Fake Speech Detection

(Conformer Model)

Aanshu Maurya

Computer Science and Engeering
Under the guidence of Mr. Shashank Yadav

04-06-2024

Outline

Introduction

Problem Statement

Dataset

Data Augumentation

Feature Extraction

Visualisation

Model

Training

Conclusion

Introduction

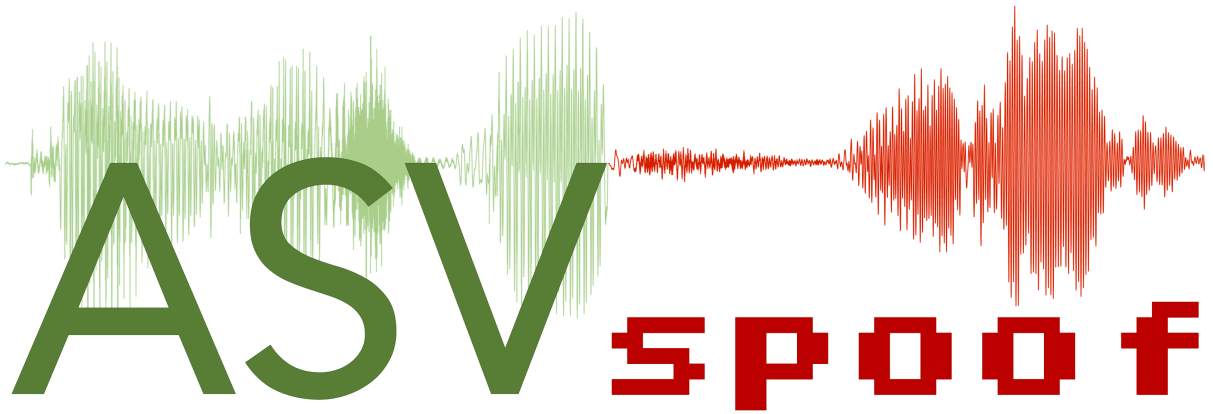
The challenge of automatically determining whether a particular voice sample is synthetic or authentic is known as fake speech detection. The importance of this problem has increased recently with the development of deep learning based voice synthesis technology. Google AI created the conformer deep learning model architecture for voice recognition. It is a powerful tool for collecting both local and global interdependence in audio data because it combines self-attention processes with convolutional neural networks (CNNs). A deep learning model architecture called Conformer was first created for automated speech recognition (ASR). Conformer achieves state-of-the-art performance on a range of ASR tasks by combining the advantages of self-attention mechanisms with convolutional neural networks (CNNs).[

Problem Statement

Creating algorithms and systems that can recognize and discriminate between real speech and artificially created speech is the task of detecting fake speech.
The difficulty is in identifying minute details and distortions in audio that set real speech apart from artificial or modified speech produced by sophisticated methods like voice conversion, deepfakes, or speech synthesis technology

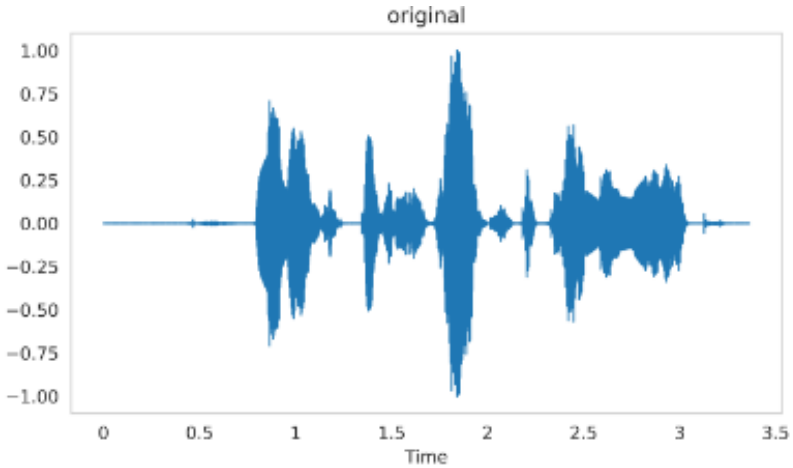
Dataset

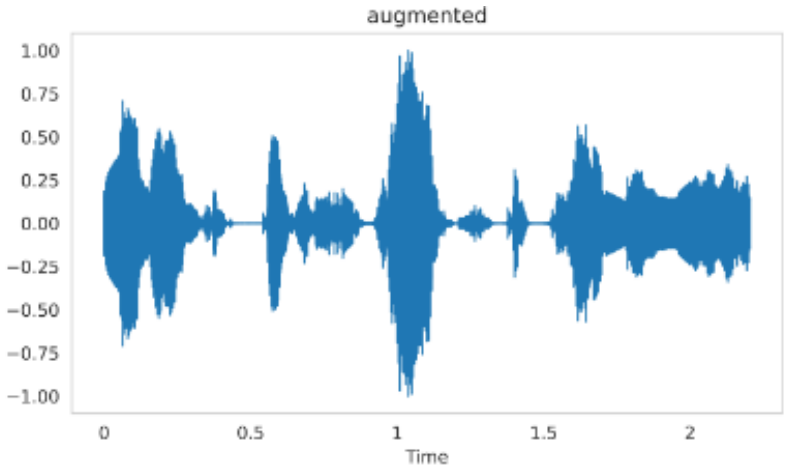
- `ASVspoof2019 Dataset :-` The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database.
  It Contains a 363k files of .flacfiles
- `ASVspoof2019 tfrecordDataset:-` It is simple Subset of the dataset.
  It contains a 30 files of .tfrecformat.



Data Augmentation

- `AudioAug:-`
- Random Noise
- Random TimeShift
- Random Crop
- `SpecAug:`
- Random TimeMask
- Random FreqMask
- Random CutMix



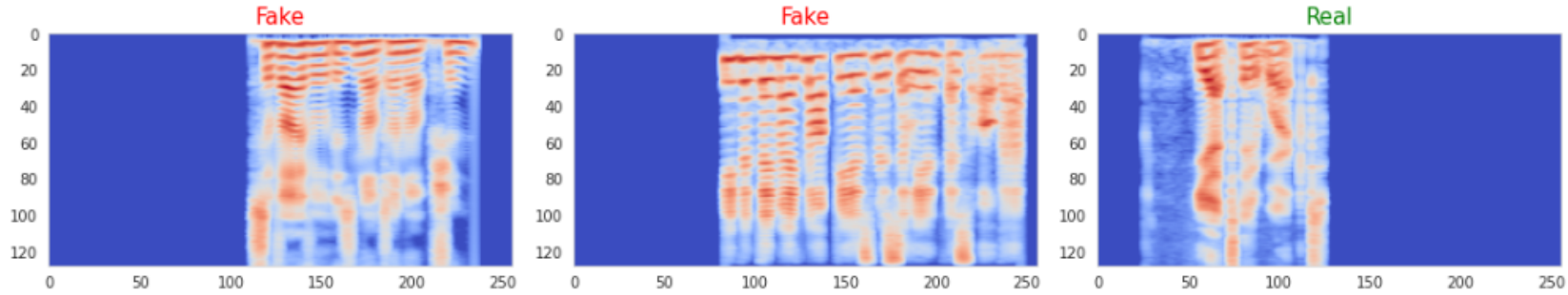


Feature Extraction

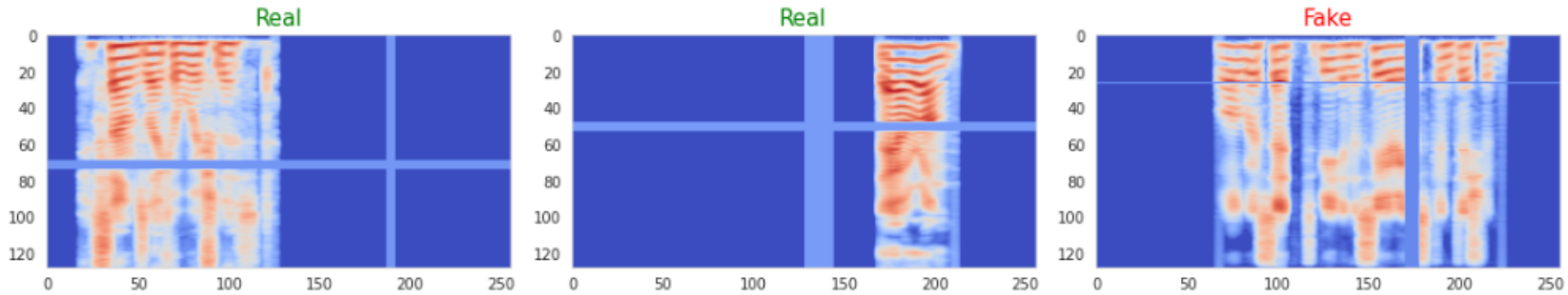
- We'll feed Mel-Spectrogram feature to our model as input.
- Spectrogram feature is extracted from audio sample using Short Time Fourier Transform (STFT).
- The fast Fourier transform is a powerful tool that allows us to analyze the frequency content of a signal.
- We can simply consider Mel-Spectrogram as 2D image which x-axis represents time and y-axis represents frequency information

Visulisation

- **spectrogram and its associate label.**
- Without Augmentation

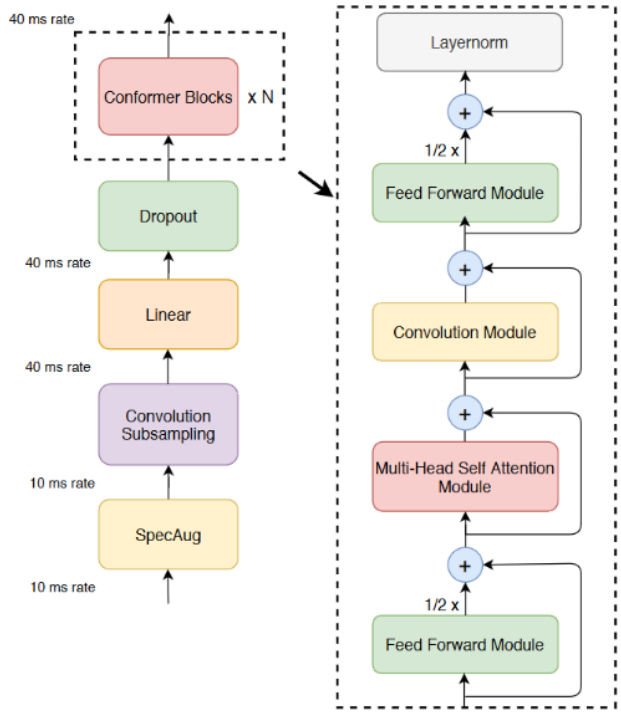


- With Augmentation



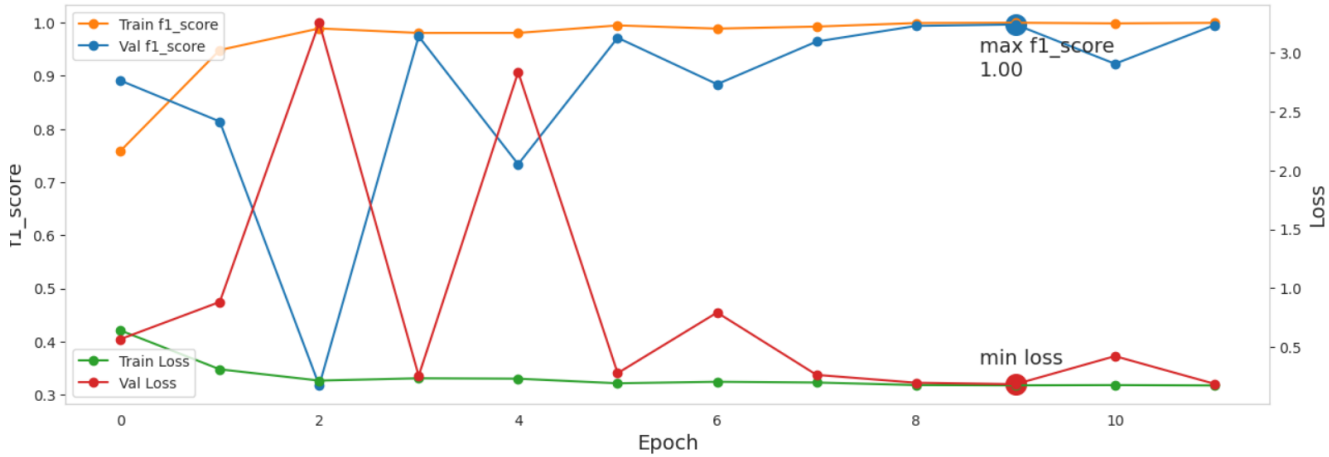
Model (Conformer)

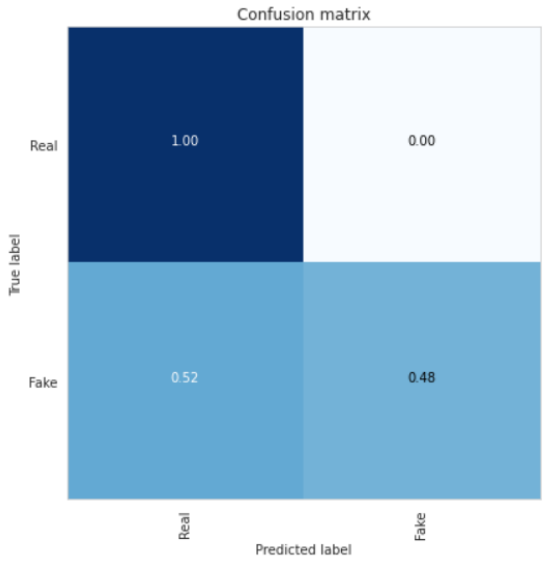
- A Neural net for speech recognition that was published by Google Brain in 2020.
- We used a transfer learning concept to train on the data.
- Here are the model architecture and Trainable parameters



Training And Matrix

- Training Accuracy: 99.67 & Testing Accuracy : 73.87
- Training Precision : 99.55 & Testing Precision :1.00
- Training Recall : 99.80 & Testing Recall: 47.75





Conclusion And Future Scope

- In the battle against audio manipulation, Conformers has become heroes.
- Studies on datasets like ASVspoof2019 showcase their impressive accuracy.
- They can be used as a Ensumble Learning as well as other techniques for the Betterment of result.
- Currenlty, It is in research field and widely used in speak verification system.