SUH_ASR@LT-EDI-ACL2022: Transformer based Approach for Speech Recognition for Vulnerable Individuals in Tamil

S. Suhasini & B. Bharathi
Department of CSE
Sri Siva Subramaniya Nadar College of Engineering
Kalavakkam - 603110
suhasinis@ssn.edu.in
bharathib@ssn.edu.in

Abstract

An Automatic Speech Recognition System is developed for addressing the Tamil conversational speech data of the elderly people and transgender. The speech corpus used in this system is collected from the people who adhere their communication in Tamil at some primary places like bank, hospital, vegetable markets. Our ASR system is designed with pre-trained model which is used to recognize the speech data. WER(Word Error Rate) calculation is used to analyse the performance of the ASR system. This evaluation could help to make a comparison of utterances between the elderly people and others. Similarly, the comparison between the transgender and other people is also done. Our proposed ASR system achieves the word error rate as 39.65%.

Keywords: Automatic Speech Recognition, Word Error Rate, Tamil speech corpus, Transformer model, Pre-trained model.

1 Introduction

In the recent days, most of the people have started using the internet through various electronic devices(Vacher et al., 2015). In such a case, the elderly people have also started using the internet through smart phones. As some of the elderly people were not educated much about the technology, they try to retrieve the information from internet using their audio message. To handle such kind of audio messages of elderly people, an acoustic model has to be designed, the model will recognize the utterance of the elderly people and extracts the output of the speech data(Fukuda et al., 2019)(Hämäläinen et al., 2015). Therefore, the output will be a text file. Based on the output of the speech, WER value will be calculated. The WER value shows the accuracy of the prediction by the model. It is identified that Automatic Speech Recognition using some standard models have not achieved a good performance(Nakajima and Aono, 2020) and also no other corpus for elderly people is larger than the Japanese Newspaper Article Sentences (JNAS), Japanese Newspaper Article Sentences Read Speech Corpus of the Aged (S-JNAS) and Corpus of Spontaneous Japanese (CSJ) corpora(Fukuda et al., 2020).

The earliest Old Tamil documents are small inscriptions in Adichanallur dating from 905 BC to 696 BC. Tamil has the oldest ancient non-Sanskritic Indian literature of any Indian language. Tamil uses agglutinative grammar, which uses suffixes to indicate noun class, number, case, verb tense, and other grammatical categories (Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021; Chakravarthi et al., 2020). Tamil's standard metalinguistic terminology and scholarly vocabulary is itself Tamil, as opposed to the Sanskrit that is standard for most Aryan languages. Tamil has many forms, in addition to dialects: a classical literary style based on the ancient language (cankattami), a modern literary and formal style (centami), and a current colloquial form (kotuntami) (Sakuntharaj and Mahesan, 2021, 2017, 2016; Thavareesan and Mahesan, 2019, 2020a,b, 2021). These styles blend into one another, creating a stylistic continuity. It is conceivable, for example, to write centami using cankattami vocabulary, or to utilize forms connected with one of the other varieties while speaking kotuntami (Subalalitha, 2019; Srinivasan and Subalalitha, 2019; Narasimhan et al., 2018). Tamil words are made up of a lexical root and one or more affixes. The majority of Tamil affixes are suffixes. Tamil suffixes are either derivational suffixes, which modify the part of speech or meaning of the word, or inflectional suffixes, which designate categories like as person, number, mood, tense, and so on. There is no ultimate limit to the length and scope of agglutination, which might result in large words with several suffixes, requiring many words or a sentence in English (Anita and Subalalitha, 2019b,a; Subalalitha and Poovammal, 2018).

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 177 - 182

Likewise, the speech corpus of different languages are addressed by many people, those corpus contains either the male or the female speech and no corpus addressed the transgender speech. But, the speech corpus released by the shared task(LT-EDI-ACL2022)contains male, female and transgender utterances, which enhance the characteristic of the acoustic model, but the number of speech utterances collected are very less compared to other elderly speech corpus. The model also need to handle challenges faced in the corpus, as this shared task speech corpus contains conversational speech data in primary locations like bank, market, hospital and public transport. As the people may have their own accent and pronunciation for conversational speech in the primary places, it is difficult to recognize the speech and the model used for recognizing standard speech cannot be used for the conversational speech corpus because it increases the WER. To address this kind of conversational speech of the elderly people a transformer model approach is used. The follow of the paper is as follows: The review of the related work is discussed in section 2, Data-set description is described in section 3, Methodology used is discussed in section 4, followed by Implementation, Observations and Discussion are described in section 5, 6 and 7 respectively. Finally, the paper is concluded with future work in section 8.

2 Related Work

Many studies have done on recognizing the elderly people speech corpus, using adaptation acoustic model for CSJ corpus which results in lowest WER values(Fukuda et al., 2020). The prosodic and spectral features are extracted for elderly people speech(Lin and Yu, 2015) and the performance of the continuous word recognition and phoneme recognition is measured from the two different age groups and the corpus is collected in Bengali language(Das et al., 2011). Additional feature analysis can also be done like loudness of the speech, sampling rate, fundamental frequency, and segmentation of the sentence. Other measures were done by identifying the pause in the sentence and measuring the duration of the pause(Nakajima and Aono, 2020). Insufficient performance is measured with low number of utterance(Fukuda et al., 2020). Increase in WER value happens if the quality of recorded speech is low(Iribe et al., 2015). E2E ASR transformer can do encoding and decoding hierarchically by combining the transformers for large context(Masumura et al., 2021). Using the Hybrid based LSTM transformer, the WER is reduced with 25.4% by transfer learning. Additionally, 13% WER is reduced by LSTM decoder(Zeng et al., 2021). Transformer model encoding and decoding can be carried with self- attention and multi-head attention layer(Lee et al., 2021). For CTC/Attention based End-To-End ASR, the transformer model is used, which result 23.66% of WER(Miao et al., 2020). End-to-End ASR system works based on transformers for streaming ASR, where an output must be generated soon after each spoken word. For the encoder, the time-restricted self-attention is used, and for the encoder-decoder attention mechanism, prompted attention is used. On the Wall Street Journal task, the novel fusion attention technique delivers a WER decrease of 16.7% compared to the non-fusion standard transformer and 12.1% compared to other authors transformer-based benchmarks.

3 Data-set Description

Tamil conversational speech data is collected from the elderly people. The speech corpus contains a total of 6 hours and 42 minutes of speech data. The recorded speech of elderly people contains how the elderly people communicate in primary locations like market, bank, shop, public transport and hospitals. It includes both male and female utterances and also this speech data is collected from the transgender people. Table 1. contains the detailed description about the collected data.

Gender	Avg-Age	Duration(mins)
Male	61	93
Female	59	242
Transgender	30	67

Table 1: Data-set Details

4 Proposed Work

In the proposed methodology, transformer model Rajaram1996/wav2vec-large-xlsr-53-tamil transformer model [1] is used. The initial part of XLSR contains a stack of CNN layers that are used to extract acoustically important features - but it is context independent - features from the raw speech

[1] https://huggingface.co/Rajaram1996/wav2vec2-large-xlsr-53-tamil

1	Target Sentence	உனக்கு என்ன கலர் பிடிச்சிருக்கு நீல கலர் ரொம்ப நல்லா இருக்குல்ல நீங்கள் உங்கள் நிறுவனத்தின் நேரடி விற்பனையாளர் உங்களுக்கு எங்கள் கிளைகள் உள்ளன வாகனத்தின்
	Predicted Sentence	உனக்கு என்னக்கலர்ப்பிடுச்சிருக்கு நீல்கல ரொம்ப நல்லவருக்குல்லாநீங்கள் உங்கள் நெருவனத்து நேரடு விற்டனையளரா உங்களுக்கு எங்கங்கு கிலைவல் உள்ளனவாகனத்தின
2	Target Sentence	அதுக்கு இன் பெட்வீன் கேப் எவ்ளோ இருக்கனும் முன்னாடியே சொல்லிடுவீங்களா ஏத்தாது டயட் லாம் போலோவ் பண்ணனுமா வஞ்சர மீன் இருக்குதா என்ன விலை வஞ்சர பீஸ் பனி தருவிங்கள நாங்க கேக்குற வெயிட்டுக்கு தருவிங்கள சுறா என விலை
	Predicted Sentence	அதுகு இன்டுற்றின் கேட்பருளங்களும் பிழ ற்கிணங்கல்லும உனுக்கிமாச னீரறீங்கா வக அதிகததில் பேற்லான் பேட சாலாதமிழமாகனனாடம் இல்லார்கே வஞ்தரர்குதா வஞ்சனைக்குல என்னகலாம உ அதிர பீப்பனி தவ்வீங்களா அத நங்க கேக்கிட ஏட்டிடுத் தரியீங்களாறாவணவள அற்கில பரிவீங்களா

Figure 1: Sample Prediction

signal. A pre-trained XLSR model maps the speech signal to a series of context representations. However, model has to recognise speech from the given dataset, it must translate this series of context representations to their corresponding transcription, which necessitates the addition of a linear layer on top of the transformer block. At a sampling rate of 16kHz, the XLSR model was pretrained using audio data from Babel, Multilingual LibriSpeech (MLS), Common Voice, VoxPopuli, and VoxLingua107. Because Common Voice has a sampling rate of 48kHz in its original form. Later, it was downsampled by fine-tuning the data to 16kHz. The parameter required to instantiate Wav2Vec2FeatureExtractor are feature_size, sampling_rate, padding_value, do_normalize and return_attention_mask. The below Figure 1. shows the sample prediction for the given corpus.

S.No.	Gender	Count	Avg WER
1	Male	9	43.8283176
2	Female	27	41.69810455

Table 2: Average WER Value for Training Data

S.No.	Gender	Count	Avg WER
1	Male	2	31.27275584
2	Female	1	43.95625294
3	Transgender	7	40.23148537

Table 3: Average WER Value for Test Data

5 Implementation

To develop an effective acoustic model using a transformer based pre-trained model. There are various transformer based pre-trained model available publicly. Here, the "Rajaram1996/wav2vec2-large-xlsr-53-tamil" pretrained model for handling Tamil speech corpus is used. This pretrained model is fine-tuned from "facebook/wav2vec2-large-xlsr-53" [2] by common voice dataset in Tamil. The model accepts the input only if the speech data is sampled at 16KHZ and it does not depend on any language model, instead it can be used directly. The XSLR is used in the model for building the wav2vec and also it experiments the cross- lingual speech data. XLSR is capable of learning the quantization of latents which is shared across languages. The speech utterance is loaded in the librosa, then it is stored in a variable and it will be tokenized using the tokenizer, which converts the audio to text and the outputs are the transcripts of the audio file which is loaded in the librosa. Once the speech recognition is done, the transcripts are stored in a separate folder. The WER(Word Error Rate) is calculated between the transcripts generated by the model and the original transcripts of the audio created by the human. Based on the WER value, the level of recognition of speech can be measured. Our speech corpus contains total of 46 audio files where it is subdivided into 1147 audio files. From 46 audio files, 36 audio files were given for training with 908 subsets and 10 audio files for testing with

[2] https://huggingface.co/facebook/wav2vec2-large-xlsr-53

S.No.	File Name	Subsets	WER Value
1	Audio-1	30	43.33907333
2	Audio-2	26	45.05577308
3	Audio-3	32	36.69085938
4	Audio-4	23	42.27066957
5	Audio-5	42	37.81365952
6	Audio-6	25	40.862304
7	Audio-7	24	47.62186667
8	Audio-8	46	35.79983696
9	Audio-9	10	35.68962
10	Audio-10	38	38.76901053
11	Audio-11	49	42.83066735
12	Audio-12	17	47.48973529
13	Audio-13	33	38.63954545
14	Audio-14	25	36.521964
15	Audio-15	2	53.0503
16	Audio-16	16	41.0687875
17	Audio-17	35	38.18157143
18	Audio-18	16	42.4245875
19	Audio-19	24	39.72085417
20	Audio-20	27	37.19958519
21	Audio-21	38	41.47868947
22	Audio-22	35	39.07802286
23	Audio-23	37	56.72666757
24	Audio-24	11	46.33136364
25	Audio-25	9	50.21177778
26	Audio-26	22	47.08117273
27	Audio-27	16	40.204475
28	Audio-28	23	45.89045217
29	Audio-29	47	50.12873617
30	Audio-30	25	36.606272
31	Audio-31	25	40.567656
32	Audio-32	16	38.5304625
33	Audio-33	16	40.3838125
34	Audio-34	16	46.8358
35	Audio-35	16	37.12355
36	Audio-36	16	42.0845

Table 4: WER values for Training Set

239 subsets. The WER value for each audio file is calculated.

6 Observations

The result contains the name of the speech data with its WER value. Similarly, for all the audio files the same process is carried out. The table also includes the details about the number of subsets that each audio file is divided into. In Table 2, the average WER value of training set audio files is calculated which holds male and female utterances. In Table 3, the average WER value of test set audio files is calculated which includes male, female and transgender utterances.

6.1 Training Results

6.2 Testing Results

S.No.	File Name	Subsets	WER Value
1	Audio-37	15	30.13258667
2	Audio-38	17	43.95625294
3	Audio-39	16	32.412925
4	Audio-40	17	37.89848235
5	Audio-41	19	42.65715789
6	Audio-42	24	43.11616667
7	Audio-43	30	37.94115667
8	Audio-44	28	36.29702143
9	Audio-45	26	44.35576154
10	Audio-46	47	39.35465106

Table 5: WER values for Testing Set

7 Discussion

From the Table 4, the experimental result says that the average WER(Word Error Rate) for the training dataset(908 audio files) is 42.23%. Similarly, Table 5, says the result of total 239 audio subset files from 10 audio files given for testing and the WER measured is 39.65%.

8 Conclusion

In order to improve the speech recognition system for recognizing the elderly people conversational speech data. An automatic speech recognition system is designed with a pre-trained model. A dataset is collected from the elderly people and transgender whose native language is Tamil. The utterance of the dataset is a Tamil language and recorded during a conversation in primary locations. As the pre-trained model used for the system is fine-tuned with common voice dataset, in future the model can trained with our own dateset and it can be used for testing, which can increase the performance.

References

R Anita and CN Subalalitha. 2019a. An approach to cluster Tamil literatures using discourse connectives. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–4. IEEE.

R Anita and CN Subalalitha. 2019b. Building discourse parser for Thirukkural. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 18–25.

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.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 61–72, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020. Overview of the track on sentiment analysis for Dravidian languages in code-mixed text. In *Forum for Information Retrieval Evaluation*, pages 21–24.

Biswajit Das, Sandipan Mandal, and Pabitra Mitra. 2011. Bengali speech corpus for continuous auutomatic speech recognition system. In *2011 International conference on speech database and assessments (Oriental COCOSDA)*, pages 51–55. IEEE.

Meiko Fukuda, Ryota Nishimura, Hiromitsu Nishizaki, Yurie Iribe, and Norihide Kitaoka. 2019. A new corpus of elderly japanese speech for acoustic modeling, and a preliminary investigation of dialect-dependent speech recognition. In *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Meiko Fukuda, Hiromitsu Nishizaki, Yurie Iribe, Ryota Nishimura, and Norihide Kitaoka. 2020. Improving speech recognition for the elderly: A new corpus of elderly japanese speech and investigation of acoustic modeling for speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6578–6585.

Annika Hämäläinen, António Teixeira, Nuno Almeida, Hugo Meinedo, Tibor Fegyó, and Miguel Sales Dias. 2015. Multilingual speech recognition for the elderly: The aalfred personal life assistant. *Procedia Computer Science*, 67:283–292.

Yurie Iribe, Norihide Kitaoka, and Shuhei Segawa. 2015. Development of new speech corpus for elderly japanese speech recognition. In *2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 27–31. IEEE.

Taewoo Lee, Min-Joong Lee, Tae Gyoon Kang, Seokyeoung Jung, Minseok Kwon, Yeona Hong, Jungin Lee, Kyoung-Gu Woo, Ho-Gyeong Kim, Jiseung Jeong, et al. 2021. Adaptable multi-domain language model for transformer asr. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7358–7362. IEEE.

Hui Lin and Yibiao Yu. 2015. Acoustic feature analysis and conversion of age speech. In *IET Conference Proceedings*. The Institution of Engineering & Technology.

Ryo Masumura, Naoki Makishima, Mana Ihori, Akihiko Takashima, Tomohiro Tanaka, and Shota Orihashi. 2021. Hierarchical transformer-based large-context end-to-end asr with large-context knowledge distillation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5879–5883. IEEE.

Haoran Miao, Gaofeng Cheng, Changfeng Gao, Pengyuan Zhang, and Yonghong Yan. 2020. Transformer-based online ctc/attention end-to-end speech recognition architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6084–6088. IEEE.

Hideharu Nakajima and Yushi Aono. 2020. Collection and analyses of exemplary speech data to establish easy-to-understand speech synthesis for japanese elderly adults. In *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 145–150. IEEE.

Anitha Narasimhan, Aarthy Anandan, Madhan Karky, and CN Subalalitha. 2018. Porul: Option generation and selection and scoring algorithms for a tamil flash card game. *International Journal of Cognitive and Language Sciences*, 12(2):225–228.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2016. A novel hybrid approach to detect and correct spelling in Tamil text. In *2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 1–6.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2017. Use of a novel hash-table for speeding-up suggestions for misspelt Tamil words. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pages 1–5.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2021. Missing word detection and correction based on context of Tamil sentences using n-grams. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 42–47.

R Srinivasan and CN Subalalitha. 2019. Automated named entity recognition from tamil documents. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–5. IEEE.

C. N. Subalalitha. 2019. Information extraction framework for Kurunthogai. *Sādhanā*, 44(7):156.

CN Subalalitha and E Poovammal. 2018. Automatic bilingual dictionary construction for Tirukural. *Applied Artificial Intelligence*, 32(6):558–567.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment analysis in Tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment lexicon expansion using Word2vec and fastText for sentiment prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Word embedding-based part of speech tagging in Tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 478–482.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2021. Sentiment analysis in Tamil texts using k-means and k-nearest neighbour. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 48–53.

Michel Vacher, Frédéric Aman, Solange Rossato, and François Portet. 2015. Development of automatic speech recognition techniques for elderly home support: Applications and challenges. In *International Conference on Human Aspects of IT for the Aged Population*, pages 341–353. Springer.

Zhiping Zeng, Haihua Xu, Yerbolat Khassanov, Eng Siong Chng, Chongjia Ni, Bin Ma, et al. 2021. Leveraging text data using hybrid transformer-lstm based end-to-end asr in transfer learning. In *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE.