SSNCSE_NLP@LT-EDI-ACL2022: Speech Recognition for Vulnerable Individuals in Tamil using pre-trained XLSR models

Dhanya Srinivasan **B. Bharathi** **D. Thenmozhi** **B. Senthil Kumar**
SSN College of Engineering
dhanya2010903@ssn.edu.in
bharathib@ssn.edu.in
theni_d@ssn.edu.in
senthil@ssn.edu.in

Abstract

Automatic speech recognition is a tool used to transform human speech into a written form. It is used in a variety of avenues, such as in voice commands, customer, service and more. It has emerged as an essential tool in the digitisation of daily life. It has been known to be of vital importance in making the lives of elderly and disabled people much easier. In this paper we describe an automatic speech recognition model, determined by using three pre-trained models, fine-tuned from the Facebook XLSR Wav2Vec2 model, which was trained using the Common Voice Dataset. The best model for speech recognition in Tamil is determined by finding the word error rate of the data. This work explains the submission made by SSNCSE_NLP in the shared task organized by LT-EDI at ACL 2022. A word error rate of 39.4512 is achieved.

1 Introduction

Speech recognition (also known as speech-to-text or Automatic Speech Recognition) is a technique used to convert human speech into a written format. It is an important tool, and has many applications, such as in mobile phones (voice commands for call routing and voice searching), customer service, emotion recognition, and more importantly, in helping disabled people. It can not only help convert words to text to assist hearing impaired people, but also aids physically impaired people in performing activities such as typing and browsing using voice commands, instead of having to manually operate a computer.

Tamil is the official language of Tamil Nadu, Puducherry, Sri Lanka and Singapore. (Chakravarthi and Raja, 2020)(Chakravarthi and Muralidaran, 2021) was the first language to be classified as a classical language in India, out of over 22 scheduled languages in India. It is also one of the oldest languages in the world, seemingly originating over 2000 years ago.

The speech recognition is achieved by considering the linguistic features of the Tamil language. The natural language processing approach is used for the speech recognition task.

The team of SSNCSE_NLP has participated in the Speech Recognition for Vulnerable Individuals in Tamil shared task, obtaining the first position with a word error rate of 39.4512.

In our paper, we have used pre-trained models designed for the Tamil language, to transcript the speech audios into tokens, eventually decoding them into text. We have made use of three pre-trained models, namely *Amrrs/wav2vec2-large-xlsr-53-tamil*[1], *akashsivanandan/wav2vec2-large-xls-r-300m-tamil-colab-final*[2] and *nikhil6041/wav2vec2-large-xlsr-tamil-commonvoice*[3].

The issue with automatic speech recognition in older adults as well as physically or mentally impaired people is that they tend to display mild dysarthric speech, or slurred speech, causing erroneous transcription of the data. Furthermore, in Tamil speaking places, people from different regions speak in non-identical dialects, accents, and speeds, and hence, the transcription of the data differs from person to person. When trained with audios from a single region, there is incapacity to accurately predict what a person from a different region is saying.

The rest of this paper is arranged as follows. Section 2 discusses the related work on

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

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

[3] https://huggingface.co/nikhil6041/wav2vec2-large-xlsr-tamil-commonvoice

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 317 - 320
May 27, 2022 ©2022 Association for Computational Linguistics

speech recognition tasks. The dataset about the shared task is described in Section 3. Section 4 outlines the features and machine learning algorithms used for this task. Results are presented in Section 5. Section 6 concludes the paper.

2 Related Work

Researchers have experimented with a few approaches to deal with speech recognition in minority languages such as Tamil recently. The writers of Voice and speech recognition in Tamil language (Kiran et al., 2017),propose the use of a Hidden Markov Model(or HMM). It is a statistical pattern matching approach which can generate speech using a number of states for each model. Since the HMM model uses positive data, it scales well and it reduces the time and complexity of the recognition process. In Design and Development of a large vocabulary, continuous speech recognition system for Tamil, the authors (Madhavaraj and Ramakrishnan, 2017) build two independent recognition systems for phone recognition (PR) and for continuous speech recognition (CSR) using deep neural networks (DNN). The DNN based triphone acoustic model is proven to yield significantly better results in CSR and PR. The authors of Speech Rate Control for Improving Elderly Speech Recognition of Smart Devices (Son et al., 2017), take the help of a convolutional neural network (CNN) to generate feature vectors to be fed into a fully connected network (FC) for frame by frame syllable transition boundary classification. Thus the syllable transition probability is calculated and the syllables are segmented. They take the help of a Synchronized Overlap- Add (SOLA) Algorithm to adjust the speech rate according to the measured ratio on a time-scale. In Transformer-Transducer: End-to-End Speech Recognition with Self-Attention (Yeh et al., 2019), the authors attempt to build a model for end-to-end speech recognition using transformer networks in neural transducer. They propose two methods, namely using VGGNet with causal convolution to incorporate positional information and reduce frame rate for efficient inference and using truncated self-attention to enable streaming for transformer and reduce computational complexity. In this paper, however, we use pre-trained XLSR models to transcript the audios.

Data Set	Instances	Running Time
Training Set	909	20 seconds
Testing Set	239	20 seconds

Table 1: Specifications of the Dataset provided

3 Dataset Analysis and Preprocessing

The data set given by the shared task organizers consists of a training set and a testing set, each consisting of 909 and 239 instances respectively (Bharathi et al., 2022). The training set contains the audio files and transcriptions of the audios in the Tamil language, whereas the testing set contains only the audio files. The audios in both the training set and the testing set contain audio recordings, each having an average running time of 20 seconds.

4 Experimental setup and Features

For feature extraction, the n-gram model is experimented upon. The three pre-trained models are each used to extract the features. All three pre-trained models are fine-tuned versions of the Facebook XLSR Wav2Vec2 model, trained using the Common Voice Dataset containing 9283 hours of audios of different languages. These models have been trained using the Tamil speech corpus in the same.

The pre-trained XLSR model maps the speech signal to a sequence of context representations. For the model to map the latter to its corresponding transcriptions, a linear layer used to classify each context representation to a token class has to be added on top of the transformer block. The output size of this layer corresponds to the number of tokens in the vocabulary, which does not depend on XLSR's pre-training task, but only on the labeled dataset used for fine-tuning. The training data is run and the transcriptions are obtained. Punctuation marks and other characters without meaning are removed from the transcriptions and all distinct letters of the training data are used to build the vocabulary (an enumerated dictionary). The

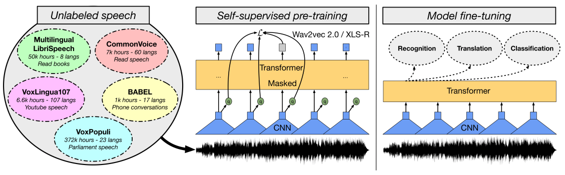

Figure 1: Working of the finetuned XLSR model (von Platen)

Word Error Rate for each Model	
Model	Testing Data
Amrrs	45.128
akashsivanandan	46.945
nikhil6041	39.4512

Table 2: Word error rate in the testing data for each pre-trained model

vocabulary saved as a .json file is used to load the vocabulary into an instance of the Wav2Vec2CTCTokenizer class.

For XLSR, the fine tuning data has a sampling rate of 16kHz. XLSRs feature extraction pipeline is fully defined as an instance of the Wav2Vec2FeatureExtractor class and the feature extractor and tokenizer are wrapped into a single Wav2VecProcessor class.

Each training audio file is loaded as a floating point time series at a sampling rate of 16000 samples per second. A data collator is defined to pad the training data to the longest sample in the batch. The pre-trained checkpoint of Wav2Vec2-XLSR is loaded and all parameters related to training are defined. The model is then trained and the word error rate is found.

The three models are trained in the above manner and used to generate input values using tokenizer and the logits are found out using the model. The tokens for the logits are predicted and decoded to find the transcriptions of the audio.

5 Observations

At a first glance at the transcriptions, it can be seen that the Amrrs model is unable to not only differentiate between when a particular word ends and another begins, but is also unable to apply the stressed consonants in many places, such as in the words எவ்வளவு and ஒவ்வொரு . This can be justified by the stress being applied for uncertain periods of time in different words. It is also ineffective at identifying vowels before stressed consonants which are mainly used only as stressed consonants and not simply as consonants, such as ந் and ங், because their sounds are almost always preceded by a vowel, hence making it indiscernible.

Looking at the transcriptions generated by the akashsivanandan model, it can be said that it is unable to distinguish between consonants of the same pronunciation sets, such as the harshly pronounced letters, feebly pronounced letters and the medially pronounced letters. This can be seen in words such as மூன்று which are transcripted as மூண்டு due to the stressed consonant ன் being misinterpreted, subsequently leading to the next letter to also be misunderstood. The occasional English word in the audios is pronounced differently compared to its Tamil transliteration, causing it to be wrongly interpreted.

The nikhil6041 model is found to produce the most accurate transcriptions out of all three models tested. It occasionally mislabels similar consonants(such as ள and ழ) and vowels (such as அ and எ) and is sometimes unable to mark the vacant spaces between two words, but for the most part, it generates transcriptions which are in the vicinity of how the words are actually pronounced in the audio. However, it does not always correspond to the actual transcription of the audio as the pronunciations differ when they are spoken or written.

6 Conclusion

The need for automatic speech recognition for vulnerable individuals is growing to be increasingly important every day. More and more of people's daily lives are made easier by technology, regardless of whether they have disabilities or not. Speech recognition technology, though popular and well refined for prominent western languages such as English, is not available easily to minorities who do not speak that language. Our motive has been to make automatic speech recognition software more easily accessible to the Dravidian population, more importantly, the Tamil speaking population. In this paper we propose to use pre-trained speech recognition models created for the Tamil language and use it to transcribe the testing audios provided by the organizers. It is noted that the nikhil6041 model yields the best results out of all the three used models. The above model can be finetuned further by obtaining and using a more extensive dataset, and training the model against a more sizable range of accents and dialects. This will lead to an overall more accurate transcription of the audios.

References

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.

Asoka Chakravarthi and Bharathi Raja. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

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.

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.

A Madhavaraj and AG Ramakrishnan. 2017. Design and development of a large vocabulary, continuous speech recognition system for tamil. In *2017 14th IEEE India Council International Conference (INDICON)*, pages 1--5. IEEE.

Guiyoung Son, Soonil Kwon, and Yoonseob Lim. 2017. Speech rate control for improving elderly speech recognition of smart devices. *Advances in Electrical and Computer Engineering*, 17(2):79--84.

Patrick von Platen. Fine-tuning xls-r for multi-lingual asr with huggingface transformers. https://huggingface.co/blog/fine-tune-xlsr-wav2vec2.

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