CUET_InferX@DravidianLangTech 2026: Shared Task on Dialect Based Speech Recognition and Classification in Tamil

Md. Ashraful Islam Semon, Jihadul Islam, Ratnajit Dhar, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2104059,u2104080,u2004008}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Tamil has a lot of internal variability, including the way it is used in casual conversations, code mixing, and phonetic differences in the way it is spoken in different regions, making it quite difficult to transcribe the spoken word and classify the dialects. In order to address these challenges, our paper presents the system developed by the CUET_InferX team for the Shared Task on Dialect Based Speech Recognition and Classification in Tamil, which was part of DravidianLangTech@ACL 2026. For Subtask 2 (ASR), our proposed system is based on a dual-architecture design that incorporates a fine-tuned Whisper-large-v3 model with Low-Rank Adaptation (LoRA) and a Wav2Vec2 XLSR-53 model, topped with a KenLM statistical language model for n-gram phonetic correction. Our ASR system resulted in a Word Error Rate (WER) of 0.54, which earned us 2nd position for Subtask 2. For Subtask 1 (Speech-Based Dialect Classification), our proposed system is based on a text-based weighted ensemble of IndicBERT, MuRIL, XLM-RoBERTa, and TamilBERT models, which is completely dependent on our ASR system's transcription outputs. Our proposed system achieved a Macro F1 score of 0.22, which earned us 9th position for Subtask 1.

1 Introduction

As voice technology grows, Automatic Speech Recognition (ASR) has also gained prominence as a key area of research. Despite recent progress in automatic speech recognition for high-resource languages, Tamil speech recognition is still difficult. This is due to dialect differences, complex structure, code-mixing, and a lack of balanced speech data, especially for spontaneous conversation.

The "Shared Task on Dialect Based Speech Recognition and Classification in Tamil: DravidianLangTech@ACL 2026" (Bharathi et al., 2026) is designed with the aim of creating a multi-dialect speech dataset (Bharathi et al., 2025) that challenges speech recognition models to perform well despite high intra-language variations. The objective is two-fold: speech classification into one of the four dialects such as Northern, Southern, Western, and Central, and accurate speech recognition into dialect-specific Tamil text.

In our case, we employed a dual-architecture acoustic system for ASR and a text-driven ensemble system for dialect classification. For ASR, we employed a fine-tuned Whisper-large-v3 model (Radford et al., 2023) using Low Rank Adaptation (LoRA) (Hu et al., 2021), along with a Wav2Vec2 XLSR-53 model (Conneau et al., 2020a), along with a statistical language model based on KenLM (Heafield, 2011). This system performed well, yielding a Word Error Rate of 0.54. For dialect classification, we employed a weighted ensemble of four text-based models IndicBERT, MuRIL, XLM-RoBERTa, and TamilBERT over the ASR-transcribed text. This system obtained a Macro F1 score of 0.22. The core contributions of our research work are as follows -

- We designed a dual architecture ASR pipeline that combines Whisper (with LoRA) and Wav2Vec2 XLSR-53 to robustly handle heavy dialect variations and code-mixed speech.
- We used a KenLM based n-gram rescoring method to maintain unnormalized dialect-specific colloquial spelling forms.
- A cascaded text-based ensemble was considered for the classification of dialects demonstrating the importance of acoustic features in the classification of dialects, in addition to textual features.

Detailed implementation information is available in the GitHub repository- https://github.com/semon87/shared-task-dialect-cuetinferx

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 201–206

2 Related Work

The domain of Tamil ASR is shifting towards end-to-end deep learning with cross-lingual transfer learning. The DravidianLangTech shared task (Bharathi et al., 2026), built upon the multi-dialect Tamil speech corpus introduced by (Bharathi et al., 2025), provides an important benchmark for evaluating ASR systems under strong intra-language variability and spontaneous speech conditions.

Self-supervised acoustic models have shown strong effectiveness for Tamil ASR. (Akhilesh et al., 2022) fine-tuned XLSR-Wav2Vec2.0 using a CTC objective and reported a WER of 0.58 on Common Voice Tamil. Similarly, (Srinivasan et al., 2022) demonstrated that fine-tuned XLSR models obtain competitive performance in shared task settings. Focusing on Whisper-based architectures, (Saranya et al., 2024) observed substantial performance degradation of pre-trained Whisper models on unconstrained dialectal speech, but reduced the WER to 61% using LoRA-based adaptation. (Acharya et al., 2025) further validated the effectiveness of parameter-efficient fine-tuning of Whisper models in shared task scenarios, reinforcing the importance of adaptation for handling speaker and dialect variability.

Beyond acoustic modeling, the role of external language models in low-resource ASR has gained renewed attention. (Liu et al., 2024) showed that incorporating n-gram language models during decoding consistently reduces word error rates in low-resource languages.

For dialect classification, acoustic representations have been shown to capture critical phonetic and prosodic cues (Saranya et al., 2024). Text-based dialect classification relies on vocabulary and context patterns. The effectiveness of multilingual transformers like XLM-RoBERTa (Conneau et al., 2020b), IndicBERT (Kakwani et al., 2020), MuRIL (Khanuja et al., 2021) and TamilBERT (Joshi, 2022) has been demonstrated for NLP in Indian languages, which have a rich morphology and code-mixing capabilities.

3 Dataset

The data for the shared task (Bharathi et al., 2025) consists of a set of speech recordings in Tamil with their corresponding ASR transcripts and dialect information. Table 1 showing the training and test data, four dialects are present: Southern, Northern, Western, and Central.

Dialect Region	Training	Test
Southern	1427	212
Northern	1696	189
Western	1126	116
Central	885	62
Total	5134	579

Table 1: Distribution of training and test samples across dialect categories.

4 Methodology

4.1 Problem Formulation

The problem formulation for Subtask 2 (ASR) entails speech transcription with respect to mapping dialectical Tamil speech into transcripts considering colloquialisms. On the other hand, the formulation of Subtask 1 for dialect classification is done by means of cascade text classification using transcripts generated from ASR for predicting the four regional dialects, i.e., Northern, Southern, Western, and Central.

4.2 Data Preprocessing

For the acoustic models, we paired speech data with their respective transcripts and dialect regions, and the audio was converted to 16 kHz mono to match the pre-trained ASR model requirements. For mitigating the dialect imbalance, we have also tried basic audio augmentation techniques like time stretching and concatenating short utterances of minority dialect classes. Although the WER slightly decreases, it was excluded from the final pipeline due to increased computational overhead and GPU memory limitations.

For the classification of the dialect, the transcripts were tokenized based on the tokenizer for the respective model. The transcripts were then padded or truncated to a maximum of 128 tokens, and the dialect labels were converted into numerical form for training.

4.3 Experimented Models

4.3.1 Automatic Speech Recognition

To handle the dialectal variations of Tamil language, we implemented a dual-architecture ASR framework. We fine-tuned *whisper-large-v3*(with LoRA) optimized by Unsloth (Daniel Han and team, 2023) library on the provided corpus of Tamil dialects. LoRA adapters were applied to the attention projection layers (`q_proj` and `v_proj`) with rank $r = 64$, $\alpha = 64$, dropout $= 0$, and no bias pa-

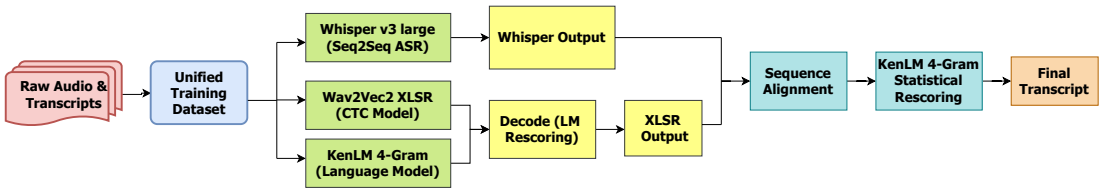

Figure 1: Proposed dual-architecture dialectal ASR pipeline

rameters, enabling computationally efficient training under limited GPU memory constraints. Whisper decoding was performed using beam search to obtain high quality transcripts. We fine-tuned Facebook's cross-lingual *wav2vec2-large-xlsr-53* model using Connectionist Temporal Classification (CTC). This model is specifically suited to handle the dialects and the phonetic variations in the speech patterns of the speakers. This model has also been fine-tuned on the provided dataset to suit the dialects and filtered too short (less than 1 sec) or too long (greater than 25 sec) audio. To further improve the transcription results, we trained a 4-gram statistical language model using KenLM on the training transcripts and used it to decode the transcripts. The KenLM decoding parameters were selected through empirical grid search on a held-out validation subset by evaluating various language-model weights (α) and word insertion bonuses (β) using WER as the selection criterion. A larger beam width was used during final decoding to improve search coverage, while smaller beams were used during parameter tuning for computational efficiency. Figure 1 shows the system architecture of ASR task.

4.3.2 Dialect Classification

In the case of dialect classification, we used a text-based approach by utilizing ASR transcripts. We have employed four different transformer-based architectures IndicBERT, MuRIL, XLM-RoBERTa, TamilBERT and trained with ASR transcripts of training data. After that, we used generated ASR transcript from Subtask 2 to classify the dialect region. Figure 2 shows the system architecture of dialect classification task.

4.4 Ensembling and Post-Processing

In ASR task, Whisper output and rescored XLSR model were combined with an edit-distance-style sequence alignment based on the Python SequenceMatcher. Each hypothesis was tokenized into Tamil words based on whitespace, without further morphological segmentation. Simply matching word

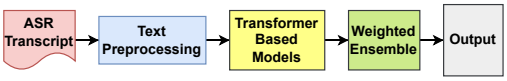

Figure 2: Proposed text-based dialect classification pipeline

Component	Key Settings
Whisper (Seq2Seq)	5 epochs, LR = 1e-4, Num of Beam = 5
XLSR (CTC)	30 epochs, Beam width = 500, α (LM weight) = 0.4, β (Word bonus) = 1.5
Language Model	KenLM 4-gram

Table 2: Hyperparameters for ASR and decoding

segments were directly copied to the combined final alignment; and conflicts due to substituted, inserted, or deleted segments were resolved with 4-gram KenLM rescoring. For each conflicting segment, normalized language model scores were computed and the segment with the higher contextual probability was used in the final alignment. A small bias toward XLSR outputs was additionally applied during conflict resolution, as XLSR showed better validation performance on dialect-specific phonetic variants.

For the classification of dialects, a four-model transformer ensemble is employed. In this approach, each model is used for the analysis of the transcript obtained from the ASR system. The contribution of each model is weighted according to its validation accuracy, and the final prediction is obtained via weighted soft voting.

4.5 Parameter Setting

Table 2 shows the best performing parameter settings for the ASR task. For dialect classification, all transformer models were trained for 15 epochs. In the weighted ensemble, XLM-RoBERTa got the highest weight of 1.4, L3Tamil-BERT followed with 1.1, IndicBERT and MuRIL both received weights of 0.9 based on their validation performance.

5 Results and Analysis

For dialect classification, the macro-averaged F1 score was used to evaluate the system's performance across the four dialect classes, while ASR performance was evaluated using the Word Error Rate (WER).

5.1 Comparative Analysis

Model	WER	CER
XLSR	0.6127	0.1779
Whisper	0.5510	0.1593
XLSR + KenLM	0.5491	0.1694
Whisper + Audio Augmentation	0.5434	0.1577
Whisper + XLSR + KenLM (Ensemble)	**0.5410**	0.1690

Table 3: Performance of different ASR systems on the test dataset

For ASR task, Table 3 shows that Whisper-large-v3 achieved a WER of 0.5510, while KenLM rescoring improved the XLSR model to 0.5491. Audio augmentation further improved Whisper performance to 0.5434. The best result came from using the Whisper + XLSR + KenLM ensemble, which achieved a WER of 0.5410.

For dialect classification task, Table 4 results suggest that the models tested on the Whisper ASR source outperform the ones tested on the ensemble ASR source. The best-performing model was the weighted ensemble model (XLM-RoBERTa, IndicBERT, MuRIL, TamilBERT) on the Whisper ASR source, with a macro F1 of 0.2209 and Figure 3 shows the confusion matrix.

5.2 Error Analysis

During sequence alignment, long compound words were often split inconsistently between the Whisper and XLSR models. This inconsistency led to localized insertion and deletion penalties in the final ensemble. In addition, the trade-off between preserving dialects within the transcription and improving overall transcription accuracy through KenLM rescoring was noted. Although KenLM rescoring reduced the overall Word Error Rate (WER) by focusing on contextually likely n-grams, it caused a normalization effect by often replacing rare dialect markers with standard vocabulary. As a result, regional linguistic features in the combined text were weakened. In contrast, the standalone Whisper outputs maintained these casual patterns more effectively. This explains why Whisper performed better in the downstream dialect classification task,

ASR Source	Model	F1
Whisper+XLSR +KenLM	IndicBERT	0.1635
	XLM-RoBERTa	0.0698
	Ensemble	0.1633
	Whisper + XLM-RoBERTa (Fusion)*	0.3839
Whisper	IndicBERT	0.1698
	XLM-RoBERTa	0.1153
	Ensemble	**0.2209**
	Whisper + XLM-RoBERTa (Fusion)*	0.4044

*Post-competition result.

Table 4: Performance of different dialect classification systems on the test dataset

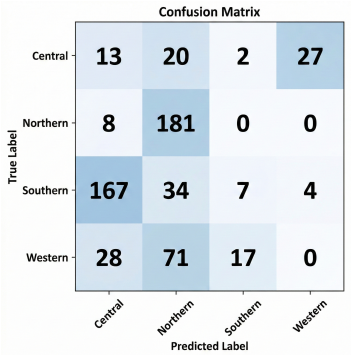

Figure 3: Confusion matrix of weighted ensemble process

even though it had a higher baseline ASR error rate.

Our primary submission focused on text-based dialect classification. This approach limited the model's ability to capture pronunciation and prosodic cues. In a post-competition experiment, a multimodal early-fusion model combined Whisper acoustic embeddings with XLM-RoBERTa text embeddings which achieved a macro F1 score of 0.40. This result shows how crucial acoustic features are for Tamil dialect classification.

6 Conclusion

In this work, we presented systems for dialect-based Tamil ASR and classification for DravidianLangTech@ACL 2026. Our Whisper XLSR KenLM ensemble achieved the best ASR performance with a WER of 0.5410. The weighted text-based dialect classification ensemble reached a macro F1 score of 0.2209. Experimental results show that transcript-only methods are not enough for strong dialect classification. Future work will focus on combining acoustic embeddings and increasing strength against ASR-induced errors.

Limitations

The ASR system can be improved by utilizing a larger and more varied speech corpus of multiple dialects and by employing sophisticated data augmentation techniques to improve its resilience against environmental noises and speech variations. For the dialect classification process, relying on the ASR transcript resulted in propagating errors and losing critical acoustic information such as intonation, prosody, and speech rates that are critical for dialect differentiation. Future studies should investigate multimodal models that combine acoustic and textual information to better capture dialect variations.

Ethical Considerations

In this study, we have developed our methodology following the highest ethical practices. By advancing dialect-aware speech recognition and classification in the Tamil language, we aim to develop speech technology that is more diverse and representative of the various dialects spoken in the regions. While there are chances of dialect identification tools being misused to profile or reinforce regional prejudices, we are committed to developing this technology for research and development purposes only. We are committed to developing artificial intelligence technology that is fair, transparent, and respectful of cultures, and to the ethical use of dialect-sensitive speech technology.

Acknowledgments

We extend our heartfelt thanks to the organizers of the Shared Task on Dialect Based Speech Recognition and Classification in Tamil at DravidianLangTech@ACL 2026 (Bharathi et al., 2026) for providing the multi-dialect corpus (Bharathi et al., 2025) and evaluation system. We also thank Kaggle for providing us with the computational resources to train our models.

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, Naples, Italy. Unior Press.

A Akhilesh, Brinda P, Keerthana S, Deepa Gupta, and Susmitha Vekkot. 2022. Tamil speech recognition using xlsr wav2vec2.0 & ctc algorithm. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

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. 2020a. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Simran Khanuja and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Zoey Liu, Nitin Venkateswaran, Eric Le Ferrand, and Emily Prud'hommeaux. 2024. How important is a language model for low-resource asr? In *Findings of*

the Association for Computational Linguistics: ACL 2024, pages 206–213, Bangkok, Thailand. Association for Computational Linguistics.

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 28448–28481. PMLR.

S. Saranya, B. Bharathi, S. Gomathy Dhanya, and Aishwarya Krishnakumar. 2024. Real-time continuous tamil dialect speech recognition and summarization. *Circuits, Systems, and Signal Processing*, 44(4):2855–2881.

Dhanya Srinivasan, B. Bharathi, Thenmozhi Durairaj, and B. Senthil Kumar. 2022. Ssncse_nlp@lt-edi-acl2022: Speech recognition for vulnerable individuals in tamil using pre-trained xlsr models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 317–320, Dublin, Ireland. Association for Computational Linguistics.