IIITK_SpeechScape@DravidianLangTech 2026: Dialect based speech recognition and classification using Speech Foundation Models and Deep Learning Techniques

G Srishtik Sekar[1]*, Harissh Ragav Dhamodaran[1]*, Kishore Shankar S[1]*, Balasubramanian Palani[2]†, R. Tharaniya sairaj[2]
[1]Department of Computer Science and Engineering, IIIT Kottayam, Kerala, India
[2]Assistant Professor, Indian Institute of Information Technology Kottayam
{srishtik23bcs220, harissh23bcs59, kishore23bcs110, pbala, sairaj}@iiitkottayam.ac.in

Abstract

Dialectal variation poses a significant challenge to Automatic Speech Recognition (ASR), particularly for low resource morphologically rich languages such as Tamil. Although widely spoken in India, Sri Lanka, and the global diaspora, Tamil exhibits substantial phonetic, lexical, and prosodic variation across dialects, complicating both dialect classification and speech recognition. In this work, we address these tasks within a unified framework. We evaluate state-of-the-art models for dialect classification, including Whisper, CLDNN, wav2vec, and wavLM, and for ASR, Whisper and a zero-shot Conformer. Whisper achieved the best performance with **0.46** Macro F1 for dialect classification and **0.57** WER for ASR. These results highlight the strong generalization capability of transformer-based foundation models across dialects and languages. The code is publicly available in github for research purpose.[1]

1 Introduction

Speech technology performance often degrades under dialectal variation. This is particularly true for Tamil, a morphologically rich classical language. Differences in pronunciation, vocabulary, and intonation make dialect-sensitive ASR challenging. Recent work has highlighted the need for multi-dialect Tamil speech corpora and dialect-aware ASR systems (Bharathi et al., 2025). Despite its importance, Tamil speech technology remains under-explored, particularly in handling dialectal diversity.

This work addresses two primary tasks:

- **Dialect Classification** – identifying the dialect of a Tamil speech utterance.
- **Automatic Speech Recognition (ASR)** – transcribing Tamil speech into text.

For dialect classification, we evaluate Whisper, CLDNN, wav2vec, and wavLM, covering both CNN-RNN hybrid and transformer-based self-supervised architectures. For ASR, we assess Whisper and a IndicConformer model for cross-dialect robustness.

Contributions of this work include:

- Evaluation of deep learning models for Tamil **Dialect Classification** and **ASR**.
- Comparison of CLDNN and transformer-based models (wav2vec, wavLM, Whisper), including zero-shot Conformer ASR.
- Best performance achieved using Whisper (Macro F1: 0.46, WER: 0.57).

2 Related Works

Recent work has specifically explored Tamil dialect speech recognition and classification. Bharathi et al. (2026) presented findings on Tamil dialect ASR and classification, highlighting the impact of regional variation on recognition performance. Automatic Speech Recognition (ASR) research has addressed dialectal diversity using unified deep learning models. Li et al. (2018) showed that a single end-to-end sequence-to-sequence model can learn shared representations across dialects by incorporating dialect information during training. In contrast, dialect classification focuses on predicting dialect labels from speech, where earlier approaches relied on engineered acoustic features; Sinha et al. (2015) demonstrated that multi-stream feature fusion improves performance, and Koolagudi et al. (2017) highlighted the effectiveness of spectral and prosodic features for distinguishing closely related South Indian languages. Recent studies emphasize robustness and fairness, with Harris et al. (2024) reporting performance disparities

*Equal contribution.
†Corresponding author.
[1]Code and resources: https://github.com/GL3MON/IIITK-SpeechScape-DravidianLangTech2026

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 268–272

across dialect and demographic groups and Jakhar et al. (2024) improving language identification in multilingual Indic ASR through a unified framework. Recent work shows a transition from handcrafted features to deep learning-based dialect modeling.

3 Methodology

3.1 Problem Definition

We address two related tasks in Tamil speech processing: Automatic Speech Recognition (ASR) and Dialect Classification (DC). Let $D = \{(x_i, y_i^{\text{asr}}, y_i^{\text{dial}})\}_{i=1}^{N}$ denote dataset consisting of N Tamil audio samples. For each sample, x_i represents the raw audio waveform, y_i^{asr} denotes its corresponding transcript, and $y_i^{\text{dial}} \in \{1, 2, \ldots, M\}$ represents the dialect label, where M is the total number of dialect classes. Here, $M = 4$ represents the dialect classes.

The DC task aims to learn a classifier $f_{\text{DC}} : x_i \rightarrow y_i^{\text{dial}}$ that predicts the dialect category of the input audio waveform. For the ASR task, the objective is to learn a mapping $f_{\text{ASR}} : x_i \rightarrow y_i^{\text{asr}}$ that transcribes spoken Tamil into text.

3.2 Data Preprocessing

All audio signals were resampled in 16 KHz mono format. Recordings with SNR $\geq$ 10 dB were retained to reduce noisy samples.

3.3 Dialect Classification Architecture

3.3.1 Speech Foundation Models

To address the speech dialect classification task, we leverage pretrained speech foundation models such as Whisper (Radford et al., 2023), WavLM (Chen et al., 2021), and wav2vec (Schneider et al., 2019) for rich speech representations. These models learn transferable speech representations from large-scale corpora.

For a given input speech signal x_i, we extract hidden representations from all encoder layers. Let $\mathbf{h}_i^{(l)} \in \mathbb{R}^{D_{\text{enc}}}$ denote the representation from the l-th layer. We compute a weighted aggregation:

$$\mathbf{h}_{\text{agg}} = \sum_{l=1}^{L} \alpha_l \mathbf{h}_i^{(l)}, \qquad (1)$$

where α_l are learnable scalar weights. The aggregated representation is then passed through a projection layer that reduces the embedding dimension from $\mathbb{R}^{D_{\text{enc}}}$ to $\mathbb{R}^{D_{\text{proj}}}$. The projected features

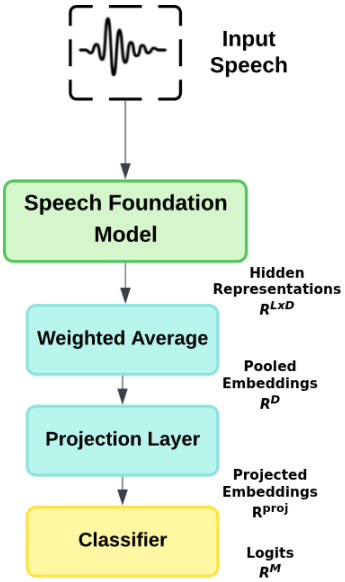

Figure 1: Overview of the proposed dialect classification framework built on top of Speech Foundation Models.

are subsequently fed into a feed-forward network to predict the dialect label.

3.3.2 CLDNN

The CLDNN architecture combines convolutional and recurrent layers to model spectral and temporal speech patterns (Ghahremani et al., 2015). Log-Mel spectrograms are processed by CNN layers for time–frequency feature extraction, followed by a bidirectional LSTM to capture contextual dependencies. A fully connected layer produces the final dialect prediction, enabling modeling of phonetic and prosodic variations. The final dialect prediction is obtained using the softmax function applied to the output logits.

3.4 Model Selection and Comparison

We systematically selected models covering traditional and transformer-based architectures. Whisper, trained on 680,000 hours of multilingual data, provides the strongest general-purpose speech representations with an encoder-decoder architecture that jointly learns speech understanding and language modeling. CLDNN serves as a baseline hybrid approach. WavLM and Wav2vec represent alternative self-supervised approaches with different learning objectives.

The results demonstrate a clear performance hierarchy: Whisper (0.46 F1) substantially outperforms CLDNN (0.33 F1, 39% improvement), WavLM (0.30 F1, 53% improvement), and

Table 1: Sample Entries from the Tamil Dialect Speech Dataset

Dialect	Duration (hrs)	Total Samples	Example Transcription (Tamil)
Central	1:08:18	885	ஆயி இந்த பொடவ மடிப்ப எடுத்து விடேன்.
Southern	2:44:30	1427	ஏல வெயிலு என்ன இன்னைக்கு இந்த பொள...
Northern	3:29:15	1696	இல்ல யூடியூப் பாத்து பண்ணது கிடையாது...
Western	1:59:59	1126	இன்னைக்கு என்னங்க ஒரே உப்பசமா...
Total	**9:22:02**	**5134**	

Wav2vec (0.26 F1). This advantage stems from: (1) *Pretraining scale*: Whisper's 680,000 hours of multilingual data enables richer representations, (2) *Architecture*: Transformer self-attention (0.30–0.46 F1) outperforms CNN-RNN (0.33 F1) by 12–40%, (3) *Pretraining objective*: Supervised ASR training (0.46 F1) outperforms self-supervised methods (0.26–0.30 F1, 35–43% difference) by learning task-specific patterns, and (4) *Robustness*: Diverse audio conditions and accents in pretraining directly transfer to dialect variation challenges. For ASR, Whisper is selected for its superior dialect classification performance, while IndicConformer is evaluated zero-shot to provide a cross-lingual transfer baseline.

3.5 Automatic Speech Recognition Architecture

3.5.1 Whisper

Whisper encoder uses multi-head self-attention to model long-range dependencies in speech representations:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (2)$$

where Q, K, and V denote the query, key, and value matrices.

We fine-tuned Whisper-v3-large using Low-Rank Adaptation (LoRA). This reduces memory usage and catastrophic forgetting. Specifically, LoRA modules are applied to the query, key, and value projection matrices, as well as to the two feed-forward transformation layers across all layers of the model.

3.5.2 IndicConformer

The Conformer architecture uses convolutional neural networks and self-attention mechanisms to effectively capture both local acoustic patterns and global contextual dependencies in the speech signals. IndicConformer follows the original Conformer design (Gulati et al., 2020). Here, the Conformer model is evaluated in a zero-shot fashion.

4 Experiments

4.1 Experimental Setup

All experiments were conducted on a server with two NVIDIA T4 GPUs. For the DC task (Speech Foundation Models), we used a per-device batch size of 2 with gradient accumulation over 4 steps (effective batch size 8). For ASR, Whisper-v3-large was fine-tuned with a batch size of 1 and 4 accumulation steps (effective batch size 4). We used the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1\times10^{-8}$). The learning rate was set to 1×10^{-4} for DC and 3×10^{-5} for ASR, with cosine decay and 0.1% warmup.

4.2 Datasets

The dataset comprises Tamil speech recordings from four major regional dialects — Southern, Northern, Western, and Central — supporting both Dialect Classification and Automatic Speech Recognition (ASR). We also used the IndicVoices dataset (Javed et al., 2024) to improve ASR.

4.3 Performance Metrics

For the DC task we report Accuracy, Precision, Recall, F1, and their macro-averaged variants:

$$\text{Acc.} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$\text{Prec.} = \frac{TP}{TP + FP}, \qquad \text{Rec.} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

$$\text{Macro-Metric} = \frac{1}{K}\sum_{k=1}^{K} \text{Metric}_k \quad (6)$$

Table 2: Dialect Classification Results

Model	Acc	Prec	Rec	Macro F1
Whisper	**0.64**	**0.49**	**0.52**	**0.46**
CLDNN	0.38	0.31	0.34	0.33
WavLM	0.37	0.30	0.32	0.30
Wav2Vec2	0.36	0.26	0.28	0.26

where K is the number of classes. For ASR, we report Word Error Rate (WER) and Character Error Rate (CER):

$$\text{WER/CER} = \frac{S + D + I}{N} \tag{7}$$

where S, D, and I denote substitutions, deletions, and insertions, and N is the total number of words (WER) or characters (CER).

5 Results and Analysis

Tables 2 and 3 show that Whisper achieves the best performance for both dialect classification and ASR due to its pretrained speech representations. Combining IndicVoices with dialect-specific data improves ASR robustness.

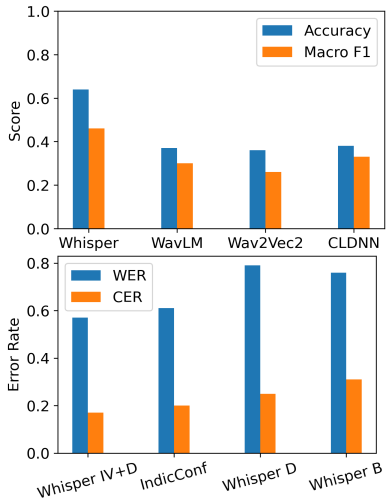

Figure 2: Dialect classification performance (top) and ASR error rates (bottom).

Whisper performs best on Southern dialects, moderately on Northern, and poorly on Central and Western due to limited training data.

6 Conclusion

In this work, we evaluated deep learning and speech foundation models for Tamil dialect classification and ASR. Whisper achieved the best performance with 0.46 Macro F1 and 0.57 WER.

Table 3: ASR Results

Model	WER	CER
Whisper (IndVoic.+Dia.)	**0.57**	**0.17**
Indic-Conformer	0.61	0.20
Whisper (Dia.)	0.79	0.25
Whisper (Base)	0.76	0.31

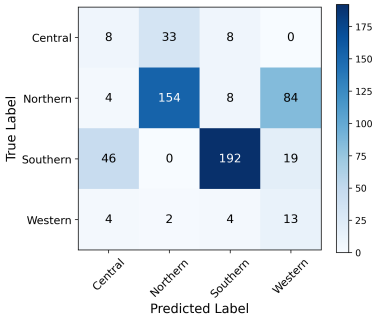

Figure 3: Confusion Matrix of Whisper for Dialect Classification.

Results show that combining general and dialect-specific speech data improves ASR robustness, while low-resource dialects remain challenging. As future work, we plan to explore a unified Whisper-based framework for both dialect classification and ASR by introducing dialect-specific tokens, enabling dialect-aware speech recognition.

References

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.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Long Zhou, Yanmin Qian, and 1 others. 2021. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *arXiv preprint arXiv:2110.13900*.

Pourdamghani Ghahremani and 1 others. 2015. Investigation of cnn□dnn and cnn□rnn architectures for speech recognition. In *Proceedings of Interspeech 2015*, pages 2019–2023.

Anmol Gulati, James Qin, Chung□Cheng Chiu, and 1 others. 2020. Conformer: Convolution□augmented transformer for speech recognition. In *Proceedings of Interspeech 2020*, pages 5036–5040.

Camille Harris, Chijioke Mgbahurike, Neha Kumar, and Diyi Yang. 2024. Modeling gender and dialect bias in automatic speech recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15166–15184, Miami, Florida, USA. Association for Computational Linguistics.

Nikhil Jakhar, Sudhanshu Srivastava, and Arun Baby. 2024. A Unified Approach to Multilingual Automatic Speech Recognition with Improved Language Identification for Indic Languages. In *Interspeech 2024*, pages 3949–3953.

Tahir Javed, Janki Nawale, Eldho George, Sakshi Joshi, Kaushal Bhogale, Deovrat Mehendale, Ishvinder Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Gandhi, Ambujavalli R, Manickam M, C Vaijayanthi, Krishnan Karunganni, and 2 others. 2024. IndicVoices: Towards building an inclusive multilingual speech dataset for Indian languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10740–10782, Bangkok, Thailand. Association for Computational Linguistics.

Shashidhar G. Koolagudi, Akash Bharadwaj, Y. V. Srinivasa Murthy, Nishaanth Reddy, and Priya Rao. 2017. Dravidian language classification from speech signal using spectral and prosodic features. *International Journal of Speech Technology*, 20(4):1005–1016.

Bo Li, Tara N. Sainath, Khe Chai Sim, Michiel Bacchiani, Eugene Weinstein, Patrick Nguyen, Zhifeng Chen, Yanghui Wu, and Kanishka Rao. 2018. Multi-dialect speech recognition with a single sequence-to-sequence model. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4749–4753.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre□training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Shweta Sinha, Aruna Jain, and S. S. Agrawal. 2015. Fusion of multi-stream speech features for dialect classification. *CSI Transactions on ICT*, 2(4):243–252.