CHMOD_777@DravidianLangTech 2026: Tamil-Adapted Whisper and MMS for Dialect Speech Recognition and Classification

Arunaggiri Pandian Karunanidhi[1], **Prabalakshmi Arumugam**[2]
[1]Micron Technology, United States
[2]Boise State University, United States
arunaggiri.karunanidhi@gmail.com, prabalakshmiarumugam410@gmail.com

Abstract

This paper describes Team CHMOD_777's system for the DravidianLangTech@ACL 2026 shared task on Tamil dialect speech recognition and classification. The task comprises two subtasks: classifying Tamil speech into four regional dialects (Northern, Southern, Western, Central) and transcribing dialectal Tamil speech to text. For dialect classification, we fine-tune MMS-1b-all with Focal Loss and weighted sampling, achieving 83.04% Macro F1 on the development set (5th out of 11 teams on the test set). For speech recognition, we fine-tune a Tamil-specific Whisper model (763M parameters), achieving 53.72% WER on the development set and 49.75% on the official test set, ranking 1st out of 13 teams. Our key finding is that domain-specific pre-training significantly outperforms larger general-purpose models: Tamil Whisper (763M) beats Whisper-large-v3 (1.5B) by 8 WER points despite having half the parameters.

1 Introduction

Tamil, spoken by over 80 million people, exhibits significant regional variation across four major dialect groups: Northern, Southern, Western, and Central. These dialects differ in phonology, vocabulary, and prosody, posing challenges for both dialect identification and automatic speech recognition (ASR). The DravidianLangTech@ACL 2026 shared task (Bharathi et al., 2026) addresses these challenges through two subtasks using the multi-dialect Tamil speech corpus (Bharathi et al., 2025).

Subtask 1 requires classifying Tamil speech into four dialect categories, while Subtask 2 requires transcribing dialectal Tamil speech to text, evaluated by Word Error Rate (WER). Our approach evaluates multiple pre-trained speech models across both subtasks. For classification, we compare MMS-1b-all (Pratap et al., 2023) (1B parameters, 1,100+ languages) against wav2vec2-large-xlsr-53 (Conneau et al., 2021) (300M, 53 languages). For ASR, we compare a Tamil-specific Whisper variant (763M) against the general-purpose Whisper-large-v3 (Radford et al., 2023) (1.5B). Our experiments demonstrate that domain-specific pre-training consistently outperforms model scale across both tasks.[1]

2 Related Work

Speech processing for Indian languages has gained momentum through shared tasks and community efforts. Bharathi et al. (2022) organized the first shared task on speech recognition for vulnerable individuals in Tamil, establishing benchmarks for Tamil ASR. The multi-dialect Tamil speech corpus (Bharathi et al., 2025) used in this shared task provides a foundation for studying dialectal variation in Tamil speech.

Self-supervised speech models have transformed the field. Wav2Vec2 (Baevski et al., 2020) learns speech representations from unlabeled audio through contrastive learning. Its multilingual extension, XLSR-53 (Conneau et al., 2021), covers 53 languages. MMS (Pratap et al., 2023) scales further to 1,100+ languages with 1B parameters. For ASR, Whisper (Radford et al., 2023) achieves strong multilingual performance through weakly supervised pre-training on 680,000 hours of audio. Community-driven efforts have produced Tamil-specific Whisper variants fine-tuned on Tamil speech corpora, which we leverage in this work.

Focal Loss (Lin et al., 2017) has been successfully applied to speech classification tasks with class imbalance. Prior work on Dravidian language processing (Jada et al., 2021; Sai Kumar et al., 2021; Premjith et al., 2022) has shown

[1]The code for this work is available at: https://github.com/Arunaggiri-Pandian/Dialect_Based_Speech_Recognition_and_Classification_in_Tamil

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 186–190

Dialect	Train	Dev	Test
Northern	1,525	171	189
Southern	1,186	241	212
Western	944	182	116
Central	716	168	62
Total	**4,371**	**762**	**579**

Table 1: Dataset distribution across four Tamil dialects.

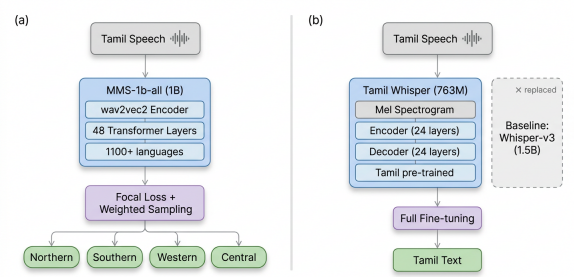

Figure 1: System architecture. (a) Dialect classification via MMS-1b-all. (b) Speech recognition via Tamil Whisper, which outperforms the 2x larger Whisper-large-v3 by 8 WER points.

that language-specific approaches consistently outperform general multilingual methods across text, speech, and multimodal tasks.

3 Methodology

3.1 Dataset

The dataset (Bharathi et al., 2025) contains Tamil speech samples across four regional dialects. Table 1 shows the distribution. Audio files are mono-channel WAV at 16 kHz sampling rate, with Tamil script transcriptions.

3.2 Subtask 1: Dialect Classification

We build a classification pipeline on top of self-supervised speech encoders. Given an input waveform, the encoder produces frame-level representations that are mean-pooled over time to obtain a fixed-dimensional utterance embedding $\mathbf{h} \in R^d$. This embedding is passed through a two-layer classification head:

$$\mathbf{y} = W_2 \cdot \text{ReLU}(W_1 \cdot \text{Dropout}(\mathbf{h}) + \mathbf{b}_1) + \mathbf{b}_2 \quad (1)$$

where $W_1 \in R^{d \times d/2}$ and $W_2 \in R^{d/2 \times 4}$. We evaluate two encoders:

- **MMS-1b-all** (Pratap et al., 2023): 1B parameters (d=1280), pre-trained on 1,100+ languages via self-supervised learning.
- **wav2vec2-large-xlsr-53** (Conneau et al., 2021): 300M parameters (d=1024), pre-trained on 53 languages.

Both encoders are fine-tuned end-to-end with Focal Loss (γ=2.0), WeightedRandomSampler for class balance, dropout (p=0.2), and early stopping on dev Macro F1 with patience of 5 epochs. We set γ=2.0 following Lin et al. (2017)'s recommended default, which down-weights well-classified dialects while focusing learning on harder dialect boundaries without overly aggressive focusing that higher gamma values can cause on small datasets. Figure 1 illustrates our system architecture for both subtasks.

Parameter	Classification	ASR
Learning rate	5e-6	3e-5
Batch size	16	16
Grad. accumulation	1	4
Loss	Focal (γ=2)	Cross-entropy
Precision	FP32	BF16
Early stop patience	5	10
Seed	42	42

Table 2: Training configuration for both subtasks.

3.3 Subtask 2: Speech Recognition

For ASR, we fine-tune Whisper encoder-decoder models using sequence-to-sequence training. The encoder processes mel-spectrogram features, and the decoder autoregressively generates Tamil script tokens. We compare two models:

- **Tamil Whisper medium**: 763M parameters, a community fine-tuned variant of Whisper-medium pre-trained on Tamil speech corpora.
- **Whisper-large-v3** (Radford et al., 2023): 1.5B parameters, the largest general-purpose Whisper model.

We fine-tune with BF16 mixed precision, gradient accumulation (effective batch size 64), AdamW optimizer (lr=3e-5), and 500 warmup steps. Label sequences exceeding 448 tokens are truncated. Table 2 summarizes configurations for both subtasks.

4 Results

4.1 Dialect Classification

Table 3 shows classification results. MMS-1b-all achieves the best dev F1 of 83.04%, outperforming wav2vec2-xlsr-53 by 17 points. Table 4 shows per-dialect F1 scores. Central dialect is easiest to classify (94.2% F1) while Northern is hardest (77.9%). Figure 2 shows the confusion matrix: Central is

Model	Dev MF1	Epoch
MMS-1b-all (tuned)	**83.04**	**7**
MMS-1b-all (initial)	79.87	5
wav2vec2-xlsr-53 (tuned)	66.38	8
wav2vec2-xlsr-53 (baseline)	57.48	10

Table 3: Dialect classification results (dev Macro F1 %).

Dialect	Precision	Recall	F1	Samples
Northern	69.4	88.9	77.9	171
Southern	81.1	76.8	78.9	241
Western	87.3	75.8	81.2	182
Central	97.5	91.1	94.2	168
Macro	**83.8**	**83.1**	**83.0**	**762**

Table 4: Per-dialect classification metrics (%) for MMS-1b-all.

cleanly separated while Northern is most often confused with Southern, reflecting their geographic and phonological proximity.

Figure 3 shows the training dynamics: MMS-1b-all reaches peak dev F1 of 83.04% at epoch 7 while train F1 continues rising past 95%, indicating rapid adaptation with some overfitting.

On the official test set, our system ranks 5th out of 11 teams with 43.09% Macro F1. Analysis of the test labels reveals a distribution shift across dialects: Southern (80.5%) and Northern (57.8%) remain stable, while Central and Western show sensitivity to the changed test distribution, indicating opportunities for more robust dialect-invariant features. Full results are in the task overview paper (Bharathi et al., 2026).

4.2 Speech Recognition

Table 5 shows ASR results. Tamil Whisper achieves 53.72% WER, outperforming Whisper-large-v3 by 8 points despite having half the parameters. Figure 4 visualizes this comparison.

Table 6 shows per-dialect WER on the dev set. A 40-point gap separates Central (28%) from Northern (69%).

On the official test set, our system achieves 49.75% WER, ranking 1st out of 13 teams. Table 7 shows per-dialect test WER. Central WER increases from 28% (dev) to 47% (test) while Northern improves from 69% to 52%, compressing all dialects into a 6-point range (Figure 5).

5 Analysis

Domain-specific pre-training is the decisive factor: Tamil Whisper (763M) outperforms Whisper-large-v3 (1.5B) by 8 WER points, and MMS-1b-all outperforms wav2vec2-xlsr-53 by 17 F1 points. Full fine-tuning outperforms LoRA by 9 WER points, and SpecAugment hurts (70.07% vs 61.65%) at this data scale. The classification dev-to-test gap (83% to 43%) reflects speaker distribution shift between splits. Central Tamil achieves the lowest WER (28% dev) due to phonological proximity to standard Tamil, while the narrower test gap (6 vs 40 points) suggests more extreme dialectal variation in the dev split.

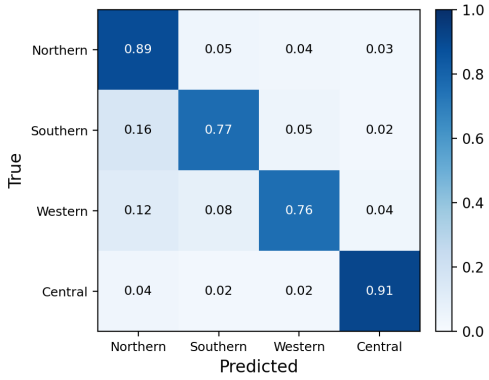

Figure 2: Normalized confusion matrix for MMS-1b-all. Central is cleanly separated; Northern is most confused with Southern.

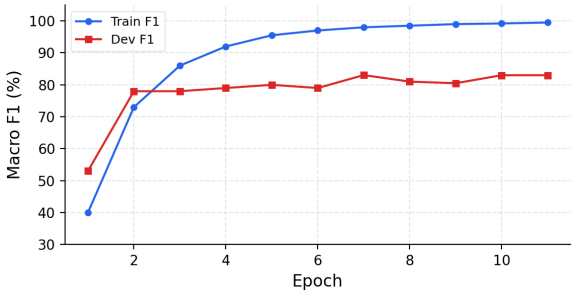

Figure 3: Dialect classification training curves for MMS-1b-all. Dev F1 peaks at epoch 7 while train F1 continues rising.

6 Conclusion

We presented a system for Tamil dialect speech processing achieving Rank 1 in ASR (49.75% WER, 13 teams). Key findings: (1) domain-specific pre-training outperforms scale (Tamil Whisper 763M beats Whisper-v3 1.5B by 8 WER points); (2) full fine-tuning outperforms LoRA by 9 points; (3) SpecAugment hurts on small datasets; (4) per-dialect WER compresses from 40 points (dev) to 6 (test).

Model	Params	WER
Tamil Whisper (full FT)	**763M**	**53.72**
Tamil Whisper (LoRA)	763M	62.93
Tamil Whisper (zero-shot)	763M	86.39
Whisper-large-v3 (best)	1.5B	61.65
Whisper-large-v3 (+SpecAug)	1.5B	70.07

Table 5: ASR results (dev WER %, lower is better).

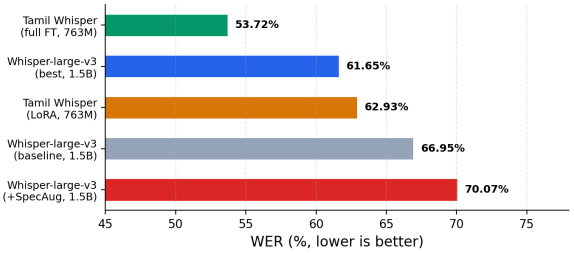

Figure 4: ASR model comparison. Tamil Whisper (763M) outperforms the twice-larger Whisper-large-v3 (1.5B) by 8 WER points.

Limitations

Dialect-adaptive decoding strategies that leverage the classification model to condition ASR output could yield further improvements. Multi-task learning jointly optimizing dialect classification and transcription is another avenue worth investigating. The significant dev-to-test gap in classification (83% to 43%) suggests potential distribution differences that merit closer examination. Incorporating dialect-specific language models may also help improve recognition for underperforming dialects.

Dialect	WER	CER	Samples
Central	28.07	7.31	168
Western	48.55	16.82	182
Southern	52.93	18.09	241
Northern	68.67	32.28	171
Overall	**53.72**	**21.61**	**762**

Table 6: Per-dialect WER and CER (%) on the dev set.

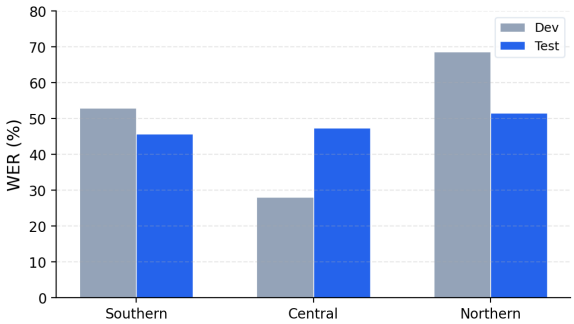

Figure 5: Per-dialect WER: dev (wide spread) vs test (6-point cluster).

Dialect	WER	CER	Samples
Southern	45.61	14.37	212
Central	47.36	13.65	62
Northern	51.50	16.42	189
Overall	**49.75**	**15.59**	**463**

Table 7: Per-dialect test WER and CER (%). Western excluded.

References

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 (NeurIPS)*.

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, Bharathi Raja Chakravarthi, Subalalitha Cn, N. Sripriya, Arunaggiri Pandian Karunanidhi, and Swetha Valli. 2022. Findings of the shared task on speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. 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. 2021. Unsupervised cross-lingual representation learning for speech processing. In *Proceedings of Interspeech*.

Pawan Kalyan Jada, D. Sashidhar Reddy, Konthala Yasaswini, Arunaggiri Pandian Karunanidhi, Prabakaran Chandran, Anbukkarasi Sampath, and Sathiyaraj Thangasamy. 2021. Transformer based sentiment analysis in Dravidian languages. In *Proceedings of FIRE 2021 (Working Notes)*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tober, Changhan Babu, Sayahna Kunber, Ali Elkahky, Zhaoheng Xue, Arya Fazel-Zarandi, Alexei Baevski, Yossi Adi, Wei-Ning Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech

technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

B. Premjith, Bharathi Raja Chakravarthi, Malliga Subramanian, B. Bharathi, K.P. Soman, V. Dhanalakshmi, K. Sreelakshmi, Arunaggiri Pandian Karunanidhi, and Prasanna Kumaresan. 2022. Findings of the shared task on multimodal sentiment analysis and troll meme classification in Dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. 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. *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

T.S. Sai Kumar, K. Arunaggiri Pandian Karunanidhi, S. Thabasum Aara, and K. Nagendra Pandian. 2021. A reliable technique for sentiment analysis on tweets via machine learning and BERT. In *2021 Asian Conference on Innovation in Technology (ASIANCON)*. IEEE.