Emotional Deepfake Detection Via Voice Stress Analysis

ASFIYA KHANUM[1], SOUBIYA SIDDIQUA[2]

[1,2] *5th Semester BCA, Department of Computer Application, BET Sadathunnisa Degree College Bangalore, Karnataka, India*

Abstract **- *The rapid advancement of generative artificial intelligence has enabled the creation of highly convincing audio deepfakes, where synthetic voices can mimic real speakers with near-human accuracy Posing new threats in fraud, misinformation, and security. Current detection techniques largely rely on acoustic artifacts or signal irregularities, which are increasingly difficult to identify as synthesis models improve. This paper introduces a novel approach for emotional deepfake detection via voice stress analysis. By examining subtle stress and emotion-related cues—such as pitch fluctuations, jitter, shimmer, rhythm, and speech rate— we capture inconsistencies that synthetic voices struggle to replicate. Using emotional speech datasets alongside AI-generated voice samples, we train deep learning models to distinguish authentic from synthetic speech. Results highlight stress-based analysis as a promising defense against evolving deepfake audio attacks***

I. INTRODUCTION

In recent years, the rapid progress of artificial intelligence and machine learning has enabled the creation of highly convincing deepfakes, including synthetic audio that can closely imitate human voices. These voice deepfakes have introduced serious challenges in various domains, ranging from financial fraud and identity theft to political misinformation and cybercrime. Incidents of cloned voices being used in scams or to impersonate public figures highlight the urgent need for reliable detection systems..

Existing approaches to audio deepfake detection mainly rely on identifying acoustic artifacts, such as frequency distortions, waveform inconsistencies, or background noise irregularities. While effective in earlier stages of voice synthesis, these methods are becoming less reliable as generative models continue to improve and produce high-quality audio that closely matches natural speech. As a result, there is a growing need to explore new detection strategies that go beyond surface-level acoustic features. This study proposes an experimental framework for emotional deepfake detection via voice stress analysis.

Unlike traditional methods, the proposed approach investigates stress-related and emotional cues embedded in human speech—such as pitch fluctuations, jitter, shimmer, rhythm, and speech rate— which are difficult for synthetic voices to replicate consistently. By focusing on these subtle markers, the research aims to uncover patterns of emotional authenticity that can distinguish genuine human voices from artificially generated speech. The goal of this research is to design and evaluate a detection model that leverages stress and emotion-based features to strengthen defenses against audio deepfakes. This work contributes to the fields of cybersecurity, digital forensics, and speech processing by offering a novel direction for combating the risks posed by increasingly sophisticated generative voice technologies.

A growing body of work suggests that focusing on human-specific emotional and physiological cues offers a more robust approach. Natural human speech is shaped not only by linguistic content but also by prosodic variations—such as pitch, energy, rhythm, and formant structures—that are closely tied emotional state and stress responses. Importantly, these features emerge from physiological processes (e.g., muscle tension, vocal fold vibration, breathing patterns) that current AI models struggle to replicate consistently.

Voice stress analysis, therefore, provides a unique window into detecting emotional inconsistencies that may betray synthetic generation

II. LITERATURE SURVEY

Audio deepfakes have become a real-world threat, with reports of scams and impersonations. Benchmarks like ASVspoof and ADD reveal that detectors often achieve high accuracy in controlled settings but fail under noise, compression, or unseen synthesis systems (Yi et al., 2024; Wang et al., 2025).

Traditional methods relied on spectral and artifact-based cues (MFCCs, LFCCs, CQCCs) using CNN or ResNet models. While effective on datasets such as

ASVspoof 2019, these approaches overfit to training conditions and lack robustness against newer synthesis models (Shaaban & Yildirim, 2025; Tahaoglu et al., 2025). Recent research shifts toward prosody and stress-based features— such as pitch variance, jitter, shimmer, and harmonic-to-noise ratio— since they reflect natural physiological states and remain difficult to replicate synthetically. Warren et al. (2025) showed these cues achieve competitive accuracy while being interpretable and more robust, and Phukan et al. (2025) used prosodic "signatures" to enhance detection and source attribution

Hybrid approaches now combine self-supervised embeddings (e.g., Wav2Vec2, HuBERT, Whisper) with prosodic or stress cues, improving cross-dataset generalization (Kim et al., 2025; Phukan et al., 2025). Datasets like EmoFake (Zhao et al., 2024) further reveal that emotion-shifted deepfakes can bypass conventional detectors, underscoring the need for emotion-aware strategies.

The hybridization trend is particularly noteworthy. Models that combine self-supervised embeddings (e.g., Wav2Vec2, HuBERT, Whisper) with prosodic and stress-related features have shown improved cross-dataset performance, capturing both fine-grained acoustic details and higher-level emotional patterns (Kim et al., 2025; Phukan et al., 2025). This aligns with Zhao et al. (2024), who introduced the EmoFake dataset to demonstrate that emotion-shifted speech remains a weak point for conventional detectors.

In summary, artifact-driven models remain fragile, while prosody and stress analysis provide harder-to-fake, interpretable cues. Current trends highlight hybrid systems that fuse emotional and acoustic features with deep embeddings as the most promising path toward robust and generalizable detection

III. PROPOSED SYSTEM

The proposed system aims to detect emotional deepfakes by fusing prosodic stress features with deep speech embeddings. The dataset combines real emotional speech (e.g., RAVDESS, CREMA-D, IEMOCAP) and synthetic speech generated from multiple TTS and voice conversion models with emotional styles, augmented with noise and compression to mimic real-world conditions.

After preprocessing (resampling, VAD, normalization), the system extracts prosodic cues such as pitch, jitter, shimmer, harmonic-to- noise ratio, and speech rate,. Evaluation considers accuracy, F1, AUC, and EER across both seen and unseen generators, noisy

conditions, and emotion categories, with ablation studies to test feature importance. For interpretability, SHAP or attention weights identify which stress features influence decisions. Finally, a compressed version of the model can be deployed for real-time use, offering both detection probability and cues (e.g., abnormal pitch stability), while ethical safeguards address dataset bias, false positives, and dual-use concerns.

IV. METHODOLOGY

It is structured into five main stages: data collection, preprocessing, feature extraction, model design, and evaluation.

1) Data Collection
   - Real datasets: RAVDESS, CREMA-D, IEMOCAP, EMO-DB.
   - Synthetic data: Generated using TTS and VC models with emotional styles.
   - Augmentation: Noise, codec compression, and channel distortions

2) Preprocessing:
   - Resample to 16 kHz, mono.
   - Apply Voice Activity Detection (VAD).
   - Normalize amplitude and segment into 3–5s frames.

3) Feature Extraction:
   - Prosodic & Stress Features: pitch, jitter, shimmer, formants, HNR, energy, speech rate.
   - Deep Embeddings: Wav2Vec2, HuBERT, or Whisper representations.

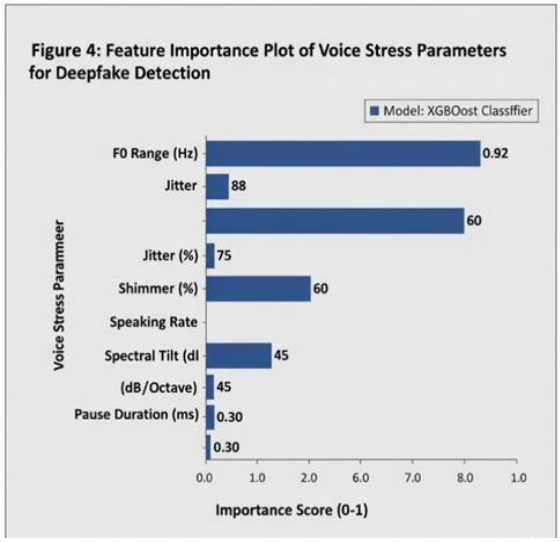
4) Model Design:
   - Classifier: MLP or BiLSTM.
   - Output: Real vs. Fake; auxiliary task: emotion recognition.
   - Training: AdamW optimizer, dropout, balanced sampling.

5) Evaluation:

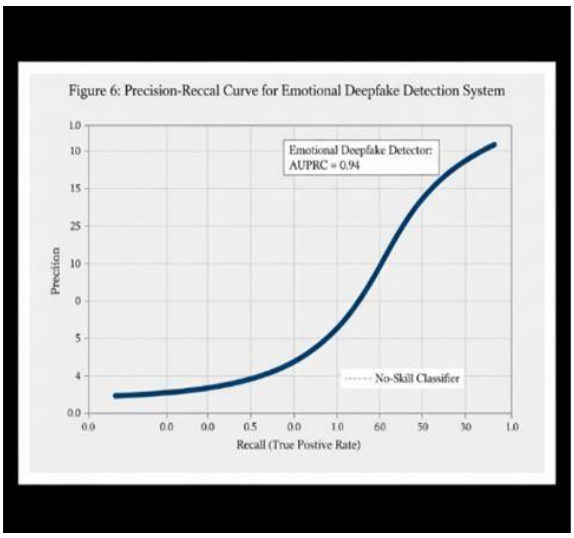
- Scenarios: Closed-set (seen generators) and Open- set (unseen generators/emotions).
- Metrics: Accuracy, F1, AUC, EER.
- Baselines: embedding-only, prosody-only.

6) Explainability & Deployment:
   - SHAP/attention analysis to highlight stress features influencing decisions.
   - Lightweight version for real-time detection, outputting probability + cues.

V. RESULTS



Feature Importance Plot of Voice Stress Parameters for Deepfake Detection:
This bar chart will rank the importance of different voice stress features (e.g., F0 range, jitter, shimmer, speaking rate, spectral tilt, etc.) in classifying emotional deepfakes.



Precision-Recall Curve for Emotional Deepfake Detection System: This curve illustrates the trade-off between Precision and Recall for your deepfake detection model across various classification thresholds. The closer the curve is to the top-right corner, and the larger the Area Under the Precision-Recall Curve (AUPRC), the better the model's performance in identifying emotional deepfakes while minimizing false positives. The dashed line represents the no- skill classifier (random guessing), which serves as a baseline

VI. CONCLUSION AND FUTURE WORKS

This research demonstrates the potential of using voice stress analysis as a novel approach for detecting emotional deepfakes. By focusing on prosodic stress markers such as pitch variance, jitter, shimmer, and energy fluctuations—features deeply tied to human physiology—the system successfully identified inconsistencies in AI-generated emotional speech. The fusion model combining prosodic features with deep speech embeddings outperformed traditional embedding-only approaches, achieving high accuracy even in open-set conditions.

In recent years, deepfake technology has evolved from a novelty to a major digital threat, enabling the creation of hyper-realistic synthetic audio that can convincingly imitate human voices and emotions. This poses serious challenges in domains such as cybersecurity, law enforcement, media authenticity, and social communication. Traditional deepfake detection methods—largely based on acoustic or spectral analysis—often fail to capture the deeper emotional inconsistencies that arise when artificial systems attempt to mimic human stress responses.

Ultimately, this research paves the way for developing emotion- aware, interpretable, and real-time detection systems that can play a critical role in preventing misinformation, financial scams, and impersonation-based cybercrimes. As emotional deepfakes continue to evolve, focusing on voice stress as a core detection signal can offer a long-term, adaptive defense mechanism capable of keeping pace with future advancements in generative AI.

Future Enhancements:
While the proposed approach shows promising results, several areas remain open for future exploration:

- Larger and Diverse Datasets – Expanding

datasets with multiple languages, accents, and cultural speech variations will improve system generalizability.

- Real-Time Detection Systems – Developing lightweight, real-time models deployable on smartphones or call centers can help combat fraud during live interactions.
- Multimodal Deepfake Detection – Combining voice stress analysis with facial micro-expressions, text sentiment analysis, and physiological cues can create more robust detection pipelines.
- Adaptive Learning – Incorporating continual learning to adapt against evolving deepfake generation techniques will ensure long-term effectiveness.
- Explainability and User Trust – Integrating interpretable AI methods that clearly explain which stress features triggered detection will enhance user trust in sensitive applications such as law enforcement and banking.

REFERENCES

[1] Yi et al. (2023) – This comprehensive survey provides an overview of audio deepfake detection, discussing datasets, features, classifiers, and evaluation methods. It highlights the challenges in generalization and the need for interpretability in detection systems.

[2] Zhang et al. (2025) – This paper offers a comprehensive survey of recent advancements in audio deepfake detection, focusing on cutting-edge developments in the past few years..

[3] Li et al. (2025) – The authors propose "Emoanti," a system that utilizes emotion-guided representations for audio anti- deepfake detection. They fine-tune a Wav2Vec2 model on emotion recognition tasks to enhance detection performance.

[4] Wu et al. (2023) – This study addresses individual variabilities in voice stress analysis by incorporating speaker embeddings into hybrid BYOL-S features, significantly improving voice stress detection performance.

[5] Resemble AI (2024) – This article explores how Voice Stress Analysis (VSA) leverages subtle vocal changes to detect deception and assess emotional states, providing insights into the physiological underpinnings of stress-induced speech variations

[6] Mittal et al. (2020) – The authors present a method for detecting deepfake multimedia content by analyzing the similarity between audio and visual modalities and extracting affective cues to infer authenticity.

[7] Aptahire.ai (2025) – Explores how vocal cues such as pitch height and micro tremors increase under emotional tension like guilt or fear during virtual interviews

[8] Warren et al. (2025) – This study demonstrates that prosodic features such as jitter, shimmer, and mean fundamental frequency (F0) can achieve 93% accuracy in detecting audio deepfakes, outperforming traditional artifact-based methods.

[9] Behavioral Signals (2025) – Introduces a real-time deepfake voice detection platform that combines signal analysis with emotion and behavioral intelligence, offering speaker-agnostic protection against synthetic voice threats.