An Approach for Identification of Speaker using Deep Learning

Syeda Rabia Arshad[1], Syed Mujtaba Haider[2], Abdul Basit Mughal[2]

[1]Department of Computing (DoC), School of Electrical Engineering and Computer Science (SEECS),

National University of Sciences and Technology (NUST) Islamabad, Pakistan

[2]Department of Computer Science School of Engineering and Applied Sciences (SEAS),

Bahria University Islamabad, Pakistan

Email: sarshad.msds20seecs@seecs.edu.pk, haiderpak77@yahoo.com, sardarbasitmughal344@gmail.com

Corresponding Author: Syeda Rabia Arshad, sarshad.msds20seecs@seecs.edu.pk

Received: 29-05-2022; **Accepted**: 25-12-2022; **Published**: 11-01-2023

Abstract: The audio data is getting increased on daily basis across the world with the increase of telephonic conversations, video conferences, podcasts and voice notes. This study presents a mechanism for identification of a speaker in an audio file, which is based on the biometric features of human voice such as frequency, amplitude, and pitch. We proposed an unsupervised learning model which uses wav2vec 2.0 where the model learns speech representation with the dataset provided. We used Librispeech dataset in our research and we achieved our results at an error rate which is 1.8.

Key Words: Speaker Identification, Deep Learning, Automatic Speaker Recognition (ASR)

1. Introduction

Identification of a speaker involves a process to recognize the person from their voice. Given a handful of voice samples, associating voices with their respective speakers using speech recognition requires identification of a person through their pitch, amplitude, and other voice characteristics. A speaker's voice has some personal traits of that speaker like their accent, rhythm, unique vocal tract shape etc. Therefore, it is possible to identify a person's voice through the computer automatically, just like how humans can do it. These voice samples can be recorded through microphone or can be telephonic conversations. Irrespective of the language, speaker identification can work on any type of human voice. However, some letters or words are not present in all of the languages, but this research can still hold because the speech signals of the voice remain the same, regardless of the language the person is speaking. Speaker recognition has many real-world applications including biometric door unlocking in security-critical places, voice-based authentication of personal smart devices like laptops, cell phones etc., in transaction security of remote paying and bank trading, in forensics to investigate if a person is suspect to be guilty of a crime, or surveillance and automatic identity tagging.

Speaker identification deals with simply associating speaker with their respective voice biometrics, speaker verification is the process of authenticating the voice when a speaker claims to be some specific person, and speaker diarization is the process of partitioning or labelling the audio segments by each speaker. This paper will mainly focus on speaker identification, and in some cases, speaker verification as well. This research can later on serve as a basis for multi-speaker identification as well, where we identify multiple speakers in an audio. In this paper, we are going to implement some deep learning models to implement speaker identification task. Deep learning has some major advantages over other conventional methods, like its representation ability through which it is able to create highly abstract embedding features out of utterances, which helps in feature extraction from the voice samples.

2. Literature review

A. Wav2Vec

Wav2Vec is a recently developed speech recognition method that masks the input of speech and helps in solving a contrastive task which is described around a quantization of the underlying interpretations which are collectively learned. This approach has been expanded using two approaches i-e the unsupervised and self supervised approach.

Deep learning models often require massive quantity of data for training. But in most cases, the training data is unlabeled. The labeled data is more difficult to come across than the unlabeled data. Which sometimes makes it hard to train the neural network as the training depends on the quantity and quality of the data. This sometimes result in reduced accuracy of the model. The self supervised learning has a theory to learn some common representations of the data from unlabeled samples of the data and to fine-tune the model to labeled data. This methodology revealed great success in Natural Language Processing and is also showing some promising results when applied to Computer Vision. The wave2vec 2.0 is a self-supervised model which is used for learning representation from the raw audio files and then encodes the audio through a multiple layer Convolutional Neural Network (CNN) which then employ further masks the spans of resulting latent speech representations. This is similar to masked language modeling where the purpose is to identify the language. The latent representations are then forwarded to a transformer network for evolving contextual representations and after that the model is trained through a contrastive task where the actual latent is to be recognized from distractors. During the training, the distinct units are learnt via gumbel softmax for representing the latent representations in contrastive tasks and after pre-training on the unlabeled data, fine-tuning is completed on the labeled data along with a Connectionist Temporal Classification (CTC) loss.[1]

The other approach of wav2vec is with unsupervised learning. Which is often known as wav2vec-U, short for wav2vec unsupervised. Most of the existing systems of speaker identification require labeled data during the training which limits the technology to the languages spoken around the globe. It is difficult to find feature-rich, labelled training dataset for each of the languages. As the name suggests, wav2vec-U model is used to train the systems by using unlabeled data. This framework leverages the self-supervised learning representations from wav2vec 2.0 to insert audio and to divide the audio into units by applying k-mean clustering technique. In this method, first of all the simple representations of audio are learned with wave2vec 2.0 on unlabeled audio, clusters in the representations are identified with simple k-means clustering technique and the next segment representations are learned by mean pooling the representations of wav2vec. Now this is sent as an input to the generator which outputs a phenome series. This is sent as an input to the discriminator, which is like the unlabeled text, for performing adversarial training.[2] The experimental findings show the capability of this model for a number of settings and languages.

B. Iterative Psuedo-Labeling (IPL)

Pseudo-Labelling recently has shown some great advancement in end-to-end speech recognition. Recent research has shifted to self- and semi-supervised learning algorithms to better utilize the effectiveness of unlabeled data. Iterative Pseudo-labeling is a semi-supervised learning model which performs repetitions of pseudo-labelling on an unlabeled data. IPL fine-tuned the existing model at each iteration used a combination of a proportion of unlabeled and labeled data. The main elements of IPL are decoding and data augmentation with a language model. This is a straight forward and a simple technique that can further scaled to huge unlabeled datasets and it can boost the performance in limited resource environments.[3]

C. Cross-lingual representation learning

Cross-lingual representation aims to learn representations generated from other languages to gain an improvement in performance. Unsupervised or pretrained representation learning does not require labeled data. It can learn from word embedding from unlabeled data as well. In this work, fine-tune of the transformer part is done instead of the parameters.[4] An approach based on vq-wav2vec was proposed in [5]. This method focused to learn discrete representations of audio segments by using self-supervised wav2vec.

An unsupervised mining of latent interpretations of speech by applying autoencoding neural networks to audio was proposed in [6]. This study aimed to discover a representation which can obtain high-level semantic content from phoneme identities. In this study, the researchers compared three variants which contain a Gaussian Variational Auto Encoder, a simple dimensionality reduction bottleneck and a distinct Vector Quantized VAE (VQ-VAE). The study focused on analysis of the the quality of learned representations in terms of the capacity to find phonetic content, autonomy of the speaker, and the expertise to recreate particular spectrogram frames accurately.

A problem agnostic speech encoder (PASE) was proposed in [7] which is a multi-task self-supervised framework for robust speech recognition. The PASE architecture is developed on a convolutional encoder paired with a quasi-recurrent neural network (QRNN) layer, an online speech deformation module and a set of personnel solving self-supervised challenges.

PASE is gone out to be considerably better standard acoustic characteristics on various speech identification tasks and showing further enhancements in end to end optimization with the objective acoustic model. PASE also offers an extraordinary amount of transfer.

Masked Predictive Coding (MPC) method was applied on unsupervised pre-trained data with transformer model [8]. MPC utilizes masked language model like structure for recognition of speech based on transformer models. The MPC improved the pre-trained model and its performance was much enhanced. Noisy student training model was used for automatic recognition of speech [9]. The study incorporated the SpecAugment, language model combination and sub-modular sampling into the noisy student training stream to change it for classification of speech. Furthermore, the researchers brought a filter for normalization which helped in gradual self-training. The gradational technique does not show much influence the high-supervised performance task Libri Speech/Libri Light, however it tends to be useful for the low-supervised-performance task Libri Speech 100-860. The researchers introduced different components in their work which were combined and matched to improve the performance of existing studies.

3. Methodology

A. The Model

For extracting the representations from the audio, we first setup a multi-layered convolutional network that takes raw audio file as input and then it extracts the latent speech representations from it which is then fed to the transformer to find the information from the whole sequence rather than for each time stamp. For creating a network that builds representation over contiguous speech representations, we created a model that consists of encoder – transformer – quantization module.

The feature encoder contains multiple blocks that comprise a temporal convolutional network which is followed by a normalization and Gaussian Error Linear Unit (GELU) activation function. The audio file which is taken as as input to the encoder is normalized by using unit variance and zero mean. Since right now we are encoding the representations on each time stamp, the stride shift of the convolutional layer of our encoder determines how many time steps we are moving forward which will be sent as input to the transformer.

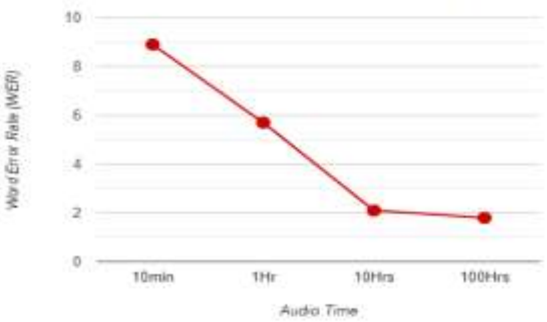
In developing the context representations of the transformer, the encoder convolutional network’s output is sent to a context network which is linked with the transformer architecture. This model uses a convolutional layer as an alternative of fixed position embedding. Convolutional layer encodes absolute positional information like which acts as relative position embedding. We add the output of the convolutional layer followed by a GELU activation to the inputs and then we apply normalization.

The output of the feature encoder discretized by quantization module to a finite set of speech representations by product quantization for self-supervised training.

B. The Dataset

The dataset that we are using is Librispeech dataset which is a large corpus of almost 1000 hours of recording in English speech. Our model learns the context representation of each time stampand self-attention module extracts the dependencies of the entire sequence end-to-end.

C. Training the model



For training of the model we masked a little proportion of time stamps in the feature of latent encoder space. We performed masking on a small proportion of the encoder outputs before sending them as input to the contextual network and substitute them with a pre-trained feature vector shared to all masked time stamps. To mask the latent representations output by the feature encoder, we simply obtained random samples without the replacing the proportion of all time stamps to be indices at start and then mask the subsequent M consecutive time steps from each of the sampled index, spans may overlap on each other.

During the training, we learned representations of audio by completing a contrastive task which is needed to identify the actual quantized latent speech representation for an already masked time stamp within a group of distractors. This is further added to a codebook diversity loss which allows the model to use the entries of codebook equally. Now we go for calculation of the contrastive loss of the model. The provided context network output, centred over the masked time stamp, the model needs to identify the actual quantized speech representation in a set of quantized representations which includes K distractors and t time-stamps. Distractors are equally sampled from other masked time stamps of the same utterance.

We optimize the model by lowering the contrastive loss, masking the channels and time during the training which avoids overfitting and helps in improving the final error, especially dealing with Libri-light subsets with a few labelled data. After training of the model, we applied fine-tuning on the learned representations on labelled data and we brought a a random output layer on the top of the Transformer to identify characters (Librispeech /Libri-light). We used Adam optimizer and a tri-state rate schedule optimizers in our research. In these optimizers, the rate of learning is updated for the initial 10% of the updates, kept same for the next of 40% and then gradually decomposition for the remaining.

We used n-gram language model in our research, where the value of n is 4. The reason for choosing 4-gram language model is that the data that our model has been trained on is very large. Apart from the pre-training, the fine tuning of the data is done on a huge corpus of LibriSpeech dataset, so for capturing most of the words in n-gram, we have to increase the value of n. So, it should be more than bigrams of trigrams.

4. Results

The results of the model are evaluated by Word Error Rate (WER) where the overall error to recognize words is considered to be a fault of the Speaker Identification model. The model is tested on the 10 minutes of LibriSpeech LV-60k dataset. The model achieved the word error rate (WER) of 18. This result was achieved through our approach on Librispeech benchmark. Moreover, we also compared the performance of previous related works in this field which were supervised, semi-supervised and unsupervised.

Table 1: Comparison of models and their word error rates(WER)

Model	Word Error Rate (WER)
Supervised	2.1
Semi-Supervised	2.0
Unsupervised (Our Research)	1.8

5. Analysis

The speaker identification model proposed in this research work performs good on the unsupervised data. However, some of the pre-training of the model is done on the labelled data in a supervised / semi-supervised fashion. The Librispeech dataset contains 10 min audio recordings to up-to 100hrs of audio recordings which helped in pre-training the model. Overall, this dataset was good for our task but we need more precise dataset with specific persons labelled along with their audio file. If we have more instances of specific people, labelled with their names then it will be helpful in the fine-tuning of our model. In this way, the model will be better able to recognize the speaker if their sample voice audio is played.

6. Conclusion and Future Work

In this work, we proposed an approach to identify the speaker based on their biometric features of voice like pitch, amplitude etc. Since this model focuses on the audio features, this model is independent of the language a person is speaking and can identify the speaker regardless of the language they are speaking. Our model achieved the Word Error Rate (WER) of 1.8.

In the future, we would want to enhance the features of this model to multi-speaker identification where the model can identify each speaker from and audio file where more than one people are speaking. And it will be able to tell the time duration from an audio file where the person’s voice is detected.

References

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael AuliJ. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, NeurIPS 2020 .

[2] Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, Michael Auli. Unsupervised Speech Recognition.

[3] Qiantong Xu, Tatiana Likhomanenko, Jacob Kahn, Awni Hannun, Gabriel Synnaeve, Ronan Collobert. ITERATIVE PSEUDO-LABELING FOR SPEECH RECOGNITION.

[4] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, Michael Auli. UNSUPERVISED CROSS-LINGUAL REPRESENTATION LEARNING FOR SPEECH RECOGNITION

[5] A. Baevski, S. Schneider, and M. Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. ICLR 2020.

[6] Chorowski, Jan, et al. "Unsupervised speech representation learning using wavenet autoencoders." IEEE/ACM transactions on audio, speech, and language processing (2019).

[7] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio. Multi-task self-supervised learning for robust speech recognition. arXiv, 2020.

[8] D. Jiang, X. Lei, W. Li, N. Luo, Y. Hu, W. Zou, and X. Li. Improving transformer-based speech recognition using unsupervised pre-training. arXiv, abs/1910.09932, 2019.

[9] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le. Improved noisy student training for automatic speech recognition. arXiv, abs/2005.09629, 2020.

Author Profile

1. **Syeda Rabia Arshad** is a research scholar currently pursuing her studies in Master of Science in Data Science at Department of Computing of National University of Science and Technology (NUST) Islamabad Pakistan. She holds a bachelor degree in Software Engineering from Bahria University Islamabad Pakistan. She has a vast experience in the field of natural language processing and deep learning.

2. **Syed Mujtaba Haider** is a graduate of Master of Science in Data Science from Bahria University Islamabad Pakistan. His area of research is text processing, computer vision and deep learning. He is currently working in a government organization.

3. **Abdul Basit Mughal** is a graduate of Master of Science in Data Science from Bahria University Islamabad Pakistan. His area of research is neural networks, computer vision and deep learning. He is currently working on freelance research projects.