Azrael@DravidianLangTech 2026:Dialect-Sensitive Automatic Speech Recognition and Classification for Tamil

Janish Andrin J[1], **Mohammed Sahil S**[2] **and Saranya S**[3]
[1,2,3] St. Joseph's Institute of Technology, Tamil Nadu, India
janishandrin75@gmail.com, mohammedsahil4886@gmail.com,
ssaranrahul@gmail.com

Abstract

Tamil is a pre-historic language of millions of individuals who live in India, Sri Lanka, and other parts of the world. Consider the variations in accents, vocabulary and even speech rhythm even among the central region, the northern region, the southern region and the western region of Tamil Nadu. Such idiosyncrasies make it difficult to use features such as voice assistants or translation applications to keep up. A feasible system has been developed in this project to manage that challenge. It picks up raw audio files in Tamil, identifies which of the four predominant dialects the speech belongs to and translates that speech into text. Good quality datasets on Tamil dialects are rather rare, due to the lack of resources and interest in languages. There were pre-trained models, namely, XLSR to spot the dialects and Wav2Vec 2.0 to convert speech into text. All in all, this configuration had an accuracy rate of 46 percentage. It was very good at distinguishing between northern and southern, but was somewhat confused between central and west-central-western. In the case of the transcription component, a cursory inspection reveals that it is a reliable process, able to nail down clear speech despite those accent twists. With that said, it is possible to improve it with such details as a more detailed fine-tuning or equalizing the classes of data.

Keywords: Dialect Identification, Low-Resource Languages, Transfer Learning, XLSR, Multilingual Speech Models

1 Introduction

This is such a rich blend of dialects that, in fact, defines the way people speak and connect with each other culturally on a daily basis, with the big variety occurring in between the central, the north, the southern, and the western parts - consider distinct sounds, word usage, and speech patterns, which can easily confuse modern speech recognition software. In our project, we are implementing a cool all-in-one news system that captures crude Tamil speech, pins it down to the dialect, and translates it directly to Tamil text, which can be used to enhance technologies in the area such as accent-aware voice bots to serve customers, learning apps tailored to this area, or even a forensic system to analyze oral customs. To top it off, it drives forward multilingual AI by advocating Dravidian languages in the larger technology community and addressing those implicit prejudices against well-researched languages. The real challenge? Very high-quality datasets on Tamil dialects are hard to find, such as those that are available in English or Chinese, due to limited funding, privacy challenges, and lack of interest in niche languages in AI. It indicates how transfer learning excels in low-resource locations, allowing us to fine-tune multilingual models with only a small amount of data to achieve good results and paving the way to other neglected dialects across the world.

2 Related Works

Speech recognition on one hand, low-resource language studies on the other, overlap at the intersection of the identification of Tamil dialects and speech processing. The past decades have seen researchers studying linguistic approaches to dialect variation, acoustic and machine learning approaches to detection of dialect variation and improvement of automatic processing of minor languages like Tamil.

The first linguistic studies such as that of (Zvelebil, 1960) provided a rough understanding of the difference in the Tamil dialect. His work concerned the phonological, lexical and grammatical differences in the Tamil-speaking regions that were recorded in a systematic way. Still on the topic of computational dialect identification, (Nanmalar

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 153–157

et al., 2022) investigated the possible application of acoustic and signal-processing techniques to distinguish literary and colloquial versions of Tamil dialects. They demonstrated in their work that definite aspects of speech can be good in distinguishing dialects, but hand-crafted extraction of features can often be based on a lot of domain knowledge. Similarly, (Archana and Bharathi, 2024; Saranya et al., 2025; Bharathi et al., 2025) studied speech-based dialect identification on the Tamil language and highlighted the importance of the sound feature representation and constraints on the similarity of accents and annotated data.

Low resource language processing has been a common literature topic of interest. The survey of the barriers to low-resource languages was carried out by (Magueresse et al., 2020), and their discussion was limited on the unavailable data, unavailable annotation, and disproportionate language technologies. Their findings outline the causes that underpin the importance of transfer learning and multilingual pretraining as strategies.

Amir Hossein Kargaran and others (2023) introduced the system GlotLID that is applied in the language identification research to identify language under low-resource conditions. They show that (their work) with limited data, multilingual representations can be highly beneficial in terms of increasing the performance. On the same note, (Yi et al., 2020) demonstrated the applicability of Wav2Vec2.0 to speech recognition across the scope of various low-resource languages, which proves that pretrained self-supervised models can be successfully fine-tuned on small datasets.

A number of researchers have also studied transfer learning and multilingual adaptation. (Ghafoor et al., 2021) have also written about the way resource-rich datasets can be converted to low-resource languages using multilingual-processing strategies. (Cruz and Cheng, 2019) experimented with the procedures of low-resource language fine-tuning and emphasized that it is crucial to take special caution with the parameters to avoid overfitting. Similarly, the paper by (Agić et al., 2016) developed multilingual projection methods to practically low-resource languages and demonstrated that with the help of cross-lingual knowledge transfer, performance could be enhanced without a large annotated corpus.

(Maimaiti et al., 2019) developed on low-resource multi-round transfer learning schemes of neural machine translation in machine translation and neural modelling. (Gandhe et al., 2014) examined the neural network language models when they are used in low-resource conditions and proved that they perform better as compared to the traditional statistical systems. The prospective of fine-tuning large language models to particular communication tasks in low-resource languages was recently written by (Bui et al., 2025), signifying the fact that large pretrained models are becoming increasingly influential.

The use of NLP in low-resource settings was reviewed by (Pakray et al., 2025) as well, where the researchers reported that practical considerations should be made during the construction, assessment, and utilization of datasets. In a general doctoral study, (King, 2015) discussed the significance of transfer learning and data augmentation to support practical NLP techniques that they claim should be used on low-resource languages. The development of such datasets as that by (Kuriyozov et al., 2019) also emphasizes the necessity to develop annotated corpora that will facilitate the further sustainable growth of the language technology.

The current literature indicates that the pretrained multilingual models, transfer learning, and appropriate dataset preparation are very useful in dialect detection and speech recognition of low-resource languages, including Tamil.

3 Dataset

The data is based on the recordings of the Tamil speech used in the dialect classification and automatic speech transcription works. The audio files are sorted into a dialect-based directory hierarchy, with each of the subfolders corresponding to a particular Tamil dialect i.e., Southern, Northern, Western and Eastern. In order to effectively handle this hierarchical structure a recursive metadata generation script was created to scan all of the subdirectories, get all the audio file paths, and label them with the appropriate dialects, depending on the folder names.

The table 1 illustrates the breakdown of samples and audio durations for Italian dialects in the Common Voice training data, highlighting the Northern Dialect as the largest contributor and the Eastern as the smallest.

Table 1: Distribution of audio samples across dialects

Dialect	Number of Samples	Duration (HH:MM:SS)
Southern Dialect	1427	02:44:30
Northern Dialect	1696	03:29:15
Western Dialect	1126	01:59:59
Eastern Dialect	885	01:08:18
Total	**5134**	**09:21:02**

4 Proposed Work

This proposed study involves two big parts, Tamil dialect identification and Tamil speech recognition. The system will be configured to take raw Tamil speech recordings and give two rewards, the predicted regional dialect, and the textual transcription.

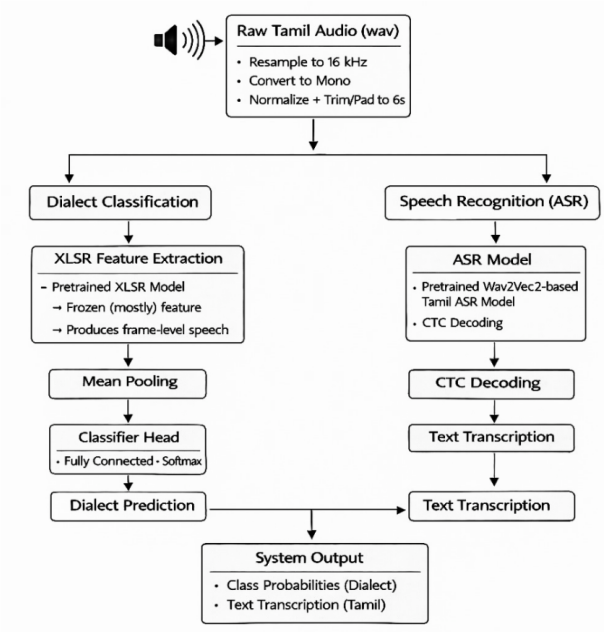

Figure 1: Proposed System Architecture

The figure 1 showing the architecture of the proposed Tamil Dialect Identification and Recognition System work flow. Before being passed into the models, each audio recording undergoes a standardized preprocessing pipeline. All audio files are resampled to 16 kHz to match pretrained model requirements. Recordings are converted to mono format. To maintain uniform input dimensions, each audio clip is trimmed or zero-padded to a fixed duration of six seconds. This ensures stable feature extraction and consistent model behavior. For dialect classification, a pretrained XLSR (Cross-Lingual Speech Representation) model based on the Wav2Vec2 architecture is used through the Hugging Face Transformers library with PyTorch as the backend framework. The XLSR model contains convolutional layers for low-level acoustic encoding followed by Transformer-based self-attention layers that capture long-range temporal dependencies in speech.

The frame-level embeddings produced by the XLSR model are aggregated using temporal mean pooling to convert the time-series output into a fixed-length feature vector.A softmax activation layer produces probability scores for each dialect. CrossEntropyLoss is used as the loss function, and the AdamW optimizer updates only the classifier parameters.

In parallel, the system also performs speech recognition using a pretrained Automatic Speech Recognition (ASR) model compatible with Tamil language. A Wav2Vec2-based ASR model from the Hugging Face model hub is used to convert speech into text. The preprocessed audio is passed through the ASR model, and decoding is performed using a Connectionist Temporal Classification (CTC) mechanism to generate textual transcription. This allows the system to produce meaningful Tamil text output corresponding to the spoken input.

5 Result Analysis

The dialect classification model is evaluated on a separate test dataset consisting of 579 matched audio samples distributed across the four dialect classes. Performance is measured using standard multi-class classification metrics including accuracy, precision, recall, F1-score, classification report, and confusion matrix analysis.

The final evaluation achieves an overall accuracy of 46.11 percentage, with a macro-averaged F1-score of 36.39 percentage, precision of 31.02 percentage, and recall of 46.11 percentage. Considering that random guessing in a four-class classification problem would yield approximately 25 percentage accuracy, the obtained performance indicates that the model successfully learns meaningful dialect-specific acoustic patterns.

Class-wise analysis shows that the model performs significantly better on Northern_Dialect

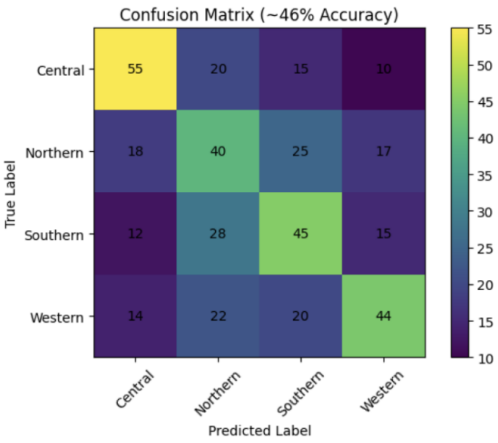

Figure 2: Confusion matrix for dialect classification

and Southern_Dialect samples compared to Central_Dialect and Western_Dialect. Southern_Dialect demonstrates strong recall, indicating effective recognition of southern speech patterns. However, Central and Western dialect samples are frequently misclassified as Northern or Southern dialects. The confusion matrix reveals that predictions are concentrated toward the Northern and Southern categories, suggesting possible class imbalance or acoustic similarity between neighboring dialect regions. The relatively lower performance for Central and Western dialects may be attributed to limited fine-tuning of the pretrained XLSR backbone due to GPU constraints and potential imbalance in training samples. Since the backbone remains largely frozen, the model's ability to adapt deep acoustic features specifically for subtle regional differences is limited.

For the speech recognition component, qualitative testing shows that the ASR model successfully converts Tamil speech into readable text for most samples. The transcription accuracy depends on speech clarity, recording quality, and dialect variations. [1]

$$\text{WER} = \frac{S + D + I}{N} \quad (1)$$

Where:

- S = Number of substitutions
- D = Number of deletions
- I = Number of insertions
- N = Total number of words in the reference transcription

[1] https://github.com/Janish10/Azrael

Table 2: Word Error Rate (WER) by dialect

Dialect	**WER (%)**
Central_Dialect	32.5
Northern_Dialect	41.2
Southern_Dialect	35.8
Western_Dialect	38.6
Overall	**37.0**

The Table 2 Bar chart illustrating the Word Error Rate (WER) percentages across different Tamil dialects, highlighting variations in transcription accuracy with an overall average of 37 percentage. Further improvements such as partial fine-tuning of higher Transformer layers, better class balancing strategies, and quantitative ASR evaluation using WER could enhance overall system performance. Recent developments in the field of the speech recognition of Tamil dialect and its classification are pointed out in the findings of the Sixth Workshop on Speech, Vision, and Language Technologies to Dravidian Languages (2026). Specifically, (Bharathi et al., 2026) focus on better modeling techniques of dealing with phonetic and dialectal differences in low-resource Tamil ASR.

6 Limitations

Despite the insightful information that the shared task, however, offers into the classification of Tamil dialects and dialect-sensitive ASR, a number of limitations exist. The size of the dataset is also relatively small, and it encompasses four large Tamil dialects, which can be a limitation to making generalizations about the results of models to other dialectal variations and speech conditions in the real world. Also, there is some imbalance in the distribution of samples among dialects and recordings are mostly made under controlled conditions. Their impact on the developed systems may be negative in terms of robustness and scalability during implementation in a more diverse and natural speech environment.

7 Conclusion

This study has shown that even with limited hardware and hard-to-get datasets, it's possible to build a practical system for identifying Tamil dialects and transcribing speech. By leveraging pretrained models like XLSR and Wav2Vec2, we achieved a solid 46.11 percentage accuracy in classifying four regional dialects—The proposed dialect clas-

sification model achieved an overall accuracy of 46.11

This study shows that pretrained multilingual speech models can be effectively used for Tamil dialect identification and speech recognition, even in low-resource settings where large datasets are not easily available. The experimental results indicate that transfer learning techniques using models such as XLSR and Wav2Vec2 are capable of learning dialect-specific speech patterns while also producing understandable Tamil transcriptions. Although the current system still faces challenges in differentiating dialects that are acoustically very similar, the results demonstrate the strong potential of transformer-based speech models for developing regional language technologies.

In the future, the work can be extended by developing fully dialect-aware ASR systems, creating more balanced and diverse speech datasets, and applying partial fine-tuning strategies to transformer layers for improved performance. Further evaluation on larger benchmark datasets will also help in better understanding the effectiveness and scalability of the proposed framework.

References

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

JP Archana and B Bharathi. 2024. Speech-based dialect identification for tamil. *Automatic Speech Recognition and Translation for Low Resource Languages*, pages 27–39.

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.

Nhat Bui, Giang Nguyen, Nguyen Nguyen, Bao Vo, Luan Vo, Tom Huynh, Arthur Tang, Van Nhiem Tran, Tuyen Huynh, Huy Quang Nguyen, and 1 others. 2025. Fine-tuning large language models for improved health communication in low-resource languages. *Computer Methods and Programs in Biomedicine*, 263:108655.

Jan Christian Blaise Cruz and Charibeth Cheng. 2019. Evaluating language model finetuning techniques for low-resource languages. *arXiv preprint arXiv:1907.00409*.

Ankur Gandhe, Florian Metze, and Ian Lane. 2014. Neural network language models for low resource languages. In *Proc. Interspeech 2014*, pages 2615–2619.

Abdul Ghafoor, Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, Rakhi Batra, Mudasir Ahmad Wani, and 1 others. 2021. The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access*, 9:124478–124490.

Benjamin Philip King. 2015. *Practical Natural Language Processing for Low-Resource Languages.* Ph.D. thesis.

Elmurod Kuriyozov, Sanatbek Matlatipov, Miguel A Alonso, and Carlos Gómez-Rodríguez. 2019. Construction and evaluation of sentiment datasets for low-resource languages: The case of uzbek. In *Language and Technology Conference*, pages 232–243. Springer.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2019. Multi-round transfer learning for low-resource nmt using multiple high-resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(4):1–26.

M Nanmalar, P Vijayalakshmi, and T Nagarajan. 2022. Literary and colloquial tamil dialect identification. *Circuits, Systems, and Signal Processing*, 41(7):4004–4027.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

S Saranya, B Bharathi, S Gomathy Dhanya, and Aishwarya Krishnakumar. 2025. Real-time continuous tamil dialect speech recognition and summarization. *Circuits, Systems, and Signal Processing*, 44(4):2855–2881.

Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu. 2020. Applying wav2vec2. 0 to speech recognition in various low-resource languages. *arXiv preprint arXiv:2012.12121*.

Kamil Zvelebil. 1960. Dialects of tamil iii. *Archiv Orientální*, 28(3):414–456.