IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Live Speech to Text Conversion (Punjabi) Using Wav2Vec2.0

Avi Aswal (aviaswal7@gmail.com)
Nikhil Aggarwal (Nikhil.agg1307@gmail.com)
Mihir Khurana (mihirkhurana.mk@gmail.com)
Ms.Akanksha Dhamija (akankshadhamija12@gmail.com)

Abstract

In this paper, we provide a live conversion of speech to textual content (for Punjabi language) and further translation of the Punjabi text into English. Our methodology is based on and takes the help of a model called 'Wav2vec 2.0' developed by Facebook AI. This is an end to end framework that allows for self-supervised learning during the modeling process of speech representation. This method has proven to be very effective and successful in Automatic Speech Recognition (ASR) even with very few available training datasets. There has, however, been a great deal of work based on this model for one language only: English. To test the model's robustness we have fine tuned the model on an Indian regional language called Punjabi. The resources available for the model are significantly less than those for English. So, we found it an appropriate way to test the model's use for other languages. Our model was further pre-trained with a Punjabi dataset in order to convert it to cross-lingual usage as we had limited resources. Based on the results, we conclude that the proposed strategy yields high accuracy using the provided dataset. We also translate the transcribed text result to English for better understanding.

1. Introduction

Using procedures, neural networks analyze and recognize patterns that are hidden within raw data. The systems consist of nodes that work similarly to the human brain. They cluster, classify and most importantly, "LEARN" overtime and tend to improve.

For example, recently LivingAI company has launched its EMO Robot that works on AI, and can recognize faces, It works on Neural Networks and tends to learn from surroundings like the face of the owner, the beats of the song playing around, the edges of the table on which it is kept so that it doesn't fall, and so on. In the research papers we read and looked upon when we were being introduced to this topic, we saw its growth in the Machine Learning Industry around 2012 and from then it has never looked back.

However, we must not forget who introduced this brilliant idea, Warren McCulloch and Walter Pitts developed the first neural network in 1943. They modeled their ideas using electrical circuits to create a simple neural network, which was a seminal paper on how neurons work.

In recent years, self-supervised methodologies have shown great results in a wide range of fields, including Natural Language Processing (NLP). In this procedure, a model is pre-trained on abundant unlabeled data and then fine-tuned on confined labeled data.
Facebook recently announced and open-sourced Wav2Vec 2.0, a new framework that lets users learn representations from raw audio by themselves. The researchers claim that with just 10 minutes of transcribed speech data, the framework can enable Automatic Speech Recognition (ASR) which is quite unbelievable and astonishing. In traditional speech recognition, audio recordings with transcriptions are used to train the models. In order for a system to work well, it needs large amounts of annotations which are only available only for a few languages. Self-supervision allows for better

systems to be built by leveraging un-annotated data.

Several self-supervised speech reconstruction approaches attempt to reconstruct the audio signal, which includes capturing the recording environment, channel noise and speaker characteristics. An alternative common approach is to train the model by asking it to predict what the speaker will say next by contrasting several alternatives.

After fine-tuning with small amounts of labeled data, the Wav2vec 2.0 model has shown excellent performance in English ASR tasks. No matter how impressive the model has proven to be, it is yet questionable whether it can be applied to other languages or not, since most experiments have been conducted to date with only English datasets, such as Librispeech.

The distinctive Gurmukhi script is used for Punjabi in India, and is especially associated with the Sikhs. This script is a member of the Indic script family, written from left to right, but it's correspondences contrast fundamentally with Devanagari, which is used to compose Hindi. Gurbani's Gurmukhi content has 35 akhar, or consonants, which are conjugated similarly to the Punjabi painted letters in order including three vowel holders and 32 consonants. Each character addresses a phonetic sound. Alphabetical order of the Gurmukhi script isn't totally equivalent to the English letters all together.

Taking inspiration from prior works performed with other languages and dialects, we propose a multi-task hierarchical fine-tuning architecture of Wav2vec 2.0 to mirror the exceptional relationship that exists in Punjabi Composing among syllables and graphemes.

Our methodology learns a bunch of discourse units, which are more limited than phonemes, to depict the discourse sound grouping. Since this set is limited, the model can't address all varieties, for example, background noise. Instead, the units urge the model to zero in on the main elements to address the speech audio. In our analysis, we find that this works better as compared to alternative methodologies on the LibriSpeech benchmark.

2. Background

Computing ability & artificial intelligence are prominently the reasons behind the advances in the area of NLP and Speech Recognition. With more and more quantities of speech statistics added together with speedy processing, speech recognition has hit a point in which its talents are almost at the equal stage with humans.
By the year 2001, speech reputation generation had attained close to 80% accuracy. For a huge part of the decade there weren't a whole lot of improvements until Google launched the Google Voice Search. Since it became an app, this positioned speech recognition into the palms of hundreds of thousands of individuals. It also changed vastly because the processing power could be offloaded to its server farms. Now not just that, Google was accumulating information from billions of searches that may assist it with predicting what an individual is actually saying. At that time Google's English Voice Search System included 230 billion words from user searches. In 2011 Apple introduced Siri which was similar to Google's Voice Search. The early part of this decade noticed an explosion in the field of different voice recognition applications. Additionally, with Amazon's Alexa, Google Home/Nest devices we've observed customers to become more open to conversing with machines.

Wav2Vec 2.0 is a pre-trained version for Automatic Speech Recognition (ASR) and was launched in September 2020 by Facebook AI. Quickly after the overall performance of Wav2Vec2 turned into testing on one of the most famous English datasets for ASR, called LibriSpeech. Facebook AI provided a multilingual version of Wav2Vec2, called XLSR. XLSR stands for cross-lingual speech representations and refers to a model's capacity to learn speech representations that are used across various languages. XLSR's successor, simply called XLS-R (referring to the "XLM-R

for Speech"), was released in November 2021 by Arun Babu, Changhan Wang, Andros Tjandra, et al. XLS-R used almost five hundred thousand hours of audio data in 128 different languages for self-supervised pre-training and comes in packets ranging from 300 million up to 2 billion variables.

Wav2vec 2.0 is a contemporary deep learning structure proposed for ASR that uses self supervision for speech illustration. Even though the original tasks focused on the English language, [8] pre-trained a new edition, referred to as Wav2vec 2.0 XLSR-53, using fifty six thousand hours of speech audio of fifty three distinctive languages. This work uses the version pre-trained by [8] to construct an ASR for Punjabi, using only openly available information and data. The resulting fine-tuned version for Punjabi and code for this research is readily available.

2.1. Pre-training Wav2vec 2.0

As based on [8], Wav2vec 2.0 is an aggregate of 3 networks, including a feature encoder, a contextual transformer, and a quantization module as mentioned in [9]. The feature encoder, that's made from a multi-layer convolutional neural network (CNN), encrypts the raw audio X and processes the latent speech representations Z. The contextual transformer, which includes a stack of remodel encoders, learns the context illustration C using the latent speech representations as input. The quantization module is used to map latent representations into the discretized space Q, selecting distinct codebook entries in a fully distinguishable way. In pre-training, a particular portion of the latent representations is arbitrarily masked before supplying them into the contextual transformer. After training, it is used to distinguish between the true quantized latent vector and the distinct latent vectors sampled from the other masked time steps. At the time of pre-training, the model learns contextualized representations from only the un-annotated speech audio data.

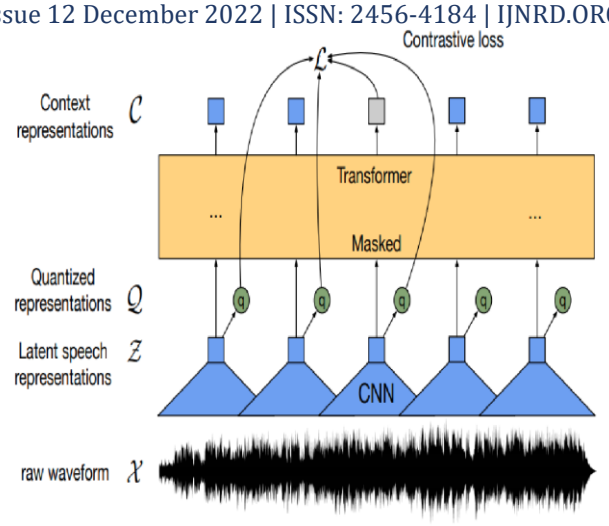

Fig 1. Wav2vec 2.0 Architecture

2.2. Fine-tuning for ASR

A randomly initialized linear layer is added using only the configured linear layer on top of the pre-trained model for ASR. This linear layer accepts the contextualized representations of the pre-trained model and shows the most expected words. In fine-tuning, connectionist temporal classification (CTC) cost, which is for sequence labeling without the need of alignment data between the output sequences and the input audio, is used in training, the linear layer & the pre-trained model. As per [8], XLS-R shows magnificent improvements over earlier state-of-the-art outcomes on speech identification, speech translation & speaker/language recognition.

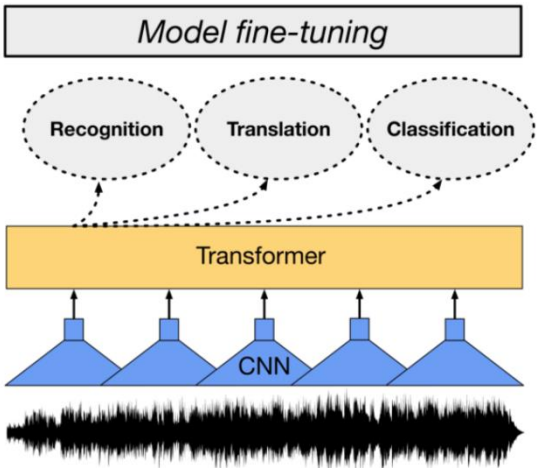

Fig 2. Fine-tuning the model

3. Literature Review

Initial readings of studies on the speech recognition field assisted us in moving towards the right path for the implementation of our project. [1] and [2] helped us acquire understanding of Natural Language Processings (NLP) and Automatic Speech Recognitions (ASR). Concepts like POS tagging, phonemes, graphemes were encountered and learned during this phase. [3] further expanded on feature extraction necessary for implementation of Neural Networks (NNs).

Our reading of the works of [4] and [5] inspired us to explore the applicability of the Wav2vec 2.0 model on languages with less resources and helped us in deciding the type of neural network we wanted to choose for the implementation of the project. Wav2vec 2.0 is a CNN based model and both, [4] and [5], have problem statements regarding regional languages. Hence, we decided to choose an Indian regional language called Punjabi for the development of our project.
Further reading and exploration suggested by [6] and [7] indicated that the Wav2vec 2.0 model achieves great accuracy for tasks related or associated with speech recognition also. Hence, our decision of choosing the Wav2vec 2.0 model seemed encouraging and made us feel confident in our choice.

As suggested by [4], we explored fine-tuning methods for languages with fewer resources and through the help of [8] gained knowledge on the process of fine-tuning using XLS-R for the Wav2vec 2.0 model. [8] Provided us with the advice of choosing Hugging Face transformers while fine-tuning our model.

To understand the working and architecture of the Wav2vec 2.0 model we referred to the works of original developers of the model in [9], [10] and [11]. These assisted us in acquiring the technical knowledge to pre-train , moreover fine-tune our model. With the help of [8] we were able to shortlist datasets that we could use to pre-train and fine-tune our model.

However, till now most of the work in speech processing has focused on supervised cross-lingual training which needs transcribed and annotated data in multiple languages as mentioned by [4]. Transcribed speech is normally much scarcer in availability than un-annotated speech and requires human intervention for labeling the data.

Unsupervised representation learning, or pre-training, does not need annotated data and has gained a lot of attention lately in computer vision after profound success in Natural Language Processing (NLP). For the latter, cross-lingual pre-training has been established to be very compelling and very efficient, especially for lower-asset languages.

In this paper, we center around the cross-lingual setting by learning portrayals on un-annotated data that generalize across languages. We proceed further on the pretraining approach of [9] which together learns contextualized speech portrayals as well as a separate lexicon of latent speech representations. The latter serves to efficiently prepare and train the model with a contrastive loss and the distinct speech representations are shared across languages. Different from recent work on unsupervised cross-lingual pre-training, we fine-tune the transformer part of our model as opposed to freezing all pre-trained representations or feeding them to a separate downhill model.

4. Methodology

4.1 Architecture

We follow the design choices described in [9]. The model contains a convolutional feature encoder f : X → Z to map raw audio X to latent speech representations $z_1, \ldots, z_T$ which are fed to a Transformer network g : Z → C to output context representations $c_1, \ldots, c_T$ ([9]). For the purpose of training the model, feature encoder representations are discretized to $q_1, \ldots, q_T$ with a quantization module Z → Q to represent the targets in the self-supervised learning objective. The quantization is based on product quantization ([9]) by choosing quantized representations from G = 2 codebooks with V = 320 entries each. The result is concatenated to obtain q. A Gumbel softmax enables choosing discrete codebook entries in a fully differentiable way. Each Z_T represents about 25ms of audio strided by 20ms, the context network architecture follows BERT, except for relative positional embeddings [9].

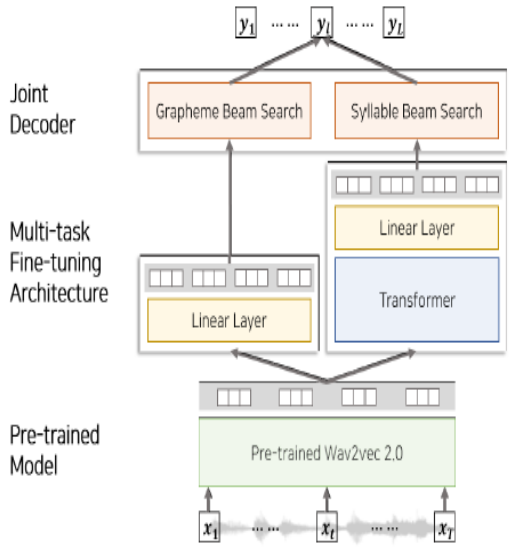

Fig 3. Proposed Model Architecture

4.2. Choosing Dataset

Multilingual LibriSpeech (MLS): an enormous dataset available in a lot of different languages. The MLS is primarily based on audiobook recordings in public domain like LibriVox10. The dataset contains a total of 6000 hours of transcribed speech data in many different languages with the most easily available language as English. The model is pre-trained on this dataset for the English language.

CommonVoice: The CommonVoice dataset is a vast multilingual corpus of read speech consisting of more than two thousand hours of speech data in 38 languages. The amount of speech data available per language ranges from 3 hours for Swedish ("low-resource") to 1350 hours for English ("high-resource").

We verify the proposed method with a Punjabi speech dataset, available on an open source platform known as Common Voice developed by Mozilla. This is a very small dataset as per today's standard and hence tests the applicability of the Wav2vec 2.0 model on smaller training datasets. The dataset comprises 3 hours of labeled and transcribed speech. The used dataset comprises 46 different voices. The version of the dataset is: 'pa-IN_3h_2021-07-21'.

4.3. Pre-training Wav2vec 2.0

To build our Punjabi adapted model we used the English Wav2vec 2.0 that was released by [9], that is pre-trained on 960 hours of Librispeech with no fine-tuning. We have further pre-trained the English model for 3 hours of Punjabi speech.

4.4. Fine-tuning for ASR

Our technique of assessing cross-lingual transfer is to further pre-train the English model with the Punjabi dataset using ASR performance with limited data. Many studies have looked at the transfer of speech representations across languages by pre-training the model on a higher-asset language, then fine-tuning to the lower-asset ones. Although a pre-trained model from an alternate language allegedly improves performance, it is not specialized to adapt the target languages. Hence, we offer further pre-training on the target data (i.e. the Punjabi dataset) after pre-training on a higher resource language (i.e. English). The strategy is based on information provided by [8] in his blog post.

5. Result

We've successfully implemented the Wav2vec 2.0 model using XLSR cross lingual version to convert Punjabi speech to English text.

The WER of our model is 54.86 %. However, the meaning of the sentence spoken is correctly conveyed to the end user in English.

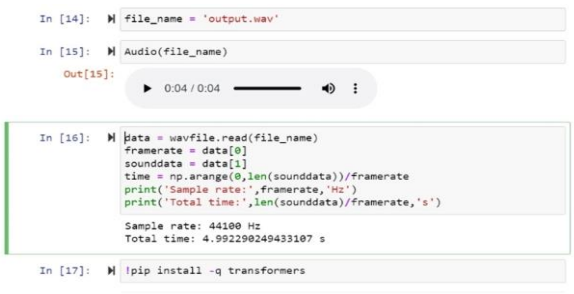

Fig 4. Converted this Audio file

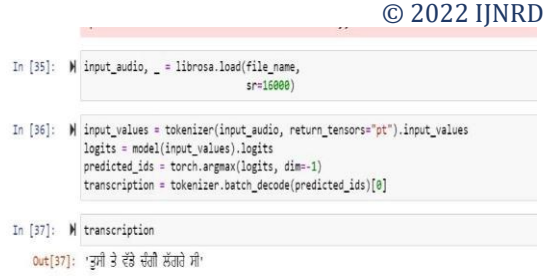

Fig 5. Into this Punjabi text

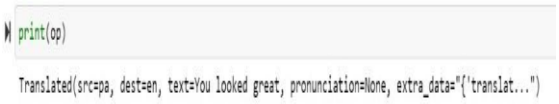

Fig 6. Translated the Punjabi text to English

6. Future Scope

Our project was developed keeping those people in mind that can speak Punjabi but are not able to write it. In future, improvements can be made on many parameters, such as the accuracy, till now we are able to achieve decent accuracy but surely there is room for improvement.

Many people in parts of Northern India have their mother tongue as “Punjabi” i.e. they learn to speak it from their families in which they are born and brought up, but only a few of them choose to learn to write it, so as an attempt to not let this language (in written form) get wiped out we developed our model.

Also, similar models can be developed for several other languages that are on their verge of extinction.

Additionally, we can think of text to speech conversion in Punjabi, purely for people who visit Punjabi speaking areas and cannot speak the native language for local communication.

Furthermore, we can develop something which does “speech to gesture” or “gesture to speech” for people with disabilities to help them understand better. We have already seen many news channels doing it, but this task can be done by machines as well.

7. References

[1] Zong, Z., and Hong, C., On application of natural language processing in machine translation, 2018.

[2] Mukherjee P., et al., Development of GUI for text-to-speech recognition using NLP, 2018.

[3] Dara, S., and Tumma, P., Feature extraction by using deep learning: A survey, 2018.

[4] Kim, J., and Kang, P., K-Wav2vec 2.0: Automatic Speech Recognition based on Joint Decoding of Graphemes and Syllables, 2021.

[5] Al Nazi, Z., and Tasmimul Huda, S. M., Byakto Speech: Real-time long speech synthesis with CNN, 2021.

[6] Fan, Z., Li, M., Zhou, S., and Xu, B., Exploring Wav2vec 2.0 on speaker verification and language identification, 2021.

[7] Pepino, L., Riera P., and Ferrer L., Emotion recognition using Wav2vec 2.0 embeddings, 2021.

[8] Patrick Von Platen, Blog on ‘Fine-tuning XLS-R for Multi-Lingual ASR with Huggingface Transformers’, 2021.

[9] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[10] Baevski, A., Schneider, S., and Auli, M. vq-wav2vec: Self-supervised learning of discrete speech representations, 2019.

[11] Baevski, A., Zhou, Y., Mohamed, A., Auli, M., and Collobert R., Wav2vec: Unsupervised Pre-training for speech recognition, 2019.

[12] Sadhu, S., et al., Wav2vec-C: A self-supervised model for speech representation learning, 2021.