Combating AI-Generated Audio Threats with Fine-tuned Foundation Models

Abstract

The ability to create realistic AI-generated voices has improved significantly with advancements in Text-to-Speech (TTS) and Voice-Conversion (VC) technology. While this brings many benefits, it also poses serious risks, including fraud, impersonation, and the spread of misinformation. A recent example involved an AI-generated voice mimicking U.S. President Joe Biden in robocalls that aimed to influence voter decisions. Incidents like this highlight the urgent need for better detection methods to distinguish real voices from AI-generated ones.

Currently, the tools used to detect AI-generated audio are not keeping up with the rapid progress of synthetic speech technology. Many existing detection models struggle to identify deepfake audio accurately, especially when tested on real-world examples beyond controlled laboratory conditions. Additionally, different detection methods use various datasets and evaluation criteria, making it difficult to compare their effectiveness.

To address these challenges, our solution explores how large AI models trained on vast amounts of speech data—known as foundation models—can improve detection accuracy. We evaluate models such as Wave2Vec2BERT, HuBERT, and Whisper, finding that these models perform better at identifying AI-generated voices across different datasets. Additionally, we explore "few-shot fine-tuning," a technique that allows these models to adapt more effectively to new deepfake audio. Our findings emphasize the need for stronger, more adaptable detection systems to keep up with evolving AI-generated voice technologies.

Key Challenges and Our Solutions

The Problem: Growing Risks of AI-Generated Audio

- AI can now generate highly realistic human-like voices.
- These fake voices can be used for fraud, impersonation, and misinformation.
- Current detection methods are not effective enough to reliably identify AI-generated voices in real-world scenarios.

Limitations of Current Detection Methods

- Existing models often work well only in controlled environments and fail in real-world applications.
- Different studies use different datasets, making it hard to compare effectiveness.

- There is no standardized way to test how well these models detect the latest deepfake technologies.

Our Approach: Improving AI-Generated Voice Detection

1. **Testing Large Foundation Models**: We evaluate AI models trained on vast speech datasets to see if they can better identify deepfake voices.
2. **Fine-Tuning for Better Detection**: We explore techniques to improve detection accuracy by training models with small samples of new deepfake voices.
3. **Using Diverse Datasets**: We test detection models on various datasets, including real-world deepfake audio from social media, to ensure practical effectiveness.
4. **Setting New Benchmarks**: We emphasize the need to test detection models against the latest AI-generated voices to ensure they remain effective.

Our study highlights the need for continuous improvement in AI-generated audio detection to combat deepfake threats. By leveraging advanced foundation models and improving detection techniques, we aim to develop more reliable solutions that can keep pace with evolving AI technology.