model incorporating a small regional adapter for Tamil dialect speech recognition and classification. Our method jointly addresses ASR and dialect identification, leveraging multi-task learning to enhance robustness to dialectal variations. Evaluation on stratified validation and hidden test sets show strong performance of our dialect-aware ASR as well as classification task. Despite the limited size and imbalance of the dataset, our proposed method demonstrates the potential of lightweight adaptation for low-resource dialectal ASR.

Ethical Considerations

This work uses publicly available speech data provided by the DravidianLangTech 2026 shared task organizers with appropriate consent procedures. No additional data is collected or redistributed. We acknowledge that dialect identification systems could potentially be misused for speaker profiling; our intended use is solely to improve speech technology accessibility for underrepresented Tamil

dialect communities.

Limitations

Our dataset is relatively small (9.22 hours) and unevenly distributed across dialects, which may limit generalization to underrepresented speakers and regions. Intra-dialect variation such as speaker age, gender, and code-switching remains difficult for the lightweight adapter to fully capture. Future work could extend the regional adapter to model finer-grained dialectal characteristics and improve cross-dialect ASR performance.

References

Zrar Khald Abdul and Abdulbasit K. Al-Talabani. 2022. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 10:122136–122158.

Muzaffar Ahmad Dar and Jagalingam Pushparaj. 2025. Machine learning and deep learning approaches for accent recognition: A review. *IEEE Access*, 13:51527–51550.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

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.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *Preprint*, arXiv:2006.13979.

Hem Das and Utpal Bhattacharjee. 2024. Assamese dialect identification using static and dynamic features from vowel. *Journal of Advances in Information Technology*, 15:306–321.

Shubham Godbole, Vaishnavi Jadhav, and Gajanan K. Birajdar. 2020. Indian language identification using deep learning. *ITM Web of Conferences*.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Tanvira Ismail and L. Joyprakash Singh. 2017. Dialect identification of assamese language using spectral features. *Indian journal of science and technology*, 10:1–7.

Shashidhar G. Koolagudi, Deepika Rastogi, and K. S. Sekhara Rao. 2012. Identification of language using mel-frequency cepstral coefficients (mfcc). *Procedia Engineering*, 38:3391–3398.

A. Madhavaraj and Ramakrishnan Angarai Ganesan. 2022. Data and knowledge-driven approaches for multilingual training to improve the performance of speech recognition systems of indian languages. *Preprint*, arXiv:2201.09494.

M. Nanmalar, S. Johanan Joysingh, P. Vijayalakshmi, and T. Nagarajan. 2024a. A feature engineering approach for literary and colloquial tamil speech classification using 1d-cnn. *Speech Commun.*, 173:103254.

M. Nanmalar, S. Johanan Joysingh, P. Vijayalakshmi, and T. Nagarajan. 2024b. A feature engineering approach for literary and colloquial tamil speech classification using 1d-cnn. *Speech Commun.*, 173:103254.

M. Nanmalar, P. Vijayalakshmi, and T. Nagarajan. 2019. Literary and colloquial dialect identification for tamil using acoustic features. *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, pages 1303–1306.

M. Nanmalar, P. Vijayalakshmi, and T. Nagarajan. 2022. Literary and colloquial tamil dialect identification. *Circuits, Systems, and Signal Processing*, 41:4004 – 4027.

Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. 2020. Massively multilingual asr: 50 languages, 1 model, 1 billion parameters. *Preprint*, arXiv:2007.03001.

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.