CrewX@LT-EDI-2025: Transformer-Based Tamil ASR Fine-Tuning with AVMD Denoising and GRU-VAD for Enhanced Transcription Accuracy

Ganesh Sundhar S[1], Hari Krishnan N[1], Arun Prasad TD[1], Shruthikaa V[1], Jyothish Lal G[1]
[1]Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
{cb.en.u4aie22017, cb.en.u4aie22020, cb.en.u4aie22004, cb.en.u4aie22047}
@cb.students.amrita.edu, g_jyothishlal@cb.amrita.edu

Abstract

This research presents an improved Tamil Automatic Speech Recognition (ASR) system designed to enhance accessibility for elderly and transgender populations by addressing unique language challenges. We address the challenges of Tamil ASR—including limited high-quality curated datasets, unique phonetic characteristics, and word-merging tendencies—through a comprehensive pipeline. Our methodology integrates Adaptive Variational Mode Decomposition (AVMD) for selective noise reduction based on signal characteristics, Silero Voice Activity Detection (VAD) with GRU architecture to eliminate non-speech segments, and fine-tuning of OpenAI's Whisper model optimized for Tamil transcription. The system employs beam search decoding during inference to further improve accuracy. Our approach achieved state-of-the-art performance with a Word Error Rate (WER) of 31.9,winning first place in the LT-EDI 2025 shared task.

Keywords: Speech Recognition, Transformer, Whisper, Adaptive Variational Mode Decomposition, Vulnerable populations, Low-resource, Dravidian language.

1 Introduction

Communication has been the cornerstone of human evolution, enabling individuals to share thoughts, emotions, and information. Human communication evolved naturally through verbal expression, with speech emerging as one of our earliest and most intuitive forms of interaction. With recent advancements in technology, a more natural way for human-computer interaction is necessary, which is satisfied with the help of speech processing techniques. This has led to the invention of Automatic Speech Recognition (ASR) (Yu and Deng, 2016) systems. The goal of an ASR system is to convert spoken language into text and it plays a crucial role in various applications such as virtual assistants, transcription services, and accessibility tools.

Despite advancements in technology, ASR still remains a complex task due to several challenges. The high variability in speech data makes it difficult for the models to generalize across different speakers. Additionally, while ASR systems for languages like English, Spanish and Mandarin benefit from extensive datasets and pretrained models, Dravidian languages like Tamil and Malayalam suffer from the scarcity of extensive & good quality annotated datasets, making it hard to build robust ASR systems. Tamil, in particular, has unique phonetic characteristics and word-merging tendencies that further complicate transcription accuracy (Akhilesh et al., 2022; Shraddha et al., 2022).

This work addresses these challenges by fine-tuning a Transformer-based ASR system for Tamil speech, specifically for vulnerable old-aged and transgender individuals, who often face difficulties in accessing essential services due to lack of literacy and familiarity with technology. To enhance transcription accuracy, our method integrates Adaptive Variational Mode Decomposition (AVMD) (Lian et al., 2018) for noise reduction and Silero Voice Activity Detection (VAD) (Team, 2021) to eliminate non-speech segments. We then, finetune OpenAI's Whisper (Radford et al., 2023) model for Automatic Speech Recognition, leading to promising results. This approach achieved state-of-the-art (SOTA) performance, ranking first in the shared task with a Word Error Rate (WER) of 31.9. This demonstrates the effectiveness of this pipeline in handling the complexities of Tamil speech for vulnerable populations (Chowdary et al., 2024).

2 Related Work

Early ASR systems relied on rule-based phonetic models and statistical methods like Hidden Markov Models (HMMs) (Levinson, 1986), Gaussian Mix-

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 11–16

ture Models (GMMs) (Gorin et al., 2014) and Dynamic Time Warping (DTW) (Nair and Sreenivas, 2008). However, these methods struggled with high variability in speech, such as speaker, accents, age, gender, background noise, and spontaneous speech patterns. This led to the emergence of traditional deep learning based models such as Recurrent Neural Networks (RNNs) (Jain et al., 2020) and Long Short-Term Memory (LSTM) (Weninger et al., 2015) networks. (Dahl et al., 2011) used Deep Neural Network and HMM based hybrid model which further improved the quality of the transcriptions generated.

But even they struggled with capturing long range dependencies. The major breakthrough in processing temporal was made with the invention of the Transformer architecture (Vaswani et al., 2017) in 2017. It introduced the Multi Head Self Attention (MHSA) mechanism which excelled in capturing both long term and short term dependencies. Hence, the Transformer-based models significantly improved transcription accuracy by learning complex patterns from large datasets. Self supervised models like Wav2Vec 2.0 (Baevski et al., 2020) reduced reliance on labeled data by learning speech representations from raw audio, significantly improving ASR for low-resource languages. OpenAI's Whisper further advanced ASR with large-scale weak supervision, enabling robust transcription, translation, and language identification across diverse datasets (Barathi Ganesh et al., 2024).

3 Dataset

Elderly individuals frequently visit essential service locations such as banks, hospitals, and administrative offices but struggle with technological tools designed to assist them. Similarly, transgender individuals often face educational barriers due to societal prejudices, making speech a primary mode of communication for accessing essential services. Hence, the dataset provided for this shared task specifically targets vulnerable elderly and transgender individuals (B et al., 2022). The dataset consists of 7.5 hours of spontaneous Tamil speech, which is split into 5.5 hours of transcribed speech for training and 2 hours of unlabeled speech for testing. The train data provided was further made into a 80-20 split for training and development. The dataset distribution is provided in Table 1

Split	Audios	Duration (hours)
Train	726	4.4
Dev	182	1.1
Test	451	2.0
Total	1,359	7.5

Table 1: Dataset distribution of Tamil ASR corpus

4 Methodology

The speech signal was initially denoised using Adaptive Variational Mode Decomposition (AVMD), which decomposes it into variational modes and reconstructs the relevant components to remove noise while preserving speech clarity. Next, Silero VAD (GRU based) was applied to eliminate silence and non-speech segments, reducing computational load due to unnecessary processing and at the same time improving transcription accuracy.The processed audio is then passed through the Whisper processor, which converts it into log-Mel spectrogram (Stevens et al., 1937) features using a standardized pipeline. Finally, the extracted features are passed to the Whisper-Tamil-Medium model for generating transcriptions, with beam search decoding (Lowerre, 1976) during testing to enhance the accuracy and reduce the WER. The overall workflow is depicted in Figure 1.

4.1 AVMD - Denoising

The dataset contains multiple audio samples, where some are of high quality while others suffer from a bit of background noise, which can affect the transcription process. Specifically, audio files in Series 3 (*Audio-3_1.wav* to *Audio-3_32.wav*) exhibit excellent clarity, whereas Series 2 (*Audio-2_1.wav* to *Audio-2_26.wav*) contains noticeable background noise. To account for this, Adaptive Variational Mode Decomposition (AVMD) was selectively applied based on a noise parameter threshold (Signal to Noise Ratio (Johnson, 2006)) , ensuring only noisy signals underwent processing. Unlike traditional Variational Mode Decomposition (VMD) (Dragomiretskiy and Zosso, 2013), which uses fixed decomposition parameters, AVMD dynamically adjusts mode selection and bandwidth constraints based on the signal's characteristics, making it more effective in separating noise components from speech. This method outperforms other denoising methods such as wavelet denoising (Luo et al., 2012) or spectral subtraction (Martin, 1994), which often introduce artifacts.

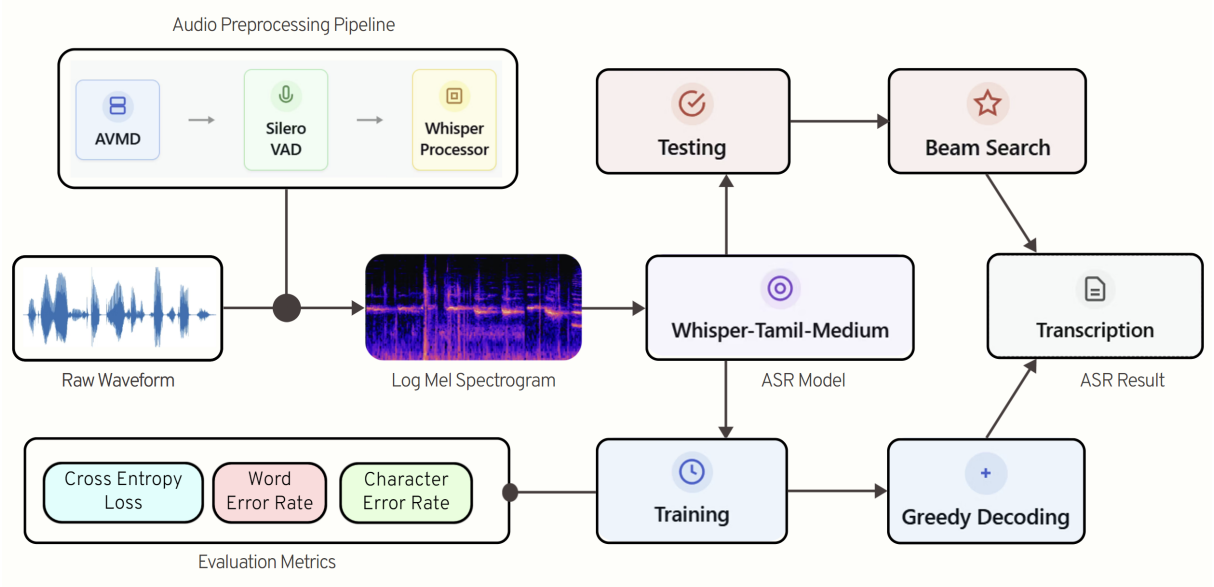

Figure 1: End to end pipeline for Tamil ASR using the Whisper-Tamil-Medium model. The raw audio is preprocessed using AVMD, Silero VAD, and the Whisper Processor, transforming it into Log Mel Spectrograms. These spectrograms serve as input to the Whisper model, which undergoes training and evaluation with greedy decoding and beam search techniques. The resulting transcriptions are assessed using cross-entropy loss, WER, and CER to measure performance.

4.2 Silero VAD

Some of the audio samples in the dataset contained long pauses, which negatively impacted ASR performance. So, any segment with silence exceeding 300 ms was removed, such as in *Audio-1_01.wav*, where a pause from 12s to 14s was eliminated. Silero VAD was chosen for this task due to its robust deep learning based Voice Activity Detection (VAD), outperforming traditional energy-based or statistical methods (Sohn et al., 1999). The implementation involved loading the pretrained Silero VAD model, detecting speech timestamps, merging overlapping or close speech segments, and extracting speech regions while preserving 300 ms of buffer before and after detected speech to compensate for potential mis-detections where speech might be partially cropped.

4.3 Whisper Finetuning

We chose the pretrained hugging face Whisper model from *"vasista22/whisper-tamil-medium"* as it achieved the SOTA performance even without preprocessing in the same shared task last year (Jairam et al., 2024). The selected model was pretrained on various tamil ASR datasets such as *IISc-MILE Tamil ASR Corpus* (A et al., 2022), *ULCA ASR Corpus*, etc. Whisper is a transformer-based ASR model that processes audio using a feature extractor that converts raw waveforms into mel spectrograms, capturing key frequency components over time. The transcriptions given were converted to tokens with the help of the Whisper Tokenizer. The model begins with a CNN-based feature extractor (O'shea and Nash, 2015) layer, where two 1D convolution layers (Conv1d) expand the input from 80 mel filter banks to 1024 channels, followed by downsampling. The processed audio features are directly fed to the Whisper encoder then passed to the decoder, both containing 24 Transformer layers for transcription generation. The decoder fetches the features from the encoder with the help of cross-attention with the encoder memory. The final linear layer maps the decoder output to the vocabulary space, generating transcriptions (B et al., 2025).

4.4 Beam Search Decoding

Beam search decoding is an exploratory search technique employed in ASR systems to identify the optimal output sequence. In contrast to greedy decoding, which chooses the maximum probability token per step, beam search tracks several potential sequences (beams) during each decoding phase, minimizing errors resulting from choices that appear optimal locally but prove suboptimal overall. It is used only during test time to refine the output by considering multiple possible candidate states. During training, it is not required since the model is optimized using teacher forcing, where the correct

target sequence is provided at each step. Training prioritizes computational efficiency, while beam search during inference ensures higher-quality predictions.

5 Experimentation

For training, the model was optimized using *AdamW* (Loshchilov and Hutter, 2017) with a learning rate of 10^{-5} and weight decay of 10^{-2} to prevent overfitting. The loss function used was *CrossEntropyLoss*, which measures the difference between predicted and actual token distributions. We employed ReduceLROnPlateau to automatically decrease the learning rate by half whenever validation loss stopped improving for two consecutive epochs, which helped maintain smooth training progress. The models were trained on a *4080 Super* GPU with 4 as the batch size. During testing, beam search decoding was employed, thus increasing transcription accuracy. The trends observed during the training and validation process are given in Figure 2.

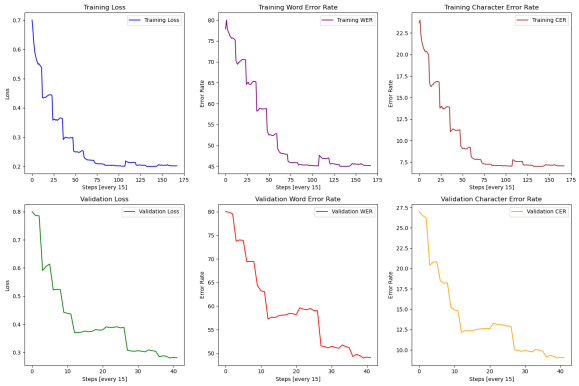

Figure 2: Training and validation performance metrics for the best-performing Tamil ASR model.

6 Result and Analysis

6.1 Word Error Rate

To evaluate the ASR pipeline built, Word Error Rate (WER) (Levenshtein et al., 1966) was used, as it is the metric specified for comparing the results in the shared task. It quantifies the transcription accuracy by comparing the words in the system's output with the reference transcript. The WER is defined as

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

where S represents the number of substitutions (incorrectly transcribed words), D denotes deletions (omitted words), I accounts for insertions (extra words added), and N is the total number of words in the original reference transcript. A lower value of WER generally indicates better performance.

6.2 Model Evaluation

The pipeline built was assessed using three different metrics: Cross Entropy Loss, Word Error Rate (WER), and Character Error Rate (CER) (Rice et al., 1995). The results of the best performing model in terms of WER are summarized in Table 2.

Metric	Train	Validation	Test
Loss	0.202	0.281	-
WER	45.212	49.095	31.9
CER	7.061	9.037	-

Table 2: Performance Metrics of the Best Model

6.3 Result Comparison

Our proposed methodology, integrating Adaptive Variational Mode Decomposition (AVMD) for denoising, Silero VAD for voice activity detection, Whisper for transcription, and Beam Search decoding, achieved a WER of 31.9 on the testing set, securing the 1st rank in the competition. The scores of the top five performing teams in the shared task are summarized in Table 3.

Rank	Team Name	WER (%)
1	CrewX	31.90
2	NSR	34.85
3	Wictory	34.93
4	JUNLP	38.42
5	SSNCSE	42.30

Table 3: Top 5 Teams scored based on Word Error Rate

7 Conclusion

In this paper, we presented a robust ASR pipeline tailored for noisy audio conditions. In conclusion, our methodology showcases the potential of combining noise-aware preprocessing and advanced decoding strategies to deliver accurate and reliable transcription in real-world, low-resource scenarios. This also provides a strong foundation for future enhancements in speech recognition under challenging conditions. **The entire implementation can be found here:** https://github.com/Ganesh2609/VulnerableSpeechASR

8 Limitations

Even though the proposed pipeline performed well in terms of the WER, it had a few noticeable limitations, which are as follows:

1. Some audio samples had severely distorted speaker voices, making them unintelligible even to human listeners. As a result, the model also struggled to transcribe such cases accurately.
2. The model occasionally merges separate Tamil words into a single word. For example, *vandhu irunthaal* is sometimes transcribed as *vanthirunthaal*, which may lead to a lesser WER.
3. The training dataset was relatively small for ASR tasks, with 908 audio samples for training and 451 for testing (approximately a 66.7-33.3 split). Creating a development set from the training data reduced the effective training size further, leading to signs of overfitting during training.

References

Madhavaraj A, Bharathi Pilar, and Ramakrishnan A G. 2022. Subword dictionary learning and segmentation techniques for automatic speech recognition in tamil and kannada. *arXiv preprint*.

A Akhilesh, P Brinda, S Keerthana, 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.

Bharathi B, Bharathi Raja Chakravarthi, Subalalitha Cn, Sripriya N, Arunaggiri Pandian, 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*, pages 339–345, Dublin, Ireland. Association for Computational Linguistics.

Bharathi B, Bharathi Raja Chakravarthi, Sripirya N, Rajeswari Natarajan, Rajalakshmi R, Suhasini S, and Swetha Valli. 2025. Overview of the fourth shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, Naples. 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. *Advances in neural information processing systems*, 33:12449–12460.

HB Barathi Ganesh, G Jyothish Lal, R Jairam, KP Soman, NS Kamal, and B Sharmila. 2024. Corepool—corpus for resource-poor languages: Badaga speech corpus. *Automatic Speech Recognition and Translation for Low Resource Languages*, pages 193–211.

Divi Eswar Chowdary, Rahul Ganesan, Harsha Dabbara, G Jyothish Lal, and B Premjith. 2024. Transformer-based multilingual automatic speech recognition (asr) model for dravidian languages. *Automatic Speech Recognition and Translation for Low Resource Languages*, pages 259–273.

George E Dahl, Dong Yu, Li Deng, and Alex Acero. 2011. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42.

Konstantin Dragomiretskiy and Dominique Zosso. 2013. Variational mode decomposition. *IEEE transactions on signal processing*, 62(3):531–544.

Arseniy Gorin, Denis Jouvet, Emmanuel Vincent, and Dung Tran. 2014. Investigating stranded gmm for improving automatic speech recognition. In *2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pages 192–196. IEEE.

Mahaveer Jain, Gil Keren, Jay Mahadeokar, Geoffrey Zweig, Florian Metze, and Yatharth Saraf. 2020. Contextual rnn-t for open domain asr. *arXiv preprint arXiv:2006.03411*.

R Jairam, G Jyothish, B Premjith, and M Viswa. 2024. Cen_amrita@ lt-edi 2024: A transformer based speech recognition system for vulnerable individuals in tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 190–195.

Don H Johnson. 2006. Signal-to-noise ratio. *Scholarpedia*, 1(12):2088.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Stephen E Levinson. 1986. Continuously variable duration hidden markov models for automatic speech recognition. *Computer Speech & Language*, 1(1):29–45.

Jijian Lian, Zhuo Liu, Haijun Wang, and Xiaofeng Dong. 2018. Adaptive variational mode decomposition method for signal processing based on mode characteristic. *Mechanical Systems and Signal Processing*, 107:53–77.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Bruce T Lowerre. 1976. *The harpy speech recognition system.* Carnegie Mellon University.

Guomin Luo, Daming Zhang, and DD Baleanu. 2012. Wavelet denoising. *Advances in wavelet theory and their applications in engineering, physics and technology*, pages 59–80.

Rainer Martin. 1994. Spectral subtraction based on minimum statistics. *power*, 6(8):1182–1185.

Nishanth Ulhas Nair and TV Sreenivas. 2008. Multi pattern dynamic time warping for automatic speech recognition. In *TENCON 2008-2008 IEEE Region 10 Conference*, pages 1–6. IEEE.

Keiron O'shea and Ryan Nash. 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

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 28492–28518. PMLR.

Stephen V Rice, Frank R Jenkins, and Thomas A Nartker. 1995. The fourth annual test of ocr accuracy. Technical report, Technical Report 95.

S Shraddha, Sachin Kumar, et al. 2022. Child speech recognition on end-to-end neural asr models. In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pages 1–6. IEEE.

Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. 1999. A statistical model-based voice activity detection. *IEEE signal processing letters*, 6(1):1–3.

Stanley Smith Stevens, John Volkmann, and Edwin B Newman. 1937. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190.

Silero Team. 2021. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. *Retrieved March*, 31:2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. 2015. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *Latent Variable Analysis and Signal Separation: 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings 12*, pages 91–99. Springer.

Dong Yu and Lin Deng. 2016. *Automatic speech recognition*, volume 1. Springer.