ETSP Report - *Detecting AI-Generated Speech: Synthetic Voice Classification*

Giuseppe Doda (*giuseppe.doda@uzh.ch*) Ronald Domi (*ronald.domi@uzh.ch*)
Arjun Roy (*arjun.roy@uzh.ch*)

December 14, 2025

Abstract

The rapid advancement of text-to-speech (TTS) and voice conversion technologies has significantly increased the realism of synthetic speech, raising serious security risks related to fraud, identity theft, and misinformation. This project focuses on the challenge of distinguishing between bona fide (human) and spoofed (AI-generated) audio by implementing and evaluating three distinct deep learning architectures.

We approach the problem from two perspectives: visual analysis of audio spectrograms and temporal analysis of acoustic features. We develop a Convolutional Neural Network (CNN) with attention mechanisms and a Vision Transformer (ViT) to process Mel-spectrograms as images, alongside a Bidirectional Long Short-Term Memory (Bi-LSTM) network that processes Mel-frequency cepstral coefficients (MFCCs) as time-series data.

Utilizing the ASVspoof 2021 dataset [2], we train and assess these models to determine the most effective approach for detecting synthetic artifacts. The study aims to provide a robust comparative analysis of these architectures, achieve acceptable performance on ASVspoof 2021 evaluation metrics, and demonstrate model robustness across different synthesis methods.

1 Introduction

Developments of the synthetic speech generation has evolved greatly in recent years. Advanced text-to-speech (TTS) and voice conversion models, such as ElevenLabs' Flash and Turbo models [3], Tortoise TTS [1], and GPT-SoVITS [4], are now capable of producing highly realistic human-like speech. While these technologies offer significant benefits for accessibility and entertainment, their ability to clone voices from as little as a few seconds of reference audio presents substantial security challenges. The normalization of these powerful tools has lowered the barrier for malicious actors to commit fraud, spread misinformation, and impersonate individuals through voice spoofing attacks.

This project addresses the critical need for automated systems capable of distinguishing between authentic (bonafide) and AI-generated (spoof) speech. The core problem is formulated as a binary classification task: given an input audio sample, the system must determine whether it was produced by a human voice or synthesized by an algorithm.

Our objective is to evaluate and compare the effectiveness of different deep learning methods for this detection task. We investigate two primary methodological approaches. The first treats audio analysis as a computer vision problem, converting raw waveforms into Mel-spectrograms to be processed by Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). This approach aims to identify visual artifacts in the frequency-time domain that are characteristic of synthesis algorithms. The second approach treats the problem as a time-series analysis task, utilizing Recurrent Neural Networks (specifically Bidirectional LSTMs) to process sequences of Mel-frequency cepstral coefficients (MFCCs). By comparing these architectures on the standardized ASVspoof 2021 dataset, we aim to identify the strengths and limitations of each modality in detecting modern speech synthesis attacks.

2 Methodology

This study employs two distinct data representations—visual spectrograms and temporal acoustic features—to drive three different deep learning architectures.

3 Experimental Setup

This is a quick overview of the adopted methods.

4 Results

This is a quick overview of the results.

5 Conclusions

This is a quick overview of the conclusions.

References

[1] James Betker. Tortoise TTS: A Multi-voice TTS System Trained with an Emphasis on Quality, 2023.

[2] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang, and J. Yamagishi. ASVspoof 2021 Challenge - Logical Access Database (1.1), 2021.

[3] ElevenLabs. Models Documentation - Flash v2.5, Turbo v2.5, and Multilingual v2, 2025.

[4] RVC-Boss. GPT-SoVITS: 1 Min Voice Data Can Also Be Used to Train a Good TTS Model, 2024.