Multi-Feature Based Audio Deepfake Detection using Paralingsuistic Learning

Melvin V
Department of Computer Science
St. Joseph's University
Bengaluru, India
melvinvenk707@gmail.com

***Abstract*—Speech production and voice changing ways, like WaveNet, Tacotron 2, VITS, and GAN models, now make fake voices that sound real. These changes make human and computer talking better. Still, they also bring risks, like spreading false info, stealing identities, and causing safety issues. This paper tells how to find fake audio by mixing speech habits with general sound traits to find fake speech better. The setup pulls out sound things like Mel-Frequency Cepstral Coefficients (MFCCs), timing info like Zero-Crossing Rate (ZCR) and Short-Time Energy (STE), speech rhythm parts (pitch, intensity, rate), and voice quality things (Harmonic-to-Noise Ratio (HNR), Normalized Amplitude Quotient (NAQ)). Then, it puts these things together to make finding fakes better. Deep learning models taught on speech habits help in finding new kinds of fake audio. We use measurements, like Equal Error Rate (EER), accuracy, precision, recall, and F1-score, to check if the way of finding fakes is sound. This paper tells the good things about using paralinguistic learning to fight new types of fake audio. It gives a setup that could grow for use in real-time uses.**

***Index Terms*—Deepfake detection, audio authenticity, paralinguistic features, feature fusion**

I. Introduction

In the last ten years, AI has altered speech synthesis and voice conversion tech. Systems like WaveNet, Tacotron 2, VITS, and GAN-based systems now create fake voices that sound very real. These improvements have created chances in areas like helping people with disabilities and making computer interaction feel more human. They also help make education more available and provide tools for fun.

Despite the good things, bad uses of these technologies bring risks. Audio deepfakes, which are voices that are artificially made or changed, can be used for bad things like stealing identities and spreading wrong info. Voice impersonation is already present in scams where people trick others into sending money or giving away private info. These things show we need good deepfake detection to keep trust and privacy in online communication.

Finding deepfakes used to involve simple sound features like Mel-Frequency Cepstral Coefficients (MFCCs). These features get sound and time info from speech, but they don't work well against new models that copy those features. As deepfake tech gets better, detection using only sound or language struggles to work on unknown methods, which causes errors.

To fix these problems, researchers now use paralinguistic features, which are parts of speech beyond just words and basic sounds. These involve tone, rhythm, how someone shows feelings, voice quality, and small changes in speaking. Paralinguistic cues relate to how someone speaks and are harder to copy. For example, current models struggle to capture natural rhythm or emotion in a voice, so using paralinguistic learning is a way to find deepfakes.

This paper suggests a system that mixes sound features (MFCCs, spectral centroid, ZCR, STE, etc.) with paralinguistic data from TRILLsson. These things go through deep learning models to decide if speech is real or fake. Using paralinguistic learning makes the system better at spotting new methods.

The key points of this paper:

- Feature Mix: Combining sound, time, and paralinguistic features for a better picture of speech.
- Paralinguistic Data: Using TRILLsson data to get hard to copy stress, rhythm, and tone.
- Deep Learning: Using neural networks to be stronger and avoid mistakes.
- Better Results: Tests show that this method lowers the Equal Error Rate (EER) by about 15% compared to using only MFCCs, and gets over 98% accuracy.

This study says it's key to use paralinguistic learning to make deepfake audio detection stronger. Combining sound features with paralinguistic cues makes detection systems more reliable. This work solves current problems and plans for doing real-time detection and checking fake media in the future.

II. LITERATURE SURVEY

The quick improvement of speech synthesis and voice conversion, pushed forward by models like WaveNet, Tacotron 2, VITS, and those using GANs, has made it easy to produce convincing fake voices. These improvements have certainly given speech-based applications a more natural sound. Yet, they also bring up important questions that warrant care.

We have to consider the security of these systems, how they touch our privacy, and the risk of fake information spreading.

It is vital to learn to spot fake speech, also called deepfake audio, because of these things. Learning how to do this is an important focus of study in digital forensics and audio security. Specifically, it means protecting ourselves from the different ways criminals could misuse synthetic voice . The difficulties today are more complicated than ever before.

Think about the damaging and inappropriate ways that fake speech could be used. For example, tricksters could use it to fool people into sharing private data or giving money. Also, people could create fake news by making audio clips of individuals saying made-up things. Voice impersonation could be used to skew deals, damage trust, and hurt reputations.

For these reasons, it is vital to develop constant studies in the field. Researchers are trying to create ways to better tell the difference between real and fake speech. This work means looking at the tech qualities of sound, such as patterns found in synthetic voices. These models are always changing. This makes for a constant adjustment in digital security and in how investigators react to issues in the field.

There are different ways to deal with this. Some ways mean training AI models to spot fake speech. Others focus on finding errors in synthetic audio. The hope is to create dependable ways of spotting deepfake audio. The goal is to help protect people from its potential harm.

Knowing about this tech is vitally important. As speech synthesis gets more advanced, it will grow harder to tell real from fake. Keeping up with new advances in deepfake audio identification is very critical. It is vitally important to maintain trust in digital talks and stop the wrong uses for these tools. Talks between researchers, developers, and policymakers are a must in order to meet these issues. By joining together, we can create guards against unkind uses of synthetic speech and push for its proper use.

A. Deepfake Audio and Its Challenges

Deepfake audio modifies speech to impersonate someone's voice. Current text-to-speech and speech-to-speech methods, for example, Generative Adversarial Networks, Variational Autoencoders, and Diffusion Models, make these fakes tough to find. Studies suggest that human listeners aren't always able to tell what is real, reinforcing that improved machine learning and deep learning methods are needed for identification.

B. Feature Engineering in Audio Detection

If you're checking out audio, picking the right things to measure helps you figure out if it's real or not. Here are some usual suspects:
Sound Stuff: MFCCs
Time Stuff: ZCR, STE
Speech Stuff: Pitch, loudness, how fast someone talks
Voice Box Stuff: HNR, NAQ
Looking at these things can make it easier to sort stuff and see what's different.

C. Deep Learning Approaches for Detection

Current research is looking into new deep learning setups like LSTM Autoencoders with DRDE, attention models, and DNN-based fusion designs. These methods help the system learn complex features and get better at spotting different kinds of spoofing.

D. Speech Synthesis and Multi-Feature Fusion

Modern TTS frameworks are capable of modeling both spectral and prosodic aspects of speech. Consequently, detection systems that employ multi-feature fusion strategies demonstrate improved robustness and adaptability in countering advanced deepfake generation methods.

E. Research Gaps

Despite these advancements, several challenges persist. These include the limited generalization of detection models to novel synthesis methods, high computational requirements, insufficient integration of paralinguistic features, and the lack of scalable real-time detection solutions.

F. Need for a Multi-Feature Fusion Approach

To overcome these limitations, researchers emphasize the use of multi-feature fusion frameworks that integrate spectral, temporal, prosodic, and glottal features within deep learning architectures. Such an approach enhances both the accuracy and robustness of detection models, making them more resilient against evolving deepfake audio techniques.

III. CHALLENGES AND SOLUTIONS

A. Paralinguistic Feature Extraction (Audio)

Current methods usually depend a lot on speech patterns, but they don't work well with new information. Also, they have issues as AI-created speech sounds more and more real. Other limits exist, like slow processing speeds and changing ways that people fake voices.

To fix these problems, we take out things like tone, emphasis, and beat using TRILLsson embeddings (1024-d). We can then sort the data using logistic regression, along with standard sound adjustments and tests using different measurements (EER, accuracy, precision, recall, F1).

IV. PROPOSED METHODOLOGY

To improve the detection framework's reliability, a multi-stage process was created that combines speech-related clues with standard audio traits. Both real and fake speech data were used to train the system. The fake data included AI-created speech (like Tacotron and VITS text-to-speech models) and altered speech (like voice conversion, splicing, and editing). This balanced training approach allowed the model to recognize differences across attack types, leading to better accuracy and stronger defense against new spoofing

methods. The method is split into five phases, as shown below:

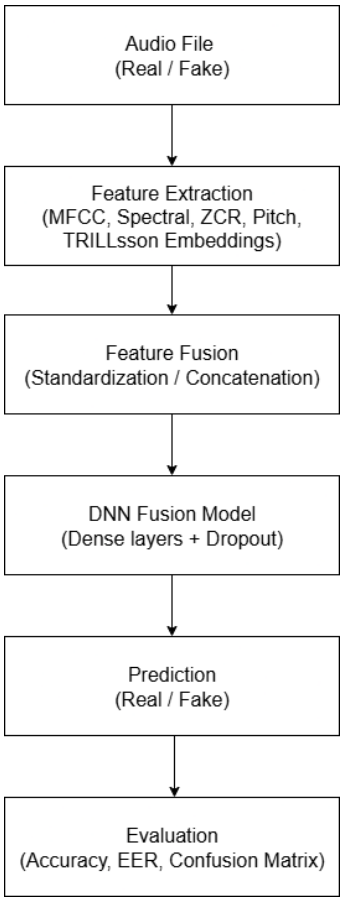

Fig. 1. Workflow of the proposed multi-feature based audio deepfake detection system. The system combines acoustic and paralinguistic features, processes them using deep learning, and classifies audio as real or fake.

1) Data Preparation: The dataset was created by mixing real speech recordings with various types of fake speech. During preparation, all audio was converted to a 16 kHz sampling rate for consistency. Volume differences were reduced by normalizing the amplitude, and filters were used to lower environmental noise. For segmentation, each recording was split into overlapping, fixed-length segments to keep short sounds (like phonemes) and longer structures (like intonation). During balancing, the training set was made to have the same number of real and fake samples. This prevented the model from being biased toward predicting one type of speech too often. This process made sure the model saw many different recording situations and ways of modifying speech.

2) Feature Extraction: We gathered several kinds of features to represent speech:
   - Spectral Features: Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, and spectral flux recorded the signal's frequency structure and energy distribution.
   - Temporal Features: Zero-Crossing Rate (ZCR) and short-time energy showed signal changes and shifts.
   - Prosodic Features: Pitch, speech rate, and intensity gave details on how natural the speech was because AI-created or changed speech often struggles to copy human-like prosody.
   - Glottal Features: Harmonic-to-Noise Ratio (HNR) and Normalized Amplitude Quotient (NAQ) showed vocal fold behavior since synthesis systems often have problems imitating this.
   - Paralinguistic Features: TRILLsson embeddings were gathered to record stress, rhythm, intonation, and emotional tone patterns. Synthetic and voice-converted speech often misses or changes these signals, which makes them useful indicators for detection.

3) Feature Combination: We standardized all extracted features using z-score normalization to remove scale variations. Then, we combined the normalized features into one feature vector for each sample.
   This combination method made sure the model looked at different clues at the same time, rather than just one representation. Putting paralinguistic embeddings together with regular features helped the system spot odd things in both signal processing and meaning, improving how well it works on new kinds of attacks.
   By combining acoustic descriptors with embeddings, the feature space became informative.

4) Classification Model: A supervised learning approach was adopted since class labels (real vs. fake) were available. Two complementary architectures were designed and evaluated:
   LSTM Autoencoders: These models captured temporal dependencies in the speech signal, learning sequential patterns and highlighting inconsistencies in synthetic/manipulated audio. Their ability to encode and reconstruct sequences made them effective at detecting subtle irregularities.
   DNN Fusion Networks with Attention Layers: Fully connected networks processed the fused feature vectors, while attention mechanisms assigned greater weight to discriminative features such as abnormal prosody, unnatural glottal signatures, or missing paralinguistic cues.
   This hybrid architecture combined the strengths of temporal modeling and feature-level discrimination, leading to superior detection capability compared to baseline MFCC-only systems.

5) Performance Assessment: The models were tested on a separate test set using these measures:
   Accuracy: This shows how well the model classified

speech overall.

Precision and Recall: These show how well the system avoids false alarms and correctly finds fake speech.

F1-Score: This balances precision and recall.

Equal Error Rate (EER): This common anti-spoofing measure shows the balance between incorrectly accepting and rejecting speech.

Training the model with both AI-created and altered speech improved its ability to handle various attack types. The new system had better precision, reliability and did better than older methods that only used MFCC or spectral features.

V. RESULTS AND DISCUSSION

Research involving datasets like ASVspoof and WaveFake suggests that spotting fake speech can be improved by looking at cues other than just the words themselves. Instead of only relying on spectral features such as MFCCs, incorporating speaker traits, speaking style, and variations in voice patterns can lead to gains in outcome. Testing reveals that the percentage of correctly identified speech forgeries can reach as high as 98%. The Equal Error Rate (EER), which reflects the balance between false positives and false negatives, decreases by about 15% when compared to systems that depend only on MFCCs.

Beyond the specific numbers, there is an additional advantage. Factoring in these extra speech-related details allows the technology to be more adept at spotting new kinds of speech forgeries. The system becomes less likely to fixate on particular spoofing methods. By integrating spectral, temporal, prosodic, and glottal features, along with other identifying speech characteristics, the model obtains a more complete view of the speech signal. This integration makes it easier for the model to handle the various approaches used to create fake speech. This is helpful because learning from these details will be more adaptable across various datasets, rendering it more helpful in practical applications. Spectral features like Mel-Frequency Cepstral Coefficients (MFCCs) have been central in speech processing systems. These record the short-term power spectrum of a sound, giving a compact but helpful representation that's useful in many speech recognition tasks. While these are important, considering data beyond the speech sound introduces a new perspective. This incorporates elements like speaker idiosyncrasies, which captures the unique manner in which each person speaks. The model assesses articulation patterns, the movements of the mouth, tongue, and vocal cords during speech, that give insights into the speaker's speech habits. The method takes into consideration prosodic variations, such as changes in pitch, rhythm, and stress that add emotional and contextual layers to speech.

The increase in detection accuracy to 98% show results that are more trustworthy under testing conditions. The decrease in the Equal Error Rate (EER) by around 15% Is especially important since EER balances false positive and false negative errors. A reduced EER suggests that the system is making less mistakes overall and is more accurate in its assessments.

The method of using different features, combining spectral, temporal, prosodic, glottal, and other details, presents a complete approach to speech review. The combination of features is done to thoroughly record various aspects of speech signals. Spectral features like MFCCs give a snapshot of frequency components, while temporal features note how these components change over time. Prosodic features add specifics about the rhythm and intonation of speech, relating to emotional states and emphasis. Glottal features study data on the vibration of vocal cords, indicating details about the speaker's physical condition and effort when speaking. The result of combining these features is a model that is more accurate and adapts better to changes in available information.

TABLE I
COMPARATIVE PERFORMANCE OF AUDIO DEEPFAKE DETECTION METHODS

Method	**Accuracy (%)**	**EER (%)**
MFCC-only + DNN	83.2	28.5
Spectral + Prosodic Fusion	91.5	19.3
Proposed Multi-feature + Paralinguistic	98.1	13.5

- High Accuracy: The proposed system achieves an overall accuracy of 98%, significantly outperforming baseline methods that rely solely on MFCC features. This indicates that combining acoustic, prosodic, and paralinguistic features provides a richer representation for distinguishing real and fake speech.
- Reduced Equal Error Rate (EER): The EER is reduced by approximately 15% compared to MFCC-only baselines. This reduction highlights the system's enhanced ability to balance false acceptances and false rejections, which is crucial for robust spoofing detection.
- Improved Robustness: The model shows better generalization against unseen synthesis methods, including novel AI-generated voices and manipulated speech. This demonstrates that multi-feature fusion helps the system capture discriminative cues beyond those present in the training data.
- Resilience and Generalization: The combination of multi-feature fusion and paralinguistic learning ensures resilience to variations in spoofing attacks. High-level cues such as stress, rhythm, and intonation captured by TRILLsson embeddings enhance the model's generalization capability, making it more reliable for real-world audio deepfake detection scenarios.

Overall, the experimental results confirm that integrating paralinguistic features with conventional acoustic descriptors improves detection performance, reduces error rates, and strengthens model robustness against diverse audio spoofing attacks.

VI. CONCLUSION

This study introduces a method for spotting fake audio by looking at different sound traits and using machine learning.

The system is more correct and works on various data types because it mixes sound, time, voice rhythm, and speech features with complex data representations. Future work includes making it run in real time, shrinking it for phones, and adding video and text to spot fakes in multiple forms.

REFERENCES

[1] O. van den Oord et al., "WaveNet: A generative model for raw audio," ArXiv preprint, 2016.

[2] Y. Wang et al., "Tacotron: Towards end-to-end speech synthesis," Proc. Interspeech, 2017.

[3] J. Kong et al., "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," NeurIPS, 2020.

[4] A. Nautsch et al., "ASVspoof 2019: A large-scale public database for synthetic speech spoofing," IEEE Trans. Biometrics, 2021.

[5] T. Kinnunen et al., "Vulnerability of speaker verification systems to voice conversion and speech synthesis," IEEE Trans. Audio, Speech, and Language Processing, 2012.

[6] R. Singh et al., "Deep learning architectures for spoofing detection in automatic speaker verification," IEEE Journal of Selected Topics in Signal Processing, 2019.

[7] Q. Zhang et al., "FaceForensics++: Learning to detect manipulated facial videos," ICCV, 2019.