CEN_Amrita@LT-EDI 2024: A Transformer based Speech Recognition System for Vulnerable Individuals in Tamil

Jairam R[1,2], Jyothish Lal G[1], Premjith B[1], and Viswa M[2]
[1]Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham,
Coimbatore, India.
[2]RBG AI Research, RBG.AI, SREC Incubation Center, Coimbatore, India.
g_jyothishlal@cb.amrita.edu

Abstract

Speech recognition is known to be a specialized application of speech processing. Automatic speech recognition (ASR) systems are designed to perform the speech-to-text task. Although ASR systems have been the subject of extensive research, they still encounter certain challenges when speech variations arise. The speaker's age, gender, vulnerability, and other factors are the main causes of the variations in speech. In this work, we propose a fine-tuned speech recognition model for recognising the spoken words of vulnerable individuals in Tamil. This research utilizes a dataset sourced from the LT-EDI@EACL2024 shared task. We trained and tested pre-trained ASR models, including XLS-R and Whisper. The findings highlight that the fine-tuned Whisper ASR model surpasses the XLS-R, achieving a word error rate (WER) of **24.452**, signifying its superior performance in recognizing speech from diverse individuals.

Keywords : Dravidian Languages, Tamil, Speech Recognition, Vulnerable Speech, Transformer, Word Error Rate (WER)

1 Introduction

Speech is the most prevalent, clear, and frequently used form of worldwide communication. Speech processing involves obtaining valuable information from voice signals, such as automatic speech recognition (ASR) (Gaikwad et al., 2010). The goal of any ASR system is to teach computers to understand human speech and carry out user-defined tasks. The method of recognizing spoken words from audio input by applying auditory features is known as speech recognition. The majority of voice recognition algorithms have been trained on languages with abundant resources, like English, German, Spanish, and so on. With languages with few resources, such as Tamil, Kannada, Malayalam, etc., this is not the case. Despite being one of the most investigated fields among researchers, speech recognition systems still face issues when it comes to the conventional learning paradigm (Li et al., 2022), which can be utilized for both resource-rich and resource-poor languages. This is still one of the most unresolved challenges among the researchers (Nassif et al., 2019). The fundamental reason for this is the various natures of speech, often known as speech variants.

Speech recognition systems face a challenge when it comes to recognizing variances in speech. These variations are caused by factors such as the speaker's age, gender, and vulnerability (Kita, 2020) and etc. There have been a number of studies (Bharathi et al., 2022; Shivakumar et al., 2016; Shraddha et al., 2022; Murali Krishna et al., 2019; Bharathi et al., 2023) that have studied various methods to address these issues. These methods include the development of corpora and the fine-tuning of pre-trained models, particularly for languages that have limited resources.

In response, the LT-EDI team gathered a tagged Tamil speech corpus from elderly and transgender vulnerable individuals who had everyday conversations in administrative offices, banks, and hospitals. Some of these individuals were also vulnerable. Therefore, the vulnerability of the speaker is the primary focus. We propose modifying the Whisper Automatic Speech Recognition (ASR) model (Radford et al., 2023) in order to improve speech recognition for those who are vulnerable. In preparation for the work that the LT-EDI team is doing on voice recognition for the Tamil-language shared initiative, we fine-tuned the pre-trained Whisper model by using Tamil datasets. With a word error rate (WER) of

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 190–195
March 21, 2024 ©2024 Association for Computational Linguistics

24.452, the '*CEN_Amrita*' team was ranked top in the classification criteria. This achievement shows that the proposed strategy may address speech differences, especially in vulnerable populations.

The following sections describe the paper's contribution: Section 2 covers relevant works; Section 3 materials and technique; Section 4 results; and Section 5 conclusion.

2 Related Works

Advancements in deep learning have significantly impacted speech processing, notably in the domain of automatic speech recognition (ASR). Transformer-based architectures like BERT and GPT (Zheng and Woodland, 2021; Fohr and Illina, 2021; Kumar et al., 2022), initially tailored for text interpretation, have been extended to capture speech sequences, leveraging contextual information to enhance accuracy. DeepSpeech (Hannun et al., 2014), a flexible open-source program employing recurrent neural networks (RNNs), stands out for its adaptability across various languages and effective training methods. Self-supervised learning models such as Wav2Vec (Baevski et al., 2020) excel at speech pattern recognition by extracting pertinent features directly from unlabeled audio data.

Attention-based models, such as Listen, Attend, and Spell (LAS) (Chan et al., 2015), change how much weight is given to inputs during decoding. This helps with accurate transcription after a lot of training on big datasets. Lightweight architectures like QuartzNet (Kriman et al., 2020) emphasize high performance while maintaining low computational demands. Hybrid models, like ESPNet (Watanabe et al., 2018), combine convolutional and recurrent networks, showing that they are good at a number of different ASR benchmarks. Recent improvements, like HuBERT (Hsu et al., 2021), build upon Wav2Vec by integrating hierarchical transformations and elevating representation learning and ASR accuracy. These improvements have made significant advancements in the field of ASR and have achieved impressive results.

However, when it comes to resource-poor languages like Tamil, existing models underperform due to the variability in speech among native speakers. To address this, recent research has focused on fine-tuning pre-trained ASR models for specific languages. Models like XLSR-wav2vec2 (Conneau et al., 2020) have been customized for Tamil speech recognition, showcasing promising results with a significantly reduced word error rate (WER) of 39.65% (Bharathi et al., 2022) and 37.71% (Bharathi et al., 2023). This customized approach aims to enhance performance in understanding and transcribing speech for languages with limited available resources.

3 Materials and Methodology

3.1 Dataset Description

The dataset used in this study is from the shared task LT-EDI@2024. This shared task aims to develop a Tamil conversational speech corpus collected from vulnerable elderly people and transgender people in Tamil. This speech corpus contains recordings that capture real-world conversions from primary sites such as hospitals, banks, and administrative offices. The corpus contains males, females, and transgender speakers and a total of 7 and a half hours of speech data. There are two phases to the dataset's release: the first phase is for training, and the second phase is for testing. The test data consists of a total of two hours of unlabeled speech, whereas the training data consists of an average of 5.5 hours of speech that has been transcribed. Table 1 describes the detailed data statistics about the train, test, and validation splits used in this work.

Dataset	Splits	Audios	Hours
Training	Train	726	5.5
	Evaluation	192	
Testing	Test	348	2
Total		1266	7.5

Table 1: Data Statistics describing the train, test and validation split.

3.2 Methodology

In this study, speech recognition was performed using two pre-trained state-of-the-art (SOTA) models, Whisper and XLS-R. Both models were trained on the Tamil corpus, and the best results were submitted for the competition. Figures 1 and 2 show schematic block diagrams of the proposed approaches.

3.2.1 Whisper ASR

Whisper (Radford et al., 2023) is a pre-trained automatic speech recognition (ASR) model trained on 680,000 hours of multilingual and multitask supervised data sourced from the web. This end-to-end transformer-based model adopts the encoder-decoder architecture. Log-Mel spectrogram features are extracted from each audio file; this feature input undergoes processing in the encoder, featuring a compact stem composed of two convolution layers with a filter width of three and the GELU activation function. Notably, the stride of the second convolution layer is set at three. Following this, sinusoidal position embeddings are added to the stem's output, paving the way for the inclusion of encoder transformer blocks. Using pre-activation residual structures, these blocks build up to a final layer normalization step for the encoder output. The learned position embeddings are then fed into the decoder, which is responsible for generating the textual output.

The Whisper model boasts various variants, including whisper-large-v1, whisper-large-v2, and whisper-large-v3. For this study, we opted for the 'Vasista22/whisper-tamil-medium' pre-trained model from the huggingface and fine-tuned it on the Tamil speech corpus. Figure 1 displays the whisper model's flow diagram.

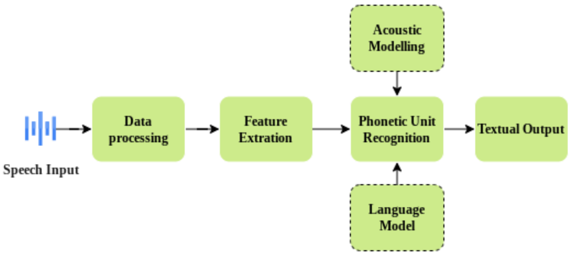


Figure 1: Flow Diagram for Whisper Model.

3.2.2 XLS-R

The XLS-R model is a multi-lingual adaptation of the Wav2Vec2 model for cross-lingual representational learning of speech. The pre-training of the model utilized more than 436,000 hours of speech data that was easily accessible to the general public. The speech data used for per-training is derived from a variety of sources, including audio books produced in 128 different languages and parliamentary proceedings. Since the model was trained on connectionist temporal classification (CTC), the Wav2Vec2CTCTokenizer should be used to decode the model's output. The model has three variations: Wav2Vec2-XLS-R-300M, Wav2Vec2-XLS-R-1B, and Wav2Vec2-XLS-R-2B. The parameters for each variant vary. For the experiments, we have utilized 'Wav2Vec2-XLS-R-300M' and fine-tuned it on the Tamil speech corpus. Figure 2 displays the flow diagram for the XLS-R model.

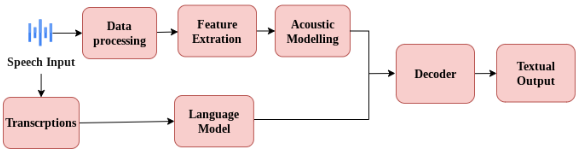


Figure 2: Flow Diagram for XLS-R Model.

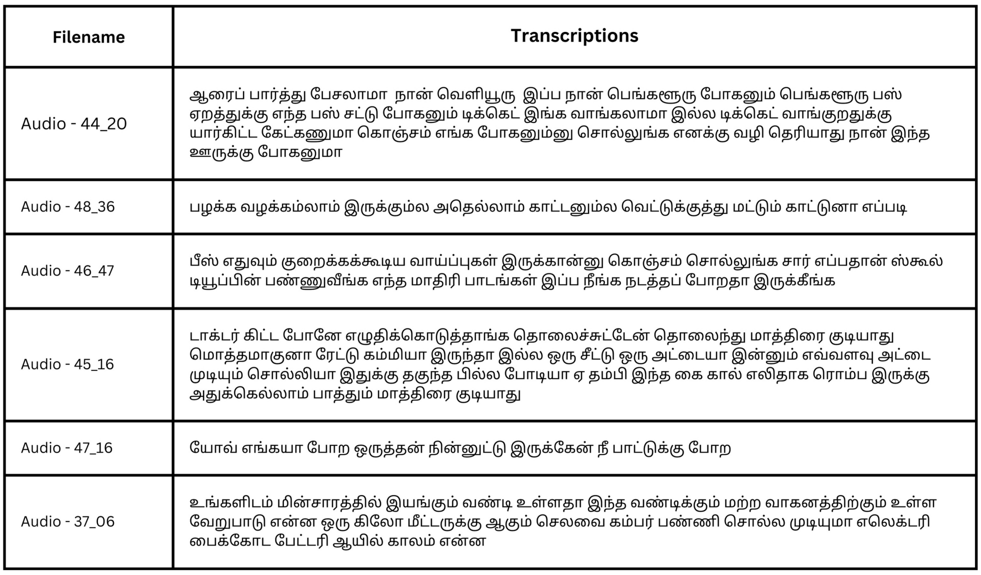
4 Results and Discussion

4.1 Experiments

The experimental setup comprises a Linux operating system, an 8-core Intel Xeon processor, 32GB of RAM, a 16GB NVIDIA T4 tensor core GPU,and CUDA 11.0. In the series of experiments employing both the whisper and XLS-R models, for the whisper part, we focused on the 'Vasista22/whisper-tamil-medium[1]' pre-trained model on Tamil. To make the audio data consistent for the Whisper model, all the audio files had to be resampled to 16 kHz, and then Log-Mel spectrogram features had to be extracted. Subsequently, we utilized the WhisperTokenizer to encode transcriptions into label IDs. To facilitate model training data preparation, we defined a data collator to handle batching and padding of the training examples.

The word error rate (WER) was an essential metric for assessing model performance during the training process. We fine-tuned the model using various combinations of hyperparameters, including learning rate, batch

[1] https://huggingface.co/vasista22/whisper-tamil-medium



Filename	Transcriptions
Audio - 44_20	ஆரைப் பார்த்து பேசலாமா நான் வெளியூரு இப்ப நான் பெங்களூரு போகனும் பெங்களூரு பஸ் ஏறத்துக்கு எந்த பஸ் சட்டு போகனும் டிக்கெட் இங்க வாங்கலாமா இல்ல டிக்கெட் வாங்குறதுக்கு யார்கிட்ட கேட்கணுமா கொஞ்சம் எங்க போகனும்னு சொல்லுங்க எனக்கு வழி தெரியாது நான் இந்த ஊருக்கு போகனுமா
Audio - 48_36	பழக்க வழக்கம்லாம் இருக்கும்ல அதெல்லாம் காட்டனும்ல வெட்டுக்குத்து மட்டும் காட்டுனா எப்படி
Audio - 46_47	பீஸ் எதுவும் குறைக்கக்கூடிய வாய்ப்புகள் இருக்கான்னு கொஞ்சம் சொல்லுங்க சார் எப்பதான் ஸ்கூல் டியூப்பின் பண்ணுவீங்க எந்த மாதிரி பாடங்கள் இப்ப நீங்க நடத்தப் போறதா இருக்கீங்க
Audio - 45_16	டாக்டர் கிட்ட போனே எழுதிக்கொடுத்தாங்க தொலைச்சுட்டேன் தொலைந்து மாத்திரை குடியாது மொத்தமாகுனா ரேட்டு கம்மியா இருந்தா இல்ல ஒரு சீட்டு ஒரு அட்டையா இன்னும் எவ்வளவு அட்டை முடியும் சொல்லியா இதுக்கு தகுந்த பில்ல போடியா ஏ தம்பி இந்த கை கால் எலிதாக ரொம்ப இருக்கு அதுக்கெல்லாம் பாத்தும் மாத்திரை குடியாது
Audio - 47_16	யோவ் எங்கயா போற ஒருத்தன் நின்னுட்டு இருக்கேன் நீ பாட்டுக்கு போற
Audio - 37_06	உங்களிடம் மின்சாரத்தில் இயங்கும் வண்டி உள்ளதா இந்த வண்டிக்கும் மற்ற வாகனத்திற்கும் உள்ள வேறுபாடு என்ன ஒரு கிலோ மீட்டருக்கு ஆகும் செலவை கம்பர் பண்ணி சொல்ல முடியுமா எலெக்டரி பைக்கோட பேட்டரி ஆயில் காலம் என்ன

Figure 3: Sample Transcriptions in Tamil from fine-tuned Whisper Model.

size, maximum steps, and optimizer selection. The most optimal results were achieved with specific hyper-parameter configurations. During training, a batch size of 4 proved effective, while for testing, a batch size of 8 was found to be optimal. The learning rate was set to $10^{-}5$, and the 'adamw_bnb_8bit' optimizer was employed to initialize the optimization process. This comprehensive approach to hyperparameter tuning and data preprocessing contributed to the success of the experiments with the Vasista22/whisper-tamil-medium model.

On the other hand, the 'Wav2Vec2-XLS-R-300M[2]' underwent fine-tuning on the Tamil speech corpus. To address sequence-to-sequence problems, typical fine-tuning of XLS-R models involves employing the connectionist temporal classification (CTC) algorithm. As a part of preprocessing, transcription texts have been cleaned by excluding special characters and developing a vocabulary from the processed transcriptions. The model anticipates input in the form of a 1-dimensional array at 16 kHz, prompting the loading and resampling of all audio files accordingly.

For the training phase, the Wav2Vec2Processor was employed to extract input values from the loaded files, encoding corresponding transcriptions into label IDs. A data collator was then developed, and the training arguments have been adjusted to aid in the training of the model. Notably, the best results were observed when the learning rate was fixed at $10^{-}4$, the number of training epochs set at 100, and the batch size established at 16. The word error rate (WER) was used as the metric to assess the model's performance during training.

4.2 Discussion

In the process of training the models, it has been observed that the fine-tuned version of the 'Vasista22/whisper-tamil-medium' model performs better than the fine-tuned version of the XLS-R-300M model when it comes to the training of the models. During the training process, it was observed that the WER for the whisper model was **71.367695**, whereas the WER for the XLS-R model was **84.6958** to begin with. When evaluating the fine-tuned model that was utilized, both the whisper and the XLS-R fine-tuned models were utilized in the process. Since the whisper model's word error rate was much smaller than the XLS-R model, we submitted the results of the whisper model to the LT-EDI team. By assessing the WER for each submission, the team will determine which outcomes are the best and rank them appropriately. The submission, which was made by our team 'CEN_Amrita',

[2] https://huggingface.co/facebook/wav2vec2-xls-r-300m

Team Name	WER (in %)
CEN_Amrita - Jairam Kanna	**24.452**
ASR_TAMIL_SSN	29.297
VIT Chennai	35.774
DRAVIDIAN LANGUAGE - Abirami Jayaraman	37.733
CUET_NLP_GoodFellows - Disco Dancer	41.031

Table 2: Speech Recognition for Vulnerable Individuals in Tamil: Published Results

employing the fine-tuned whisper model, has been ranked first among the other participants, with a WER of **24.452** for the testing dataset. Table 2 describes the published results. Figure 3 describes the submitted sample Tamil transcriptions obtained from evaluating the fine-tuned whisper model.

5 Conclusion

In this work, we utilized a pre-existing, pre-trained model such as Whisper and XLS-R to improve the efficiency of an automatic speech recognition (ASR) system for understanding conversational speech from elderly and transgender Tamil speakers. This work is carried out as part of participation in the shared task of speech recognition for vulnerable individuals in Tamil. On the Tamil conversational speech corpus, the pre-trained models, such as whisper and the XLS-R model, have been fine-tuned and compared to one another. The results of all the experiments indicate that the fine-tuned version of whisper ASR models performs better than the XLS-R model, which has a word error rate (WER) of 24.452.

References

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.

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Cn, Sripriya Natarajan, Rajeswari Natarajan, S Suhasini, and Swetha Valli. 2023. Overview of the second shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 31–37.

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Cn, N Sripriya, 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.

William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. 2015. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Dominique Fohr and Irina Illina. 2021. Bert-based semantic model for rescoring n-best speech recognition list. In *INTERSPEECH*.

Santosh K Gaikwad, Bharti W Gawali, and Pravin Yannawar. 2010. A review on speech recognition technique. *International Journal of Computer Applications*, 10(3):16–24.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Sotaro Kita. 2020. Cross-cultural variation of speech-accompanying gesture: A review. *Speech Accompanying-Gesture*, pages 145–167.

Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. 2020. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6124–6128.

CS Ayush Kumar, Advaith Maharana, Srinath Murali, B Premjith, and Soman Kp. 2022. Bert-based sequence labelling approach for dependency parsing in tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 1–8.

Jinyu Li et al. 2022. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1).

P Murali Krishna, R Pradeep Reddy, Veena Narayanan, S Lalitha, and Deepa Gupta. 2019. Affective state recognition using audio cues. *Journal of Intelligent & Fuzzy Systems*, 36(3):2147–2154.

Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. 2019. Speech recognition using deep neural networks: A systematic review. *IEEE access*, 7:19143–19165.

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.

KM Shivakumar, KG Aravind, TV Anoop, and Deepa Gupta. 2016. Kannada speech to text conversion using cmu sphinx. In *International Conference on Inventive Computation Technologies (ICICT)*, volume 3, pages 1–6.

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

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*.

Zhang Chao Zheng, Xianrui and Philip C Woodland. 2021. Adapting gpt, gpt-2 and bert language models for speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 162–168.