SANBAR@LT-EDI-2023:Automatic Speech Recognition: vulnerable old-aged and transgender people in Tamil

S Saranya & B Bharathi
Computer Science and Engineering Department
Sri Siva Subramaniya Nadar College of Engineering
Kalavakkam - 603110
saranyascse@ssn.edu.in & bharathib@ssn.edu.in

Abstract

An Automatic Speech Recognition systems for Tamil are designed to convert spoken language or speech signals into written Tamil text. Seniors go to banks, clinics and authoritative workplaces to address their regular necessities. A lot of older people are not aware of the use of the facilities available in public places or office. They need a person to help them. Likewise, transgender people are deprived of primary education because of social stigma, so speaking is the only way to help them meet their needs. In order to build speech enabled systems, spontaneous speech data is collected from seniors and transgender people who are deprived of using these facilities for their own benefit. The proposed system is developed with pretraind models are IIT Madras transformer ASR model and akashsivanandan/wav2vec2-large-xls-r-300m-tamil model. Both pretrained models are used to evaluate the test speech utterances, and obtained the WER as 37.7144% and 40.55% respectively.

1 Introduction

The earliest known Old Tamil inscriptions are found in Adichanallur and date back to the period between 905 BC and 696 BC. These inscriptions provide valuable insights into the early stages of the Tamil language.Tamil employs an agglutinative grammar system. This means that suffixes are attached to words to convey various grammatical features, such as noun class, number, case, verb tense, and other categories. Tamil's historical significance and linguistic features make it a fascinating language with a rich cultural heritage. Now a days all the people are using internet for their everyday needs. Especially old-aged and transgender who are deprived of education due to prejudice of social. So speech is only tool to do their own needs. In the present situation all people are using smartphones to do their daily needs without any direct visits to bank, hospital, etc. The integration of a voice and speech recognition system in Tamil would be highly advantageous for native Tamil users who encounter difficulties with default foreign languages on their smartphones (Kiran et al., 2017). It would elevate their user experience, facilitate accessibility, boost productivity, uphold the language's preservation, and encourage the creation of localized applications. In (Madhavaraj and Ramakrishnan, 2019), two approaches were employed to improve automatic speech recognition (ASR) for Gujarati, Tamil, and Telugu languages. The first approach involved data-pooling and phone mapping, which combined the data by mapping phones from the source languages to the target language. The second approach utilized a multi-task deep neural network (DNN) with a modified loss function to train the ASR model using the pooled data. This technique achieved relative reductions in the WERs. (Kwon et al., 2016)Because of the impact of speech articulation and speaking tendencies, older individuals tend to exhibit slower speech rates, longer pauses between syllables, and slightly reduced speech clarity. Smart devices are now not only extensively used by the younger generation but also by seniors, both indoors and outdoors throughout the day. Thus, we concluded that implementing a speech-recognition interface within a smart device could enhance its user-friendliness for the elderly and transgender. This feature would be particularly valuable for senior citizens during critical scenarios, including emergencies or situations where they might be physically restricted due to a traumatic event. Numerous seniors encounter uncertainty when attempting to utilize the provided devices meant to aid them. Similarly, transgender individuals, due to societal discrimination, are often deprived of access to primary education, leaving speech as their sole means of addressing their requirements. The information pertaining to natural speech is gathered from elderly and transgender individuals who are unable to avail themselves of

LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion, pages 155–160, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_023

such assistance.

The paper is organized as follows: the following section gives a detailed description of pretrained models used for our proposed work. Section 2 discusses the related work previously done for our current work. Section 3 detailed description of the data set used for this work. Discusses on the recognition system and the various approaches used for this work. Section 4 explains in detail the pretrained models used for the proposed work. Section 6 discusses on the results. Section 7 analysis on the performance of each pretrained model. Section 8 concludes the paper and discusses the areas of further improvement.

2 Related Work

The ASR system, which involves fine-tuning a pretrained Wav2Vec2.0 XLSR model with CTC (Connectionist Temporal Classification), has been successfully developed. This system demonstrates the ability to recognize speech samples and provide accurate transcriptions. The average Word Error Rate (WER) achieved by the system is 0.58, indicating a relatively low rate of transcription errors. Similarly, the Character Error Rate (CER) is 0.11(Akhilesh et al., 2022). In this paper, (Rojathai and Venkatesulu, 2014), have PAC features were extracted from input speech samples,the extracted features are Energy entropy, Zero crossing rate and short time energy. The extracted PAC features were trained by ANFIS system. The process of recognition performance is validated based on test words.This proposed method gave better results with different noise levels compared to previous methods. In this work (Martin et al., 2015) Show the features extraction based on English phonemes and language-independent inferred phones (IPs). The Tamil language-independent inferred phones (IPs) are achieved better performance. But it has less number of speakers, So increasing number of speakers to the model training to avoid over-fitting. (Thamburaj et al., 2021), were presented the Deep Neural Network and Membrane Bio Reactor design for Tamil. A single vowel sound were linked with Five different mono phones. The transcription was linked with LM (Language Model). So that it will decrease the impact of domain difference in AM. The target of this work is preprocess the data in order to remove noise and improve BRNN-SOM classification method scheme gains high-accuracy. SGf method was use for removing noise present in the speech samples. We achieve highest SNR values. Preprocessing technique(Lokesh et al., 2019). (S and N, 2017), show Reduce the feature vector dimension reduction using Linear Discriminant Analysis(LDA). LDA method is perform great when compared to all other well known dimensionality reduction methods like, PCA, MDS, LLE. Deep Speech architecture is used by (Changrampadi et al., 2022) and achieves 24% WER. The accuracy depends on the training language model. Tamil dialect recognition required to be regionally based spoken Tamil data to build independent Tamil recognition system. (Nivetha S, 2020), explain the speech recognition system used Random forest algorithm to identify isolated Tamil word. Both MFCC and LPC features are extracted from speech data. This algorithm gave better result and took less time for training model building. In this work, (Radha et al., 2012) show the HMM based model used to build the speaker independent isolated Tamil words recognition. Its achieved 88% accuracy and 0.88 WER. But data set size is minimal only 2500 words are used in this recognition system. In this task, (Chakravarthi, 2020) author used 20,198 Tamil comments collected from YouTube. Its includes women in STEM, LGBTIQ issues, COVID-19, India China war and affairs of Dravidian from YouTube comments. This data has two or more languages used by a single speaker. The data set is code mixed. Naive Bayes, KNN, and SVM, logistic regression used to train the model and use held-out test set to evaluate the trained model. Evaluate the trained model. The result is shown using precision, recall, F1-score. Findings of the automatic speech recognition for vulnerable individuals are given in (Bharathi et al., 2022). (S and B, 2022) (B et al., 2022), have pretrained Rajaram1996/wav2vec-large-xlsr-53-tamil transformer model used for transformer based ASR for Vulnerable Individuals in Tamil. In this pretained model gave 39.65% WER.

3 Data Set Description

Tamil speech utterances are collected from the old-aged people and transgender whose mother tongue is Tamil. The recorded speech utterances of old-aged people and transgender contains how those people communicate in primary locations like bank, hospitals and administrative office.The data set contains 51 Speakers of literates, illiterates elders and transgenders. People who have their primary education till sixth grade are considered literates while collecting data. The duration of corpus is 7 hours and 30 minutes. It is ensured that no audio recorded from an individual is less than 5 minutes. No interruption or overlap of other person voices in the audio other than the speaker's audio. The speech files in the directories are in the WAV format. The sampling rate of the speech utterances are 44kHz. The speech corpus with 5.5 hours of transcribed speech will be released for the training, and 2 hours of speech data will be released for testing. Table 1. shows that detailed description about the collected speech utterances. More information about data collection is explained in (B et al., 2023).

Speakers	Literate	Illiterate	Total
Male	4	9	13
Female	7	24	31
Transgender	3	4	7

Table 1: Detailed Description of speech corpus

4 Proposed Work

Our proposed work, transformer model of IIT Madras and transformer model akashsivanandan/wav2vec2-large-xls-r-300m-tamil-colab-final are used. The ASR model for Indian languages described in the transformer ASR model is created by IIT Madras follows a espnet.nets.pytorch_backend.e2e_asr_transformer, E2Eself-attention mechanism architecture. The following stages are performed in ESPnet transformer architecture.

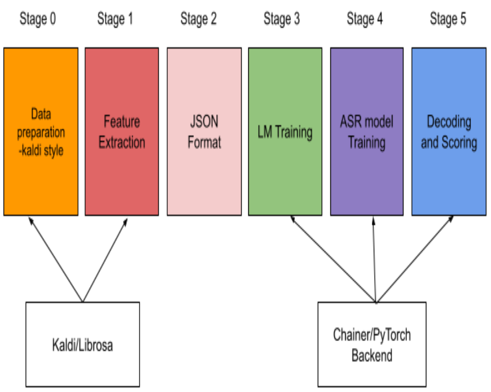


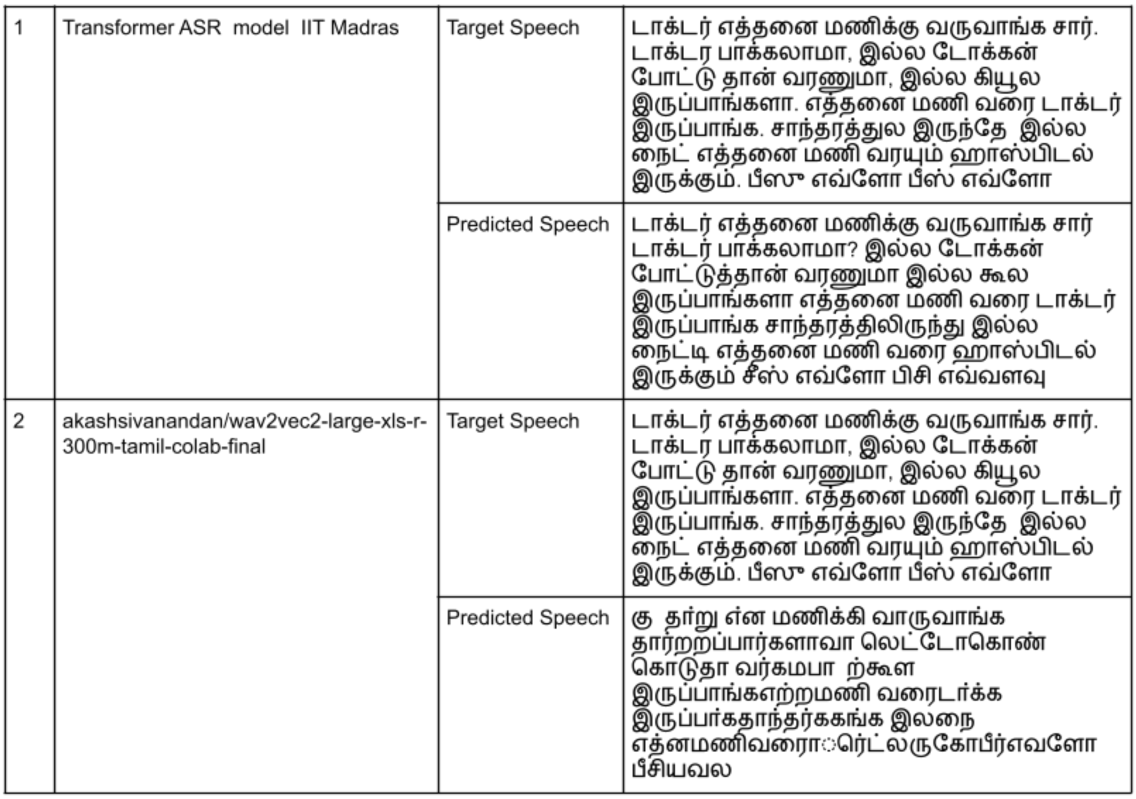
Figure 1: Architecture of ESPnet Transformer Model

- Stage 0: Preprocessing the speech data including transcription and dictionary creation.
- Stage 1: Extract the features from the speech data using Kaldi toolkit.
- Stage 2: Create a JSON configuration file to specify the details of the data preparation. This file should include paths to of the audio and transcript files, as well as other parameters like sampling rate, the language of the text, etc, are dumped.
- Stage 3: Training the language model.
- Stage 4: Train the acoustic model using the preprocessed data. Include the details such as the number of epochs, batch size, learning rate, etc.
- Stage 5: Evaluate the trained model on the test speech utterances.

The transformer model akashsivanandan/wav2vec2-large-xls-r-300m-tamil model is a variant of the Wav2Vec2 architecture that has been fine-tuned specifically for the Tamil language. The model has been trained in an unsupervised learning manner, it means the model doesn't need explicit transcriptions. The model was pre-trained using speech signal from multiple sources including Babel, Multilingual LibriSpeech (MLS), Common Voice, VoxPopuli, and VoxLingua107. The sampling rate of The original Common Voice dataset is 48kHz. However, training the XLSR model, To ensure compatibility with the XLSR model's 16kHz sampling rate, the Common Voice data was resampled and adjusted from its original 48kHz sampling rate to the desired 16kHz sampling rate. This downsampling process was implemented to align the data with the model's requirements.

5 Implementation

The one of the pretrained model used in our proposed work is "akashsivanandan/wav2vec2-large-xls-r-300m-tamil". This is model is fine tuned the XLSR model using the common voice Tamil speech corpus. The other model used in the proposed system is "IIT Madras transformer ASR model". The IIT Madras transformer ASR model is created using ESPnet transformer model:E2Eself-attention mechanism employs Language model



1	Transformer ASR model IIT Madras	Target Speech	டாக்டர் எத்தனை மணிக்கு வருவாங்க சார். டாக்டர பாக்கலாமா, இல்ல டோக்கன் போட்டு தான் வரணுமா, இல்ல கியூல இருப்பாங்களா. எத்தனை மணி வரை டாக்டர் இருப்பாங்க. சாந்தரத்துல இருந்தே இல்ல நைட் எத்தனை மணி வரயும் ஹாஸ்பிடல் இருக்கும். பீஸூ எவ்ளோ பீஸ் எவ்ளோ
		Predicted Speech	டாக்டர் எத்தனை மணிக்கு வருவாங்க சார் டாக்டர் பாக்கலாமா? இல்ல டோக்கன் போட்டுத்தான் வரணுமா இல்ல கூல இருப்பாங்களா எத்தனை மணி வரை டாக்டர் இருப்பாங்க சாந்தரத்திலிருந்து இல்ல நைட்டி எத்தனை மணி வரை ஹாஸ்பிடல் இருக்கும் சீஸ் எவ்ளோ பிசி எவ்வளவு
2	akashsivanandan/wav2vec2-large-xls-r-300m-tamil-colab-final	Target Speech	டாக்டர் எத்தனை மணிக்கு வருவாங்க சார். டாக்டர பாக்கலாமா, இல்ல டோக்கன் போட்டு தான் வரணுமா, இல்ல கியூல இருப்பாங்களா. எத்தனை மணி வரை டாக்டர் இருப்பாங்க. சாந்தரத்துல இருந்தே இல்ல நைட் எத்தனை மணி வரயும் ஹாஸ்பிடல் இருக்கும். பீஸூ எவ்ளோ பீஸ் எவ்ளோ
		Predicted Speech	கு தா்று ஏ்ன மணிக்கி வாருவாங்க தார்றறப்பார்களாவா லெட்டோகொண் கொடுதா வர்கமபா ற்கூள இருப்பாங்களற்றமணி வரைடர்க்க இருப்பர்கதாந்தர்ககங்க இலநை எத்னமணிவரைரா◌ெர்ட்லருகோபீர்எவளோ பீசியவல

Figure 2: Sample recognised text using proposed system

has a 12-layer encoder and a variable decoder. Each encoder layer incorporates self-attention and 2048 units of a feed-forward neural network and ReLU activation. The decoder comprises six layers for Indian languages, each featuring self-attention and a 2048-unit feed-forward network. For Voxforge, there is a single layer with a 1024-unit feed-forwaProceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusionrd network. 104.5 hours of Tamil Speech data used for training, and unified processor was used to generate lexicon.

The model akashsivanandan/wav2vec2-large-xls-r-300m-tamil pretrained model used for testing the spoken Tamil speech data. This pretrained model is tuned by facebook/wav2vec2-large-xlsr-r-300m-tamil using common voice Tamil data. Wav2Vec2 uses a combination of convolutional neural networks (CNNs) and self-attention mechanisms to learn speech representations from raw audio waveform. By CNNs, it can extract local acoustic features, while self-attention mechanisms enable it to capture long-range dependencies in the audio data. It can cognizant and process Tamil speech data efficacious. The model take advantage of a fusion of CNN and self-attention mechanisms to learn powerful representations from raw audio waveform. It has a larger capacity, allowing it to capture complex patterns in the audio data. In this model learning rate is 0.0003, training batch size and testing batch size are 16 and 8. optimizer: Adam with betas=(0.9,0.999) and epsilon=1e-08 are used for training. The speech corpus has 239 speech files for testing. The test speech data is given as input to the both the pretrained system. Once the speech recognition is completed the output text transcription stored in a individual file. Finally WER is calculated based on what transcription got from recognition and target transcription of the speech data which is used for testing.

6 Results

The word error rate (WER) is calculated using the following equation,

$$WER = \left(\frac{S + I + D}{N}\right) \tag{1}$$

where as,

- S is the number of substitutions

- D is the number of deletions
- I is the number of insertions
- N is the number of words in the reference transcriptions

S.no	Model	WER
1	Transformer model of IITM	37.71
2	akashsivanandan/wav2vec2-large-xls-r-300m-tamil-colab-final	40.55

Table 2: Performance of the proposed system using test utterances

7 Discussion

From Table 2, both the pretrained model results are shown. The IIT Madras transformer ASR model WER is 37.71%. akashsivanandan/wav2vec2-large-xls-r-300m-tamil-colab-final model WER is 40.55%. IIT Madras transformer ASR model gave a better result compared to another one. akashsivanandan/wav2vec2-large-xls-r-300m-tamil-colab-final does not have language model. However IIT Madras transformer ASR has a language model (LM). LM is used to boost recognition system should help the acoustic model. For example ”Enkengu” word is correctly predicted by a model which has LM, in other hand ”Enkengu” word predicted as ”Enkanku”. It shows that LM and region-based spoken Tamil data are used to develop a better recognition ASR system.

8 Conclusions

An automatic speech recognition system is developed with a pretrained fine tune models which is available publicly. The speech data is collected from old-aged people and transgender whose mother tongue is Tamil. The speech data contains how the people access primary location in day to day life. Evaluate the test speech samples using pretrained models and calculated WERs. Going forward, increase speech data and create our own training model with more region-based Tamil speech data. This will enhance the performance of the proposed system.

References

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 Chinnaudayar Navaneethakrishnan, Sripirya N, Rajeswari Natarajan, Suhasini S, 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 Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Bharathi B, Dhanya Srinivasan, Josephine Varsha, Thenmozhi Durairaj, and Senthil Kumar B. 2022. SSNCSE_NLP@LT-EDI-ACL2022:hope speech detection for equality, diversity and inclusion using sentence transformers. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 218–222, Dublin, Ireland. Association for Computational Linguistics.

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, 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*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

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).

R Kiran, K Nivedha, T Subha, et al. 2017. Voice and speech recognition in tamil language. In *2017 2nd International Conference on Computing and Communications Technologies (ICCCT)*, pages 288–292. IEEE.

Soonil Kwon, Sung-Jae Kim, and Joon Yeon Choeh. 2016. Preprocessing for elderly speech recognition of smart devices. *Computer Speech Language*, 36:110–121.

S Lokesh, Priyan Malarvizhi Kumar, M Ramya Devi, P Parthasarathy, and C Gokulnath. 2019. An automatic tamil speech recognition system by using bidirectional recurrent neural network with self-organizing map. *Neural Computing and Applications*, 31:1521–1531.

A Madhavaraj and AG Ramakrishnan. 2019. Data-pooling and multi-task learning for enhanced performance of speech recognition systems in multiple low resourced languages. In *2019 National Conference on Communications (NCC)*, pages 1–5. IEEE.

Lara J Martin, Andrew Wilkinson, Sai Sumanth Miryala, Vivian Robison, and Alan W Black. 2015. Utterance classification in speech-to-speech translation for zero-resource languages in the hospital administration domain. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 303–309. IEEE.

Gayathri S Nivetha S, Rathinavelu A. 2020. Speech recognition system for isolated tamil words using random forest algorithm. *International Journal of Recent Technology and Engineering (IJRTE)*, 9:2431–2435.

V Radha et al. 2012. Speaker independent isolated speech recognition system for tamil language using hmm. *Procedia Engineering*, 30:1097–1102.

S Rojathai and M Venkatesulu. 2014. Noise robust tamil speech word recognition system by means of pac features with anfis. In *2014 IEEE/ACIS 13th International Conference on Computer and Information Science (ICIS)*, pages 435–440. IEEE.

Suhasini S and Bharathi B. 2022. SUH_ASR@LT-EDI-ACL2022: Transformer based approach for speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 177–182, Dublin, Ireland. Association for Computational Linguistics.

Sundarapandiyan S and Shanthi N. 2017. Automatic speech recognition system of tamil language using linear discriminant analysis. *International Journal of Recent Technology and Engineering (IJRTE)*, 6:6298–6301.

Kingston Pal Thamburaj et al. 2021. A process of developing an asr system for malay and tamil languages. *Design Engineering*, pages 731–741.