DLRG@DravidianLangTech 2026: Dual-Purpose Whisper Adaptation for Tamil Dialect Identification and Dialectal Speech Recognition

Gulisetty Abhinav, Tanisha Nanda, Ramesh Kannan R and Ratnavel Rajalakshmi
School of Computer Science and Engineering,
Vellore Institute of Technology(VIT), Chennai, India
Corresponding Author: rajalakshmi.r@vit.ac.in

Abstract

This paper describes our system for the shared task on Dialect Based Speech Recognition and Classification in Tamil at DravidianLangTech@ACL 2026. We participate in both Subtask 1 (Dialect Identification) and Subtask 2 (Dialectal ASR). Our approach leverages a single Tamil-adapted Whisper Medium model as a unified foundation for both tasks. For dialect classification, we repurpose the Whisper encoder as a feature extractor via mean pooling and full encoder fine-tuning with a lightweight classification head, achieving 73.4% accuracy. For dialectal ASR, we apply Low-Rank Adaptation (LoRA) to the full encoder-decoder architecture with SpecAugment-based data augmentation, achieving a Word Error Rate (WER) of 0.55. Our experiments reveal that unfreezing the pre-trained encoder is critical for dialect discrimination, boosting accuracy from 52.78% to 73.4%. The code is publicly available.[1]

1 Introduction

Tamil, one of the classical Dravidian languages, is spoken by over 80 million people across diverse geographical regions exhibiting rich sociolinguistic and dialectal variations spanning four major dialect groups: Northern, Southern, Western, and Central (Chakravarthi et al., 2026). These variations pose substantial challenges for speech technologies, particularly Automatic Speech Recognition (ASR) and dialect identification, as models trained on standard Tamil often fail to capture the subtle phonetic, prosodic, and lexical differences that distinguish regional varieties.

The shared task on Dialect Based Speech Recognition and Classification in Tamil at DravidianLangTech@ACL 2026 (Chakravarthi et al., 2026; Bharathi et al., 2026) addresses these challenges by providing a curated dialectal speech corpus comprising 9.22 hours of training data and 2.05 hours of test data, with both spontaneous and read speech from native speakers across the four dialect groups. The task is organized into two subtasks: (1) speech-based dialect classification and (2) automatic speech recognition for dialectal Tamil.

Previous editions of the DravidianLangTech workshop have advanced research on various NLP tasks for Dravidian languages, including sentiment analysis, hate speech detection (B et al., 2024; Rajalakshmi et al., 2025), and abusive language detection (Rajiakodi et al., 2025; Rajalakshmi et al., 2024, 2022, 2021). However, dialect-level speech processing for Tamil remains relatively unexplored (Bharathi et al., 2025), motivating this shared task.

Our key contribution is the dual-purpose adaptation of a single foundation model [2], for both generative (ASR) and discriminative (classification) tasks. For Subtask 1, we repurpose the Whisper encoder as a dialect-aware feature extractor via full fine-tuning and mean pooling. For Subtask 2, we apply parameter-efficient fine-tuning via LoRA (Hu et al., 2022) to the full sequence-to-sequence architecture.

2 Related Work

The Transformer architecture (Vaswani et al., 2017) has driven recent advances in speech processing. OpenAI's Whisper (Radford et al., 2023) is a large-scale speech model trained on 680,000 hours of multilingual data, demonstrating strong zero-shot performance. Self-supervised models such as wav2vec 2.0 (Baevski et al., 2020) have also shown effectiveness for low-resource speech tasks. However, dialect-level processing remains challenging due to high acoustic similarity between dialects and scarcity of labeled dialectal data.

Low-Rank Adaptation (LoRA) (Hu et al., 2022) injects trainable rank-decomposition matrices into

[1] https://github.com/abhinavgulisetty/tamil-dialect-asr

[2] vasista22/whisper-tamil-medium

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 232–236

frozen pre-trained weights, substantially reducing trainable parameters. This approach is particularly beneficial for low-resource scenarios where full fine-tuning risks overfitting (Dettmers et al., 2023).

The DravidianLangTech workshop series has been instrumental in advancing computational research for Dravidian languages. The DLRG team has a sustained participation history, including offensive language identification on code-mixed Tamil (Rajalakshmi et al., 2021), abusive comment detection (Rajalakshmi et al., 2022), and multimodal tasks (Rajalakshmi et al., 2024, 2025). On the speech side, Bharathi et al. (2025) present a multi-dialect speech corpus for Tamil. The current shared task (Chakravarthi et al., 2026; Bharathi et al., 2026) is the first to systematically benchmark both dialectal speech recognition and dialect identification for Tamil.

3 Task and Dataset Description

The Tamil Dialect Speech Dataset captures linguistic differences across major dialect regions of Tamil Nadu (Bharathi et al., 2025). It comprises both spontaneous and read speech produced by native speakers from four dialect groups: Northern, Southern, Western, and Central. Speakers include both male and female participants across various age groups, with all recordings captured at 16 kHz in natural acoustic environments. The training set consists of 9.22 hours and the test set of 2.05 hours of manually transcribed speech.

Subtask 1 requires classifying each audio sample into one of the four dialect categories. **Subtask 2** requires transcribing dialectal Tamil speech into text.

4 System Description

Our architecture employs Whisper Tamil medium, a pre-trained model for automatic speech recognition (ASR) and speech translation (Vasista, 2023), a variant of OpenAI's Whisper Medium (Radford et al., 2023) fine-tuned for Tamil, as the shared foundation for both subtasks. Figure 1 illustrates the overall system design. The full pipeline was implemented using HuggingFace Transformers (Wolf et al., 2020).

4.1 Model 1: Dialect Classification

Architecture. We discard the Whisper decoder and utilize only the encoder. The encoder's `last_hidden_state`—a tensor of 1024-dimensional feature vectors per audio frame—is aggregated via **mean pooling** across the time dimension into a single fixed-length embedding. This embedding is fed into a classification head consisting of a Dropout layer ($p = 0.1$) followed by a Linear layer projecting to four target classes.

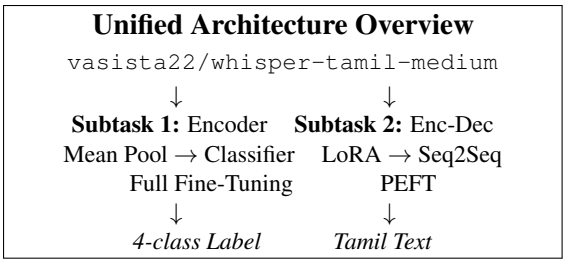

Figure 1: Overview of the dual-purpose architecture. Both subtasks share the same pre-trained Whisper Tamil model but use different adaptation strategies.

Data processing. The standalone Whisper encoder expects exactly 30 seconds of audio input. We configure the processor with padding set to `max_length`, ensuring every audio file is zero-padded to exactly 3000 Mel frames.

Full encoder fine-tuning. Our initial approach of freezing the encoder yielded only 52.78% accuracy. Because the four Tamil dialects share substantial linguistic overlap, the pre-trained standard Tamil features were insufficient. We therefore conducted **full fine-tuning** of the encoder, allowing self-attention to shift toward prosodic and tonal variations unique to regional dialects, with a constrained learning rate of 1×10^{-5} to prevent catastrophic forgetting. Training used a batch size of 8 for 5 epochs with Cross-Entropy loss.

4.2 Model 2: Dialectal ASR

LoRA adaptation. The ASR system uses the full Whisper encoder-decoder, processing 80-channel log-Mel spectrograms to auto-regressively generate text. We apply LoRA (Hu et al., 2022) to the attention and feed-forward modules (`q_proj`, `k_proj`, `v_proj`, `out_proj`, `fc1`, `fc2`) with rank $r = 64$, $\alpha = 128$, and dropout 0.05, yielding ∼69.2M trainable parameters (∼8.3% of total).

Data augmentation. We implement **SpecAugment** (Park et al., 2019) within the data collator, randomly masking frequency bands (up to 15 bins) and time steps (up to 35 frames). Ground-truth transcripts are stripped of punctuation and normalized.

Strategy	Val Acc.	Val Loss	Test Acc.
Frozen Encoder	52.78%	—	—
Unfrozen (Epoch 5)	98.93%	0.0310	**73.4%**

Table 1: Dialect classification results comparing frozen vs. unfrozen encoder strategies.

Custom data collator. We engineer a custom collator that preserves labels and manually constructs `decoder_input_ids` by replacing -100 masking tokens with pad tokens and prepending the forced prompt sequence (BOS, Language, Task tokens).

Training. Training uses FP16 mixed precision with batch size 16, gradient accumulation of 2 (effective batch size 32), `adamw_torch_fused` optimizer (Loshchilov and Hutter, 2019), and cosine learning rate scheduling with peak rate $5{\times}10^{-4}$. During inference, greedy decoding generates up to 225 tokens with `language="tamil"` and `task="transcribe"`.

5 Results and Analysis

5.1 Subtask 1: Dialect Classification

Table 1 shows the comparison between frozen and unfrozen encoder approaches.

With the frozen encoder, the classifier plateaued at 52.78%, only modestly above the 25% random baseline, indicating that standard Tamil acoustic embeddings are insufficient for dialect discrimination. After full fine-tuning, validation accuracy reached 98.93% and test accuracy reached **73.4%**.

The gap between validation and test accuracy illustrates the generalization challenge of fine-tuning high-capacity models on small datasets. The encoder can memorize acoustic environments and speaker-specific characteristics in training. Nevertheless, 73.4% on unseen speakers demonstrates that the adapted embeddings capture genuine regional phonological features.

5.2 Subtask 2: Dialectal ASR

The ASR model achieved a WER of **0.55** on the test set (Table 2). While standard ASR benchmarks target lower WER, this score is characteristic of dialectal transcription where dialects lack standardized orthography. The WER largely reflects an **orthographic mismatch** between phonetically accurate outputs and formal spelling conventions in the ground truth, rather than a failure in acoustic modeling.

Metric	Value
Word Error Rate (WER)	0.55

Table 2: Dialectal ASR performance on the test set.

6 Discussion

Unified foundation architecture. The dual-purposing of a single foundation model demonstrates how a single pre-trained acoustic space can be efficiently manipulated for both generative and discriminative tasks in a low-resource setting, without building separate pipelines.

The frozen vs. unfrozen gap. Our experiments expose a critical limitation: standard Tamil acoustic embeddings are insufficient for dialect discrimination. Unfreezing enabled the self-attention layers to restructure toward dialect-specific acoustic cues, yielding a performance leap from 52.78% to 73.4% (test).

Orthographic penalty in WER. Standard WER is a suboptimal metric for dialectal ASR when dialects lack standardized orthography. Many "errors" represent valid phonetic transcriptions that differ from the formal ground-truth spelling.

7 Conclusion

We presented the DLRG team's system for dialect-based speech recognition and classification in Tamil at DravidianLangTech@ACL 2026. Our approach demonstrates that a single pre-trained Whisper model can be effectively adapted for both dialect identification (73.4% accuracy) and dialectal ASR (0.55 WER) through different fine-tuning strategies. Key findings include: (1) full encoder fine-tuning is essential for dialect classification, boosting accuracy from 52.78% to 73.4%; (2) LoRA provides an effective parameter-efficient adaptation for dialectal ASR; and (3) standard WER is limited as a metric for dialectal transcription due to orthographic variation. Future work will explore ensemble approaches combining multiple self-supervised speech models and dialect-specific decoding strategies.

Limitations

Our work has several limitations. First, the training corpus is relatively small (9.22 hours), which constrains the model's ability to generalize across speakers and acoustic conditions. The substantial

gap between validation accuracy (98.93%) and test accuracy (73.4%) for dialect classification suggests overfitting to speaker-specific and environment-specific cues in the training data.

Second, we evaluate only a single base model (`whisper-tamil-medium`). Comparing against other self-supervised speech models such as wav2vec 2.0 or HuBERT could reveal whether our findings on encoder unfreezing generalize across architectures.

Third, the WER metric used for Subtask 2 does not distinguish between genuine transcription errors and orthographic mismatches arising from dialectal variation. A phoneme-level evaluation or a dialect-normalized WER would provide a more accurate assessment of ASR quality.

Finally, our system does not incorporate any explicit dialect-conditioning mechanism for the ASR subtask. Jointly leveraging dialect identification outputs to guide ASR decoding could improve transcription quality.

Ethical Considerations

Our work involves processing speech data from speakers of regional Tamil dialects. We used only the dataset provided by the shared task organizers, which was collected with appropriate consent from participants. We did not collect any additional personal data.

Speech technologies that favor standard dialects over regional varieties risk marginalizing speakers of underrepresented dialects. Our work aims to mitigate this by explicitly developing models for dialectal Tamil. However, we acknowledge that classification errors could lead to misrepresentation of a speaker's dialect identity, and ASR errors on dialectal speech could reduce the utility of downstream applications for dialect speakers.

All experiments were conducted using publicly available pre-trained models and open-source libraries. The code and model configurations are publicly released to support reproducibility.

Acknowledgments

We thank the organizers of the DravidianLangTech@ACL 2026 shared task for providing the Tamil Dialect Speech Dataset and evaluation framework.

References

Premjith B, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Prashanth Karnati, Sai Rishith Reddy Mangamuru, and Chandu Janakiram. 2024. Findings of the shared task on hate and offensive language detection in Telugu codemixed text (HOLD-Telugu)@DravidianLangTech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55, St. Julian's, Malta. Association for Computational Linguistics.

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, Virtual. Curran Associates Inc.

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*, Vienna, Austria. 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.

Bharathi Raja Chakravarthi and 1 others. 2026. Overview of the shared task on dialect based speech recognition and classification in Tamil: DravidianLangTech@ACL 2026. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Vienna, Austria. Association for Computational Linguistics. To appear.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115, New Orleans, LA, USA. Curran Associates Inc.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shanan Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, pages 2613–2617, Graz, Austria. ISCA.

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*, pages 28492–28518, Honolulu, Hawaii, USA. PMLR.

Ratnavel Rajalakshmi, Ankita Duraphe, and Antonette Shibani. 2022. DLRG@DravidianLangTech-ACL2022: Abusive comment detection in Tamil using multilingual transformer models. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 207–213, Dublin, Ireland. Association for Computational Linguistics.

Ratnavel Rajalakshmi, Ramesh Kannan, Meetesh Saini, and Bitan Mallik. 2025. DLRG@DravidianLangTech 2025: Multimodal hate speech detection in Dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 376–380, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Ratnavel Rajalakshmi, Saptharishee M, Hareesh S, Gabriel R, and Varsini Sr. 2024. DLRG-DravidianLangTech@EACL2024 : Combating hate speech in Telugu code-mixed text on social media. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 140–145, St. Julian's, Malta. Association for Computational Linguistics.

Ratnavel Rajalakshmi, Yashwant Reddy, and Lokesh Kumar. 2021. DLRG@DravidianLangTech-EACL2021: Transformer based approach for offensive language identification on code-mixed Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 357–362, Kyiv. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan P, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the shared task on abusive Tamil and Malayalam text targeting women on social media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 671–681, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Sai Vasista. 2023. Fine-tuned whisper model for Tamil automatic speech recognition. `https://huggingface.co/vasista22/whisper-tamil-medium`. Hugging Face Model Hub.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, Long Beach, CA, USA. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.