Here’s the requested breakdown:

* **Author(s):** Lim, Suk-Young and Lee (2022)
* **Dataset Name:** LJSpeech (base dataset for Tacotron)
* **Method:** Detection of AI-generated speech using spectrogram analysis
* **Accuracy:** 97-99%
* **Model Name(s):** Convolutional Neural Networks (CNNs), Long-Short-Term Memory (LSTM) networks

2.

Here’s the breakdown for the study by Mcuba et al.:

- \*\*Author(s):\*\* Mcuba et al.

- \*\*Dataset Name:\*\* Not explicitly mentioned, but likely related to chromagrams, spectrograms, mel-spectrum, and mel-frequency cepstral coefficients.

- \*\*Method:\*\* Convolutional Neural Networks (CNNs) applied to images generated from chromagrams, spectrograms, mel-spectrum, and mel-frequency cepstral coefficients.

- \*\*Accuracy:\*\* 85.91%

- \*\*Model Name(s):\*\* VGG-16 CNN with Adadelta optimizer

3.

Here’s the breakdown for the study by Conti et al.:

- \*\*Author(s):\*\* Conti et al.

- \*\*Dataset Name:\*\* ASVSpoof2019 dataset

- \*\*Method:\*\* Transfer learning from Speech Emotion Recognition (SER) for neural speech detection

- \*\*Accuracy:\*\* Not explicitly mentioned in the summary

- \*\*Model Name(s):\*\* Speech Emotion Recognition (SER) model with transfer learning

4.

The dataset is available in two forms:

1. \*\*Raw Audio Files:\*\* Found in the "AUDIO" folder, divided into "REAL" and "FAKE" directories. Filenames like "Obama-to-Biden" show which speaker's real voice (Obama) was converted to another's (Biden).

2. \*\*Extracted Features:\*\* Found in the "DATASET-balanced.csv" file, containing features extracted from one-second audio windows. The dataset is balanced through random sampling.

The total dataset size is about 4 GB.

### Objective

The primary objective of "DEEP-VOICE" is to develop a real-time detection system for AI-generated speech, focusing on DeepFake voice conversion. This system aims to improve accuracy and speed in differentiating between real and synthetic voices, enhancing the security and integrity of voice communications. By leveraging advanced machine learning models, the project will address the growing risks posed by the misuse of AI-driven voice conversion technologies in areas like fraud and impersonation.

### Conclusion & Future Direction

In conclusion, the "DEEP-VOICE" system demonstrates the potential of machine learning models to detect AI-generated speech effectively, especially in real-time scenarios. This work contributes to safeguarding communication channels by improving the ability to identify synthetic voices and reducing the risks associated with DeepFake voice technology.

For future work, the system could be enhanced by integrating more sophisticated models that learn from diverse datasets, including real-world audio samples from various sources. Additionally, expanding the scope to detect not just AI-generated voices but also hybrid or partially manipulated audio could further enhance detection capabilities. Collaboration with industries affected by DeepFake voice misuse, such as cybersecurity and telecommunications, will also help refine and implement these solutions in practical applications.

Here’s an introduction and problem statement for "DEEP-VOICE: Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion":

---

### Introduction

With the rapid development of artificial intelligence (AI) and machine learning, voice conversion technologies have advanced significantly. DeepFake voice techniques now allow AI systems to mimic human speech with high accuracy, making it difficult to distinguish between real and AI-generated voices. While this technology has applications in entertainment, voice assistants, and language translation, it also poses serious risks, such as misuse for fraud, impersonation, or other malicious purposes. The ability to detect AI-generated voices in real time is becoming increasingly important to combat these threats and maintain trust in audio communications.

### Problem Statement

Current methods for detecting AI-generated or DeepFake voices are often inadequate, particularly in real-time applications. The increasing sophistication of voice conversion techniques makes it harder to distinguish between genuine and synthetic speech. This presents a significant risk to security, privacy, and the integrity of voice communications. An effective, real-time solution for detecting AI-generated speech is crucial to mitigate these risks and prevent the misuse of voice conversion technologies.

**Conclusion**

In conclusion, the "DEEP-VOICE" system demonstrates the potential of machine learning and deep learning models to detect AI-generated speech effectively, especially in real-time scenarios. This work contributes to safeguarding communication channels by improving the ability to identify synthetic voices and reducing the risks associated with DeepFake voice technology.

For future work, the system could be enhanced by integrating more sophisticated models that learn from diverse datasets, including real-world audio samples from various sources. Additionally, expanding the scope to detect not just AI-generated voices but also hybrid or partially manipulated audio could further enhance detection capabilities. Collaboration with industries affected by DeepFake voice misuse, such as cybersecurity and telecommunications, will also help refine and implement these solutions in practical applications.

Abstract

The rapid advancement of artificial intelligence (AI) and machine learning has led to significant improvements in voice conversion technologies, enabling the creation of highly realistic DeepFake voices. While these technologies offer valuable applications in entertainment, voice assistants, and language translation, they also pose serious risks, including fraud, impersonation, and other malicious activities. Current methods for detecting AI-generated voices, especially in real-time, are often insufficient. This paper presents "DEEP-VOICE," a system designed to detect AI-generated speech in real-time with enhanced accuracy and speed. By utilizing advanced machine learning models, the system aims to distinguish between real and synthetic voices, thereby improving the security and integrity of voice communications. The project demonstrates the effectiveness of machine learning in identifying AI-generated voices and highlights the potential for further advancements by incorporating more sophisticated models and expanding detection to hybrid or partially manipulated audio. These improvements could significantly bolster defense against the misuse of DeepFake voice technologies in critical industries such as cybersecurity and telecommunications.