Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 10, October 2021: 2273-2281

Research Article

Self-Supervised Model for Speech Tasks with Hugging Face Transformers

G.kumari [1], A.M.Sowjanya[2]

Abstract

For many years, speech recognition has been a focus of research. Automatic speech recognition (ASR) is the process for converting a speech signal into its corresponding sequence of words or other linguistic entities using algorithms implemented in a device. As our work and life are becoming integrated with mobile devices, such as tablets and smartphones (e.g., Amazon Alexa , Siri, Google Now, and Cortana), speech recognition technology has quickly become one of the most popular modes of communication.The arrival of this new trend is attributed to the significant progress made in several areas like high computing power and powerful deep learning models, leading to dramatically lower error rates in speech recognition systems. In this regard, our research is focused on reducing the error rate by using a self-supervised model for Speech Tasks. This paper presents the XLS-R model for multi-lingual speech representation learning based on wav2vec 2.0. XLS-R's new model learns basic speech units in order to answer a self-supervised task. The model is trained by predicting correct speech units for masked parts of the audio, while simultaneously learning what those units should be. The XLS-R model is fine-tuned by using Connectionist Temporal Classification (CTC), which is a technique used to train neural networks to solve sequence-to-sequence problems, such as automatic speech recognition (ASR) and handwriting recognition.We have used a common voice corpus in the Turkish language. This model performs well and the word error rate (WER) is significantly decreased.

Keywords:XLS-R,Automatic Speech Recognition, Connectionist Temporal Classification,Huggingface transformer,Wav2Vec2,training loss,validation data, Word Error Rate

[1]G.kumari, Department of Computer Science and Engineering, Andhra University college of Engineering(A),Visakhapatnam.

[2]A.M.Sowjanya, Department of Computer Science and Engineering, Andhra University college of Engineering(A),Visakhapatnam.

1. Introduction

In human-computer interaction, Automatic Speech Recognition (ASR) is quickly becoming mainstream. Whether composing a text message on WhatsApp, playing music, or even text-to-speech or speech-to-voice services with virtual personal assistants, most tools today offer a speech recognition option for several types of dictation activities. Even though many of the techniques have only recently gained popularity, deep learning has been applied to the sequence portion of the ASR task, replacing the HMM for many of the techniques and moving towards end-to-end models for speech recognition. Many modern methods, such as attention and RNN, have found their way into ASR since their emergence. With massive datasets, the incorporation of sequence-to-sequence architectures allows the models to learn the acoustic and linguistic dependencies directly from the data, resulting in higher quality.

Speech recognition, also known as speech-to-text, refers to a machine or program's ability to recognize words spoken aloud and convert them into readable text. The technology behind speech recognition is based on research in the fields of computer science, linguistics, and computer engineering. Many modern devices and text-focused programs include speech recognition functions to make device use easier or hands-free. Speech recognition is a technique for recognizing words in spoken language. Speech recognition systems process and interprets spoken words before converting them to text using sophisticated algorithms. Following these four steps, a computer algorithm converts the sound recorded by a microphone into written language that computers and humans can understand:

- analyze the audio
- break it into parts;
- Generate it to a computer-readable format.; and
- Apply an algorithm to match it to the best text representation.

Different speech patterns, speaking styles, languages, dialects, accents, and phrasings are trained into the software algorithms that process and organize audio into text. Speech recognition systems employ two types of models to meet these requirements:

- **Acoustic models**. These represent the interaction of linguistic units of speech and audio signals.
- **Language models**. In this case, sound sequences are matched with word sequences to identify words with similar sounds.

The use of speech recognition software in healthcare applications allows doctors to capture notes in real-time into medical records. Emotion recognition analyzes vocal characteristics to recognize what emotion is being expressed.

The speech recognition [1] field had been dominated by shallow architecture, namely Hidden Markov model (HMM) with each state characterized by a Gaussian Mixture Model (GMM) until the recent rise of deep learning. The next generation of speech recognition require solutions to many new technical challenges under diverse deployment environments to be successful. Researchers had expected for a long time that this would require complex and carefully engineered variations of GMM-HMMs [2] and acoustic features suitable for them. Audio and Speech processing breakthrough developments have been driven by transformers. With the release of the state-of-the-art Natural Language Processing library Transformers v4.30, Hugging Face has expanded its capabilities to incorporate automatic speech recognition with one of the leading methods developed by Facebook called the Wav2Vec2. Deep learning models have benefited from large amounts of labeled training data.

2. Literature Review

M. Ravanelli et al.[3] proposed a simplified Light Gated Recurrent Units(Li-GRU) architecture, and the standard Gated Recurrent Units (GRU)for speech recognition.The computational complexity is reduced by using this model. Titouan Parcollet et al.[4] incorporated multiple feature views into a quaternion-valued convolutional neural network (QCNN) for sequence-to-sequence mapping with the Connectionist Temporal Classification (CTC) model on the Timit benchmark dataset. The results obtained in terms of phoneme error rate (PER) while using fewer learning parameters were promising. Luo et al.[5] developed the Audio Feature Fusion-Attention based CNN and RNN (AFF-ACRNN)model which is a novel utterance-based deep neural network that combines CNNs and LSTMs to produce a set of representative features called Audio Sentiment Vector (ASV) that can be used to analyze sound utterances to find expressions of sentiment. Cai and co-authors [6] proposed a multimodal emotion recognition model from speech and text. The Bi-LSTM (bidirectional long short-term memory) network was implemented using textual features that are captured, and deep neural networks were used to classify the fusion features. Schneider et al.[7] discussed a novel application of unsupervised pre-training for speech recognition using wav2vec on the Wall Street Journal(WSJ) test set, they achieved 2.43% WER. Uddin et al.[8] proposed a robust method for recognizing emotions using audio speech as input for machine learning. Audio data are needed to learn emotion recognition systems independent of individuals and Mel-frequency cepstral coefficients (MFCC) are calculated as features. Alexei et al. [9] developed Vq-wav2vec to study discrete representations of audio samples using a self-supervised prediction task using BERT pretraining archives better results on Timit corpus. According to Ardila et al. [10], the Common Voice dataset is a large dataset trained on audio that contains multiple language annotations for speech recognition. Wang et.al [11] proposed a unified pre-training approach for learning speech representations with both labeled and unlabeled data(UniSpeech), in which supervised phonetic Connectionist Temporal Classification(CTC) learning and Multi-task self-supervised contrastive learning is performed with phonetically-aware self-response.Wu et.al [12] investigated performance-efficiency trade-offs in pre-trained models for automatic speech recognition (ASR) on wav2vec 2.0.

3. Research Methodology

Our study focuses on the multilingual setting by learning representations that generalize across languages on unlabeled data. The proposed approach extends the pretraining approach in which both contextualized speech representations and a discrete vocabulary of latent speech representations are jointly learned. With XLS-R's self-supervised pre-training[13], almost half a million hours of audio data were used in 128 languages, with sizes ranging from 300 million to two billion parameters. There is a significant difficulty in obtaining labeled data, particularly in the speech recognition domain, where it takes thousands of hours of transcription to achieve acceptable performance for more than 6,000 different languages.

Recent years have seen the emergence of self-supervised learning as a paradigm for learning general data representations from unlabeled examples and fine-tuning the model with labeled data. In computer vision, this has proven to be particularly successful for natural language processing. Using self-supervised learning, Wav2Vec2 can recognize speech in a broad range of languages and dialects by generating training sets from unlabeled data. The model accepts a raw speech signal in any language as input. Audio data is input to a multilayer 1-D Convolutional neural network that generates audio representations. Deep learning algorithms are used to select the degrees of freedom between these latent representations and a chart of the audio parameters. In this case, about half of the audio representations are masked as they pass through the transformer. This can be achieved by predicting these masked vectors based on output from the transformer. This is accomplished by using the contrastive loss function.

For downstream tasks like emotion recognition and speaker identification, the model is fine-tuned with labeled data after pretraining on unlabeled speech. In XLS-R, which is similar to BERT's [14] masked language model objective, feature vectors are randomly masked before being passed to a transformer network during self-supervised pre-training. On top of the pre-trained network, a single linear layer is added to fine-tune the model using labeled audio downstream tasks, such as speech recognition and translation.

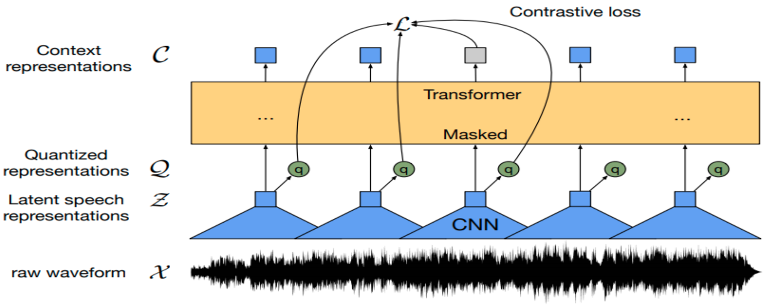


Fig 1: Adopted XLSR approach.[15]

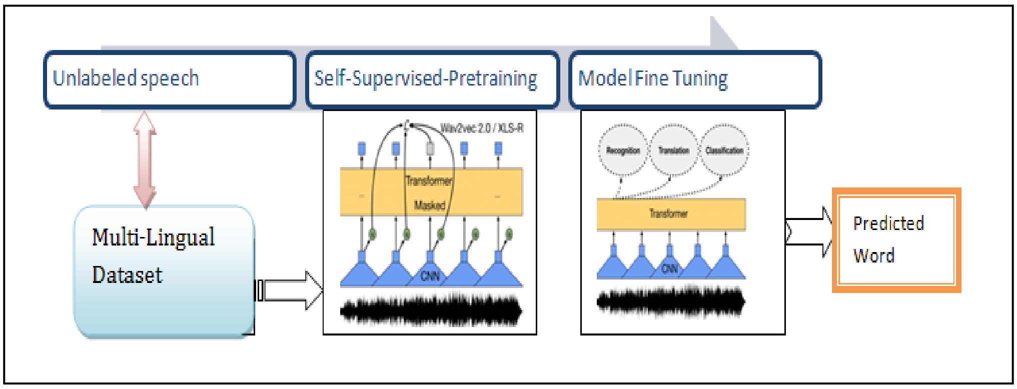


Fig 2: Self-Supervised Pretraining model on speech data.

Our model is fine-tuned with the low-resource ASR dataset of Common Voice[16], which contains only around 4 hours of validated training data. Connectionist Temporal Classification (CTC), an algorithm used for training neural networks for sequence-to-sequence problems such as ASR and handwriting recognition has been used to fine-tune XLS-R.

Some of the functions used in our model are listed below:

a) Wav2Vec2ForCTC: As the Wav2Vec2 model has been trained on 16 kHz audio, our input sample is resampled to 16 kHz. Finally, we tokenize the input and set several tensors for each input.

b) Wav2Vec2Tokenizer: Pre-training on unlabeled data can provide surprisingly good results for speech processing. The Wav2Vec2 system can improve automatic speech recognition in a much wider range of languages and domains with less annotation, yet still achieve outstanding results according to the State of Art.

c) PyTorch: Python Torch assists in the rapid prototyping of neural networks. Using touch audio, we downloaded and represented the dataset.

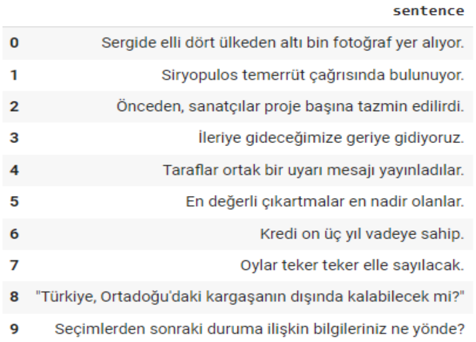
d) LibROSA: It is commonly used for digital signal processing (DSP) and feature extraction in audio analysis.

4. Experimental Setup

a) Dataset: The CommonVoice dataset contains over two thousand hours of reading speech in 38 languages.

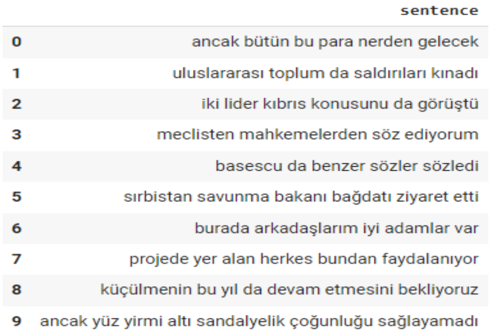
b) Preprocessing and Feature vector:ASR models can translate speech into text, so we need a feature extractor for converting the speech into the model's input format, e.g. a feature vector, and a tokenizer for converting the model's output format into text. As a result of this, transformers offer the XLS-R model with a tokenizer, called Wav2Vec2CTCTokenizer, and a feature extraction process, called Wav2Vec2FeatureExtractor.

A few sample transcriptions have been generated for common voice data observing what needs to be preprocessed shown in Fig 3. Preprocessing eliminates specific characters and normalizes the text to lower case to eliminate ambiguity shown in Fig 4.



	sentence
0	Sergide elli dört ülkeden altı bin fotoğraf yer alıyor.
1	Siryopulos temerrüt çağrısında bulunuyor.
2	Önceden, sanatçılar proje başına tazmin edilirdi.
3	İleriye gideceğimize geriye gidiyoruz.
4	Taraflar ortak bir uyarı mesajı yayınladılar.
5	En değerli çıkartmalar en nadir olanlar.
6	Kredi on üç yıl vadeye sahip.
7	Oylar teker teker elle sayılacak.
8	"Türkiye, Ortadoğu'daki kargaşanın dışında kalabilecek mi?"
9	Seçimlerden sonraki duruma ilişkin bilgileriniz ne yönde?

Fig 3: Random samples for Common voice Transcriptions



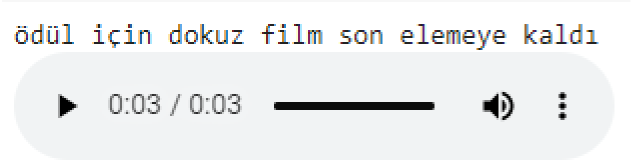
	sentence
0	ancak bütün bu para nerden gelecek
1	uluslararası toplum da saldırıları kınadı
2	iki lider kıbrıs konusunu da görüştü
3	meclisten mahkemelerden söz ediyorum
4	basescu da benzer sözler sözledi
5	sırbistan savunma bakanı bağdatı ziyaret etti
6	burada arkadaşlarım iyi adamlar var
7	projede yer alan herkes bundan faydalanıyor
8	küçülmenin bu yıl da devam etmesini bekliyoruz
9	ancak yüz yirmi altı sandalyelik çoğunluğu sağlayamadı

Fig 4: Removed special charaecters and normaized to lower case

A Wav2Vec2FeatureExtractor requires the following parameters:

- feature_size, sampling_rate, padding_value, do_normalize ,return_attention_mask.

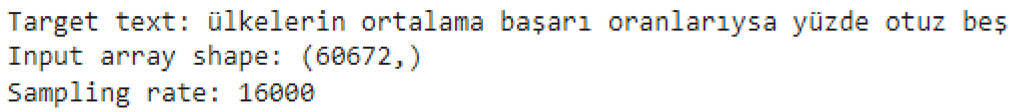
To better understand the dataset and verify that the audio was correctly loaded the below sample shown in Fig 5.



```
ödül için dokuz film son elemeye kaldı
```

Fig 5: Audio transcrptions generated on turkish language.

In Figure 6, the shape of the speech input, its transcription, and the corresponding sampling rate is displayed.



```
Target text: ülkelerin ortalama başarı oranlarıysa yüzde otuz beş
Input array shape: (60672,)
Sampling rate: 16000
```

Fig 6: Shape of the speech input and its transcriptions.

c) Training :

The input length is considerably longer than the output length for XLS-R, in contrast to other NLP models.As such the training batches are padded dynamically. Pretraining the XLS-R model and fine-tuning the parameters is necessary.

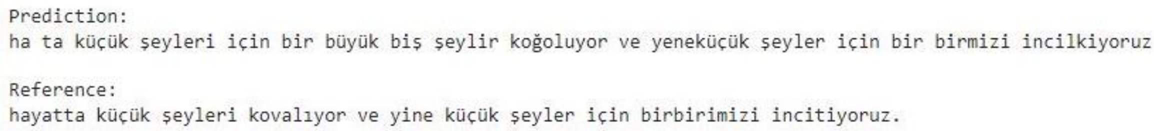
5. Results

On unlabeled data, we used a pretrained model trained on XLS-R with a set of Wav2Vec2 on 300M parameters. Word error rates (WER)[17] has been used as a metric for determining the quality of automatic speech recognition. When recognized speech is translated into text form, some words may be left out or mistranslated. Table 1 shows that training loss and validation loss are reduced as the model is trained giving a small word error rate.

Step	Training Loss	Validation Loss	Wer
400	3.76	0.625	0.69
800	0.37	0.405	0.456
1200	0.177	0.412	0.408
1600	0.122	0.399	0.394
2000	0.095	0.408	0.371
2400	0.073	0.394	0.351
2800	0.058	0.369	0.331
3200	0.045	0.359	0.324

Table 1: Performance measures on Transformers Model

Fig 7 shows that the final predicted words are fairly close to the transcriptions given by the pre-trained model for Voice Database.



```
Prediction:
ha ta küçük şeyleri için bir büyük biş şeylir koğoluyor ve yeneküçük şeyler için bir birmizi incilkiyoruz

Reference:
hayatta küçük şeyleri kovalıyor ve yine küçük şeyler için birbirimizi incitiyoruz.
```

Fig 7: Predicted transcriptions

6. Conclusion

This work is based on an XLS-R pre-trained model using common voice datasets. Our results demonstrate significant improvements over previous state-of-the-art results on speech recognition. Our method is equally effective for learning multi-lingual representations in an unsupervised fashion. We examined one model pre-trained on a language and were able to predict the model's performance using a word error rate that decreased after fine-tuning the transformers.

References

[1] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” vol. 10, 01 2010.

[2] Li, J., Deng, L., Gong, Y, Haeb-Umbach, R., 2014. An overview of noise-robust automatic speech recognition. IEEE/ACMTrans. Audio Speech Lang. Process. 22 (4), 745-777.

[3] M. Ravanelli, P. Brakel, M. Omologo and Y. Bengio, "Light Gated Recurrent Units for Speech Recognition," in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 2, no. 2, pp. 92-102, April 2018, doi: 10.1109/TETCI.2017.2762739.

[4] Titouan Parcollet, Ying Zhang, Mohamed Morchid, Chiheb Trabelsi, Georges Linarès, et al.. Quaternion Convolutional Neural Networks for End-to-End Automatic Speech Recognition. Interspeech 2018, Sep 2018, HYDERABAD, India. pp.22-26, ff10.21437/Interspeech.2018-1898ff.ffhal-02107611f. arXiv:1806.07789

[5] Luo, Ziqian et al. "Audio Sentiment Analysis by Heterogeneous Signal Features Learned from Utterance-Based Parallel Neural Network." *AffCon@AAAI* (2019). https://doi.org/10.29007/7mhj.

[6] Cai, Linqin & Hu, Yaxin & Dong, Jiangong & Zhou, Sitong. (2019). Audio-Textual Emotion Recognition Based on Improved Neural Networks. Mathematical Problems in Engineering. 2019. 1-9. 10.1155/2019/2593036.

[7] Schneider, Steffen & Baevski, Alexei & Collobert, Ronan & Auli, Michael. (2019). wav2vec: Unsupervised Pre-Training for Speech Recognition. 3465-3469. 10.21437/Interspeech.2019-1873.

[8] Uddin, Md. Zia & Nilsson, Erik. (2020). Emotion recognition using speech and neural structured learning to facilitate edge intelligence. Engineering Applications of Artificial Intelligence. 94. 103775. 10.1016/j.engappai.2020.103775.

[9] Alexei Baevski,Steffen Schneider, Michael Auli(2020). VQ-WAV2VEC: SELF-SUPERVISED LEARNING OF DISCRETE SPEECHREPRESENTATIONS.*P.No 1-12*, arXiv:1910.05453v3 .

[10] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., & Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. *LREC*. arXiv:1912.06670v2 [cs.CL] 5 Mar 2020.

[11] Wang, Chengyi & Wu, Yu & Qian, Yao & Kumatani, Kenichi & Liu, Shujie & Wei, Furu & Zeng, Michael & Huang, Xuedong. (2021). UniSpeech: Unified Speech Representation Learning with Labeled and Unlabeled Data. arXiv:2101.07597v2 [cs.CL]

[12] Wu, Felix & Kim, Kwangyoun & Pan, Jing & Han, Kyu & Weinberger, Kilian & Artzi, Yoav. (2021). Performance-Efficiency Trade- offs in Unsupervised Pre-training for Speech Recognition.

[13] Babu, Arun & Wang, Changhan & Tjandra, Andros & Lakhotia, Kushal & Xu, Qiantong & Goyal, Naman & Singh, Kritika & Platen, Patrick & Saraf, Yatharth & Pino, Juan & Baevski, Alexei & Conneau, Alexis & Auli, Michael. (2021). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale.

[14] Hourigan, T.,& Murray, L. (2010). Using blogs to help language students to develop reflective learning strategies: Towards a pedagogical framework. *Australalasian Journal of Educational Technology, 26*(2), 209-225.

[15] Conneau, Alexis & Baevski, Alexei & Collobert, Ronan & Auli, Michael. (2020). Unsupervised Cross-lingual Representation Learning for Speech Recognition. arXiv:2006.13979v2 [cs.CL] 15 Dec 2020.

[16] https://huggingface.co/datasets/common_voice

[17] Park, Youngja & Patwardhan, Siddharth & Visweswariah, Karthik & Gates, Stephen. (2008). An Empirical Analysis of Word Error Rate and Keyword Error Rate. 2070-2073. 10.21437/Interspeech.2008-537.

[18] Shukla, Rachit. (2020). Keywords Extraction and Sentiment Analysis using Automatic Speech Recognition.

[19] https://huggingface.co/blog/fine-tune-xlsr-wav2vec2

[20] Chen, Y.-C., Yang, S.-. wen ., Lee, C.-K., See, S., and Lee, H.-. yi ., "Speech Representation Learning Through Self-supervised Pretraining And Multi-task Finetuning", <i>arXiv e-prints</i>, 2021.