Automatic Speech Recognition in Hindi using XLSR - Wav2Vec 2.0

Deepthi Sudharsan., Harish K., Asmitha U., Roshan Tushar S.
Center for Computational Engineering and Networking, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India,

Abstract Speech recognition is one of the common application of speech processing since it is a faster alternative to capture speech compared to manual typing and interpretation. After the boom of transformers for NLP tasks, the popularity of transformers in other areas such as speech processing, computer vision etc., started becoming popular. Wav2Vec2 is one of the most popular transformers for Automatic Speech Recognition (ASR). It uses a semi-supervised approach (based on the idea of contrastive learning) that owes to its good performance. It is the first ASR to be included in Transformers. In this research, XLSR-Wav2Vec 2.0 will be tried on the Common Voice Hindi dataset for performing ASR.

1 Introduction

Speech recognition is the ability to recognize and identify spoken words and phrases. To recognize a speech signal, the signal is first broken down into individual sounds and using various algorithms the next probable word is found and the speech is interpreted. Its ability to capture speech faster makes it one of the powerful applications of speech processing and hence, it is incorporated in many day-to-day applications. Over the past few years, a lot of research has been done around speech recognition in English and other non - Indic languages. Speech recognition of Indic languages is still at its youth due to the lack of Gold standard dataset of Indic languages especially Dravidian languages and accurately trained pre-trained models for these languages. This paper talks about the automatic speech recognition of Hindi language using pre-trained transformers such as XLSR - Wav2Vec2. The XLSR - Wav2Vec2 pre-trained transformer was designed for performing cross-lingual automatic speech recognition tasks. The XLSR-53 model is highly sought out for its

Deepthi Sudharsan., Harish K., Asmitha U., Roshan Tushar S.
Center for Computational Engineering and Networking, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India

capability to handle over 53 languages. The model performed with a word error rate (WER) of 73.201856 on the Common Voice Hindi dataset used.

2 Related Works

Since the research in the area of automatic speech recognition for Indic Languages is still at its youth, there is a need for pre-trained models that can solve low-resource speech recognition tasks in languages such as Hindi, Tamil etc.. Wav2Vec 2.0 was applied to various low-resource languages in [4] and it was found to yield good results on the low-resource languages as well but in some works such as [7], it was noticed that the small model did not yield any results (may have needed more unlabelled data to train) whereas the large model proved to be successful in predicting the audio recordings in Arabic and achieved a Word Error Rate of 24.40% . [2] tackled the problem of learning new language representations while utilizing existing model knowledge using the Wav2Vec 2.0 model. In 2021, [6] proposed a weakly supervised multilingual representation learning framework, called cross-lingual self-training (XLST) to Learn Multilingual Representation for Low Resource Speech Recognition and in the same year, [5] came up with the idea of fusing a pre-trained acoustic encoder (Wav2Vec2.0) and a pre-trained linguistic encoder (BERT) into an end-to-end ASR model for Low-Resource Speech Recognition. [3] also suggested pre-training a simple multi-layer convolutional neural network optimized via a noise contrastive binary classification task Unsupervised Pre-training for Speech Recognition.

3 Methodology

3.1 Dataset Description

The dataset used for this work is the Common Voice Corpus 6.1 Hindi dataset that was released in late 2020. The 20 MB dataset consists of 292 MP3 clips in the training data and over 127 Hindi clips for the testing data shown in Table.1. The dataset also includes demographic metadata like age, gender, accent etc. Over 71% of the entries are male and 2% of entries are females with a large proportion of entries (over 63%) belonging to participants between the ages of 19 and 29. Out of the total 0.8 recorded hours in the dataset, around 0.54 recorded hours are for validation.

Split	Number of Clips
Train	292
Test	127

3.2 Training and Classification

For the purpose of this work, Facebook AI's XLSR - 53 pre-trained on the Wave2Vec 2.0 platform has been utilized. The model was first proposed in [1] and its architecture can be seen in .1.

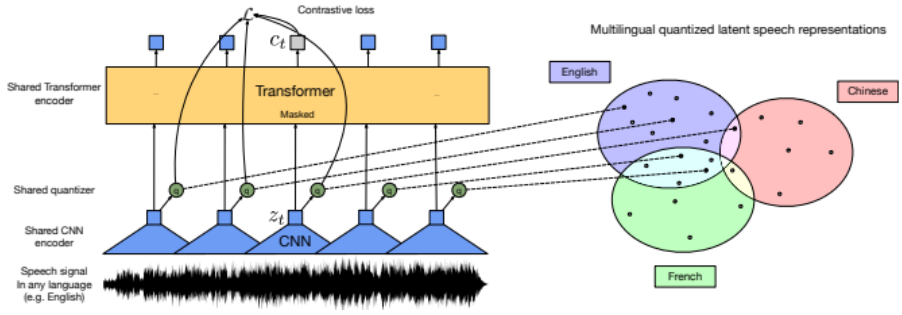

Fig. .1: The wave2vec 2.0 - XLSR approach

It is known that for speech signal input unlike text input data is continuous in nature and needs to be made discrete. Firstly, the input data with a large number of samples is passed through the CNN to get the latent feature representations. By doing so, there is dimensionality reduction as only the important feature representations are considered. This is the encoding stage. The input has been encoded. This encoded output is discretized by means of quantization and this contributes towards the overall loss. The encoded output is also fed through a transformer that makes use of convolutional networks that help in attaining the contextual information and the self attention of the transformer focuses on the positional embeddings. The output is a sequence of probabilities for the different classes. In order to perform decoding if required, a simple 4-gram model goes a long way.

4 Results and Conclusion

In this work, the self-supervised model, XLSR - 53 was explored for the Common Voice Corpus 6.1 Hindi dataset. It can be seen that the overall, the model performed reasonably well with a WER of 73.201856%. In the future, the work can further be expanded for code-mixed speech signals like "Tanglish" etc. and further hyper-tuning can be performed to further improvise minimizing the error rate.

References

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Kessler, S., Thomas, B., & Karout, S. (2022). An adapter based pre-training for efficient and scalable self-supervised speech representation learning. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3179–3183.

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). Wav2vec: Unsupervised pre-training for speech recognition, 3465–3469. https://doi.org/10.21437/Interspeech.2019-1873

Yi, C., Wang, J., Cheng, N., Zhou, S., & Xu, B. (2020). *Applying wav2vec2.0 to speech recognition in various low-resource languages*.

Yi, C., Zhou, S., & Xu, B. (2021). Efficiently fusing pretrained acoustic and linguistic encoders for low-resource speech recognition. *IEEE Signal Processing Letters*, *28*, 788–792.

Zhang, Z.-Q., Song, Y., Wu, M.-H., Fang, X., & Dai, L.-R. (2021). *Xlst: Cross-lingual self-training to learn multilingual representation for low resource speech recognition*.

Zouhair, T. (2021). Automatic speech recognition for low-resource languages using wav2vec2: Modern standard arabic (msa) as an example of a low-resource language.