DRAVIDIAN LANGUAGE@ LT-EDI 2024:Pretrained Transformer based Automatic Speech Recognition system for Elderly People

Abirami Jayaraman , Aruna Devi Shanmugam , Dharunika Sasikumar & B. Bharathi
Department of CSE
Sri Sivasubramaniya Nadar College of Engineering,
Tamil Nadu, India
abirami2210382@ssn.edu.in
aruna2210499@ssn.edu.in
dharunika2210459@ssn.edu.in
bharathib@ssn.edu.in

Abstract

In this paper, the main goal of the study is to create an automatic speech recognition (ASR) system that is tailored to the Tamil language. The dataset that was employed includes audio recordings that were obtained from vulnerable populations in the Tamil region, such as elderly men and women and transgender individuals.

The pre-trained model Rajaram1996/wav2vec2-large-xlsr-53-tamil is used in the engineering of the ASR system. This existing model is fine-tuned using a variety of datasets that include typical Tamil voices. The system is then tested with a specific test dataset, and the transcriptions that are produced are sent in for assessment. The Word Error Rate is used to evaluate the system's performance. Our system has a WER of 37.733.

1 Introduction

The goal of this research project is to create an Automatic Speech Recognition (ASR) system that is specifically designed to serve vulnerable populations in the Tamil-speaking community, such as the elderly and transgender people. This shared task's scope entails improving speech recognition skills to enable certain demographic groups to access facilities including banks, hospitals, and shopping centers. The problem is closing the information gap between the elderly, who might not know how to use the tools at their disposal, and the transgender population, which experiences educational inequalities as a result of prejudice from society.

It is highlighted how important speaking is to these groups as their main form of communication because it is essential to meeting their everyday requirements. The study's dataset is made up of spontaneous speech samples that were obtained from transgender and elderly people who have trouble using different facilities. A two-hour speech dataset is provided for testing, and a training set of 5.5 hours of transcribed speech is also accessible.

Even while technological developments have made it easier for people to access electronic devices in a wider range of industries, educational barriers still pose a barrier for the elderly and transgender people. Even though these people use electronics, the fact that they primarily rely on audio messages highlights the need for better speech recognition models in order to accommodate their distinct speech patterns and deliver accurate responses. These people provide audio input to the method described in this paper, which converts it into corresponding Tamil transcripts. The Word Error Rate (WER) is used to measure the accuracy of these transcripts, taking into account the inherent difficulty that comes with different accents.

The structure of the report offers a thorough understanding of the research process. In Section 2, relevant literature is examined in detail, emphasizing gaps in the field and current knowledge. The dataset is thoroughly described in Section 3, along with its makeup and significance to the goals of the research. While Section 5 describes the implementation procedure, Section 4 covers the technique used in the development of the ASR system.

Key findings from the research are summarized in Section 6, which also offers insights on the study's performance and difficulties. A thorough discussion is started in Section 7 by interpreting the findings and placing them within the larger context of the study goals. The paper's main conclusions are summarized in Section 8, which also highlights the ongoing development of ASR systems for vulnerable populations and moves into a discussion of possible directions for further research.

The last section, Section 9, offers a thorough summary of the academic background that underpinned this project and is devoted to the reference articles that were examined during the research pro-

cess. Essentially, this work lays the groundwork for future advancements and improvements in this crucial area by advancing ASR technologies that meet the particular communication demands of vulnerable groups.

2 Related work

Voice recognition has advanced significantly over the past ten years, mostly due to the quick development of deep learning methods. Ten years ago, most researchers concentrated on using deep learning to extract audio data. Subsequently, these data were combined with hidden Markov models, which were popular at the time. But things have changed, and modern approaches now use more advanced recurrent neural networks (RNNs), like Long Short-Term Memory (LSTM) networks, in place of more conventional Gaussian mixture models. Notably, large models with higher parameterization, such Contextnet and Conformer, have shown improved voice recognition accuracy.

Convolutional neural networks (CNNs) are used by the authors of a study titled "Speech Rate Control for Improving Elderly Speech Recognition of Smart Devices" to overcome the difficulties in geriatric speech recognition on smart devices.

In a different study, "TransformerTransducer: End-to-End Voice Recognition with Self-Attention" [6], the authors want to use transformer networks in the neural transducer architecture to create an end-to-end speech recognition model. In order to incorporate positional information and reduce frame rates, their method integrates VGGNet with causal convolution, maximizing computing efficiency through reduced self-attention.

The preprocessing of data is essential to building reliable models. The authors stress the significance of labeled data for training models and discuss methods that entail using both labeled and unlabeled speech data to train large-scale models. An example of self-supervised learning is Wav2vec [3], which uses contrastive learning for feature learning and unlabeled speech data for training.

The effectiveness of deep learning in machine translation and speech recognition is recognized, highlighting its natural language comprehension abilities. This impact is being felt in a variety of domains, as scientists are investigating neural methods to comprehend code semantics and spot weak points. Notably, studies on low-resource speech recognition methods for minority languages have been conducted, and attempts have been made to improve accuracy by means of data augmentation.

The Rajaram1996/wav2vec-large-xlsr-53-tamil transformer model is used by the authors of [5],[4] to identify Tamil speech utterances made by susceptible individuals. To calculate the word mistake rate, the model takes into account variables like the number of utterances and the quality of the.wav file. This concept is primarily intended to promote inclusivity for marginalized people by increasing accessibility to regional languages.

In this work, audio files are transcribed using a pre-trained XLSR model, and word error rates are computed as a result. The focus is on using cutting-edge deep learning techniques to improve speech recognition systems' inclusivity and accuracy, especially when it comes to disadvantaged populations. The following parts provide a thorough analysis of relevant literature, dataset specifications, methodology, implementation details, outcomes that have been seen, and a thorough discussion. A few suggestions for future directions in this ever-evolving field of study are included in the paper's conclusion. The reference section lists the articles that were consulted for this research project, giving the provided conclusions a strong basis.

3 Dataset Description

The dataset given to this shared task [1] is an Tamil conversational speech recorded from the elderly people whose average age is around 61 for male, 59 for female and 30 for transgender people .A total of 6 hours and 42 minutes is collected from the elderly people. 46 audio files were recorded and each audio file is split into many subsets as transformer model does not support the large audiofiles. The speech is recorded with a sampling rate of 16KHZ. The audio files from Audio - Audio-10, to Audio-35 are used for training (duration is approximately 5.5 hours) [2] and Audio - 37 to Audio - 48 are used for testing (duration is approximately 2 hours).

4 Proposed Work

To construct their automatic speech recognition system, the researchers used the Rajaram1996/wav2vec2-large-xlsr-53-tamil pre-trained transformer model. This model is an advanced speech recognition system designed by Facebook AI specifically for the Tamil language, and it is based on the Wav2Vec2 architecture. Wav2Vec2 is a self-supervised learning technique

that builds representations that capture important information about audio features by utilizing massive amounts of unlabeled voice data. The model's fundamental architecture is based on transformers, which have proven to be incredibly useful in a variety of natural language processing applications. Transformers improve the model's ability to do nuanced analysis by allowing it to effectively capture long-range dependencies in audio inputs.

The model is trained by subjecting it to a sizable corpus of Tamil speech-containing monolingual and multilingual data. Pre-training enables a thorough grasp of the fundamental structure and characteristics of speech data by teaching the model to predict masked or distorted chunks of the input audio. After pre-training, the model is fine-tuned utilizing labeled data customized for particular downstream tasks, like Tamil keyword detection or transcription. Through this process of fine-tuning, the model can be made to adjust to the specifics of a given speech recognition task.

The model gains from the multilingual character of its pre-training data even though it was particularly trained on Tamil. Because of the large corpus of words it has been trained on, it can process a wide variety of words and phrases, which makes it ideal for tasks like transcription or speech recognition. The training set of the model and its fine-tuning methodology are purposefully created to capture the unique phonetic, phonological, and grammatical characteristics of the Tamil language. This careful process improves the model's capacity to identify and translate Tamil speech.

Business-wise, the use of cutting-edge models such as Rajaram1996/wav2vec2-large-xlsr-53-tamil highlights a dedication to utilizing cutting-edge technologies in voice recognition system development. The model is able to extract meaningful representations from unlabeled data thanks to the innovative use of a self-supervised learning strategy. Applying transformers, which are well-known for their effectiveness in natural language processing, shows a deliberate architectural decision to improve the model's analytical powers.

In line with industry best practices, the focus on fine-tuning for certain downstream applications guarantees that the model is tuned for the subtleties of Tamil speech recognition. Given that the model can accommodate a wide vocabulary, it can be used as a flexible solution for transcription or speech recognition tasks that need to cover a wide range of languages.

To sum up, the researchers' careful selection of the model, training process, and fine-tuning technique shows a dedication to creating a reliable and adaptable automatic speech recognition system that is suited to the nuances of the Tamil language. This strategy, which is based on cutting-edge technologies and best practices from the industry, presents the system as a useful tool for companies looking for precise and flexible voice recognition solutions. The ongoing development of these models has implications for various applications in various industries and offers potential for the area of natural language processing as a whole.

5 Implementation

We have harnessed an efficient model, leveraging a pre-trained transformer-based architecture named Rajaram1996/wav2vec2-large-xlsr-53-tamil. This particular model, a derivative of facebook/wav2vec2-large-xlsr-53, is specialized for Tamil and fine-tuned using the Common Voice dataset. To operate seamlessly, this model mandates a 16 KHz sampling rate for voice input. Our assessments have utilized LT-EDI's dataset to evaluate the model's efficacy.

The core functionality revolves around loading voice utterances into the library, storing them as variables, and tokenizing them via a dedicated tokenizer. This transformation pipeline is instrumental in converting audio signals into textual representations. Our meticulous approach involves a thorough comparison between these transcribed texts and the original audio transcripts. This critical alignment allows us to calculate the Word Error Rate (WER), a metric that reflects the fidelity of voice recognition thereby used to quantify the accuracy and precision of the model's voice recognition capabilities. This approach, rooted in the XLSR (Cross-Lingual Speech Representation) framework, extends its capabilities to cross-lingual speech data, showcasing the model's adaptability across languages. The derived WER provides a robust assessment of the model's proficiency in voice-to-text transcription. By using the WER as our benchmark, we gain deeper insights into the model's performance, and efficiency for affirming its prowess in transforming spoken words into accurate text.

S.No	File Name	Number of Sub-sets	WER
1	Audio-10	38	40.84
2	Audio-11	49	41.03
3	Audio-12	17	35.26
4	Audio-13	33	39.99
5	Audio-14	25	34.72

Table 1: WER Values for Training Set used for testing

6 Evaluation of Results

The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence. The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level. The task's evaluation measure is based on the WER (Word Error Rate) computed between the original transcriptions of the given audio and the transcribed text.

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

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. Table 1 provides insights to some of the transcribed statements using the training data.

1	Targeted Sentence	அடுத்து எப்போ வரணும்.வந்தா எப்போ பாக்கலாம் .இல்ல ஏதும் வேற ஏதும் மருந்து மாதர வாங்கணுமா ஊசி ஏதும் போடணுமா .திருப்பி கட்டாயம் வரணுமா என்னனு சொல்லுங்க மேடம் பணம் டெபாசிட் பண்ணனும் பிள்ளைங்களுக்கு ரெண்டு பிரிவா
	Predicted Sentence	பநிக பாற்றறக்கமே நாடத்துயர்ப வற வந்தாயப்ப பாகலா லல இதும யாரது மார்தநது மத்ற வாங்கனமா ஊஷி யரும் போடணுமா திரப்பிகட்டாயம் வரணுமா எ நான் சொல்லுங்கறம் உர பணங்ககும் தபழ்ஷ்டுபணம்பலேயலுகேஅரெண்டுபெரிவா அ
2	Targeted Sentence	தெரிஞ்சவங்க உள்ள இருகாங்க பா . சொந்தக்காரங்க .அவங்கள பார்க்கணும் . பாக்க வந்துருக்கன் .எந்த இதுல இருகாங்க .அவங்கள இப்போ பார்க்க முடியுமா .எந்த டயத்துல வந்தா பார்க்கலாம். ஏதும் அவங்களுக்கு வாங்கிட்டு போலாமா
	Predicted Sentence	பேய தெரம்பங்க உள்ளார்காங்கபா ஐந்துதரம்ம அங்மள பாக்கனர் பாக்காம்பருகேன் எண்டவா தலர்காங்க அஅவங்களப் ப பாக்கமுடியுமா எலேஎந்தத டேயட்டுக்கே வந்தா பாகவான் நஏத மாவங்களுக் வாய்ந்து போழமா அலவவேதுவ

Figure 1: Sample predicted sentences

8 Discussion

The number of test speech utterances are 295. From the total number of 295 audio subset files from 10 audio files which is given for testing and the WER measured is 37.73. We ranked fourth position in shared task competition.

9 Conclusions

Conversational speech data is utilised to improve the speech recognition system's capacity to detect elderly people. A trained model is used to construct an automatic speech recognition system. A dataset collection is focusing on older adults and transgender people who use Tamil as their first language. The dataset's utterance was taken during a conversation in a major site in Tamil. Because the system's pre-trained model was enhanced using a common speech dataset, the model might be trained using our own dataset and tested in the future, which could improve performance.

References

[1] Bharathi B, Bharathi Raja Chakravarthi, Subalalitha Cn, Sripriya N, Arunaggiri Pandian, and Swetha Valli. Findings of the shared task on speech recognition for vulnerable individuals in Tamil. In Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar, editors, *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 339–345, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[2] Bharathi B, Bharathi Raja Chakravarthi, Sripirya N, Rajeswari Natarajan, Suhasini S, and Swetha Valli.

Overview of the third shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, Malta, March 2024. European Chapter of the Association for Computational Linguistics.

[3] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*, 2019.

[4] Varsha Balaji, Archana Jp, and B Bharathi. Cse_speech@ lt-edi-2023automatic speech recognition vulnerable old-aged and transgender people in tamil. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 204–208, 2023.

[5] S Suhasini and B Bharathi. Asr_ssn_cse@ ltedi-2023: Pretrained transformer-based automatic speech recognition system for elderly people. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 161–165, 2023.

[6] Ching-Feng Yeh, Jay Mahadeokar, Kaustubh Kalgaonkar, Yongqiang Wang, Duc Le, Mahaveer Jain, Kjell Schubert, Christian Fuegen, and Michael L Seltzer. Transformer-transducer: End-to-end speech recognition with self-attention. *arXiv preprint arXiv:1910.12977*, 2019.