Dialectmind@DravidianLang Tech 2026:Zero-Shot Dialectal Tamil Automatic Speech Recognition Using a Large Pretrained Conformer Model

K Gayathri[1], **B Bharathi**[2]
[1] St. Joseph's Institute of Technology, Tamil Nadu, India
[2]Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India
gayathrimsec.8@gmail.com, bharathib@ssn.edu.in

Abstract

The low-resource dialectal Automatic Speech Recognition (ASR) in languages like Tamil is a critical issue because of phonological differences, lack of labeled data and because of the differences in the acoustic of speech patterns among regions. This paper will introduce a dialect-conscious Tamil ASR model that is trained on the Conformer-CTC-BPE-Large framework via the NVIDIA NeMo framework. This model is an integration of convolutional subsampling, multi-head self-attention, and Connectionist Temporal Classification (CTC) decoding along with a BPE tokenizer to make possible both efficient end-to-end speech recognition. The system is tested on the audio recordings of dialectal Tamil, in which mono-channel audio normalization and batch transcription are used. Our findings indicate that even using large pretrained Conformer models, dialectal ASR tasks are successfully implemented even in zero-shot. Transcriptions generated are examined and the challenges associated with the dialectal differences and acoustic models, and we comment on the possible future directions of enhancing data-efficient adaptation in low-resource speech recognition.

Keywords: Tamil ASR, Whisper Large-V3, Multilingual Transformer, Low-Resource Languages, WER, Language-Guided Decoding.

1 Introduction

Recent developments in Automatic Speech Recognition (ASR) have been driven by deep neural network models trained on large multilingual datasets, such as wav2vec 2.0 and Whisper. Although these models are effective with languages that have abundant data, it nevertheless faces challenges with low-resource and dialectal languages. Tamil, with over 80 million speakers in India, Sri Lanka, and worldwide, draws attention to these problems due to its agglutinative morphology and phonetic variability and absence of annotated data of dialectal speech. The fact that there are differences in pronunciation, rhythm and vocabulary in different dialects, like Madurai, Coimbatore, Chennai and Jaffna, compounds the problem, making it more difficult to generate reliable ASR systems. This paper evaluates the performance of the Conformer-CTC-BPE-Large model in a zero-shot setting, in the NVIDIA NeMo architecture. The Conformer model utilizes convolutional layers, as well as self-attention, to capture both local and global contextual dependencies, which may give it the ability to work across languages. Other important steps such as turning an audio into a mono audio, normalizing the sampling rate and standardizing output are also outlined to be important in ensuring accuracy in transcription. This study provides a reproducible baseline and highlights key challenges and opportunities for improving ASR in Tamil and other low-resource South Asian languages.

2 Related Work

Tamil speech processing has evolved alongside broader advances in automatic speech recognition (ASR). Early systems primarily relied on rule-based and linguistically informed modeling approaches. For example, morpheme-based language modeling techniques demonstrated that Tamil's rich morphological structure could improve recognition accuracy (Saraswathi and Geetha, 2007). Broader surveys on Indian language speech recognition also highlighted early acoustic and language modeling challenges (Hemakumar and Punitha, 2013). Tamil speech recognition has gone through rule based and statistical methodology of speech recognition to the modern deep learning-based methods (Saraswathi and Geetha, 2007; Changrampadi et al., 2022). Recent work is concerned with dialect recognition, and dialect-conscious speech processing based on acoustic and neural models (Nanmalar et al., 2019, 2025). Nevertheless, Tamil

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 227–231

dialectal ASR is still difficult because there is a phonological variation and little labeled data (Nagarajan et al., 2024).

With the emergence of machine learning, research shifted toward data-driven and end-to-end architectures. Transformer-based and neural ASR systems for Tamil showed improved robustness compared to traditional pipelines (Changrampadi et al., 2022). Multilingual and low-resource approaches further demonstrated the benefits of shared representations across related languages.

Dialect identification has received increasing attention in recent years. Acoustic feature-based literary and colloquial Tamil classification was explored by Nanmalar et al. (2019), while later work applied 1D-CNN models for improved dialect discrimination (Nanmalar et al., 2025). Prosodic and speech-based dialect modeling was further examined by Archana and Bharathi (2024). Cross-lingual dialect recognition across Tamil and Telugu was investigated by Raaghavendran et al. (2025), highlighting shared regional speech patterns.

Recent efforts have also focused on dialect-aware system development and corpus construction. A multi-dialect Tamil speech corpus was introduced to support ASR benchmarking (Bharathi et al., 2025). Real-time dialect-aware recognition and summarization systems were proposed by Saranya et al. (2025), demonstrating practical deployment interest. Recent work has explored dialectal Tamil speech recognition and classification within Dravidian language technologies, highlighting the importance of modeling dialectal variation in speech systems (Bharathi et al., 2026). These studies emphasize the role of dialect-aware approaches for improving ASR performance in low-resource Tamil speech settings.

Recent survey and review papers emphasize ongoing challenges, including limited annotated dialectal corpora, dataset imbalance, and computational constraints (Nagarajan et al., 2024). These limitations continue to restrict large-scale supervised dialect modeling.

Unlike prior studies that rely on dialect-labeled training data or handcrafted feature engineering, the present work evaluates a pretrained multilingual ASR system in a zero-shot setting for dialectal Tamil. This approach aims to explore transfer learning–based dialect awareness under low-resource conditions.

3 Dataset

The data is composed of four large groups of Tamil dialects which include the Southern dialects, Northern dialects, Western dialects and Eastern dialects. Among these, the Northern dialect has the largest number of training samples (1696), and a total length of some 3 hours and 29 minutes. The Southern dialect is next in number with 1427 samples and a little less overall duration. The Western dialect has 1126 samples and the Eastern dialect has the least number of samples amounting to 885 and approximately an hour of recorded speech. The differences in the sample size and the overall time of different dialects suggest that the imbalance between classes in the dataset is moderate and can affect the learning process of speech recognition models. Specifically, the Eastern dialect that contains the smallest number of data points possibly can be prone to transcription errors or worse recognition stability. Another difference in the model performance could be variation in speech duration in addition to the difference in the sample count. Dialects of longer total length offer more acoustic variation, such as various speakers, speaking styles and patterns of pronunciation. This can aid the model in generalization of such dialects. On the other hand, exposure time can be short, and that will limit exposure to phonetic variability, which can harm strength. In sum, the data is diverse enough to assess dialect-sensitive speech recognition and also represents the real-life conditions of dialect-sensitive speech recognition, including data imbalance and dialect low-resource situation.

4 proposed work

4.1 Conformer Architecture

The Conformer architecture combines multi-head self-attention and convolutional layers to capture both global context and local acoustic features in speech signals. This hybrid approach achieves better performance than standalone Transformer or CNN models on benchmarks such as LibriSpeech. In the NVIDIA NeMo framework, the Conformer-CTC-BPE-Large model uses Connectionist Temporal Classification (CTC) and Byte Pair Encoding (BPE) for efficient end-to-end speech recognition. CTC enables the model to learn speech-to-text alignment without frame-level labeling, while BPE improves recognition of complex and rare words by breaking them into smaller subword units, making it especially effective for morphologically rich

languages like Tamil.

4.2 Tamil ASR

Previous studies on Tamil ASR had been done with the primary emphasis on standard Hidden Markov Model (HMM)-based models and then moved to deep learning-based acoustic model designs. The studies form substantial foundations towards Tamil speech recognition, which dealt with phoneme modeling, language modeling, and dataset development, under resource-constrained environments. The focus of most of this work, however, was on standard or limited-domain speech of Tamil. Much more recently, multilingual models In larger scale, there are Whisper and Conformer-based systems built using NVIDIA NeMo which have shown high multilingual behavior. Although successful in high resource and cross-lingual situations, these models have not as yet been tested on dialectal Tamil corpora in a systematic and holistic manner. This discontinuity suggests that particular research should be done on how well such pretrained architectures will adapt to other Tamil dialects without any explicit adaptation.

4.3 Conformer-CTC-BPE-Large

The Conformer-CTC-BPE-Large model is a speech recognition architecture that is trained based on NVIDia NeMo. It is constructed on top of stacked Conformer blocks which are specially developed to balance global contextual modeling with local feature extraction. Each block has four principal ones, viz., (1) a feed-forward module, (2) a multi-head self-attention layer, (3) a convolution module, which uses depthwise separable convolutions, and (4) a second feed-forward module, which has a post-normalization half-step residual connection, which completes the typical feed-forward sandwich structure. This hierarchical structure allows the model to simultaneously acquire a global and a local expression of speech signals.

4.4 CTC Decoder and BPE Tokeniser

The representations of the encoder are decoded with the help of a linear projection layer, followed by a log-softmax operation producing frame-level posterior probabilities of each subword unit. Inference Greedy decoding works by taking the most probable token at each time-step and applying CTC collapsing rules to the resulting transcription to give the final transcription. The model undergoes training using a tokenizer, a Byte Pair Encoding

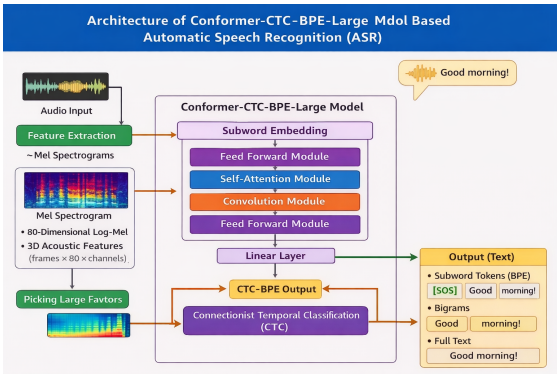

Figure 1: Architecture of Conformer-CTC-BPE-Large model BASED Automatic Speech Recognition (ASR)

(BPE) encoder, on the source corpus of the model, and decodes and encodes text in units of reusable sub words. This method implicitly includes morphological structure and is particularly useful for agglutinative Tamil, in which the terms of complex and inflected forms of words are made easier and the number of vocabulary is not large.

Component	Specification
Architecture	Conformer-CTC
Tokeniser	Byte Pair Encoding (BPE)
Encoder Layers	18
Model Dimension	512
Attention Heads	8
Conv. Kernel	31
Input Features	80-dim log-Mel filterbanks
Framework	NVIDIA NeMo

Table 1: Conformer-CTC-BPE-Large model specifications.

5 Methodology

The proposed system follows an end-to-end approach for dialectal Tamil speech recognition using the Conformer-CTC-BPE-Large architecture. First, the input Tamil speech recordings in WAV format are preprocessed by converting multi-channel audio into mono format to ensure compatibility with the model. Log-Mel spectrogram features are then extracted from the audio signal, capturing the time–frequency characteristics of speech. These acoustic features are passed through a convolutional subsampling layer to reduce temporal resolution and computational complexity. The processed features are fed into the Conformer encoder, which combines feed-forward layers, multi-head self-attention, and convolution modules to model both global contextual information and local acoustic patterns. The encoder output is projected

through a linear layer, and Connectionist Temporal Classification (CTC) is applied to generate frame-level predictions without requiring explicit alignment between audio and text. During inference, greedy CTC decoding selects the most probable subword sequence. Finally, the Byte Pair Encoding (BPE) tokenizer reconstructs the predicted subword units into complete textual transcriptions. The entire process is performed in a zero-shot setting without Tamil-specific fine-tuning, and audio files are transcribed individually to handle variable-length inputs efficiently.

6 Experimental Setup

Hardware. Experiments have been run on Google Colab on the NVIDIA GPUs (T4/A100) in CUDA 12.1 and PyTorch 2.x.

Dataset. The data was collected as dialectal Tamil speech records in the WAV format on the Google Drive. They only picked valid files of type wav, not hidden or system generated. The study closely resembled a zero-shot qualitative testing environment since not all the recordings would have ground-truth transcriptions, involving transcription behavior between various dialects. The source code available here.[1]

Evaluation Metric. The main measure of evaluation used in samples when it is possible to refer to transcriptions of references was Word Error Rate (WER). WER is a measure of the disparity between the predicted and the reference transcription and it is calculated as:

$$\text{WER} = \frac{S + D + I}{N} \tag{1}$$

and where S, D and I are the number of substitutions, deletions, and insertsions and N is the overall length of the reference transcription.

7 Results and Analysis

Metric	Value
Word Error Rate (WER)	0.60

Table 2: Overall WER on labeled dialectal Tamil samples in the zero-shot setting.

The results are shown in the Table 2, The proposed dialectal Tamil ASR system achieved a WER of,0.60 (zero-shot setting). We note that this shows the ability of the pretrained Conformer-CTC-BPE-Large architecture to generate intelligible transcriptions without any task-specific fine-tuning on Tamil speech data. Nevertheless, the fact that WER is relatively large indicates the difficulties related to the dialectal variation, such as pronunciation differences, regional lexis, and phonemic structure. Moreover, transliteration or substitution errors could be made by using a BPE vocabulary that has mainly been trained on other languages to process Tamil speech. Irrespective of these shortcomings, the findings indicate that big pretrained speech models can be used as a baseline of dialect-conscious Tamil ASR and a starting point of further enhancements with the help of fine-tuning and dialect-specific data adaptation.

8 Conclusion

In the current paper, we present a dialect-based evaluation scheme of Tamil Automatic Speech Recognition (ASR) using the Conformer-CTC-BPE-Large model within the context of NVIDIA Nemo. The proposed pipeline encourages multi-channel audio normalization, GPU based inference, and structured transcription output. The results of The zero-shot results highlight the challenges associated with dialectal Tamil, including variation of vocabulary and acoustic variation. However, the good representational potential of the Conformer architecture provides a good foundation on which future fine-tuning and adaptation can be accomplished. Overall, this work offers a reproducible baseline and encourages future research in low resource and dialect sensitive ASR for Tamil and other South Asian languages.

References

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

[1] http://lisindia.ciil.org/Tamil/Tamil_vari.html.

enhancing tamil automatic speech recognition. *Circuits, Systems, and Signal Processing*, pages 1–19.

Mohamed Hashim Changrampadi, A Shahina, M Badri Narayanan, and A Nayeemulla Khan. 2022. End-to-end speech recognition of tamil language. *Intelligent Automation & Soft Computing*, 32(2).

G Hemakumar and P Punitha. 2013. Speech recognition technology: a survey on indian languages. *International Journal of Information Science and Intelligent System*, 2(4):1–38.

Sureshkumar Nagarajan and 1 others. 2024. Advances in speech and text processing for dravidian language: A comprehensive review. In *2024 9th International Conference on Communication and Electronics Systems (ICCES)*, pages 1715–1722. IEEE.

M Nanmalar, S Johanan Joysingh, P Vijayalakshmi, and T Nagarajan. 2025. A feature engineering approach for literary and colloquial tamil speech classification using 1d-cnn. *Speech Communication*, 173:103254.

M Nanmalar, P Vijayalakshmi, and T Nagarajan. 2019. Literary and colloquial dialect identification for tamil using acoustic features. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, pages 1303–1306. IEEE.

G Naveen Raaghavendran, P Jaswanth, S Shreenithi, Sri Vaishnavi JV, and KR Bindu. 2025. Dialect recognition in tamil and telugu: An integrated approach. In *2025 3rd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, pages 638–643. IEEE.

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.

SS Saraswathi and TVG Geetha. 2007. Language models for tamil speech recognition system. *IETE Technical Review*, 24(5):375–383.