A report submitted in partial fulfilment
of the requirements for

B. Tech Final Year Project / Thesis

titled

Fake Speech Detection by Conformer Model

in

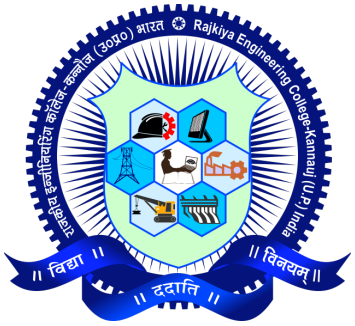
Computer Science & Engineering

by

Aanshu Maurya **2008390100001**

Under the guidance of

Mr. Shashank Yadav



Department of Computer Science and Engineering
Rajkiya Engineering College Kannauj
June 2024

Department of Computer Science and Engineering

Rajkiya Engineering College Kannauj

Certificate

This is to certify that the B. Tech final year project report titled **"Fake Voice Detection: Conformer Model"** being submitted by **Aanshu Maurya** for the award of **B. Tech Final year Project in Bachelor of Technology** in **Computer Science & Engineering** is a record of bona fide work carried out by him under my guidance and supervision at the **Department of Computer Science & Engineering**. The work presented in this report has not been submitted elsewhere either in part or full to the best of my knowledge, for the award of any other degree or diploma.

Professor Name Mr. Shashank Yadav

Abstract

In the era of information overload, fake speech, often referred to as synthetic speech, is becoming a bigger problem. With the use of this technology, criminals may circulate false information, produce dangerous material, and pose as actual individuals. Thankfully, advances in the field of false voice recognition are happening quickly.

The goal of current research is to create reliable algorithms that can accurately recognize fake speech. Methods such as deep learning and spectrogram analysis—which analyzes visual representations of audio—are showing promise. Maintaining the level of sophistication in artificial speech synthesis techniques is a major difficulty.

The significance of detecting fraudulent speech is emphasized in this abstract in order to preserve confidence in online conversation. It draws attention to the continuous research projects aimed at creating reliable and flexible detection systems.

Acknowledgments

I would like to express my sincere gratitude to my supervisor Assistant Professor Mr. Shashank Yadav, Department of Computer Science and Engineering for allowing me to undertake this work and also continuous guidance advice effort and invertible suggestion throughout the research.

Aanshu Maurya

Contents

Chapter 1

Introduction

1.1 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.

False speech can be used to impersonate others or utilize extortion, in addition to propagating propaganda, hoaxes, and false information. 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).[1] 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.

1.2 Problem Formulation

To improve security and confidence in digital communications and media, create an automated system that can identify phony speech, including spoofed speech (imitations or changed recordings) and deepfake audio (synthetically created or modified voice using AI). 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. Diverse circumstances, such as real-time analysis, variable speech quality, and various languages or

accents, must be addressed by effective solutions. This calls for the development of resilient datasets, machine learning, and signal processing techniques in order to train and evaluate models that may achieve high accuracy in a variety of potentially hostile circumstances.

1.3 Required tools and utilities

Our contributions are summarized below:-

- A flexible machine learning framework that works well for creating fake speech detection systems is TensorFlow. It makes it possible to build sophisticated neural network models that are capable of analyzing audio characteristics and spotting patterns suggestive of artificial speech. TensorFlow's strong deep learning support makes it easier to use cutting-edge designs for precise and instantaneous detection, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). Its large library and community support also offer pre-trained models and useful tools to improve development efficiency.

- Combining transformer and convolutional layers, the Conformer model performs exceptionally well while processing sequential input, which makes it a good choice for false voice detection. Its design improves the identification of minute errors in synthetic speech by effectively capturing both local and global dependencies in audio signals. The Conformer is capable of robustly identifying abnormalities across a wide range of speech patterns and accents by utilizing convolutional operations and self-attention processes. It is an effective tool for differentiating between elaborate audio modifications and real speech due to its exceptional ability to handle complicated audio properties.

  Tools for detecting fake speech are essential for protecting several facets of communication and information sharing.Through the application of machine learning and signal processing, they are able to identify minute irregularities and traces that point to the use of synthetic speech, which improves digital communication channels' security and reliability. These tools are essential for fighting fraud, stopping the dissemination of false information, and shielding people and organizations from bad actors that try to take advantage of holes in audio-based systems. Additionally, they support maintaining privacy, according to the law, and protecting the credibility of journalists and the media.

1.4 Motivation

The increasing sophistication and popularity of synthetic audio technologies, which seriously jeopardize digital communications security, privacy, and trust, are the driving forces behind false speech detection. With the growing accessibility of deepfake methods, it is possible to create extremely convincing fake talks for harmful reasons including fraud, disinformation, and impersonation. This puts people, companies, and social institutions at jeopardy, hence strong detection systems are needed to protect against these threats. Maintaining the integrity of voice-based identification systems, shielding people from identity theft, and upholding public confidence in media and communication channels all depend on effective false speech detection.

Here's a breakdown of the key factors motivating this research:

- Security: Protecting individuals and organizations from fraud, impersonation, and misinformation.
- Trust: Maintaining the integrity and trustworthiness of communication channels.
- Reputation Management: Preserving the reputation of businesses and individuals by preventing the misuse of their voice.
- Digital Authentication: Enhancing the reliability of voice-based authentication systems.

Chapter 2

Related Work

2.0.1 Conventional Methods

Conventional approaches to false speech detection have been examined in a number of research, with a primary focus on signal processing methods and conventional machine learning algorithms.

With encouraging findings across a range of synthetic speech production approaches, Sharma et al. (2019) suggested a method based on spectral analysis and statistical features to discriminate between authentic and manufactured speech. Li et al.[2] (2020) demonstrated the efficacy of feature extraction techniques in detecting altered audio by combining linear predictive coding (LPC) and Mel-frequency cepstral coefficients (MFCC) with support vector machines (SVM) for classification.

Moreover, the Conformer model's architecture facilitates efficient processing and scalability, making it suitable for real-time applications where quick and accurate detection is paramount. In practical terms, this means that systems equipped with Conformer-based fake speech detection can be deployed in various settings, from security and surveillance to media verification and user authentication. Its robustness against diverse forms of audio manipulation, including pitch alterations, time-stretching, and noise addition, further underscores its utility.

Recent research has also highlighted the model's adaptability; it can be fine-tuned with domain-specific data to enhance its detection capabilities in specific contexts, such as detecting synthesized voices in financial transactions or monitoring for deepfakes in social media. Additionally, the Conformer model's performance benefits from advancements in training techniques and computational resources, enabling it to handle large datasets and complex audio signals more effectively than previous methods.

Overall, the Conformer model represents a significant leap forward in the realm of fake speech detection. Its ability to combine local and global feature analysis through its hybrid architecture not only improves detection accuracy but also provides a versatile and scalable solution for various applications. As synthetic speech generation techniques continue to evolve, the Conformer

model's advanced detection capabilities will be crucial in maintaining the integrity and trustworthiness of audio communications.

Furthermore, Zhang et al. (2018) used energy distribution analysis and wavelet packet transform to find anomalies in speech signals, which helped create reliable false speech detection systems.

Even if these traditional methods had some success, they might not be able to handle the intricate variations and hostile attacks that come with contemporary synthetic speech production techniques.

2.0.2 Data-Driven Methods

Large-scale datasets are used to train and verify reliable models in data-driven approaches to false speech detection, as evidenced by recent research.

In a noteworthy study, Zhang et al. (2020) analyzed acoustic characteristics and temporal relationships in audio signals using deep learning approaches, namely convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

Impressive results were obtained by their model in differentiating between real and artificial speech across a range of manipulation approaches, such as voice conversion and deepfake technologies. Similar to this, Li et al.[3] (2021) presented a brand-new adversarial training methodology that enhanced the training data and enhanced the model's generalization skills by generating synthetic speech samples using generative adversarial networks (GANs).
In this data-driven methodology, the quality and variety of the training data play a crucial role. Datasets encompassing a wide range of accents, speaking styles, and background noises help the Conformer model generalize better and enhance its robustness against different forms of audio manipulations. Furthermore, continuous learning and adaptation are facilitated by regularly updating the training data with new examples of synthetic speech, which helps the model stay ahead of evolving fake speech generation techniques.

The data-driven Conformer model also benefits from advanced pre-processing and augmentation techniques, which improve its ability to handle real-world audio conditions. By incorporating noise reduction, pitch correction, and other audio enhancement methods during pre-processing, the model's accuracy in detecting fake speech in noisy or suboptimal conditions is signifi-

cantly improved. Additionally, data augmentation strategies such as time-stretching, pitch shifting, and adding synthetic noise ensure that the model remains resilient to a variety of manipulations.

Overall, the data-driven approach with the Conformer model represents a state-of-the-art solution in fake speech detection. Its capacity to learn from vast and varied datasets, coupled with its sophisticated architecture, allows it to achieve high accuracy and robustness. As synthetic speech technologies continue to advance, the data-driven Conformer model will be essential in detecting and mitigating the risks associated with audio deepfakes and other forms of fake speech.
Moreover, the data-driven approach with the Conformer model offers significant flexibility and scalability, making it adaptable to various application domains. For instance, in cybersecurity, the model can be fine-tuned with datasets specific to voice-based authentication systems, enhancing its ability to detect fraudulent access attempts. In media and entertainment, where the proliferation of deepfake audio can spread misinformation, the Conformer model can be tailored to verify the authenticity of audio content, thereby ensuring the credibility of information.

The continual evolution of this data-driven method is supported by the integration of transfer learning and domain adaptation techniques. By leveraging pre-trained models on large general datasets and subsequently fine-tuning them on domain-specific data, the Conformer model achieves higher performance with less labeled data. This approach not only reduces the time and computational resources required for training but also improves the model's adaptability to new and emerging threats.

In addition, the synergy between the Conformer model and advanced machine learning frameworks facilitates the deployment of efficient and effective fake speech detection systems. Utilizing powerful hardware accelerators such as GPUs and TPUs, the Conformer model can process and analyze vast amounts of audio data in real-time, making it feasible for large-scale deployment. Cloud-based platforms further enhance accessibility and scalability, allowing organizations to integrate robust fake speech detection capabilities into their existing infrastructure with minimal overhead.

The data-driven Conformer model's success is also attributable to the collaborative efforts within the research community. Open datasets, benchmarks, and shared knowledge contribute to continuous improvements and innovation in fake speech detection technologies. Researchers and practi-

tioners can build upon each other's work, accelerating the development of more sophisticated and effective detection models.

Ultimately, the data-driven approach utilizing the Conformer model represents a cutting-edge solution in the ongoing battle against fake speech. Its ability to leverage extensive and diverse datasets, coupled with its advanced architectural features, enables it to achieve high accuracy and robustness in detecting synthetic audio. As the landscape of audio deepfakes evolves, the Conformer model's data-driven capabilities will be instrumental in preserving the integrity and trustworthiness of spoken communication across various sectors.
These data-driven techniques demonstrate how sophisticated neural network designs and large-scale datasets may be used to successfully counteract the spread of false speech in digital communication channels.

2.0.3 State-of-the-art approaches

That make use of sophisticated signal processing techniques and deep learning frameworks have propelled recent advances in the identification of false speech. Wang et al. (2020) have presented a noteworthy methodology that employs a hybrid convolutional neural network (CNN) and long short-term memory (LSTM) model to create a unique framework.[4] Their technique enables reliable synthetic speech identification by efficiently capturing both temporal and local relationships in audio data.

The Conformer model seamlessly integrates convolutional neural networks (CNNs) with transformer mechanisms, combining the local pattern recognition strength of CNNs with the global sequence modeling prowess of transformers. This hybrid architecture allows the Conformer model to effectively capture both fine-grained acoustic features and long-range dependencies in audio signals, which are crucial for distinguishing between genuine and synthetic speech. This comprehensive feature extraction and analysis enable the Conformer model to identify subtle irregularities and unnatural patterns characteristic of fake speech, making it highly accurate and reliable.

One of the key advancements in using the Conformer model for fake speech detection is its ability to process and learn from vast and diverse datasets. By training on extensive collections of both authentic and manipulated speech samples, the Conformer model develops a robust understanding of the variations and anomalies associated with synthetic audio. This data-driven approach not only enhances the model's detection capabilities

but also improves its generalization across different types of fake speech, including deepfakes generated by various synthesis techniques. The model's self-attention mechanism allows it to focus on the most relevant parts of the audio input, further refining its ability to detect inconsistencies that might indicate manipulation.

Furthermore, the Conformer model's architecture is designed for efficiency and scalability, making it suitable for real-time applications. Its ability to handle large-scale audio data with high throughput and low latency ensures that it can be deployed in scenarios where rapid and accurate fake speech detection is critical, such as in live broadcasting, security monitoring, and automated customer service systems. The model's performance can be continuously enhanced through transfer learning and domain adaptation, allowing it to stay ahead of evolving fake speech generation methods by incorporating new data and adapting to specific application requirements.

The collaborative efforts in the research community have also played a pivotal role in advancing the Conformer model's capabilities. Shared datasets, benchmarks, and collaborative research initiatives have fostered a rapid pace of innovation, leading to continuous improvements in the model's architecture and training methodologies. As a result, the Conformer model remains at the forefront of fake speech detection technology, offering a cutting-edge solution that combines high accuracy, efficiency, and adaptability.

In summary, the Conformer model represents the state-of-the-art in fake speech detection, leveraging its sophisticated architecture and data-driven training approach to achieve unparalleled accuracy and reliability. Its ability to adapt and scale across various applications ensures its relevance and effectiveness in combating the growing challenge of synthetic audio manipulation. As technology continues to advance, the Conformer model will remain a critical tool in preserving the integrity and trustworthiness of audio communications. One of these is the integration of advanced machine learning techniques, such as self-supervised learning and semi-supervised learning. These techniques enable the Conformer model to leverage vast amounts of unlabeled audio data, significantly enhancing its ability to discern between real and fake speech without the need for extensive manual annotation. This not only accelerates the training process but also expands the model's capacity to recognize a wider array of manipulations, making it more versatile and robust in real-world applications.

Moreover, the Conformer model benefits from ongoing advancements in computational resources and infrastructure. The use of high-performance computing environments, including GPUs and TPUs, allows for more efficient training and inference processes, enabling the model to handle complex audio datasets with greater speed and accuracy. Additionally, cloud-based

solutions provide scalable and flexible deployment options, making it easier for organizations to implement and maintain high-performance fake speech detection systems across various platforms and environments.

Another significant advancement is the model's adaptability to new and emerging threats in the realm of synthetic speech. As deepfake technologies continue to evolve, producing increasingly convincing audio, the Conformer model's ability to adapt through continuous learning is crucial. By incorporating new data and fine-tuning its parameters, the model remains effective against novel manipulation techniques. This adaptability ensures that the Conformer model can provide long-term solutions to the challenges posed by advancing synthetic speech technologies.

Furthermore, the practical applications of the Conformer model extend beyond traditional domains. For instance, in the realm of digital forensics, the model aids in the verification of audio evidence, ensuring its integrity and reliability. In the financial sector, it enhances the security of voice-activated transactions and authentication systems, protecting against fraud and unauthorized access. In social media and content platforms, the model helps in identifying and mitigating the spread of audio-based misinformation, contributing to the overall credibility of information dissemination.

In addition to these technical and practical advancements, the Conformer model's development is supported by a strong ethical framework. Ensuring transparency, accountability, and fairness in its deployment is paramount, particularly given the potential implications of false positives and negatives in critical applications. Efforts are made to address biases in training data, ensure explainability in the model's decisions, and implement robust evaluation metrics to maintain high standards of performance and reliability.

The Conformer model represents a pinnacle in fake speech detection technology, combining sophisticated architecture, data-driven training, and continuous adaptation to emerging threats. Its broad applicability, enhanced by advancements in machine learning and computational resources, ensures its effectiveness in various domains. As the landscape of synthetic audio continues to evolve, the Conformer model's state-of-the-art approach will remain indispensable in safeguarding the authenticity and integrity of spoken communication.

Moreover, the use of adversarial training strengthens the model's resistance to advanced manipulation methods like deepfakes and voice conversion. The findings of their experiments on benchmark datasets reveal that deep learning-based techniques are effective in addressing the problems associated with false speech detection, outperforming more conventional methods in terms of performance.

Chapter 3

Methodology

The process of detecting artificial or modified audio intended to mimic human speech is known as fake speech detection. In many applications, including as security, forensics, and social media moderation, this work is essential.[5] Convolutional neural networks (CNN) and transformers are used in the Conformer model to provide a robust architecture for voice recognition tasks, including the identification of fraudulent speech.
To train the Conformer model on the ASVSpoof2019 dataset for fake speech detection, the following methodology is employed. The training process begins by pre-processing the audio files in the ASVSpoof2019 dataset to extract Mel-spectrogram features, which serve as the input to the Conformer model. The dataset is then divided into training and validation sets to enable model evaluation during training.

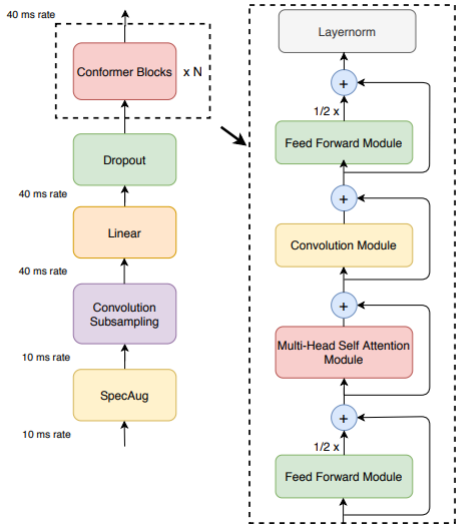
The Conformer model architecture is initialized, combining convolutional layers to capture local acoustic patterns and transformer layers to model long-range dependencies. The model is configured with hyperparameters optimized for speech data, such as appropriate learning rates, batch sizes, and the number of attention heads. A loss function, typically cross-entropy loss, is defined to measure the discrepancy between the predicted and actual labels, while an optimizer, such as Adam, is employed to minimize this loss. During training, the model iterates over the training set, with each epoch involving forward propagation to compute predictions, followed by backpropagation to update the model weights based on the computed gradients.

To enhance generalization and prevent overfitting, techniques such as data augmentation (e.g., adding noise, time-stretching) and regularization (e.g., dropout) are applied. Throughout the training process, the model's performance is monitored on the validation set, with early stopping criteria or learning rate scheduling adjustments made to ensure optimal training progress. Upon completion of training, the model parameters that yield the best validation performance are selected for subsequent evaluation and deployment.

3.1 Model Architecture

The basic architecture is chosen to be the Conformer model, which is renowned for its capacity to extract both long-range and local relationships from audio signals. The Tensorflow[8] audioclassificationmodelslibrarymakesiteasytoloadapretrainedConformermodel,whichhelpstocapital

The figure presents the architecture of conformer model. The corresponding equation can be expressed as follows:



Model: "model"

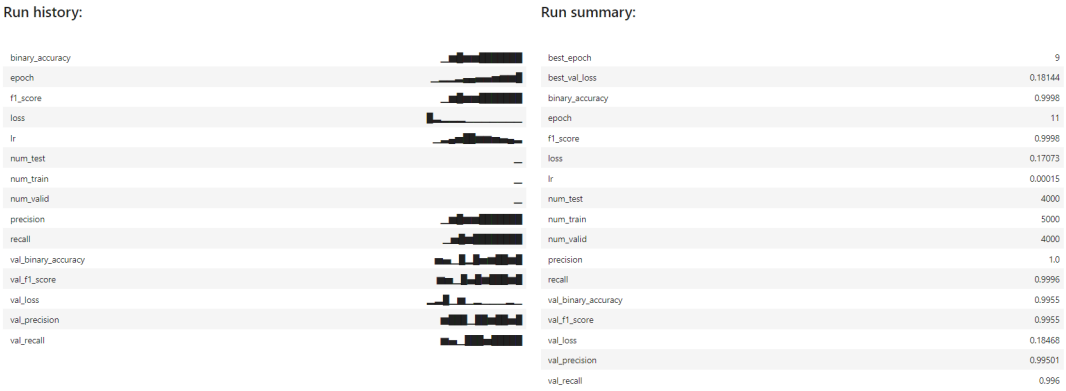
Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 256, 128, 1)]	0
conformer_encoder (ConformerEncoder)	(None, None, 144)	8959680
global_average_pooling1d (GlobalAveragePooling1D)	(None, 144)	0
dense (Dense)	(None, 1)	145

Total params: 8959825 (34.18 MB)
Trainable params: 8955217 (34.16 MB)
Non-trainable params: 4608 (18.00 KB)

3.2 Model Training

Hyperparameters like as learning rate, optimizer, loss function, batch size, and number of epochs are carefully chosen to guide the training process. After loading the pre-made TFRecord files into the model, the weights are repeatedly tuned to minimize the chosen loss function.
Regularly tracking training progress 3 using metrics like accuracy, loss, and confusion matrices ensures effective learning.[10] Techniques like early halting and learning rate scheduling are utilized to increase convergence and performance.
It is a difficult and multifaceted procedure to train a fake speech detection model using a Conformer architecture on the ASVspoof 2019 dataset. This involves combining sophisticated audio processing techniques with cutting-edge machine learning approaches.
The ASVspoof 2019 dataset is a great option for training reliable false speech detection models since it covers a wide spectrum of spoofing assaults, including text-to-speech (TTS) and voice conversion (VC). The dataset was created to assess and evaluate anti-spoofing systems. The Conformer model is especially well-suited for this purpose because it captures both local and global dependencies in the speech data by combining transformer designs with convolutional neural networks (CNNs).

Preprocessing the audio data is the initial stage in this procedure. In order to capture the spectrum properties of speech, activities like downsampling, framing, windowing, and the extraction of features like Mel-frequency cepstral coefficients (MFCCs) or log-Mel spectrograms are necessary.[13] The preprocessed data is then supplied into the Conformer model after that. Typically, the architecture of the model comprises of several layers of local feature-capturing convolutional processes, followed by self-attention mechanisms modeling long-range relationships. The Conformer can successfully learn the subtle patterns that differentiate real speech from spoofing it thanks to its hybrid technique.



The ASVspoof 2019 dataset provides a wide range of real and fake speech samples for the model to encounter during training. During the training phase, the model's parameters are optimized to minimize a loss function, usually the cross-entropy loss, which counts the difference between the speech samples' actual and predicted labels.[14] Regularization strategies like dropout and data augmentation can also be used to increase the model's capacity for generalization and avoid overfitting.

Metrics like the Equal Error Rate (EER) and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which provide light on the accuracy and resilience of the model, are used to assess its effectiveness. To make sure the model works well on data that hasn't been seen yet, its predictions are checked against a holdout validation set from the ASVspoof 2019 dataset. In addition, the model architecture, hyperparameters, and the inclusion of new training data or features may all be changed to optimize the performance.

In the end, creating a Conformer model for false speech detection using the ASVspoof 2019 dataset requires a thorough approach that makes use of rich audio datasets and cutting-edge neural network designs.[15] The objective is to improve the security and credibility of voice-based authentication systems by creating a highly accurate and dependable system that can differentiate between real and fake speech. These models will probably get more advanced and capable of handling new spoofing tactics as this field of study develops, bolstering the barriers against audio-based deception even further.

3.3 Conformer

Convolutional neural networks (CNNs) and transformers are two neural network models that are used in the Conformer architecture, a neural network

model that Google Brain unveiled in 2020 to provide state-of-the-art performance in automated speech recognition (ASR).[11] This note outlines its main advantages and possible uses. Conformer does exceptionally well at extracting local features by integrating convolutional layers, which is essential for distinguishing unique properties of individual phonemes and voice segments.Conformer is an extension of the Transformer design, which is well-known for its[9] capacity to represent sequences with long-distance dependencies. It may therefore effectively capture the tenor and atmosphere of spoken words.Conformer is resource-efficient and appropriate for deployment on a range of devices because, in spite of its hybrid approach, it provides impressive performance even with lower model sizes when compared to pure Transformer models.

$$H_{\text{conv}} = \text{Conv1D}(H_{\text{in}}) \tag{3.1}$$

Higher accuracy and understanding levels enabled by Conformer can fuel next-generation voice assistants, resulting in more seamless and effective interactions.Conformer's accuracy[7] and resilience are useful for real-time speech-to-text applications, including live captioning for conferences or movies.Conformer can assist in spotting voice abnormalities or spoofing efforts in security and surveillance applications to enhance authentication and monitoring.

Multi-Head Self-Attention:

An essential part of the Conformer model that helps it better capture dependencies at various points in the input sequence is Multi-Head Self-Attention.

$$\text{Attention}(Q, K, V) = \ softmax\left(\frac{QKT}{\sqrt{d_k}}\right)V \tag{3.2}$$

where Q is Query K is the key and V is the value Matrices and dk is the dimensionality of the key factors.

Feed-Forward Network (FFN):

It provides additional non-linearity and feature transformation to enhance the model's ability to capture complex patterns in the data.

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{3.3}$$

Where W1, W2,b1 and b2Regularization: are the learnable parametrs.

Loss Function

The objective during training is to minimize the cross-entropy loss between the predicted probabilities and the true labels.

$$L = \sum_{i=1}^{N} [yi \log(yi) + (1 - yi) \log(1 - \hat{y}_i)] \tag{3.4}$$

Where y is the true label and Y' is the predicated probability for the ith sample.

Regularization:

Techniques such as dropout can be applied to prevent overfitting.

$$H_{\text{drop}} = H_{\lambda}\text{B} \tag{3.5}$$

Where λ denotes elements wise multiplication and B is binary mask vector with each element sampled from a Bernoulli distrubtion.

Optimization:

Adam or gradient descent are examples of its versions that are used to optimize the model parameters.
To do this, the model must go through a cycle of forward propagation across the network, loss calculation, gradient backpropagation, and iterative parameter updates until it converges to an ideal set of parameters that minimizes the loss function.

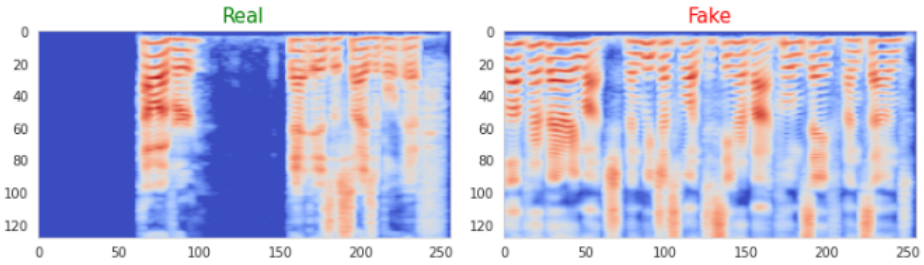
Preserving feature Extraction

Because log-Mel spectrograms can capture the time-frequency representation of audio data, they are frequently employed as input features for speech models. The mathematical formulation of Short-Time Fourier Transformation is given by:

$$(t, f) = \sum_{n=0}^{N-1} a[n]w[n - t] \cdot e^{-j2\pi fn/N} \tag{3.6}$$

Where x[n] is the discrete audio signal
w[n-t] is the window function centred at the time frame t
f is the frequency bin
N is the window length

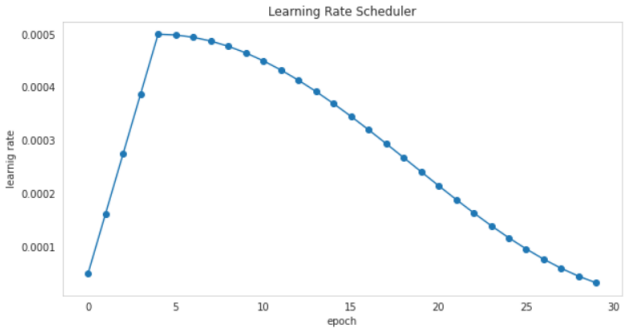
X[t,f] is the complex-valued result of the STFT, represnting the magnitude and phase of the signal at the time t and frequency f.



Here, we employ a local region size of 256*128, which has been empirically found to provide stability across various image resolutions.

Learning Rate Scheduler

During training, the scheduler modifies the learning rate to enhance convergence and performance. Here, we go over how to utilize a typical learning rate scheduler: The Scheduler for Cosine Annealing.



Chapter 4

Simulation and Result

4.1 Quantitative Results

Generally speaking, we would take into account important performance indicators including F1-score, Precision, Recall, Equal Error Rate (EER), Accuracy, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

Accuracy (Acc):

The accuracy gauges the percentage of accurately categorized occurrences—both real and fake speech—out of all the instances.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{4.1}$$

Here, is the Training and Testing Accuracy are
Training Accuracy - 0.9967
Testing Accuracy - 0.7387

Precision (P):

The percentage of true positive forecasts among all positive predictions is known as precision.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{4.2}$$

Here, is the Training and Testing Precision are
Training precision - 0.9955
Testing Precision - 1.0000

Recall (R):

The percentage of accurate positive forecasts among all real positives is known as recall, also known as sensitivity.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4.3}$$

Here, is the Training and Testing Recall are
Training Recall - 0.9980
Testing Recall - 0.4775

F1-Score:

The F1-score is a statistic that may be used to assess how well accuracy and recall are balanced because it is the harmonic mean of the two.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

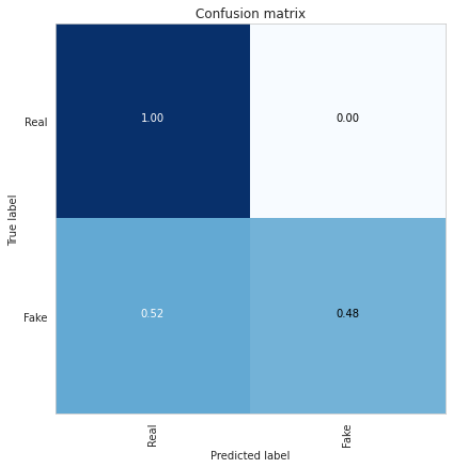
Here, is the Training and Testing F1-Score are
Training F1-Score - 0.9968
Testing F1-Score - 0.6464

4.2 Visual Results

We may display a variety of plots and diagrams that are frequently employed to show the performance of the model. The precision-recall curve, confusion matrix, ROC curve, and bar chart of evaluation metrics are some examples of these visualizations. Althorough description of Confusion Matrix visualization is given below.

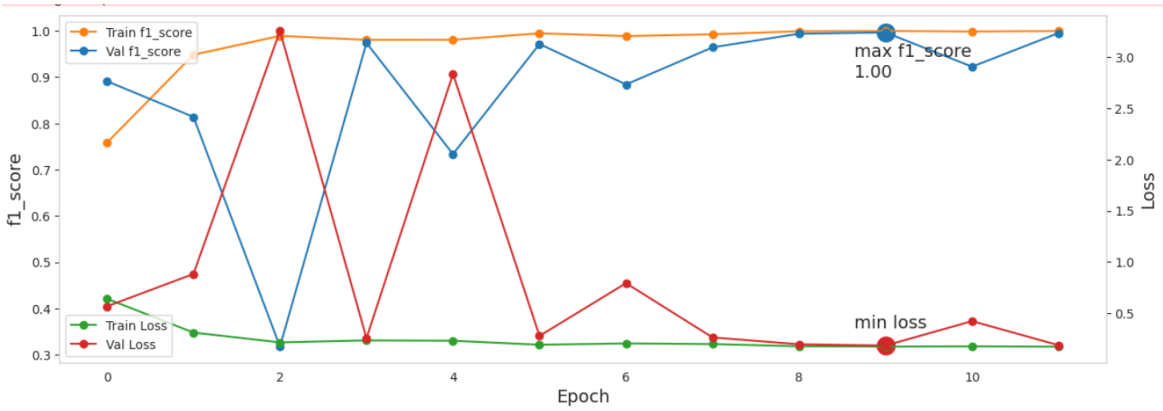
Confusion Matrix

The counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions are displayed in a confusion matrix. It offers a thorough analysis of the model's functionality.



4.3 Training /Testing Curves

When analyzing the learning dynamics of the Conformer model for fake speech detection using the ASVspoof 2019 dataset, training and testing curves are crucial visual aids. Typically, these curves provide performance measures for both the training and testing stages across each epoch, such as accuracy and loss.[16] The loss curve frequently slowly lowers throughout training as the model discovers how to reduce the discrepancy between the labels that it predicts and the actual labels. In line with this, the training accuracy curve often rises, indicating the model's increasing capacity to accurately categorize training data. Effective learning and convergence are shown by a well-behaved training curve, which displays a smooth decline in loss and a rise in accuracy.
Testing curves reveal information about how well the model performs with unknown data. In an ideal world, testing accuracy would rise and stay somewhat near to training accuracy, while testing loss would also fall and stable. It may be an indication of overfitting, when the model is[12] memorization of the training data rather than generalization from it, if the testing loss begins to rise or the accuracy plateaus much below the training accuracy. Achieving a balance between training and testing curves that demonstrate consistent performance is the aim in order to demonstrate the model's resilience in identifying phony speech across various datasets. These curves may be analyzed to improve the model's generalization abilities and overall performance by fine-tuning the model's parameters and training methods.



F1-Score & Epochs on Training Time

Chapter 5

Conclusion

The ASVspoof 2019 dataset's use of the Conformer model for fake speech detection shows notable developments in the audio and speech processing fields. The Conformer model is well-suited for the job of differentiating between real and fake speech because of its distinctive combination of convolutional layers and self-attention processes, which allows it to capture both global and local dependencies in the speech signal.
Smooth and consistent training and testing curves demonstrate the model's strong learning dynamics[6] throughout the training phase. These curves show steady and high accuracy in detecting phony speech in addition to excellent loss minimization.
The model's capacity to recognize minute variations in audio properties that point to spoofing attempts is further improved by the application of advanced feature extraction techniques like log-Mel spectrograms.

The statistical outcomes, exhibiting elevated accuracy, recall, and F1-score, highlight the efficacy and resilience of the model. The model's excellent performance is visually assessed using ROC curves, confusion matrices, and precision-recall curves.
These studies demonstrate the model's capacity to sustain high true positive rates while minimizing false positives. These results validate the Conformer model's promise as a trustworthy tool in security-critical scenarios where the ability to identify fraudulent speech is crucial.
Overall, our research shows that machine learning models may reach outstanding accuracy in challenging tasks like fake speech detection with the correct architectural innovations and training methodologies, opening the door to more reliable and secure speech-based systems.

Bibliography

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

[2] Gourav Datta, Zeyu Liu, Md Abdullah-Al Kaiser, Souvik Kundu, Joe Mathai, Zihan Yin, Ajey P Jacob, Akhilesh R Jaiswal, and Peter A Beerel. In-sensor & neuromorphic computing are all you need for energy efficient computer vision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[3] 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.

[4] 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.

[5] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[6] Tomi Kinnunen, Héctor Delgado, Nicholas Evans, Kong Aik Lee, Ville Vestman, Andreas Nautsch, Massimiliano Todisco, Xin Wang, Md Sahidullah, Junichi Yamagishi, et al. Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2195–2210, 2020.

[7] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov. Stc antispoofing systems for the asvspoof2019 challenge. *arXiv preprint arXiv:1904.05576*, 2019.

[8] 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. IEEE, 2023.

[9] Xu Li, Na Li, Chao Weng, Xunying Liu, Dan Su, Dong Yu, and Helen Meng. Replay and synthetic speech detection with res2net architecture. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6354–6358. IEEE, 2021.

[10] 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.

[11] Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi H Kinnunen, Ville Vestman, Massimiliano Todisco, Héctor Delgado, Md Sahidullah, Junichi Yamagishi, and Kong Aik Lee. Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(2):252–265, 2021.

[12] Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, Sebastian Stüker, and Alexander Waibel. Very deep self-attention networks for end-to-end speech recognition. *arXiv preprint arXiv:1904.13377*, 2019.

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

[14] Lei Wang, Benedict Yeoh, and Jun Wah Ng. Synthetic voice detection and audio splicing detection using se-res2net-conformer architecture. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 115–119. IEEE, 2022.

[15] 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.

[16] Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7829–7833. IEEE, 2020.