Report paper

Fake Speech Detection

KCS753

Submitted By

Aanshu Maurya

2008390100001

Email

2008390100001@reck.ac.in

Supervisor Name

Mr. Shashank yadav

Assistant Professor

Department of Computer Science and Engineering

Rajkiya Engineering College, Kannauj

Uttar Pradesh-209732

1 Objective

The main objective of this project to provide a Fake Speech Detection(Conformer) by using audio with the help of tensorflow.".

2 Introduction

The challenge of automatically determining whether a particular voice sample is synthetic or authentic is known as fake speech detection. [6] The importance of this problem has increased recently with the development of deep learning-based voice synthesis technology. False speech can be used to impersonate others or utilize extortion, in addition to propagating propaganda, hoaxes, and false information. [7]
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). [4]
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 [3] advantages of self-attention mechanisms with convolutional neural networks (CNNs).
This paper aims to assess Conformer's efficacy in identifying phony speech.

According to recent studies, Conformer is also a useful tool for detecting fraudulent speech. Speech audio may be utilized to teach Conformer discriminative characteristics that allow it to differentiate between actual and phony speech.

3 Related Work

Research on fake voice detection is expanding quickly, and new methods and models are constantly being put out. Conformer models have garnered increasing attention in the field of phony voice detection in recent years. [5] Transformer models that are particularly made for audio processing are known as conformer models. For a range of audio tasks, including as language identification, speaker identification, and voice recognition, conformer models have proven to be successful.

In 2020, Google AI researchers presented one of the first experiments to employ Conformer models for phony voice detection. When the researchers ran Conformer models on a collection of artificial and real speech samples, they fared better than other cutting-edge models. [2] In a different study, University of Southern California researchers created a method for identifying fraudulent speech in social media videos using Conformer models. Over 90% accuracy was attained by the algorithm on a dataset consisting of both fake and real social media videos. [1]

The fact that synthetic speech is always changing presents one of the biggest obstacles to false speech identification. There is constant development of new synthesis and voice conversion models, and these models are getting more complex. The fact that phony speech may be blended in with actual speech to evade detection presents another difficulty. We call this spliced speech. The current methods for detecting false speech are still in the early stages of research and cannot yet accurately identify speech that has been spliced.

4 Timeline

- August-
  a) Reviewing Research Papers and Previous work done.
  b) Collection of the dataset.

- September- Working on the model.

- December- Documentation on the completed model along with Report.

5 Platform

Virtual - Kaggle and Jupyter Notebook.

6 Dataset and performance metrics

ASVspoof 2019 Dataset:-

This is the biggest and most complete fake speech dataset currently accessible. Over 10,000 real speech samples and over 40,000 artificial speech samples may be found in the collection. A number of various methods, such as text-to-speech synthesis, voice conversion, and replay assaults, were used to create the false speech samples.

We'll compute Accuracy, Precision, Recall and F1_Score for both valid and test data. Finally, we'l plot confusion_matrix to get better insight about model's performance regarding FP and FN.

References

[1] Teck Kai Chan and Cheng Siong Chin. Lightweight convolutional-iconformer for sound event detection. *IEEE Transactions on Artificial Intelligence*, 2022.

[2] Changfeng Gao, Gaofeng Cheng, Ta Li, Pengyuan Zhang, and Yonghong Yan. Self-supervised pre-training for attention-based encoder-decoder asr model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1763–1774, 2022.

[3] Igor Gitman, Vitaly Lavrukhin, Aleksandr Laptev, and Boris Ginsburg. Confidence-based ensembles of end-to-end speech recognition models. *arXiv preprint arXiv:2306.15824*, 2023.

[4] Jiahong Li, Chenda Li, Yifei Wu, and Yanmin Qian. Robust audio-visual asr with unified cross-modal attention. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[5] Yunpeng Li, Qinya Zhang, and Guanyu Li. Improving tibetan end-to-end speech recognition with transfer learning. In *Journal of Physics: Conference Series*, volume 2560, page 012050. IOP Publishing, 2023.

[6] Bao Thang Ta, Tung Lam Nguyen, Dinh Son Dang, Dang Linh Le, and Van Hai Do. A multi-task conformer for spoofing aware speaker verification. In *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE)*, pages 306–310, 2022.

[7] Qing Wang, Jun Du, Hua-Xin Wu, Jia Pan, Feng Ma, and Chin-Hui Lee. A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1251–1264, 2023.