ASR_TAMIL_SSN@ LT-EDI-2024: Automatic Speech Recognition system for Elderly People

S. Suhasini
Department of CSE
R. M. D. Engineering College
ssi.cse@rmd.ac.in

B. Bharathi
Department of CSE
Sri Sivasubramaniya Nadar College of Engineering
bharathib@ssn.edu.in

Abstract

The results of the Shared Task on Speech Recognition for Vulnerable Individuals in Tamil (LT-EDI-2024) are discussed in this paper. The goal is to create an automated system for Tamil voice recognition. The older population that speaks Tamil is the source of the dataset used in this task. The proposed ASR system is designed with pre-trained model akashsivanandan/wav2vec2-large-xls-r-300m-tamil-colab-final. The Tamil common speech dataset is utilized to fine-tune the pre-trained model that powers our system. The suggested system receives the test data that was released from the task; transcriptions are then created for the test samples and delivered to the task. Word Error Rate (WER) is the evaluation statistic used to assess the provided result based on the task. Our Proposed system attained a WER of 29.297%.

1 Introduction

This shared task tackles a difficult area in Tamil automatic speech recognition system for vulnerable elderly and transgender individuals. To take care of their daily necessities, elderly people go to important places including banks, hospitals, and administrative offices. Many elderly folks are not aware of how to use the tools provided to help people. Similar to how transgender persons lack access to basic schooling due to societal discrimination, speech is the only channel that can help them meet their demands. The data on spontaneous speech is collected from elderly and transgender people who are unable to take benefit of these amenities (Fukuda et al., 2019; Hämäläinen et al., 2015). 2 hours of speech data will be made available for testing, while the speech corpus containing 5.5 hours of transcribed speech will be made available for the training set. Recently, the majority of people have begun using various electronic devices to access the internet. In this situation, the elderly people have also started using smart phones to access the internet (Vacher et al., 2015). Some elderly people attempt to acquire information from the internet using their audio message because they are not well-versed in technology. An acoustic model must be created to handle these types of audio messages from elderly individuals; the model will identify their speech and extract the output of the speech data. As a result, a text file will be the output. The speech's output will be used to determine the WER value. The WER number demonstrates how accurately the model predicted the outcome . 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). It has been determined that Automatic Speech Recognition using some standard models has not achieved a good performance (Nakajima and Aono, 2020). It is challenging to identify conversational speech in public settings since each person may have their own accent and pronunciation. Additionally, the methodology for identifying standard speech cannot be applied to the conversational speech corpus because it raises WER. A transformer model technique is utilised to treat this type of older people's conversational speech. The paper is organised as follows: Section 2 discusses the examination of related literature, Section 3 describes the data set, Section 4 discusses the methodology, Section 5 describes the implementation, Section 6 describes the observations, and Section 7 discusses the discussion. Section 8 of the essay concludes with a section on future research.

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 294–298
March 21, 2024

2 Related Work

There have been numerous studies on recognising the speech of elderly persons using the adaption acoustic model for CSJ corpus (Fukuda et al., 2020), which yields the lowest WER values. The performance of continuous word identification and phoneme recognition is measured from the two distinct age groups, and the corpus is collected in Bengali (Das et al., 2011). Prosodic and spectral properties are retrieved for senior people speech.The exploration of additional features (Lin and Yu, 2015), such as the speech's volume level, sampling frequency, fundamental frequency, and sentence segmentation, is also possible. Other measurements were locating the pause in the sentence and calculating how long it lasted (Nakajima and Aono, 2020). Low number of utterances is a sign of inadequate performance. If the recorded voice quality is poor (Iribe et al., 2015), the WER value rises. By integrating the transformers for a broad context (Masumura et al., 2021), the E2E ASR transformer can perform encoding and decoding in a hierarchical manner. The WER is decreased by 25.4% via transfer learning when using the hybrid-based LSTM transformer (Zeng et al., 2021). Additionally, the LSTM decoder lowers WER by 13%. Self-attention and a multi-head attention layer (Lee et al., 2021) can be used to carry out the encoding and decoding of transformer models. The transformer model is utilised for CTC/Attention based End-To-End ASR, and it produces a WER of 23.66% (Miao et al., 2020). Transformers for streaming ASR are the foundation of the end-to-end ASR system, where an output must be produced quickly after each spoken word. Time-restricted self-attention is employed for the encoder, and prompted attention is used for the encoder-decoder attention mechanism. The innovative fusion attention technique results in a WER reduction of 16.7% on the Wall Street Journal test compared to the non-fusion standard transformer and 12.1% compared to other authors' transformer-based benchmarks. Findings of the automatic speech recognition for vulnerable individuals are given in (S and B, 2022) (B et al., 2022), have used transformer models used for transformer based ASR for Vulnerable Individuals in Tamil.

3 Dataset 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 (Bharathi et al., 2022). 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: Dataset Details

The below Figure 1. shows the sample prediction for the given corpus.

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 கேட்பருளஞ்களும் பிழ ற்கிணங்கல்லும உனுக்கிமாச வீரறீங்கா வக அதிகததில் பேற்லான் பேட சாலாதமிழமர்கனனாடம் இல்லார்கே வஞ்தரர்குதா வஞ்சனைக்குல என்னகலாம உ அதிர பீப்பனி தவ்வீங்களா அத நங்க கேக்கிட ஏட்டிடுத் தரியீங்களாறாவணவள அற்கில பரிவீங்களா

Figure 1: Sample Prediction

4 Proposed Work

In our proposed system, the pretrained transformer model akashsivanandan/wav2vec2-large-xls-r-300m-tamil-colab-final[1] is used. The pretrained model "https://huggingface.co/akashsivanandan/wav2vec2-large-xls-r-300m-tamil-colab-final" is based on the Wav2Vec2 architecture and specifically trained for the Tamil language. Wav2Vec2[2] is a state-of-the-art speech recognition model developed by Facebook AI. It utilizes a self-supervised learning approach, where it learns from large amounts of unlabeled speech data to build representations that capture meaningful information about the audio. The model is based on the transformer architecture, which has proven to be highly effective for various natural language processing tasks. Transformers

[1] https://huggingface.co/akashsivanandan/wav2vec2-large-xls-r-300m-tamil-colab-final
[2] https://huggingface.co/docs/transformers/model_doc/wav2vec2

enable the model to efficiently capture long-range dependencies in the audio input. The model is pretrained on a large corpus of multilingual and monolingual data containing Tamil speech. During pretraining, it learns to predict masked or distorted portions of the input audio, which helps it understand the underlying structure and features of the speech data. After pretraining, the model undergoes fine-tuning using labeled data for specific downstream tasks. Fine-tuning allows the model to adapt to a particular speech recognition task, such as transcription or keyword spotting, in this case, for Tamil language. Although the model is specifically trained for Tamil, it benefits from the multilingual nature of its pretraining data. It can understand and process speech from various languages, making it useful for multilingual applications or tasks involving code-switching. The model has been trained on a vast vocabulary, enabling it to handle a wide range of words and phrases. This makes it suitable for tasks that involve transcribing or recognizing speech with diverse vocabulary. The model's training data and fine-tuning procedure focus on capturing the unique characteristics of the Tamil language, including its phonetics, phonology, and syntax. This enhances its ability to accurately recognize and transcribe Tamil speech.

5 Implementation

Efficient acoustic model can be created based on a pre-trained transformer model as there are many publicly accessible transformer-based pre-trained models. Here, the ”https://huggingface.co/akashsivanandan/wav2vec2-large-xls-r-300m-tamil-colab-final” pretrained model for handling Tamil speech corpus is used. This pretrained model is fine-tuned from ”facebook/wav2vec2-large-xlsr-53” [3] by common voice dataset in Tamil. The model can be used directly and only accepts input if the voice data is sampled at 16 KHz. It is independent of any language model. The model for creating the wav2vec uses the XSLR (Cross-Lingual Speech Representation), which additionally tests cross-lingual speech data. The quantization of latents, which is common to all languages, can be learned by XLSR. The voice utterance is loaded into the library, saved in a variable, and tokenized using the tokenizer. This process converts the audio to text, and the results are transcripts of the audio file that is loaded into the library. The transcripts are kept in a separate folder after the voice recognition process is complete. Between the transcripts produced by the model and the actual transcripts of the audio written by humans, the WER (Word Error Rate) is determined. The degree of voice recognition can be calculated using the WER value.

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

6 Evaluation of Results

The evaluation metric used by the task to test the results submitted by us is based on the WER computed between the ASR hypotheses submitted by the participants and the ground truth of human speech transcription.

WER (Word Error Rate) = (S + D + I) / N

where,
S = No. of substitutions
D = No. of deletions
I = No. of insertions
N = No. of words in the reference transcription

Word Error Rate (WER) is a commonly used metric in Automatic Speech Recognition (ASR) systems because it provides a straightforward and intuitive measure of the performance of the system. WER is calculated by comparing the recognized words from the ASR system to the reference (ground truth) transcription and counting the number of errors, including substitutions, insertions, and deletions.

7 Observation

The name of the speech data and its WER value are included in the result. Similar to this, the same procedure is used for all audio files. The number of subgroups into which each audio file is divided is also listed in the table. The test set audio files' average WER value, which comprises utterances from men, women, and transgender people, is determined in Table 2.

S.No.	Gender	Count	Avg WER
1	Male	1	33.091
2	Female	2	43.054
3	Transgender	7	40.331

Table 2: Average WER Value for Test Data

S.No.	File Name	Subsets	WER Value
1	Audio-37	15	39.227
2	Audio-38	17	37.872
3	Audio-39	16	46.916
4	Audio-40	17	43.915
5	Audio-41	19	16.792
6	Audio-42	24	17.511
7	Audio-43	30	22.308
8	Audio-44	28	21.545
9	Audio-45	26	31.871
10	Audio-46	47	28.243
11	Audio-47	56	39.192
12	Audio-48	56	22.175

Table 3: WER values for Testing Set

8 Discussion

From the Table 2, the experimental result says that the average WER for the testing dataset. The number of test speech utterances are 351. Similarly, Table 3, says the result of total 351 audio subset files from 12 audio files which is given for testing and the WER measured is 29.297%. We ranked second position in shared task competition.

9 Conclusion

The voice recognition algorithm is able to recognize older people better because to the utilization of conversational speech data. An automatic speech recognition system is developed using a trained model. A dataset pertaining to older folks and transgender individuals who speak Tamil as their mother tongue is being gathered. The utterance in the dataset was recorded in Tamil during a primary site discussion. In the future, the model may be trained using our own dataset and used for testing, which could increase performance, as the pretrained model of the system was refined using a shared speech dataset.

10 Future Work

In Future, instead of using the pretrained model we can fine tune the model with our custom dataset. Fine-tuning an Automatic Speech Recognition (ASR) system with a custom dataset is a promising approach to improve system performance in specific domains or applications, where end-to-end ASR architectures can be used, which directly map input audio to output transcriptions without intermediate steps. This can simplify the training pipeline and potentially improve performance, especially when dealing with limited custom datasets.

References

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.

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.

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.

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.