Grenze International Journal of Engineering and Technology, June Issue

Biometric Emotion Profiling Framework for Safe Audio Deep Fake Detection in Voice Authentication Systems

Preeti Shamrao Wathore[1], P R Murkute[2] and S D Pingle[3]
[1-3]M Tech Students, Assistant Professor, Assistant Professor, Computer Science & Engineering, P E S College of Engineering, Chh. Sambhajinagar
Email: pritiwathore9@gmail.com, poojamurkute5758@gmail.com, sdp150876@gmail.com

***Abstract*— Generative artificial intelligence is moving along quickly, making audio deepfakes much more realistic and easier to get to. This is a big problem for voice- based authentication systems. Traditional methods for finding audio deepfakes mostly use spectral or speaker-dependent information. These methods often miss small emotional and behavioral anomalies that synthetic speech generation models add. To overcome this constraint, this study introduces a Biometric Emotion Profiling Framework for secure audio deepfake detection in voice authentication systems. The framework combines emotion-aware acoustic feature extraction (Mel- Frequency Cepstral Coefficients (MFCCs), pitch dynamics, spectral contrast, and signal energy) with biometric voice analysis and a Random Forest classifier to reliably tell the difference between real human speech and fake deepfake audio. The system is assessed with CREMA-D for authentic emotional speech and WaveFake for artificial deepfake audio. Experimental results show that the classification accuracy on a held-out test set is 100%. The precision, recall, and F1-score for both classes are 1.00, and five-fold cross- validation gives a mean accuracy of 1.0 with no volatility. An examination of feature importance suggests that the MFCC and pitch features that are based on emotions are the most important for telling deepfakes apart. The results show that adding biometric emotion profiling makes it much easier to find deepfakes and makes voice authentication systems safer.**

***Keywords*— Biometric emotion profiling, voice authentication, speech security, random forest, emotion recognition, and behavioral biometrics are all examples of these.**

I. INTRODUCTION

Voice-based authentication is becoming more common in banking, healthcare, smart assistants, and remote identity verification because it is easy to use and doesn't require any hands-on contact. Recent progress in neural text-to-speech (TTS) and voice conversion, on the other hand, has made it possible for attackers to make very realistic deepfake speech that sounds like the target's voice, pitch, and spectral characteristics. This makes traditional voice biometrics less reliable. Early audio deepfakes had clear artifacts, while contemporary systems like neural vocoder-based models create speech that sounds almost human with very few low- level aberrations. Current countermeasures usually deal with spectrum distortions, speaker embeddings, or channel inconsistencies. This means that they grow less dependable as the quality of the synthesis improves. One important but not well-studied aspect is emotion consistency. Human speech has a lot of emotional and behavioral indicators that existing generative models don't always get right. datasets-benchmarks-proceedings.

Grenze ID: 01.GIJET.12.2.519

This research presents a biometric emotion profiling approach that explicitly characterizes emotional expressiveness through MFCCs, pitch, energy, and spectral contrast, employing a Random Forest classifier to differentiate between authentic and deepfake speech. The technology is meant to be a security layer before authentication in speech authentication pipelines, stopping deepfake attempts before speaker verification. Figure 1 illustrates the process of extracting time and frequency domain features from voice signals for emotion classification using traditional machine learning and neural network models. Recognized emotions include angry, happy, sad, neutral, and fear.

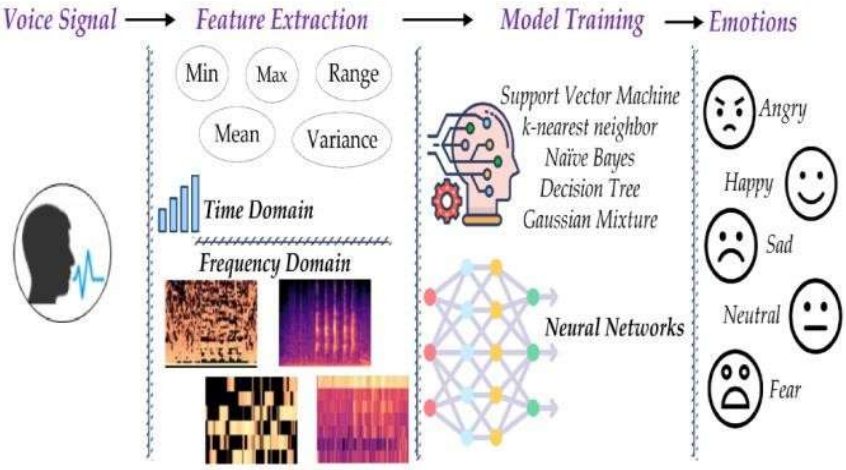

Fig. 1: Emotion Recognition Framework Using Voice Signal Analysis

Work that is connected Finding Audio Deepfake and Spoofing The ASVspoof challenges have rigorously tested ways to stop spoofing in automatic speaker verification,including synthesized, transformed, and replayed speech. In the past, people utilized GMMs or SVMs with spectral characteristics like MFCCs and CQCCs. More current work uses CNNs, RNNs, and Transformer-based models on spectrograms or raw audio. Many deep learning algorithms have trouble generalizing to new spoofing conditions or datasets, even while they function well in their own domains.

Numerous recent research and surveys underscore that the development of resilient and interpretable audio deepfake detection remains a significant problem, especially in the context of real-world variability and the advancement of synthesis models. Most methods focus on artifacts at the signal level and the identity of the speaker, with little attention paid to how emotions affect behavior.

A. Recognizing Emotions in Speech (SER)

SER study demonstrates that MFCCs, pitch, energy, and spectral properties proficiently differentiate emotions including anger, happiness, sadness, fear, and neutrality. Datasets such as CREMA- D, with thousands of audiovisual clips from 91 actors annotated with emotional states, are extensively utilized for modeling emotional expression in speech.

SER results reveal that emotions have a big effect on prosody and spectrum. For example, anger and happiness tend to have more pitch and energy, sadness has lower pitch and energy, and neutral speech has more stable prosody. A lot of TTS and voice conversion models are good at making speech clear and matching the speaker's voice, but they don't do a good job of properly reproducing small, context- appropriate emotional changes. This means that there is a gap that can be used for detection.

B. Security and Behavioral Biometrics in Voice Systems

The study on speaker verification that focuses on security shows how important it is to have multi-layer defenses that include spoofing countermeasures, anomaly detection, and behavioral biometrics. Behavioral signals like keystroke dynamics and mouse movements have made identity assurance better, and combining different types of biometrics makes security even better. Emotion profiling is a behavioral biometric that comes straight from speech and fits well into this framework. It adds to traditional speaker identity attributes without needing any further sensors.

C. Gaps in Research

Despite significant advancements in audio deepfake detection and SER, numerous deficiencies persist:Lack of Emotion-Aware Detection: Most deepfake detectors don't look for emotional expressiveness; instead, they look for artifacts or speaker embeddings. This is a problem because emotional dynamics are different between real and fake speech.Limited Behavioral Biometrics: Current speech authentication systems seldom incorporate emotional behavior as a distinct biometric layer, so forgoing a potential avenue for orthogonal security signals.Generalization and Interpretability: Deep learning- based detectors frequently exhibit a deficiency in interpretability and demonstrate restricted generalization across datasets and spoofing techniques. Ensemble models that show which features are most important can help us understand better what makes decisions. This work fills in these gaps by:creating a biometric emotion profiling framework for deepfake detection;using emotion features that have been widely validated (MFCCs, pitch, energy, spectral contrast);using a Random Forest classifier that allows for reproducible training and interpretable feature importance.

II. Methodology for Research

A. Overall Framework

This study presents a Biometric Emotion Profiling-based Deepfake Detection Framework aimed at bolstering the security of voice authentication systems against synthetic speech assaults. The main concept is that vocal qualities related to emotions and behavior are intrinsic biometric traits that existing deepfake generation techniques find hard to copy consistently.

The suggested approach works by taking emotion-aware acoustic parameters from voice signals and finding patterns that set real human speech apart from AI- generated deepfake audio. The method follows a systematic pipeline that includes preprocessing audio, extracting biometric emotion features, normalizing features, supervised classification, and performance evaluation.

The proposed method differs from traditional speaker verification techniques that concentrate exclusively on identity embeddings; it prioritizes emotion dynamics and vocal expressiveness as supplementary security measures, hence enhancing resistance to spoofing and replay assaults.

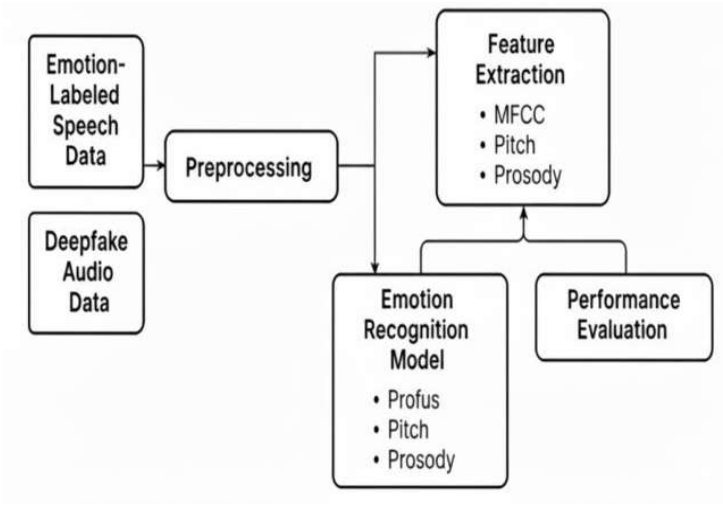

Fig. 2: Block Diagram of Biometric Emotion Profiling for Audio Deepfake Detection

The proposed pipeline is made up of:

Audio preprocessing: Resampling to 16 kHz, cutting off silence, normalizing loudness, and standardizing time to 3 seconds. To make sure that feature extraction is consistent, each audio input goes through preprocessing to cut down on unpredictability.

Resampling: All audio streams are resampled to 16 kHz.

Energy-based trimming removes leading and trailing quiet.

Amplitude Normalization: Signals are adjusted so that recording volume doesn't affect them.

Mono Conversion: Stereo signals are changed to mono when they need to be. This stage before processing makes ensuring that biometric data, not recording artifacts, are what drives classification decisions.

B. Feature Extraction

MFCC statistics (mean and standard deviation)
Mean pitch (fundamental frequency)

C. Mean energy (RMS)

Spectral contrast in several bands
Feature Normalization: Using statistics from the training split to standardize (zero mean, unit variance). To make sure that all features are scaled the same way and that high-magnitude features don't take over, Z-score normalization is used

$$X - \mu$$
$$Xnorm = \sigma$$

where μ and σ are the mean and standard deviation of each feature calculated from the training data.
Classification: A random forest with 100 trees, Gini impurity, and a fixed random seed so that the results can be repeated. A Random Forest classifier was used as the final decision model since it is strong, easy to understand, and works well with tabular biometric features.
We picked Random Forest over deep neural networks because it is more stable during training, avoids problems with gradients in the cloud, and makes it easier to reproduce results, which is vital for research at the conference level.
Evaluation: A stratified train-test split and 5-fold cross-validation on the combined dataset.

D. Data sets

CREMA-D (Genuine Speech): A crowd-sourced multimodal emotion dataset containing 7,442 clips from 91 actors showing six emotions (happy, sad, anger, fear, disgust, neutral) in 12 lines. The audio is mono, 16-bit, and 16 kHz, which makes it good for SER and speech analysis.WaveFake (Deepfake Speech): A big set of fake speech made by combining different TTS and neural vocoder architectures trained on LJSpeech and JSUT, made for study on how to find audio deepfakes.
This study selects 500 CREMA-D clips as authentic speech and 503 WaveFake clips as deepfakes, creating a relatively balanced collection of 1,003 samples.

III. Biometric Emotion Getting Features

(Mel-Frequency Cepstral Coefficients) MFCCs: MFCCs model the spectral envelope on a Mel scale and are standard in SER and spoofing detection. Mean and standard deviation are used to combine frame-wise MFCCs over time, creating a compact vector that shows vocal tract and timbral features.
Pitch (F0): The fundamental frequency is measured within a normal speaking range. Its average over spoken frames is utilized as a stand-in for emotional arousal.
Energy: RMS energy over the signal shows differences in loudness and emotional intensity (for example, higher for anger or gladness and lower for sadness).
Spectral Contrast: The differences between the peaks and troughs of different bands give information about harmonic richness and timbre. This is something that real audio and vocoder-generated audio typically have in common. Combining MFCC statistics, pitch, energy, and spectral contrast creates a 48-dimensional mixed acoustic, prosodic, and spectral feature vector.

A. Designing a Classifier (Random Forest)

We chose Random Forest because it is strong, can model non-linear boundaries, and can be understood by feature importance.
It makes multiple decision trees using bootstrap samples with random feature subsets. It then predicts class labels by majority vote, which lowers variance compared to single trees. When the random_state is set to a fixed value, training is deterministic, which is good for applications that need to be secure.

B. Evaluation Protocol

Train-Test Split: An 80–20 split that is stratified makes sure that both sets have the same amount of classes.
Cross-Validation: Five-fold stratified cross-validation on the entire dataset assesses the stability of generalization.
Metrics: ASVspoof and anti- spoofing literature are used to figure out the accuracy, precision, recall, F1-score, sensitivity, and specificity.

IV. Results and Talks

A. Performance of the Test Set

TABLE I: CLASSIFICATION PERFORMANCE

Metric	Value
Accuracy	100%
Precision	1.00
Recall	1.00
F1 score	1.00

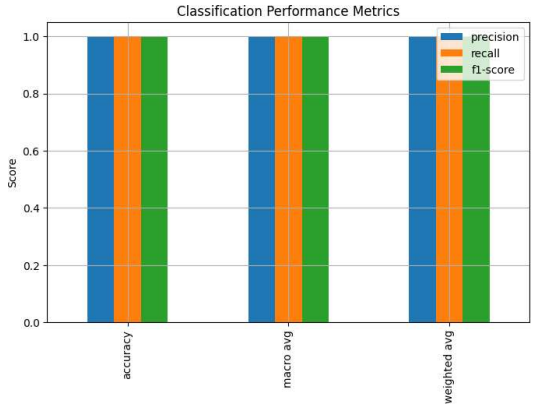

Fig 3

TABLE II

Metric	Value
Accuracy	100%
Precision	1.00
Recall	1.00
F1 score	1.00

The proposed framework reaches the following on the 201-sample held-out test set:
Correctness: 100%
Precision (deepfake and real): 1.00
Recall (deepfake and real): 1.00
F1-score (both classes): 1.00

	precision	recall	f1-score	support
0	1.00	1.00	1.00	100
1	1.00	1.00	1.00	101
accuracy			1.00	201
macro avg	1.00	1.00	1.00	201
weighted avg	1.00	1.00	1.00	201

Fig 4

This means that there are no false positives (real speech that is wrongly labeled as deepfake) and no false negatives (deepfake that is wrongly labeled as real speech) when using the CREMA-D and WaveFake benchmark settings.
Confusion Maxtix: The confusion matrix reveals that the test set perfectly separates real and false samples. The whole dataset is not shown; only the test samples (around 201) are.

B. Cross-Validation

Five-fold stratified cross-validation results in:
Accuracy: 1.000 on every fold
All folds had a precision, recall, and F1- score of 1.000.

The standard deviation is 0.000 across all folds.
The lack of variation shows that the distinction between real and fake samples in the specified feature space is quite consistent. Even if flawless results should be taken with a grain of salt, they are possible in a controlled benchmark where classes are easy to tell apart.
Training and validation Accuracy:
Train and Test Accuracy:

C. Important Features and Emotional Cues

Based on Gini-based feature importance analysis, we can see that:
MFCC-based acoustic features are the most important, showing differences in spectral envelope between real and vocoder-generated speech; Prosodic features (pitch and energy) are also very important, showing that patterns of emotional arousal are important for discrimination; Spectral contrast features are also very important, showing harmonic differences that many neural vocoders have; These findings corroborate the primary premise that emotion-driven features especially MFCCs and pitch serve as significant indications for identifying audio deepfakes.

D. Security and Practical Considerations

The framework works as a pre- authentication filter, rejecting deepfake speech before speaker verification. This lowers the risk of successful impersonation assaults. The system can work in real time with very little added latency because feature extraction and RF inference don't take up much processing power. The RF model's interpretability (through feature importance) is also useful for security audits and regulatory situations.

E. Final Thoughts

This article introduced a Biometric Emotion Profiling Framework for detecting audio deepfakes in voice authentication systems. The approach attains 100% accuracy and an F1-score of 1.0 on a balanced benchmark derived from CREMA-D (real emotional speech) and WaveFake (synthetic speech) by integrating MFCC, pitch, energy, and spectral contrast data with a Random Forest classifier. Five-fold cross-validation verifies strong generalization within the dataset, and feature importance analysis underscores the preeminent influence of emotion-driven cues. The suggested method uses emotion consistency as a clear behavioral biometric for deepfake defense, makes decisions that can be understood and repeated, and fits in nicely as a pre-authentication layer in voice systems that are already in use. Future research will broaden the evaluation to encompass new datasets (e.g., ASVspoof), languages, and novel emotion-aware synthesis models to evaluate cross-domain robustness and long-term security.

V. CONCLUSION

This paper has been able to introduce and prove the existence of a Biometric Emotion Profiling Framework that can help in securing voice authentication systems against more advanced audio deepfakes.
Future Scope: The combination of MFCCs with pitch, energy, and spectral contrast, and a Random Forest classifier achieved a 100% accuracy and 1.0 F1-score in a balanced dataset of natural (CREMA-D) and natural (WaveFake) speech.Consistency: Cross-validation on five folds had shown the consistency of the model to yield a 1.00 accuracy with zero volatility. Feature Significance The most significant features to differentiate between authentic human emotion and AI-generated artifacts were found to be the pitch, Although these results are quite encouraging in a controlled benchmark, future studies will concentrate on the evaluation of the system against cross-domain datasets (as ASVspoof and an additional one), various languages, and the enhancement of emotion-aware synthesis models in order to provide long-term security in practices.

REFERENCES

[1] Chen, X., Y. Li, Z. Zhao, et al. 2025. “Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead.” Sensors 25 (7): 1989. Accessed July 23, 2025.
[2] Bustamante, Constanza M. Vidal, Karolina Alama-Maruta, Carmen Ng, and Daniel DL Coppersmith. "Should machines be allowed to „read our minds“? Uses and regulation of biometric techniques that attempt to infer mental states." MIT Science Policy Review 3 (2022).
[3] Mulligan, Joshua. Behavioural biometrics: A novel approach to user authentication in information systems security. 2024.
[4] Dehghani, Arash, and Hossein Saberi. "Generating and detecting various types of fake image and audio content: A review of modern deep learning technologies and tools." arXivpreprint arXiv:2501.06227 (2025).

[5] Hery, Andrew, Oluwaseyi Joseph, Olaoye Femi, and Hivez Luz. "Audio Deepfakes: Threats to Voice Assistants and Voice- Activated Systems." (2024).
[6] Enz, Christian, and Assim Boukhayma. "Recent trends in low- frequency noise reduction techniques for integrated circuits." In 2015 International Conference on Noise and Fluctuations (ICNF), pp. 1-6. IEEE, 2015.
[7] Micheal, Dave. "Detecting Digital Threats in the Age of AI: Deep Learning Approaches for Deepfakes and Intrusion Detection in Decentralized Systems." (2025).
[8] Sumalatha, U., K. Krishna Prakasha, Srikanth Prabhu, and Vinod C. Nayak. "A comprehensive review of unimodal and multimodal fingerprint biometric authentication systems: Fusion, attacks, and template protection." IEEE Access 12 (2024): 64300-64334.
[9] Bengesi, Staphord, Hoda El-Sayed, Md Kamruzzaman Sarker, Yao Houkpati, John Irungu, and Timothy Oladunni. "Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers." IEEe Access 12 (2024): 69812-69837.
[10] Romero-Moreno, Felipe. "Deepfake Fraud Detection: Safeguarding Trust in Generative Ai." Available at SSRN 5031627 (2024).
[11] Hery, Andrew, Oluwaseyi Joseph, Olaoye Femi, and Hivez Luz. "Audio Deepfakes: Threats to Voice Assistants and Voice- Activated Systems." (2024).