COMILLAS
UNIVERSIDAD PONTIFICIA

ICAI

# BACHELOR'S DEGREE IN MATHEMATICAL ENGINEERING AND ARTIFICIAL INTELLIGENCE

# BACHELOR FINAL THESIS

## AUTOMATIC DETECTION OF AI GENERATED AUDIOS

**Author: Victoria García Martínez-Echevarría**

**Supervisor: Rafael Palacios Hielscher**

**Co-Supervisor: Gregorio López López**

Madrid, June 2025

**COMILLAS PONTIFICAL UNIVERSITY**
ICAI School of Engineering
Bachelor's Degree in Mathematical Engineering and AI

---

I hereby declare, under my own responsibility, that the Project entitled **Automatic Detection of AI Generated Audios**, submitted at the School of Engineering – ICAI of Comillas Pontifical University during the 2024/25 academic year, is my own work, original and unpublished, and has not been previously submitted for any other purpose.

This Project is not plagiarized from any another work, either in whole or in part, and any information taken from other sources is properly cited.

Signed by: **Victoria García Martínez-Echevarría** Date: June 10, 2025

Project submission authorized by

**PROJECT SUPERVISORS**

Signed by: **Rafael Palacios Hielscher** Date: June 10, 2025

Signed by: **Gregorio López López** Date: June 10, 2025

# ACKNOWLEDGMENTS

Many thanks to my family for their unconditional support and encouragement throughout my academic journey.

To my friends, who have accompanied me during this process, for motivating, supporting and inspiring me to keep going.

And to my supervisors, for their guidance, patience, and insightful advice every step of the way.

**COMILLAS PONTIFICAL UNIVERSITY**
ICAI School of Engineering
Bachelor's Degree in Mathematical Engineering and AI

**Automatic Detection of AI Generated Audios**

Author: Victoria García Martínez-Echevarría
Supervisors: Rafael Palacios Hielscher, Gregorio López López
Collaborating entity: ICAI – Comillas Pontifical University

# 1. Summary

This project investigates the automatic detection of AI-generated speech, a task of growing importance due to the rise of audio deepfakes and their misuse in impersonation attacks. Two reference models from the ASVspoof 2019 Challenge were retrained on a filtered version of the Logical Access partition: the first model follows a CNN-based binary classification approach, while the second adopts a one-class learning strategy. Additionally, a human classification experiment was conducted to compare human and machine performance accross multiple spoofing techniques. Results show that the one-class model outperforms humans and all other models in detecting synthetic speech, particularly when facing unseen spoofing techniques.

**Keywords:** AI-generated speech, spoofing detection, residual CNN, one-class learning

# 2. Introduction

In recent years, synthetic voice technologies—such as text-to-speech (TTS) and voice conversion (VC)— have become highly realistic, and are now capable of mimicking human voices with remarkable accuracy. While offering legitimate applications (e.g., accesibility tools and virtual assistants), these advancements also pose significant security risks, as they may fuel voice-phishing attacks [1] and biometric spoofing threats [2]. Shared challenges like ASVspoof [3] supply benchmark datasets and baseline models to promote research in this area, as many existing countermeasures struggle to generalize to unseen spoofing techniques or real-world scenarios. This project aims to improve the robustness of detection systems by eliminating duration-based biases, comparing different audio features, retraining reference models, and evaluating both human and machine performance in the classification task.

# 3. Project Definition

The main objective of this project is to evaluate the automatic detection of AI-generated speech by retraining and analyzing two reference models from the ASVspoof 2019 Challenge: the residual CNN proposed by Alzantot et al. [4], using Mel-spectrograms, MFCCs, and CQCCs; and the one-class ResNet architecture by Zhang et al. [5], trained on LFCC features. The Logical Access partition of the ASVspoof 2019 dataset was filtered to include only audio clips between 2 and 4 seconds long, maintaining class balance while eliminating potential biases related to duration.

The models are assessed on the training, development, and evaluation subsets of the filtered dataset to measure their generalization capabilities, particularly against spoofing techniques not seen during training. In addition, an external test set with high-quality synthetic samples from PlayHT [6], ResembleAI [7], and LOVO [8], was used to evaluate the models' adaptability to new spoofing methods.

Human performance was also studied through an online server [2], where participants were asked to classify audio samples as real or synthetic. A total of 1,080 responses were collected to compare human classification accuracy with that of the trained models.

# 4. Description of the Models

Two models were reimplemented:

- **Residual CNN (Alzantot et al., [4])**: A deep residual convolutional neural network that processes audio features such as spectrograms, MFCCs, and CQCCs to classify audio samples as real or synthetic. The model stacks residual blocks with batch/group normalization, dropout, and leaky-ReLU activations, followed by two fully connected layers for binary classification (Figure 1).
- **One-Class ResNet (Zhang et al., [5])**: A one-class learning model that uses a ResNet-18 architecture to learn the distribution of real speech samples, aiming to detect synthetic audios as anomalies. The model employs a One-Class Softmax loss function to concentrate the embeddings of real speech while pushing away those of synthetic audios, and it is trained exclusively on LFCC features (Figure 2).

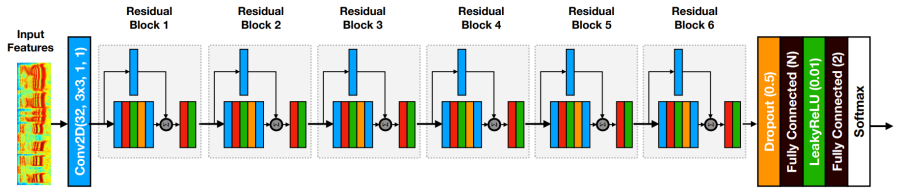


Figure 1: Model architecture proposed by Alzantot et al. [4].

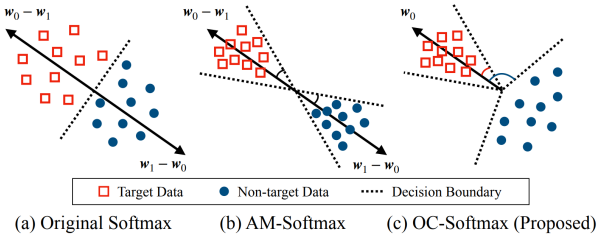


Figure 2: Embedding space division in the One-Class ResNet model [5].

# 5. Results

The bar chart in Figure 3 compares the accuracy of the four selected models (one per feature type) accross every dataset subset. While all of them achieved perfect or near-perfect accuracy on the training data and only slightly lower on the development set, the evaluation split—which contains unseen spoofing techniques—clearly ranks them: the LFCC model leads as the most robust, followed by the MFCC model, then the CQCC model, and lastly the spectrogram model, whose performance drops significantly. The external Clara samples reinforce this trend: LFCC and MFCC tie at 87.5% accuracy, spectrograms decline to 75%, and CQCCs drop to a low level (50%), confirming that one-class training generalises best to modern, unseen attacks.

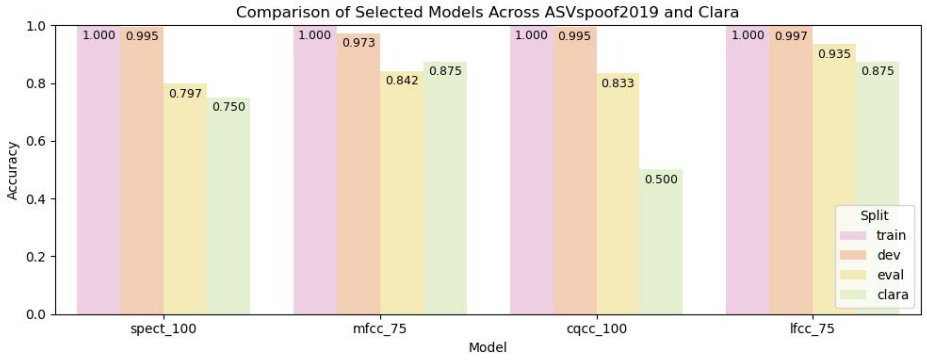


Figure 3: Comparison of accuracy across models on the ASVspoof 2019 and *Clara* datasets.

In regards of the collected human classification data, listeners achieved an overall accuracy of 69.9% on the ASVspoof subset, correctly identifying 73.4% of bona fide samples but only 67.8% of spoofed ones. Their performance dropped sharply on the *Clara* subset, where global accuracy fell to 39.8%: they still recognised genuine speech with high reliability (93.1%), but misclassified nearly 80% of synthetic clips (only 20.3% correct).

# 6. Conclusion

As seen in the results, the one-class model outperformed humans and all other models in detecting synthetic speech, particularly when dealing with unseen spoofing techniques. The residual CNN model, while still effective, struggled more with generalization and was less robust to novel attacks. Humans, on the other hand, were especially vulnerable to modern AI-generated voices, often mistaking them for real speech. These findings highlight the need for automated detection systems—as they result to be more reliable than human listeners—and the importance of developing robust models that can adapt to new spoofing techniques. Future work could explore hybrid models (e.g., ensembles) or data augmentation strategies, among other approaches, to further improve generalization.

# References

[1] S. E. Griffin and C. C. Rackley, "Vishing," in Proceedings of the 5th Annual Conference on Information Security Curriculum Development, ser. InfoSecCD '08. New York, NY, USA: Association for Computing Machinery, 2008, pp. 33–35. [Online]. Available: https://doi.org/10.1145/1456625.1456635

[2] C. Palacios-Castrillo, R. Palacios, R. Gesteira-Miñarro, A. Chávez-Macías, and G. López, "Analysis of the Security and Privacy of Smart Personal Assistants with Real and Synthetic Voices," Journal of Information Security and Applications (), vol. Under Review, 2025.

[3] A. Nautsch, X. Wang, N. Evans, T. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech," Feb. 2021. [Online]. Available: http://arxiv.org/abs/2102.05889

[4] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep Residual Neural Networks for Audio Spoofing Detection," in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2019-September. International Speech Communication Association, 2019, pp. 1078–1082.

[5] Y. Zhang, F. Jiang, and Z. Duan, "One-class Learning Towards Synthetic Voice Spoofing Detection," Oct. 2020. [Online]. Available: http://arxiv.org/abs/2010.13995

[6] PlayHT, "PlayHT: AI Voice Generator and Text-to-Speech Platform," accessed: 2025-05-06. [Online]. Available: https://play.ht/

[7] ResembleAI, "Resemble AI: AI Voice Generator and Deepfake Detection for Enterprise," accessed: 2025-05-06. [Online]. Available: https://www.resemble.ai/

[8] LovoAI, "LOVO AI: AI Voice Generator and Text-to-Speech Platform," accessed: 2025-05-06. [Online]. Available: https://lovo.ai/

**COMILLAS PONTIFICAL UNIVERSITY**
ICAI School of Engineering
Bachelor's Degree in Mathematical Engineering and AI

---

**Automatic Detection of AI Generated Audios**

Autora: Victoria García Martínez-Echevarría
Directores: Rafael Palacios Hielscher, Gregorio López López
Entidad colaboradora: ICAI – Universidad Pontifica Comillas

# 1. Resumen

Este proyecto investiga la detección automática de voces generadas por inteligencia artificial (IA), una tarea de creciente importancia debido al auge de los *audio deepfakes* y su uso indebido en ataques de suplantación de identidad. Para ello, se han reentrenado dos modelos de referencia aplicados en el *ASVspoof 2019 Challenge* sobre un subconjunto de la partición *Logical Access*: el primer modelo sigue un enfoque de clasificación binaria basado en redes convolucionales (CNNs), mientras que el segundo adopta una estrategia de *one-class learning*. Además, se ha llevado a cabo un experimento de clasificación humana para comparar el rendimiento entre personas y modelos frente a múltiples técnicas de *spoofing*. Los resultados muestran que el modelo*one-class* supera tanto a los humanos como al resto de modelos en la detección de voces sintéticas, especialmente frente a técnicas de *spoofing* desconocidas.

**Keywords:** voces generadas por IA, detección de *spoofing*, redes convolucionales residuales (*residual CNNs*), *one-class learning*

# 2. Introducción

En los últimos años, las tecnologías de voz sintética, como la conversión de texto a voz (*text-to-speech*, TTS) y la conversión de voz (*voice conversion*, VC), han alcanzado un nivel de realismo notable, siendo capaces de imitar voces humanas con gran precisión. Aunque ofrecen aplicaciones legítimas (como herramientas de accesibilidad y asistentes virtuales), estos avances también suponen riesgos significativos para la seguridad, ya que pueden facilitar ataques de *voice-phishing* [1] y ataques de suplantación de identidad [2]. Iniciativas como la competición *ASVspoof* [3] proporcionan conjuntos de datos de referencia y modelos base para promover la investigación en este ámbito, pues muchos de los sistemas de detección existentes tienen dificultades para generalizar a técnicas de *spoofing* desconocidas o a escenarios del mundo real. Este proyecto busca mejorar la robustez de los sistemas de detección eliminando sesgos relacionados con la duración, comparando distintas representaciones de audio, reentrenando modelos de referencia y evaluando el rendimiento tanto humano como automático en la tarea de clasificación.

# 3. Definición del proyecto

El objetivo principal es evaluar la detección automática de voces generadas por IA mediante el reentrenamiento y análisis de dos modelos de referencia del *ASVspoof 2019 Challenge*.

Por un lado, se ha implementado la red convolucional residual propuesta por Alzantot et al. [4], utilizando espectrogramas, MFCCs y CQCCs como características de entrada; y por otro, la arquitectura ResNet-18 de Zhang et al. [5], entrenada exclusivamente con características LFCC, siguiendo un enfoque de *one-class learning*. El conjunto de *Logical Access* del dataset *ASVspoof 2019* se ha filtrado para incluir únicamente clips de audio de entre 2 y 4 segundos, manteniendo el equilibrio entre clases y eliminando posibles sesgos relacionados con la duración.

Los modelos se evaluaron en los subconjuntos de entrenamiento, desarrollo y evaluación del dataset filtrado, con el fin de medir su capacidad de generalización, especialmente frente a técnicas de *spoofing* no vistas durante el entrenamiento. Además, se utilizó un conjunto de prueba externo con muestras sintéticas de alta calidad generadas por PlayHT [6], ResembleAI [7] y LOVO [8], para evaluar la adaptabilidad de los modelos a nuevos métodos de *spoofing*.

El rendimiento humano también se estudió a través de una plataforma *online* [2], donde se pidió a los participantes que clasificaran fragmentos de audio como reales o sintéticos. Se recopilaron un total de 1,080 respuestas para comparar la precisión de clasificación humana con la de los modelos entrenados.

# 4. Descripción de los modelos

Se han reimplementado dos modelos:

- **Residual CNN (Alzantot et al., [4])**: una red neuronal convolucional profunda con bloques residuales, que procesa características del audio como espectrogramass, MFCCs y CQCCs para clasificar las muestras como reales o sintéticas. El modelo concatena bloques residuales que incluyen normalización por lotes o por grupos (*batch/group normalization*), *dropout* y activaciones *leaky-ReLU*, seguidos de dos capas lineales para realizar la clasificación binaria (Figure 4).
- **One-Class ResNet (Zhang et al., [5])**: un modelo de *one-class learning* que utiliza una arquitectura ResNet-18 para aprender la distribución de las muestras de voz real, para detectar los audios sintéticos como anomalías. El modelo emplea una función de pérdida *One-Class Softmax* para concentrar los *embeddings* de las voces reales y alejar los de las sintéticas, y se entrena exclusivamente con características LFCCs (Figure 5).

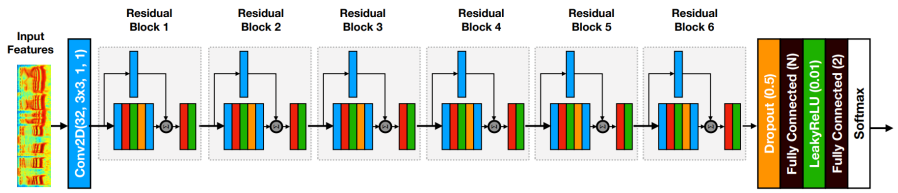


Figure 4: Arquitectura del modelo propuesto por Alzantot et al. [4].

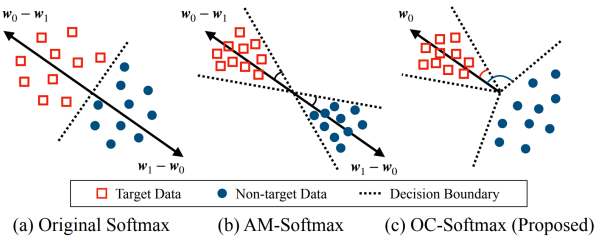


Figure 5: Espacio de embeddings del modelo One-Class ResNet [5].

# 5. Resultados

El gráfico de barras de la Figure 6 compara la *accuracy* de los cuatro modelos seleccionados (uno por cada tipo de pre-procesado de audio), evaluados en cada subconjunto del dataset. Si bien todos ellos lograron una precisión perfecta o casi perfecta en los datos de entrenamiento y solo ligeramente inferior en el conjunto de desarrollo, el conjunto de evaluación—que contiene técnicas de *spoofing* no vistas durante el entrenamiento—establece un claro orden: el modelo con LFCC destaca como el más robusto, seguido del modelo con MFCC, luego el modelo con CQCC, y finalmente el modelo con espectrogramas, cuya precisión cae significativamente. El conjunto externo de muestras *Clara* refuerza esta tendencia: LFCC y MFCC empatan en un 87.5% de precisión, espectrogramas descienden al 75% y CQCCs caen a un nivel muy bajo (50%), confirmando que el entrenamiento *one-class* generaliza mejor a ataques modernos y desconocidos.

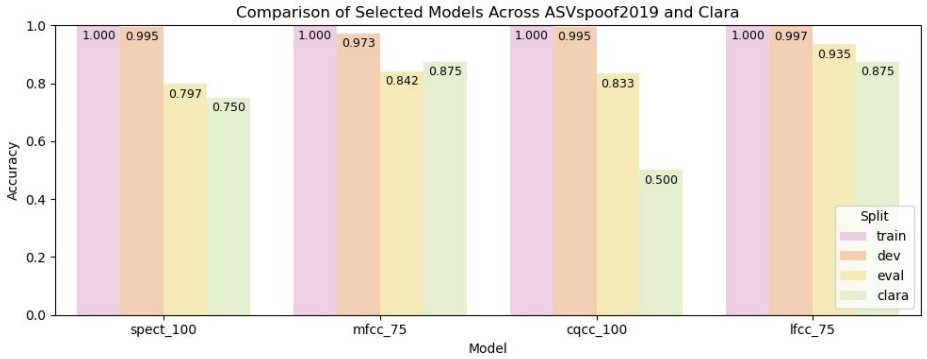


Figure 6: Comparison of accuracy across models on the ASVspoof 2019 and *Clara* datasets.

En cuanto a los datos recogidos sobre la clasificación humana, los participantes lograron una precisión global del 69.9% en el subconjunto de ASVspoof, identificando correctamente el 73.4% de las muestras reales (bona fide), pero solo el 67.8% de las sintéticas (*spoofed*). Su rendimiento cayó drásticamente en el subconjunto *Clara*, donde la precisión global se redujo a 39.8%: aunque seguían reconociendo con alta fiabilidad los audios reales (93.1%), clasificaron erróneamente casi el 80% de los clips sintéticos (solo 20.3% correctos).

# 6. Conclusiones

En conclusión, el modelo de *one-class learning* superó tanto a los humanos como al resto de modelos en la detección de voces sintéticas, especialmente al enfrentarse a técnicas de *spoofing* desconocidas. Los humanos, por su parte, fueron especialmente vulnerables a las voces generadas con técnicas modernas, confundiéndolas a menudo con voces reales. Estos hallazgos subrayan la necesidad de crear sistemas automáticos de detección—pues resultan ser más fiables que los oyentes humanos—y la importancia de desarrollar modelos robustos que puedan adaptarse a nuevas técnicas de *spoofing*. Futuras líneas de trabajo podrían explorar modelos híbridos (por ejemplo, *ensembles*) o estrategias de *data augmentation*, entre otras, para mejorar aún más la capacidad de generalización.

# Referencias

[1] S. E. Griffin and C. C. Rackley, "Vishing," in Proceedings of the 5th Annual Conference on Information Security Curriculum Development, ser. InfoSecCD '08. New York, NY, USA: Association for Computing Machinery, 2008, pp. 33–35. [Online]. Available: https://doi.org/10.1145/1456625.1456635

[2] C. Palacios-Castrillo, R. Palacios, R. Gesteira-Miñarro, A. Chávez-Macías, and G. López, "Analysis of the Security and Privacy of Smart Personal Assistants with Real and Synthetic Voices," Journal of Information Security and Applications (), vol. Under Review, 2025.

[3] A. Nautsch, X. Wang, N. Evans, T. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech," Feb. 2021. [Online]. Available: http://arxiv.org/abs/2102.05889

[4] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep Residual Neural Networks for Audio Spoofing Detection," in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2019-September. International Speech Communication Association, 2019, pp. 1078–1082.

[5] Y. Zhang, F. Jiang, and Z. Duan, "One-class Learning Towards Synthetic Voice Spoofing Detection," Oct. 2020. [Online]. Available: http://arxiv.org/abs/2010.13995

[6] PlayHT, "PlayHT: AI Voice Generator and Text-to-Speech Platform," accessed: 2025-05-06. [Online]. Available: https://play.ht/

[7] ResembleAI, "Resemble AI: AI Voice Generator and Deepfake Detection for Enterprise," accessed: 2025-05-06. [Online]. Available: https://www.resemble.ai/

[8] LovoAI, "LOVO AI: AI Voice Generator and Text-to-Speech Platform," accessed: 2025-05-06. [Online]. Available: https://lovo.ai/

# Contents

# 1 Introduction

Over the past decade, artificial intelligence (AI) has achieved remarkable progress in producing synthetic content accross various domains, including text, images, and audio. In particular, advancements in speech synthesis and audio generation have enabled the creation of realistic artificial voices that can accurately mimic human patterns. Techniques such as Text-to-Speech (TTS) and Voice Conversion (VC)—often powered by deep neural networks and transformers—can now produce audio samples that closely replicate the characteristics and nuances of human speech [1]. These advances have led to a wide range of beneficial applications such as accesibility tools, personalized virtual assistants, and language learning platforms, among others.

Nevertheless, as the quality of AI-generated audio continues to improve and become indistinguishable from real human speech, it also raises significant concerns regarding authenticity, trustworthiness, and possible misuse. These may pose a threat to privacy, trust, and security in digital communication. Audio deepfakes may be used in impersonation attacks, with the potential to undermine biometric authentication systems and deceive voice-based personal assistants. Recent studies—such as those by Gao (2022) [2]—and challenges like the ASVspoof competition [3], [4] have highlighted the need for robust countermeasures than can detect whether a given audio sample is real or synthetic.

This project focuses on the automatic detection of AI-generated audios, with the goal of improving the resilience of voice-based systems against spoofing attacks. It builds upon the work of previous participants in the ASVspoof Challenge 2019, particularly Alzantot et al. (2019) [5] and Zhang et al. (2020) [6], who developed baseline models using a variety of audio features including spectograms, Mel-Frequency cepstral coefficients (MFCCs), constant Q cepstral coefficients (CQCCs), and linear-frequency cepstral coefficients (LFCCs). In this work, the aforementioned models are retrained on a filtered subset of the Logical Access partition of the ASVspoof 2019 dataset, using only audio samples with durations between two and four seconds. The objective is to avoid duration-based biases and to evaluate the models' ability to generalize to unseen spoofing techniques, as well as to compare their performance with that of human listeners in a classification task.

## 1.1 Context and Motivation

Voice-based authentication has increasingly become a key component of digital security in sectors like banking, telecommunications, and smart home systems [7]. Its convenience and usability, however, come with significant vulnerabilities. Malicious actors are now using AI-generated audio to carry out social engineering scams known as "vishing" (voice phishing) [8], deploying voice deepfakes that convincingly resemble the voices of legitimate users. Due to the rapid growth and increasing accessibility of synthetic voice tools, even a short audio sample is now sufficient to produce highly realistic forgeries.

Cybercriminals are exploiting these weaknesses in voice biometric systems—especially in services where voice verification is often used to authenticate customers—to launch targeted attacks. As highlighted by Forrest (2024) [9], voice deepfakes can bypass biometric security measures, thus undermining trust in digital transactions. Recent research by Alali and Theodorakopoulos (2025) [10] has shown that criminals use partial deepfakes in real-world scams, deceiving both humans and automated systems with alarming success rates.

As awareness of these threats grows, research on audio deepfake detection has expanded considerably. A wide range of techniques have been proposed, from traditional signal processing methods to deep learning architectures [11]. Despite this progress, several challenges remain unresolved—particularly the ability to generalize to unseen attack types or different audio quality levels. Many of the suggested systems show strong performance on controlled benchmarks, but deteriorate significantly when faced with real-world scenarios or novel spoofing techniques. This highlights the importance of developing robust, flexible detection methods that can operate effectively beyond lab conditions.

## 1.2 Goals

The primary goal of this project is to train, evaluate, and compare several models for the automatic detection of AI-generated audio speech, focusing mainly on their robustness and generalization. Unlike previous studies that took in the entire ASVspoof 2019 dataset [4] and relied on audio samples of any length, this work restricts the dataset to a filtered subset of the Logical Access partition, using only audio clips between two and four seconds long. This approach aims to remove potential biases related to audio duration, in order to develop models that solely attend to acoustic and spectral features—rather than superficial characteristics—when classifying audio samples as real or synthetic.

To achieve this goal, the project is structured around the following specific objectives:

- Retrain existing ASVspoof models on the filtered dataset. Two reference architectures from the ASVspoof 2019 Challenge are adopted: one based on deep residual convolutional neural networks (CNNs) and another based on one-class learning. These models are adapted and retrained using the Logical Access (LA) subset of the dataset, after applying a duration filter to only include samples between two and four seconds long.
- Compare model performance accross three axes:
    1. Original benchmarks: Evaluate how well the retrained models replicate or deviate from the original ASVspoof 2019 Challenge results obtained by their respective authors, especially in training and development stages.
    2. Generalization capacity: Assess the models' ability to perform effectively on new spoofing techniques not seen during training, by testing them on the evaluation subset of the ASVspoof 2019 dataset.

3. Human classification: Conduct a human classification experiment in which participants are asked to classify audio samples as real or synthetic, allowing for direct comparison between human and machine performance.

- Analyze model behavior on audio samples generated with different voice cloning techniques. A set of external audio files—including deepfakes generated by third-party tools (PlayHT [12], ResembleAI [13], and LOVO [14])—is used to evaluate model adaptability to new spoofing methods beyond the scope of the ASVspoof 2019 dataset.

- Conduct a feature-wise comparison. As the models rely on different audio representations—spectrograms, MFCCs, CQCCs, and LFCCs—this project explores how each feature type affects classification performance and generalization.

- Investigate potential human vulnerabilities. The results of the listening experiment will also disclose whether humans are more prone to false positives or false negatives, and whether certain types of attacks are more likely to deceive human perception.

## 1.3 Project Structure

This document is structured into six main sections, each addressing a key aspect of the project. Section 2 presents an extensive overview of current research and state-of-the-art approaches in synthetic speech detection, revising recent developments and challenges in the field that have motivated this work. The detailed methodology is covered in Section 3, which is further divided into three subsections: Section 3.1 outlines the general pipeline and describes the audio features used, Section 3.2 provides an in-depth explanation of the implemented neural network architectures, and Section 3.3 details the human classification experiment and its design.

Section 4 describes the experiments conducted to evaluate the models. It covers the datasets used (Section 4.1), the configuration and setup of the models (Section 4.2), as well as the training and evaluation procedures (Section 4.3). Then, Section 4.4 presents a comprehensive evaluation of the models' performance accross multiple datasets, including the ASVspoof 2019 evaluation set and external audio samples, and compares the results with human classification performance. Section 5 analyzes the major outcomes obtained from the models and human participants, highlighting the strengths and weaknesses of each approach and showcasing the most relevant findings. Finally, Section 6 summarizes the main contributions of the project, discusses its limitations, and suggests potential pathways for future research to improve the development of robust systems for AI-generated audio detection.

The aim of this document is to provide both a technical overview of the implemented models and an accessible reference for understanding the challenges and advancements in the field of synthetic speech detection.

# 2 Related Work

Synthetic speech detection has emerged as a critical area of research in recent years due to the rapid development of high-quality text-to-speech (TTS) and voice cloning technologies [15], [16]. Modern AI-generated voices can closely replicate the characteristics of human speech, leading to security risks (e.g., spoofing voice authentication systems) and misinformation concerns. In light of this, the speaker recognition community has established shared benchmarks—such as the ASVspoof challenges—to promote the design and deployment of effective countermeasures against synthetic speech attacks. The ASVspoof 2019 challenge in particular provided a comprehensive, large-scale dataset of both logical access (LA) attacks (i.e., synthetic or converted speech) and physical access (PA) replay attacks, created with state-of-the-art TTS and voice conversion (VC) systems [17]. Research since around 2017 has increasingly focused on distinguishing bona fide human speech from such AI-generated voices, building on these standarized datasets.

Early approaches to synthetic speech detection often relied on compact signal models and handcrafted acoustic features. For example, the inaugural ASVspoof 2015 challenge [3] applied Gaussian mixture models (GMMs) using features like Mel-frequency cepstral coefficients (MFCCs) as baseline detectors. While the methods could recognize some obvious signal distortions, their performance against more sophisticated attacks was limited. In the ASVspoof 2019 logical access track, a GMM with linear frequency cepstral coefficients (LFCCs) achieved an Equal Error Rate (EER) of 13.54%, and even a stronger CQCC-GMM baseline (using constant-Q cepstral coefficients) only reached an EER of 11.04% [18]. Such results showcased the need for more powerful classifiers and richer feature representations as synthetic voices became more realistic. Over the past few years, there has been a dramatic shift in this field towards deep learning-based methods that significantly outperform traditional approaches [15].

Modern state-of-the-art systems predominantly use deep neural networks (DNNs) to automatically learn discriminative speech representations. Convolutional neural networks (CNNs) are widely used to process spectograms and other time-frequency features, capturing subtle patterns that can differentiate between real and synthetic speech better than handcrafted features. For instance, the first adoptions of deep models already showed promise: a CNN-RNN hybrid model by Zhang et al. (2017) [19] achieved then-best results on the ASVspoof 2015 dataset by combining convolutional feature extraction with recurrent layers for temporal modeling. Later studies explored more complex architectures, including residual networks (ResNets) or lightweight CNNs with gating and recurrent units. An example of the latter is the Light Convolutional GRU-based RNN proposed by Gómez-Alanis et al. (2019) [20], which achieved an EER of 6.28% on the ASVspoof 2019 LA evaluation set. Compared to GMM baselines, these deep models demonstrated a significant performance boost, with EERs dropping about an order of magnitude in known scenarios.

One notable line of research has focused on one-class learning approaches and anomaly detection techniques. Instead of training a binary classifier on real and synthetic samples, these methods only learn the characteristics of real (bona fide) speech, aiming to detect spoofed speech as deviations from this learned distribution. Zhang et al. (2021) [6] introduced a one-class learning framework that uses a ResNet-18 backbone and a novel loss function (One-Class Softmax) to concentrate the embeddings of real speech while pushing away those of synthetic samples. This approach excelled at detecting unseen attack types: without any data augmentation, it achieved an Equal Error Rate (EER) of 2.19% on the ASVspoof 2019 LA evaluation set, surpassing all prior single-model systems on this benchmark. Other recent architectures have integrated innovative network modules, such as attention mechanisms [21] or channel-wise gated Res2Nets [22], to further enhance detection performance. These studies demonstrate the community's trend towards specialized deep architectures (often ensembles of CNNs, ResNets, transformers, etc.) finely tuned for identifying synthetic speech patterns.

The choice of input features is also a key factor that has evolved over time. Classic low-level representations such as Mel-frequency cepstral coefficients (MFCCs), linear predictive cepstral coefficients (LPCCs), or spectral centroid features were common in early systems [23]. In later challenges, more effective features were introduced to target the specific characteristics of synthetic audios. The ASVspoof 2019 baseline models, for instance, integrated two other types of cepstral coefficients: constant-Q (CQCCs) and linear-frequency (LFCCs), paired with GMM classifiers [17]. Then, as deep learning gained traction, many systems now feed raw audio waveforms or spectrograms directly into neural networks. Time-frequency representations of high resolution, such as Short Time Fourier Transform (STFT) magnitude spectrograms, Mel-spectrograms, or Constant Q Transform (CQT) spectograms [24], allow CNN-based models to learn the relevant feature filters automatically. Overall, the field has seen a shift from handcrafted features to automated feature learning, though hybrid approaches (combining multiple types) are still common.

Multiple datasets have been released to support research and benchmarking in this area. The ASVspoof series provides the foundational collection: ASVspoof 2015 [3] focused on common voice conversion and text-to-speech attacks of that era, while ASVspoof 2019 [4] greatly expanded the scale and diversity of attacks by adding new techniques. In particular, the ASVspoof 2019 Logical Access (LA) partition includes spoofed and bona fide utterances from 107 speakers (male and female) accross a wide range of both TTS and VC algorithms, all derived from the Voice Cloning Toolkit (VCTK) corpus [25]. Beyond the ASVspoof data, other datasets have aimed to cover different languages and more diverse generation methods. The Fake-or-Real (FoR) dataset [26] is one remarkable example: it contains over 87,000 synthetic samples generated by a variety of modern TTS systems, and more than 110,000 genuine utterances. This collection includes highly natural-sounding fakes and is sufficiently large to train complex deep models. Also, the CFAD (Chinese Fake Audio Detection) dataset [27] provides a Mandarin Chinese benchmark with 12 different audio generation methods, and its audios include further augments with real-world degradations like background noise.

Another important aspect in the literature is the evaluation of human ability to detect synthetic speech. Surprisingly, even as algorithmic detectors have improved, studies show that human listeners often struggle to detect AI-generated voices. For instance, a recent experiment with over 500 participants found that humans correctly identified deepfake speech only about 73% of the time on average [28]. This was true even when comparing English and Mandarin speech, and providing a brief training exposure to deepfakes only resulted in a slight improvement. Namely, high-quality synthetic voices can fool humans almost one out of four times. In light with these findings, this project also assessed human performance in detecting AI-generated audios and confirmed that human accuracy was consistently lower than that of the automatic models tested. Such observations underscore why automated detection is crucial: current countermeasure models can achieve far higher accuracy than untrained humans, especially on known spoofing techniques.

# 3 Methodology

To investigate the automatic detection of AI-generated audios under controlled conditions, this project adopts and retrains reference models from the ASVspoof 2019 Challenge, focusing on the Logical Access (LA) partition of the dataset. Particularly, the data preprocessing stage was modified to filter the dataset and only include audio samples with durations between two and four seconds. This constraint was introduced to eliminate potential biases linked to audio length and to ensure that the models learn to discriminate based exclusively on acoustic properties. A key decision was to compare four distinct types of audio features—spectrograms, MFCCs, CQCCs, and LFCCs—each tested independently to isolate their influence on classification performance.

In addition to evaluating these models, a complementary experiment was designed to explore human perception in the same classification task. By deploying an interactive interface, it was possible to collect responses from participants who were asked to categorize audio samples as either real or AI-generated. This setup provides insight into the kinds of errors humans tend to make and how they compare to the performance of the automatic models [29].

## 3.1 Technical Overview

The automatic detection pipeline starts with data preprocessing and then applies feature extraction to obtain substantial audio representations. Each of these representations serves as input to a classification model trained to distinguish between real and synthetic speech. This work considers four different types of audio features, each obtained from the same audio samples: spectrograms, Mel-frequency cepstral coefficients (MFCCs), linear frequency cepstral coefficients (LFCCs), and constant-Q cepstral coefficients (CQCCs). These features vary in their signal representation approaches, frequency resolution, sensitivity to noise, and their capacity to capture synthetic speech artifacts.

The spectrograms used in this project correspond to the logarithmic magnitude of the Short-Time Fourier Transform (STFT) of the audio signal. Specifically, the STFT is computed on Hamming windows [30] of size 2048 with 25% overlap. The magnitude of each frequency component is then calculated and converted to the logarithmic scale. This results in a time-frequency matrix that captures the detailed spectral characteristics of the input waveform. Unlike handcrafted features (such as MFCCs or CQCCs) this format is relatively raw; however, it exploits the representation learning capability of deep neural networks, which are able to extract higher-level features from the spectrogram within their hidden layers [5].

Mel-frequency cepstral coefficients (MFCCs) are computed by first applying the STFT to the audio signal, and the obtained frequency spectrum is then mapped—through a filter bank—onto a Mel scale that approximates the way humans perceive sound. Lastly, a discrete cosine transform (DCT) is applied to decorrelate the resulting coefficients [31]. In this work, the first 24 coefficients are extracted to represent each audio frame [5]. MFCCs provide a compact representation that highlights relevant frequency bands. Yet, they may be less sensitive to the subtle distortions introduced by advanced speech synthesis techniques, being this a limitation that can reduce effectiveness under noisy or varying channel conditions.

Constant-Q cepstral coefficients are based on the Constant-Q Transform (CQT), which uses geometrically spaced frequency bins instead of the regularly spaced bins used by the STFT. This method provides higher frequency resolution at lower frequencies and higher temporal resolution at higher frequencies, aligning well with the human perception of pitch and timbre. To compute CQCCs, the CQT is first applied to the audio signal, followed by the calculation of a power spectrum and its conversion to the logarithmic scale. Next, a uniform resampling is performed, and finally, a DCT is applied to produce the cepstral coefficients [5]. CQCCs have been proved to outperform traditional cepstral coefficients in multiple spoofing detection challenges by successfully highlighting synthesis artifacts that appear both in low and high frequencies [32], [33].

Linear frecuency cepstral coefficients (LFCCs) share a similar computational process with MFCCs but differ mainly in the use of a linear frequency scale rather than the Mel scale. This linear spacing ensures uniform resolution accross the entire frequency spectrum, better preserving high-frequency details that may carry revealing signs of synthetic audio. LFCCs have demonstrated stronger perfomance in spoofing detection tasks, particularly against voice conversion attacks that introduce anomalies in the high-frequency range [33].

Each of these feature types offers specific advantages and trade-offs. Spectrograms retain rich temporal dynamics and detailed spectral information but require more computational resources and larger training datasets. MFCCs and LFCCs offer compact representations with varying frequency resolutions, balancing efficiency and classification performance. CQCCs stand out for their frequency scaling similar to human perception and their ability to detect subtle artifacts.

## 3.2 Model Design

This project implements and compares two distinct models for synthetic speech detection, both inspired by high-ranking systems in the ASVspoof 2019 Challenge. Each model uses different input features and architectural designs, allowing for a comparative analysis on how different representations and learning approaches affect generalization to unseen spoofing techniques.

The first model is based on the system proposed by Alzantot et al. (2019) [5] for the ASVspoof 2019 competition. It employs deep residual convolutional neural networks (ResNets) trained on three types of input features—spectrograms, MFCCs and CQCCs—which are represented as two-dimensional matrices (symbolizing time and frequency), reshaped as image-like tensors, and processed through a series of residual blocks. Each residual block (see complete structure in Figure 7) contains convolutional layers followed by normalizations (batch for spectrograms and group for MFCCs and CQCCs), leaky-ReLU activations, and dropout layers (to avoid overfitting), along with skip connections that facilitate gradient flow during training and prevent issues like vanishing gradients.

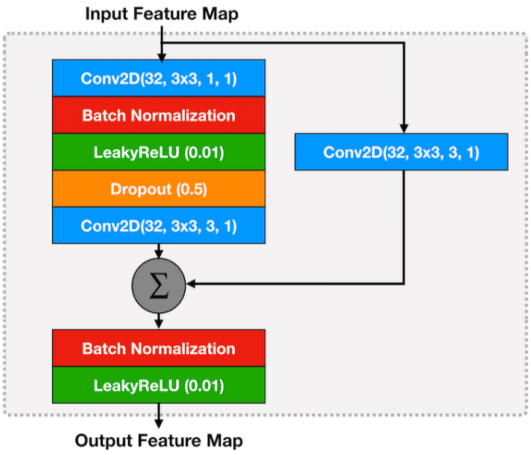


Figure 7: Detailed architecture of each residual block [5].

Before the final classification layer (see global architecture in Figure 8), the output of the last residual block is flattened and fed directly into two fully connected layers that produce the final binary prediction, indicating whether the input corresponds to human or spoofed speech. During training, the model minimizes a standard cross-entropy loss between predicted class labels (bona fide vs. spoof) and ground truth.

While the three variants of this model (trained on spectrograms, MFCCs and CQCCs) share an almost identical structure, they differ in the shape of their input tensors due to the specific dimensionality of each feature type. Consequently, the number of units in the first fully connected layer after the last block depends on the type of input feature considered.

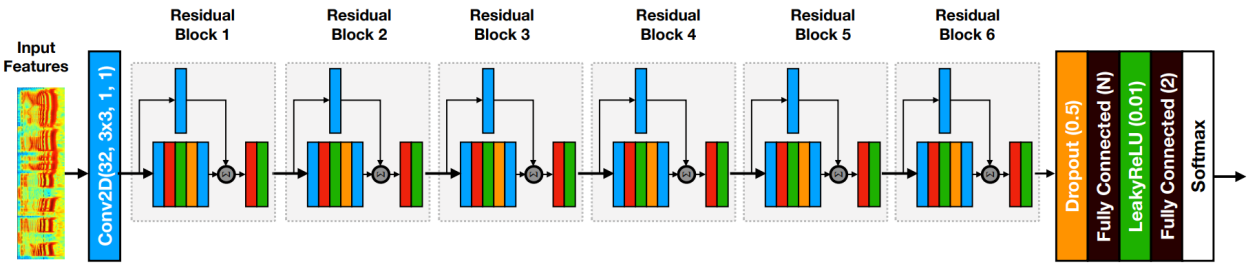


Figure 8: Model architecture proposed by Alzantot et al. [5].

The second model follows the one-class learning approach presented by Zhang et al. (2020) [6]. Instead of treating spoof detection as a binary classification problem, this method focuses exclusively on modeling the bona fide class and aims to separate unseen spoofing attacks in the embedding space. This is achieved by applying a ResNet-18 backbone to LFCC audio features as input. The model is trained with an innovative OC-Softmax (One Class Softmax) loss function—also proposed in [6]—which encourages compact clustering of bona fide embeddings while simultaneously pushing away potential spoofed samples in inference stages. Figure 9 illustrates this by showing the embedding space and how clusters are formed when applying different loss functions.

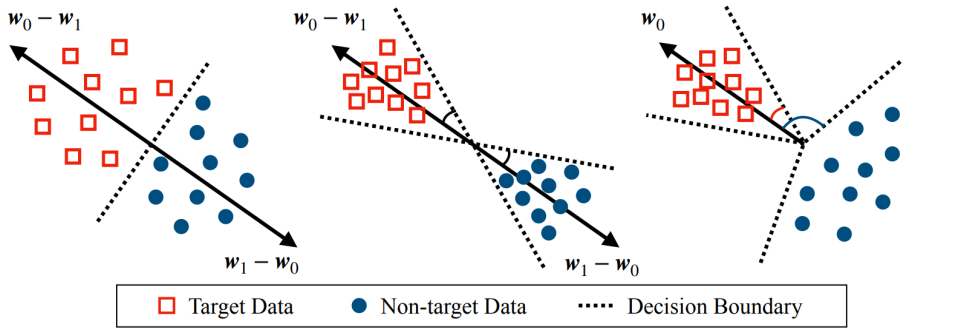


Figure 9: Embedding space division applying three different loss functions [6].

The OC-Softmax loss function applies a cosine-based projection of embeddings onto a unit hypersphere, introducing an angular margin to penalize deviations from the class center. This makes the system especially robust against unseen attack types, which is a truly beneficial property for the ASVspoof evaluation set as it includes spoofed samples generated with algorithms not present in the training set.

Both models are trained on a subset of the ASVspoof 2019 Logical Access dataset, containing audio clips between 2 and 4 seconds long. Evaluation is performed on the corresponding filtered evaluation set to ensure consistency in input length.

## 3.3 Human Data Collection

To enhance the evaluation of the trained detection systems, a custom interactive server developed in a previous project [29] was used to gather human assessments regarding whether a series of speech samples were real or AI-generated. This platform presented participants with audio clips and asked them to classify each clip by selecting one of four options: Human, Sounds Human, Sounds AI, or AI. This set of choices allowed participants to express uncertainty when differentiating between real and synthetic speech. Figure 10 below shows a screenshot of the interface as displayed to participants.

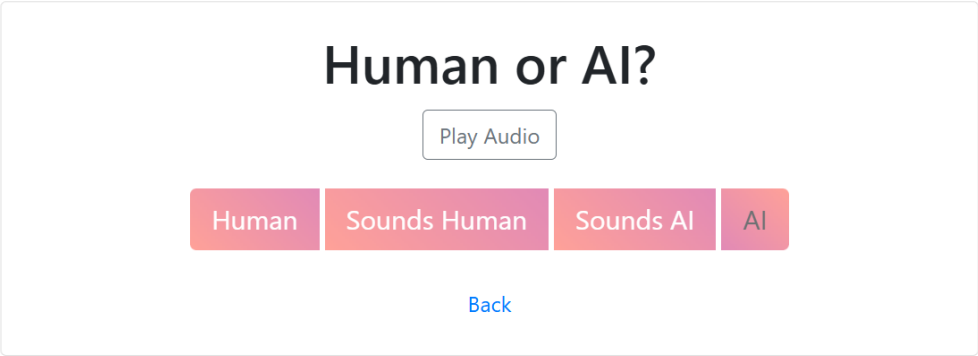


Figure 10: Screenshot of the user interface designed to collect human judgements.

The server contained a total of 701 audio samples randomly selected from the filtered ASVspoof 2019 evaluation set, restricted to utterances between two and four seconds long. Additionally, eight external audio samples generated using third-party synthesis tools [29]—PlayHT [12], ResembleAI [13], and LOVO [14]—were included to evaluate performance on voices generated by techniques outside the ASVspoof dataset.

After conducting this experiment, participants completed 108 sessions in total, each of them consisting of 10 audio clips presented sequentially. Accross all sessions, 972 responses were collected for ASVspoof evaluation audios and 108 responses for the external audio samples, which adds up to a total of 1080 responses.

For analysis purposes, the four-category response was simplified into a binary decision by merging the "Sounds Human" option with "Human", and the "Sounds AI" option with "AI". This approach maintained consistency with the automated detection task, which relies on a binary classification task of real vs. synthetic speech. The collected data provides valuable insights into how well humans can distinguish between real and AI-generated voices, what misclassification errors they tend to make, and how their performance compares to that of the automatic models.

# 4 Experiments

The experiments in this project consist of retraining spoofing detection models on a filtered subset of the ASVspoof 2019 Logical Access dataset, only containing audio clips between two and four seconds long. The trained models are then evaluated on both development and evaluation partitions of said subset, as well as on an independent set of external audio samples generated by third-party tools. The main goal is to assess the models' generalization capacity to classify unseen spoofing techniques and to compare their performance with that of human listeners in a classification task.

## 4.1 Dataset

The ASVspoof 2019 challenge is a broadly adopted benchmark for synthetic speech detection. Its Logical Access (LA) set has a large collection of both real (bona fide) and spoofed audio samples. The latter are generated using a variety of advanced text-to-speech (TTS) and voice conversion (VC) techniques, which makes it a suitable dataset for training and evaluating spoofing detection models. The data is split into three partitions—training, development, and evaluation—and each of them contains a diverse set of speakers, balanced between male and female [17].

A crucial preprocessing step in this work was to filter the dataset to only include audio clips with durations between two and four seconds. This decision was made to eliminate potential biases where models might attend to superficial temporal or length-based features rather than focusing on the acoustic properties of the audio samples. Table 1 shows that although the absolute number of samples in each partition is reduced, the overall class balance remains similar to the one of the original dataset.

| Set | Class | Original Count | Original % | Filtered Count | Filtered % |
|---|---|---|---|---|---|
| train | SP | 22800 | 89.83% | 13208 | 87.63% |
| | BF | 2580 | 10.17% | 1865 | 12.37% |
| dev | SP | 22296 | 89.74% | 12350 | 88.2% |
| | BF | 2548 | 10.26% | 1653 | 11.8% |
| eval | SP | 63882 | 89.68% | 33401 | 86.17% |
| | BF | 7355 | 10.32% | 5360 | 13.83% |

Table 1: Counts and ratios (SP-BF) before and after filtering the data by duration.

The original dataset presents a wide range of utterance lengths, with some samples extending beyond 10 seconds in some cases, and a notable skew towards shorter durations. Figure 11 shows the distribution of audio lengths by class for each partition of the dataset, both before and after filtering. Once the filtering is applied, the distribution of audio lengths becomes more uniform and tightly constrained within the targeted range, effectively removing outliers and reducing variance.

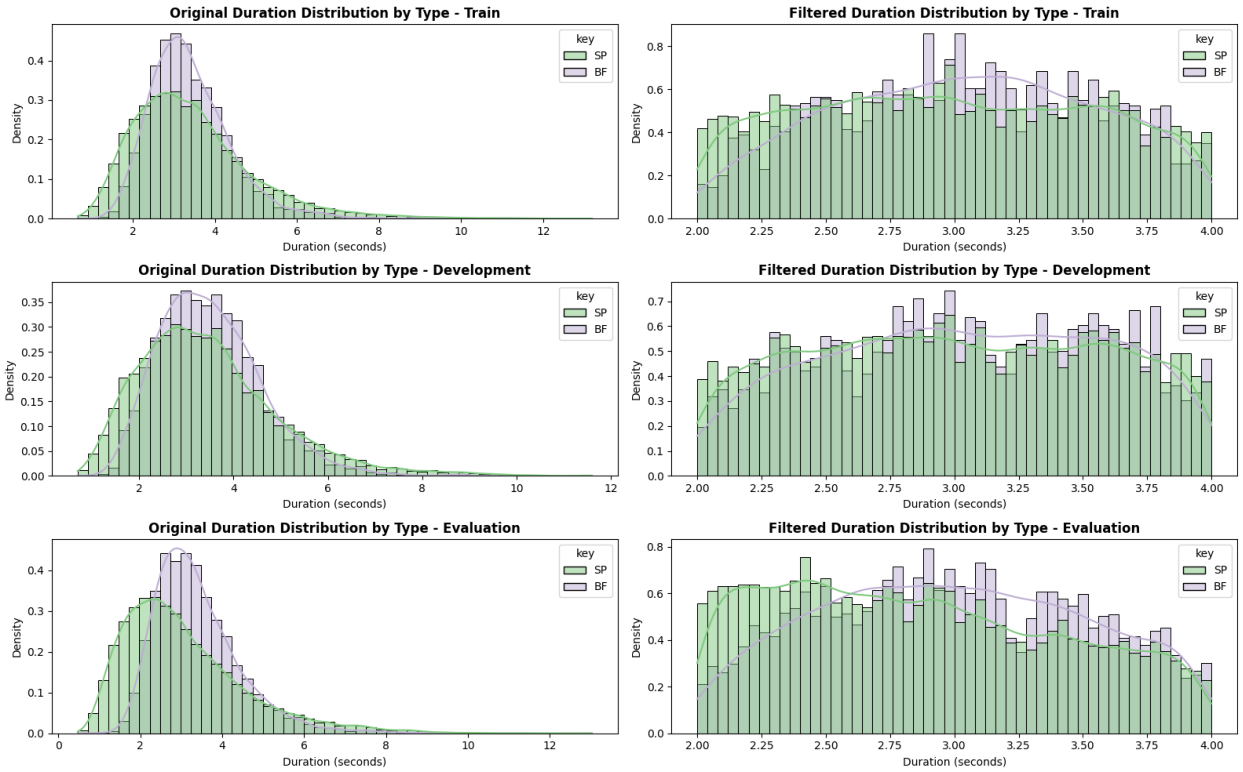


Figure 11: Distribution of audio lengths by class for each partition, before and after filtering.

In addition to the ASVspoof 2019 data, this project incorporates an external dataset composed of eight audio samples generated using third-party voice synthesis systems (two generated with PlayHT [12], two with ResembleAI [13], two with LOVO [14], and two authentic human utterances). As described in [29], these audios were originally used to evaluate the vulnerability of Smart Personal Assistants (e.g., Amazon Alexa) to realistic synthetic speech. In the context of this work, these audio samples are repurposed to evaluate classifier robustness to unseen spoofing methods and to study how well automatic models—trained exclusively on ASVspoof 2019—can generalize to modern, real-world AI-generated voices with different acoustic characteristics.

## 4.2 Setup and Configuration

The preprocessing and data handling stages were implemented in separate modules corresponding to each model architecture. In both cases, these scripts perform loading, batching, and basic audio transformations, in order to prepare inputs to be fed into the models.

Feature extraction was performed externally using MATLAB scripts—specifically for CQCCs and LFCCs—which involved computationally intensive processes and resulted in large `.mat` files. This approach aligns with the original papers' methodologies and ensures accurate calculation of these specialized features.

The values of the training hyperparameters were set to those recommended as optimal by the original authors of the models. For the first model, Alzantot et al. (2019) [5] suggested using a learning rate of 0.0001 and a batch size of 32, while Zhang et al. (2020) [6] recommended a learning rate of 0.0003 (with 50% decay every 10 epochs) and a batch size of 64 for the second model. In terms of experimentation, the only variable that was modified was the number of training epochs, as multiple versions of each model were trained to observe their performance differences and identify the best-performing configurations.

In some cases, early stopping was implemented with a considerably small threshold ($\epsilon = 1 \times 10^{-100}$). For instance, the MFCC model trained for 91 epochs (see Section 4.3) was stopped early due to this condition. However, the internal configurations and structures of the models—such as optimizers, schedulers, and other architectural details—were kept consistent with the original designs, ensuring that any differences in performance could be attributed to the number of epochs rather than to changes in model setup.

The experiments were conducted on a server provided by the Comillas Pontifical University, equipped with GPU acceleration through the JupyterHub platform (`https://jupyterhubdgx.comillas.edu`). Although specific hardware details such as GPU type and number of cores were not disclosed, the platform's resources were sufficient to support efficient training of the models.

Regarding software dependencies, the training pipeline utilized Python along with deep learning frameworks such as PyTorch, supported by libraries like NumPy and SciPy for numerical computations, and Librosa and SoundFile for audio processing. The feature extraction step required MATLAB toolboxes suited for audio analysis, particularly for computing cepstral coefficients.

## 4.3 Training and Validation

For each feature type (spectrograms, MFCCs, CQCCs, and LFCCs), several models were trained with different number of epochs to identify the most effective configurations.

As aforementioned, the proposed training pipelines were barely modified to adapt to the filtered dataset, keeping the original architectures and hyperparameters as close as possible to the ones described in the papers. In terms of loss functions, the first model used a weighted cross-entropy loss to account for the class imbalance in the training data. Thus, the ratio of weights assigned to the bona fide and spoofed classes is 9:1 respectively [5], given that the training set contains approximately 90% spoofed samples and 10% bona fide samples (see exact percentages in Table 1). For the second model, the tailored OC-Softmax loss function was employed, which focuses only on the genuine speech distribution and aims to separate unseen spoofing attacks in the embedding space [6].

In terms of evaluation, the primary performance metrics considered were accuracy and Equal Error Rate (EER). The former reflects the proportion of correctly classified samples, while the latter is calculated by identifying the point where the false acceptance rate (FAR) and false rejection rate (FRR) are equal. These metrics are commonly used in spoofing detection tasks and provide a comprehensive view of the models' performance across both classes. It should be noted, however, that the t-DCF (tandem-Detection Cost Function) [34] suggested by the ASVspoof 2019 organizers was not used in this project. The t-CDF metric, designed for speaker verification tasks, incorporates factors like user identity, which were not relevant to this work's focus on detecting whether an audio is real or spoofed.

The models performed considerably well on the training and development sets; notwithstanding, performance on the evaluation set dropped, as discussed in the next section.

## 4.4 Performance Analysis

The performance of the models is evaluated from three different perspectives: on the training and development sets, on the evaluation set, and on the external audio samples generated with third-party tools. In the last two scenarios, the models are tested on data not seen during training and generated with techniques not used for generating training samples, therefore allowing for a more realistic assessment of their generalization capabilities. The metrics observed in this section are the accuracy—to measure the proportion of correctly classified instances— and the Equal Error Rate (ERR)—to compare the performance of the original models.

The first analysis is based on the performance of the models on the training and development sets. As expected, the accuracy is higher in the training set than in the development set for all models (Figure 12). However, for most models, the difference in accuracy between the two sets is barely noticeable (see numerical values in Table 2), except when using MFCCs as input features. Additionally, it appears that increasing the number of epochs does not notably affect accuracy, indicating that the model is not learning more with additional epochs (nor is it overfitting). But again this trend does not hold for the models trained on MFCCs, where increasing the number of epochs actually results in a drop in development accuracy.

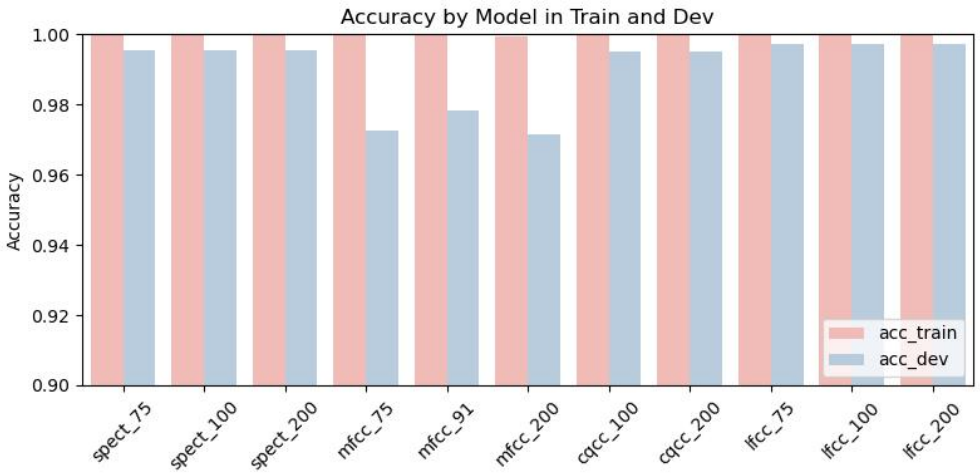


Figure 12: Accuracy bar plot of each model on the training and development sets.

The perfect training accuracy can be attributed to the models' ability to memorize the underlying details of the training samples. This suggests that, after a certain point, the models are not learning new patterns but rather focusing on the characteristics of the training set.

The EERs calculated for each model (shown in Table 2) are generally higher than those reported in the original papers for the development stage. This is not surprising given the reduced dataset size after the filtering process (training on a smaller dataset may result in less specialized models). Furthermore, it is worth noting that the original models might have indirectly incorporated the duration of the audio samples as a feature, which could have contributed positively to their performance on the original dataset. While it is difficult to determine the exact impact of this hypothesis, it is a plausible factor to consider.

| Model | Train accuracy | Dev accuracy | Dev EER | Original Dev EER |
|---|---|---|---|---|
| **spect_75** | 1.000000 | 0.995358 | 0.004344 | 0.0011 |
| **spect_100** | 1.000000 | 0.995358 | 0.004263 | 0.0011 |
| **spect_200** | 1.000000 | **0.995430** | 0.002929 | 0.0011 |
| **mfcc_75** | 0.999668 | 0.972577 | 0.043822 | 0.0334 |
| **mfcc_91** | 0.999602 | **0.978290** | 0.040631 | 0.0334 |
| **mfcc_200** | 0.999536 | 0.971649 | 0.090397 | 0.0334 |
| **cqcc_100** | 0.999934 | 0.995001 | 0.006728 | 0.0001 |
| **cqcc_200** | 1.000000 | **0.995215** | 0.008567 | 0.0001 |
| **lfcc_75** | 1.000000 | 0.997143 | 0.007395 | 0.0020 |
| **lfcc_100** | 1.000000 | **0.997286** | 0.005475 | 0.0020 |
| **lfcc_200** | 1.000000 | 0.997143 | 0.007314 | 0.0020 |

Table 2: Accuracy and EER values of each model on the training and development sets.

The evaluation set, as it contains spoofed audio generated using techniques different from those used to generate the training set, shows a decrease in accuracy for all models (Table 3). The spectrogram-based models suffer the greatest drop in performance, while the LFCC model maintains the highest accuracy. This proves the superior generalization ability of the one-class approach in handling spoofing techniques not seen during training.

When comparing model performance to a baseline simplistic model that classifies all audios as spoofed, the models shaded in yellow in Table 3 perform worse than the baseline, which would achieve an accuracy of around 86% due to the class distribution in the evaluation set (see Table 1). The only model that surpasses this baseline is the LFCC model (shaded in green), in all three versions (75, 100, and 200 epochs).

The EERs of the models on the evaluation set are also higher than those reported in the original papers, following the same trend observed in the development set.

| Model | Eval Accuracy | EER | Original Eval EER |
|---|---|---|---|
| **spect_75** | 0.788447 | 0.139396 | 0.0968 |
| **spect_100** | **0.797348** | 0.125166 | 0.0968 |
| **spect_200** | 0.780940 | 0.137352 | 0.0968 |
| **mfcc_75** | 0.842393 | 0.147221 | 0.0933 |
| **mfcc_91** | 0.832538 | 0.133189 | 0.0933 |
| **mfcc_200** | **0.843941** | 0.136575 | 0.0933 |
| **cqcc_100** | **0.833338** | 0.104786 | 0.0769 |
| **cqcc_200** | 0.757462 | 0.119310 | 0.0769 |
| **lfcc_75** | **0.934573** | 0.036393 | 0.0219 |
| **lfcc_100** | 0.931426 | 0.035325 | 0.0219 |
| **lfcc_200** | 0.932948 | 0.034899 | 0.0219 |

Table 3: Accuracy and EER values of each model on the evaluation set.

Lastly, the accuracy of the models when evaluated on the external dataset (referred to as *Clara*, being the name of the person who recorded those audios) [29] is shown in Figure 13. This dataset contains eight audio samples, two of which are real human recordings and the remaining six are AI-generated voices. The spoofing techniques used to generate these audios—PlayHT [12], ResembleAI [13], and LOVO [14]—are not present in the ASVspoof dataset, allowing for a more realistic evaluation of the models' generalization capabilities.

Overall, the accuracies achieved on this set are relatively low, with the worst-performing models being those based on CQCCs features. Similarly to the evaluation set, the LFCC models outperform the others, reaching an accuracy of 87.5% on all versions. Still, it must be noted that the limited size of this dataset also contributes to the reduced performance, without this being a definitive indicator of the models' generalization capabilities.

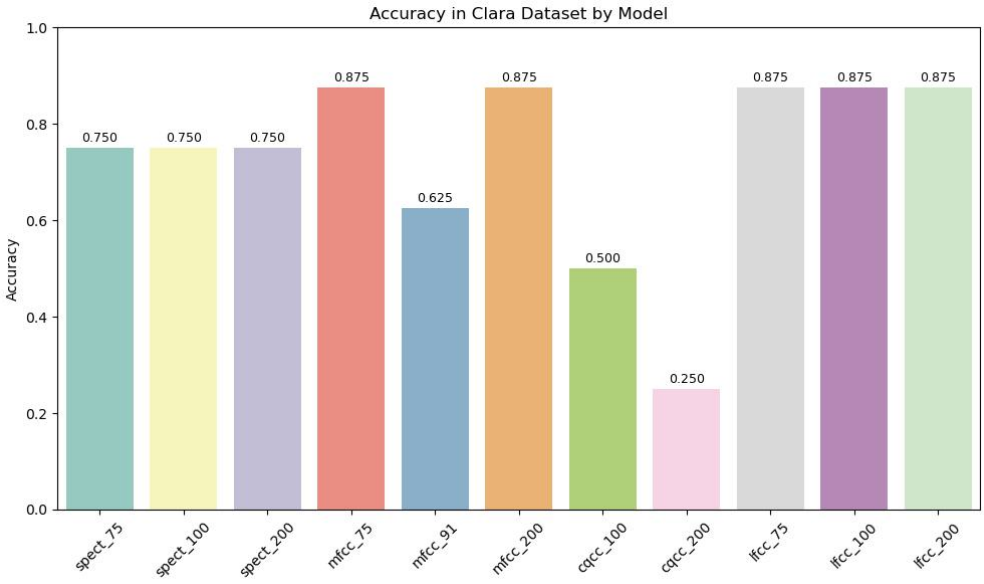


Figure 13: Accuracy of each model on the external dataset.

These results highlight that while the trained models perform well on familiar techniques, their performance drops significantly when faced with new spoofing methods. This further emphasizes the difficulty of developing robust synthetic speech detection systems that can handle a wide range of spoofing methods, especially those not represented in the training data.

# 5 Results

As this project collects performance data from two distinct sources—the retrained spoofing detection models and the human judgements gathered through the interactive server— a specific analysis of the results obtained in each case is presented below.

First, aiming at a more concise comparison accross model types, a single representative version was selected for each input feature type. This selection was guided by the development and evaluation performance presented in Section 4.4, considering as well the number of training epochs (preference was given to models that reached strong performance in fewer epochs). Accordingly, the selected versions are: 100 epochs for the spectrogram-based model, 75 epochs for MFCCs, 100 epochs for CQCCs, and 75 epochs for LFCCs.

The accuracy achieved by these models on each partition of the ASVspoof 2019 dataset (train, development, and evaluation), together with their performance on the external *Clara* set, is shown in Figure 14. Although all models reached perfect accuracy on the training set, performance decreased slightly on the development set and more notably on the evaluation and external sets (where the models are tested on unseen spoofing techniques).

Among the four, the LFCC model showed the best performance accross all partitions but especially on the evaluation set, suggesting superior generalization. The MFCC model followed closely, matching the LFCC model on the *Clara* set. In contrast, the spectrogram-based model struggled most on unseen attacks, and the CQCC model, while adequate on ASVspoof data, fell to chance level (50 %) on the external samples—suggesting that this feature representation is less suited to detecting modern AI-generated speech.

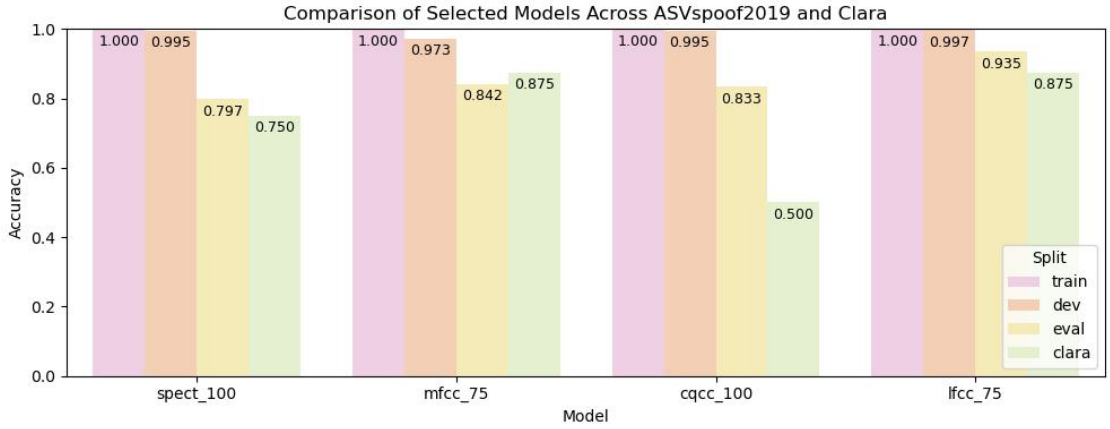


Figure 14: Comparison of accuracy across models on the ASVspoof 2019 and *Clara* datasets.

Furthermore, given that the spoofed audios were generated using different methods, it was deemed interesting to explore which of the techniques present in the ASVspoof 2019 evaluation set were more challenging for the models to detect. To do so, the accuracy of each of the four selected models was calculated independently for each spoofing technique, and the results were then averaged to obtain a mean accuracy per method (Figure 15).

The techniques A17 and A18 stand out as the most difficult to detect, with mean accuracies of 19.7% and 51.0%, respectively. According to the ASVspoof2019 documentation [17], both A17 and A18 rely on advanced voice conversion (VC) systems that do not require parallel training data, and are considered highly deceptive in spoofing contexts.

A17 was built on a VAE-based voice conversion pipeline that replaces the traditional vocoder with a generalized direct waveform modification method [35]. This method was rated as one of the most effective in the Voice Conversion Challenge 2018, due to its capacity to closely replicate the spectral structure of natural speech [36]. Conversely, A18 implements a non-parallel VC framework grounded in i-vector PLDA-based (Probabilistic Linear Discriminant Analysis) speaker representation. It leverages transfer learning from a speaker verification system to perform voice conversion through a regression-based mapping in the i-vector space [37].

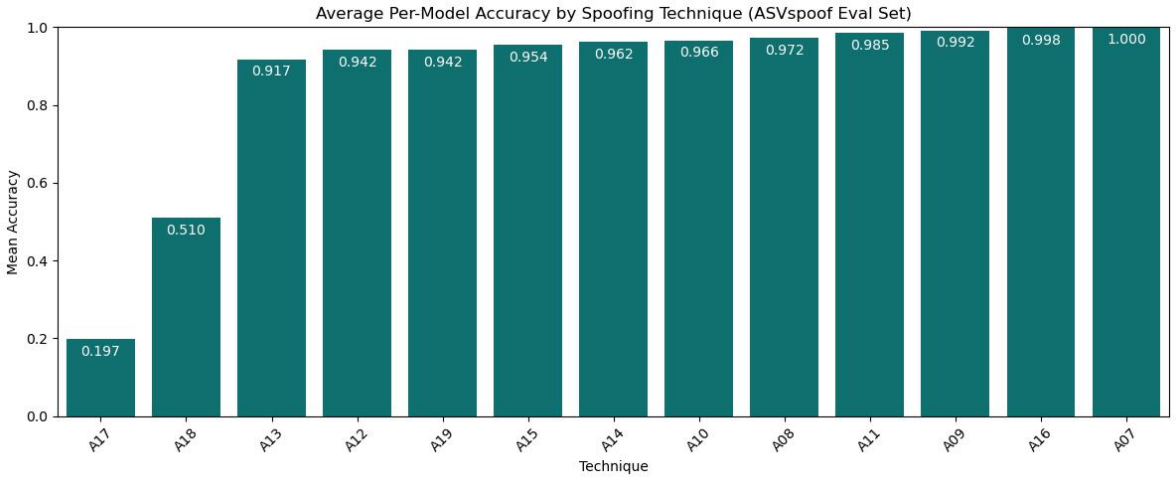


Figure 15: Mean accuracy per spoofing method on the evaluation set.

A similar procedure was applied to the external dataset. Table 4 reports the accuracy of each selected model for each spoofing technique present in the *Clara* dataset (PlayHT, ResembleAI, and LOVO). All models correctly classified samples from PlayHT and LOVO, but performance dropped for Resemble AI, especially for the CQCC model, which failed to classify any of its samples correctly. On average, Resemble AI was the most deceptive technique in this dataset.

PlayHT uses a transformer based Text-to-Speech (TTS) pipeline with an Adaptive Speech Contextualizer (ASC) that captures conversational characteristics. Text and audio prompts are converted into Mel-spectograms, and then rendered into high-quality waveforms by a neural vocoder [12]. ResembleAI enables zero-shot voice cloning through an encoder-decoder architecture: it extracts a speaker embedding from a short audio sample, then conditions a TTS model to generate speech accordingly and a neural vocoder (e.g., WaveNet) synthesizes the final waveform, using transfer learning for fast adaptation [13]. LOVO employs a latent diffusion-based TTS framework refined with CLAP (Contrastive Language-Audio Pretraining) [38] embeddings, generating both continuous Mel-spectrograms and discrete acoustic tokens that a vocoder transforms into realistic speech accross hundreds of voices and languages [14].

In parallel, human performance data was collected through a custom interactive platform [29], with the aim of gathering insights into how accurately humans can differentiate between genuine and AI-generated speech, particularly when exposed to spoofing techniques both seen and unseen in standard datasets. Participants were presented with a series of ten audio samples in each session and were asked to determine whether each clip sounded human or AI-generated.

| Model | PlayHT | ResembleAI | LOVO |
|---|---|---|---|
| spect_100 | 1.0 | 1.0 | 1.0 |
| mfcc_75 | 1.0 | 0.5 | 1.0 |
| cqcc_100 | 1.0 | 0.0 | 1.0 |
| lfcc_75 | 1.0 | 0.5 | 1.0 |
| **Accuracy** | 1.0 | 0.5 | 1.0 |

Table 4: Accuracy of each model on the *Clara* dataset per spoofing technique.

A total of 709 unique audio files were uploaded to the server, including 701 clips from the filtered ASVspoof 2019 evaluation set and 8 external samples generated using third-party text-to-speech (TTS) services (PlayHT [12], ResembleAI [13], and LOVO [14]). The ASVspoof subset consisted of 255 bona fide and 446 spoofed samples, while the external *Clara* dataset included 2 bona fide and 6 spoofed examples. In total, 80 participants took part in 108 sessions, providing 1,080 individual human responses—972 corresponding to ASVspoof samples and 108 to *Clara*. Below are the accuracies observed accross the different datasets, both overall and divided by subset (Table 5).

| Set | Global Acc. | AI Acc. | Human Acc. | AI Prop. | Human Prop. |
|---|---|---|---|---|---|
| Entire dataset | 0.6685 | 0.6239 | 0.7487 | 63.75% | 36.25% |
| ASVspoof 2019 | 0.6986 | 0.6780 | 0.7339 | 63.62% | 36.38% |
| Clara (external) | 0.3981 | 0.2025 | 0.9310 | 75.00% | 25.00% |

Table 5: Global and per-class accuracy, together with the proportion of AI and human samples in each dataset.

To interpret the obtained data, the analysis was structured into three perspectives: first, a global view that aggregates all user responses across the entire dataset (i.e., all audio samples uploaded to the platform), represented by plots in blue tones; and then, a breakdown of the responses by dataset, distinguishing between the ASVspoof subset (orange tones) and the *Clara* subset (green tones). This separation—applied already in Table 5—enables a more precise interpretation of user performance, isolating trends that may be linked to the origin and generation method of each audio sample.

Following said structure, the first step in the analysis was to assess overall human performance across all audio samples. As shown in the confusion matrix in Figure 16, humans tended to perform relatively well at identifying real audio, with a true positive rate of 62.4%. Overall accuracy was 66.85% (Table 5, first row), indicating that humans correctly classified approximately two-thirds of the samples. However, a large proportion of spoofed audios were misclassified as human, resulting in a false negative rate of nearly 37.6%.

The overall error rate (ER) was 33.15%, indicating that humans made mistakes in about one-third of the cases. When normalizing the confusion matrix by the number of responses per class, the misclassification rate for AI-generated audio as human reached 37.6%, while the rate for human audio misclassified as AI-generated was notably lower at 25.1% (see bar plot in Figure 16).

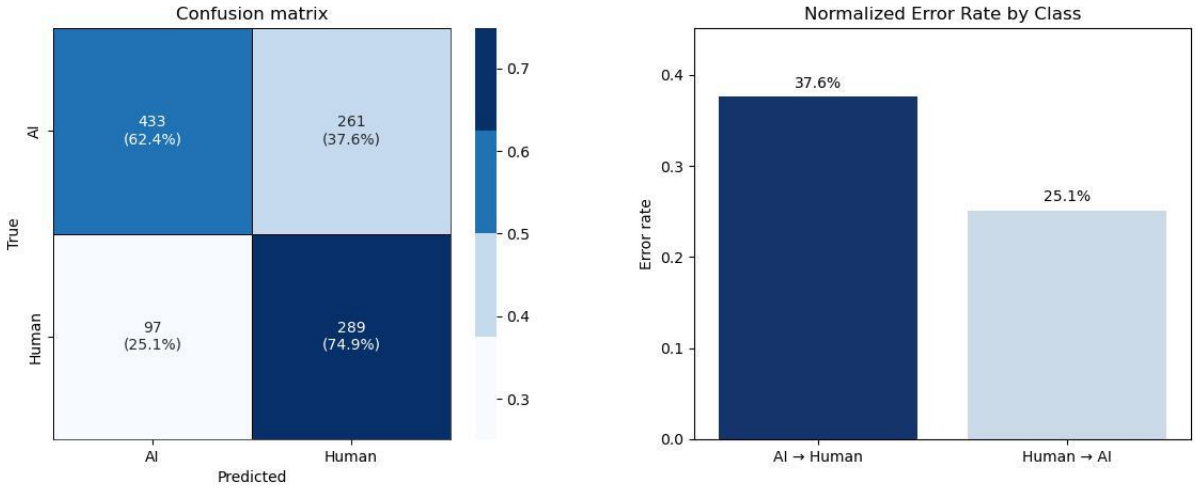


Figure 16: Global human performance confusion matrix and by-class error rates.

When the data is split by dataset, as shown in the confusion matrices in Figure 17, human performance appears to vary significantly. For the ASVspoof subset, the accuracy on bona fide samples was considerably high (73.4%), whereas the accuracy on spoofed audios was slightly lower, reaching 67.8% (Table 5).

In contrast, for the *Clara* dataset, which includes only eight audio samples, human performance was markedly worse. The global accuracy on ASVspoof responses was 69.9%, compared to just 39.8% on the *Clara* subset, highlighting that listeners are more than 1.5 times more accurate when judging traditional spoofing techniques than when faced with modern AI-generated voices. Despite the small size and although the accuracy on bona fide samples was notably higher than in other cases (93.1%), the false negative rate was strikingly high, with 79.7% of spoofed audios misclassified as human. This indicates that the newer spoofing methods used in PlayHT, Resemble AI, and LOVO are particularly challenging for human listeners.

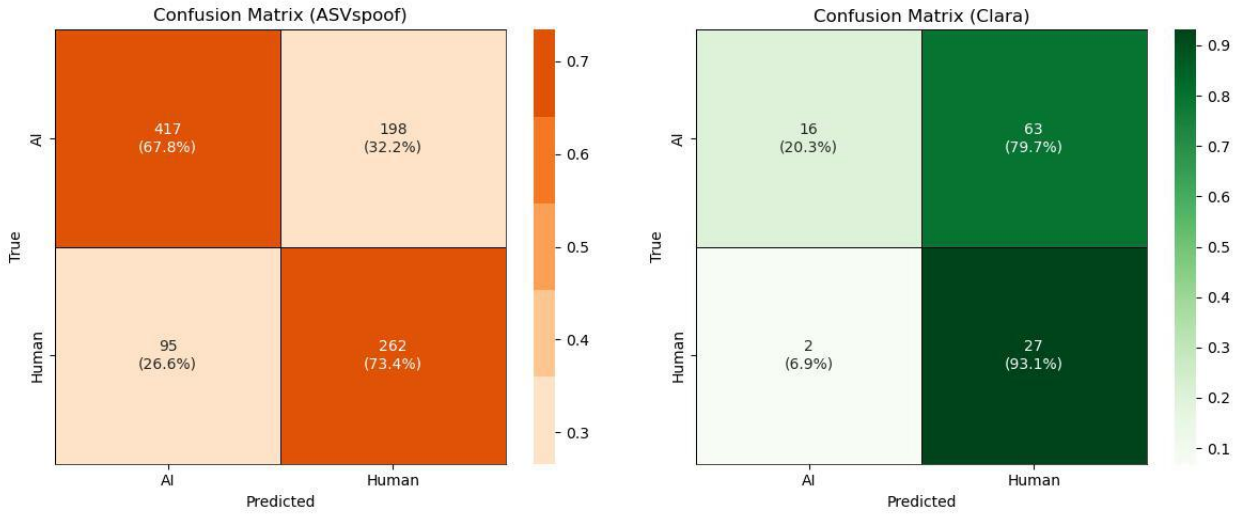


Figure 17: Human performance confusion matrices by dataset.

The difference between ASVspoof and *Clara* datasets becomes even more evident when observing the error rates by class in each subset (normalized by the number of responses per class). For ASVspoof (Figure 18, left), 32.2% of AI responses were mislabeled as human, while the reverse (human responses misclassified as AI) was 26.6%. However, in the *Clara* dataset (Figure 18, right), the proportion of spoofed audios misclassified as human skyrocketed to 79.7%, while the proportion of human audios misclassified as spoofed was only 6.9%. This startling assymetry suggests that, even though human listeners are generally cautious when identifying bona fide samples, they are significantly more prone to accepting highly realistic spoofed audios as genuine.


Figure 18: Normalized error rates by class for each dataset.

Lastly, a method-specific analysis was conducted to identify which spoofing techniques were most challenging for participants to detect. As illustrated in Figure 19, modern text-to-speech (TTS) approaches from PlayHT, ResembleAI, and LOVO ranked among the most successful at deceiveing listeners, with accuracies of 6.7%, 27.3%, and 29.6% respectively. Notably, the ASVspoof technique A10 breaks into this top-three list in second place, with human accuracy of only 26.6%.

The A10 method consists of an end-to-end neural TTS system that builds upon Tacotron 2 [39] by incorporating speaker-adaptive transfer learning. First, a sequence-to-sequence model generates Mel-spectrograms from input text or phonemes. Then, a separately trained speaker encoder extracts a speaker embedding from a short audio sample, which is used to condition the Tacotron 2 model to reproduce that speaker's vocal characteristics. Finally, a WaveRNN neural vocoder converts these spectrograms into waveforms [17].

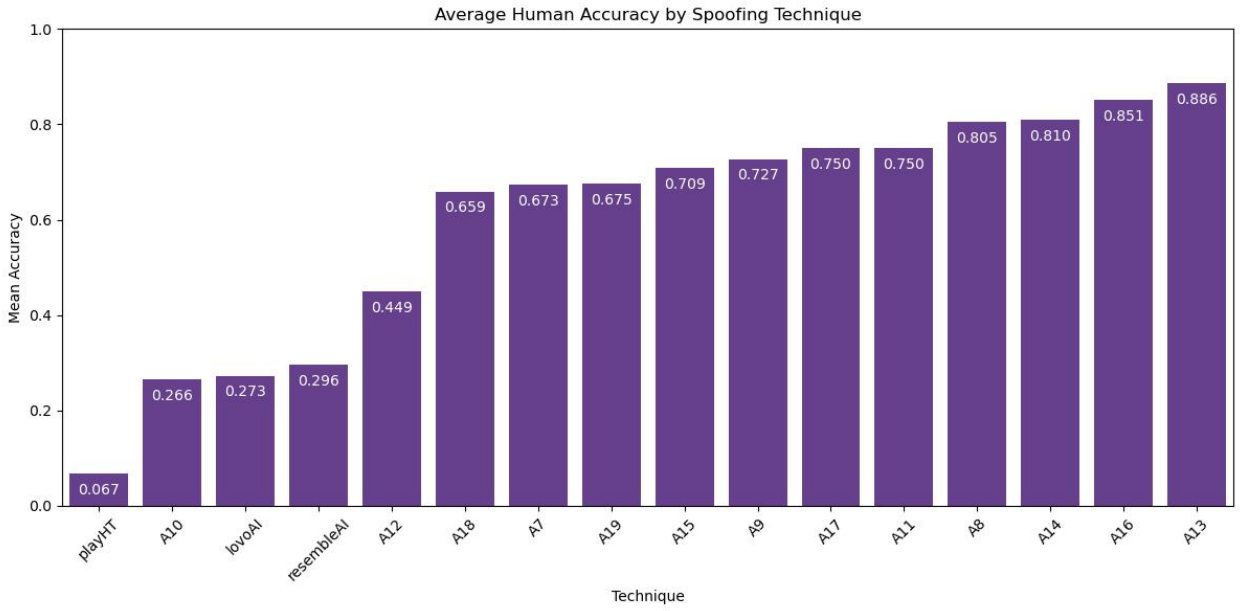


Figure 19: Mean accuracy per spoofing method by humans.

In summary, the results reveal that humans and automated detectors are each vulnerable to different aspects of synthetic speech: humans struggle most with high-quality, modern AI systems, whereas models tend to perform worse on unseen or highly adaptive spoofing techniques. Overall, the selected detection models consistently outperformed human listeners, underscoring the imminent need for robust algorithmic countermeasures against the evolving threat of synthetic speech. As voice synthesis continues to advance and new spoofing methods emerge, effective automated detection systems will remain essential to safeguard against the risks posed by deepfake audio technologies.

# 6 Conclusion and Future Work

Throughout this project, various spoofing detection models retrained and analyzed using a subset of the ASVspoof 2019 dataset, constrained to short audio clips (between two and four seconds long). The models were evaluated accross multiple datasets, including training, development, and evaluation partitions of ASVspoof, as well as an external set of audio samples generated by third-party voice cloning technologies (PlayHT, ResembleAI, and LOVO). Additionally, human performance was assessed through an interactive platform, where participants were asked to classify audio samples as either human or AI-generated. This exhaustive analysis provided valuable insights into the strengths and weaknesses of current spoofing detection models, as well as the challenges faced by human listeners in distinguishing between genuine and synthetic speech.

The LFCC-based model, trained using a one-class learning approach, notably demonstrated superior performance accross different testing scenarios, particularly when faced with unseen spoofing techniques. Conversely, models based on spectrograms, MFCCs, and CQCCs—despite achieving near-perfect accuracy during training and development— displayed significant performance drops on the evaluation set and especially on the external dataset. These results emphasize the importance of generalization in spoofing detection systems, thus highlighting the one-class approach as remarkably effective for this task.

Interestingly, human listeners showed significant difficulty in accurately distinguishing between genuine and AI-generated audio, particularly with newer synthesis methods. They exhibited strong biases towards classifying realistic synthetic speech as human, as evidenced by the markedly higher error rates when evaluating modern TTS systems. Moreover, the spoofing techniques that deluded automated models were generally distinct from those most deceptive to humans, indicating differences in how each group evaluates audio quality and authenticity.

As future work, several promising directions can be explored to enhance spoofing detection systems. Firstly, given their proven robustness and potential for generalization, the one-class learning approach should be further investigated along with other methods that leverage embeddings and latent space representations, such as those derived from autoencoders or variational autoencoders. Secondly, developing hybrid models that combine the strengths of different feature extraction techniques (for instance, both handcrafted and advanced neural features) could lead to improved performance across diverse spoofing methods. Additionally, a broader range of input features should be considered, potentially incorporating raw audio waveforms, learned embeddings, or phase-based features. Finally, integrating data augmentation and adversarial training strategies could help models become more resilient to emerging spoofing techniques, thus enhancing their ability to generalize to new, unseen methods.

# References

[1] H. Barakat, O. Turk, and C. Demiroglu, "Deep Learning-based Expressive Speech Synthesis: A Systematic Review of Approaches, Challenges, and Resources," Dec. 2024.

[2] Y. Gao and B. S. Bioengineering, "Audio Deepfake Detection Based on Differences in Human and Machine Generated Speech," Tech. Rep., 2022.

[3] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge," Tech. Rep. [Online]. Available: http://www.festvox.org/

[4] A. Nautsch, X. Wang, N. Evans, T. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech," Feb. 2021. [Online]. Available: http://arxiv.org/abs/2102.05889

[5] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep Residual Neural Networks for Audio Spoofing Detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-September. International Speech Communication Association, 2019, pp. 1078–1082.

[6] Y. Zhang, F. Jiang, and Z. Duan, "One-class Learning Towards Synthetic Voice Spoofing Detection," Oct. 2020. [Online]. Available: http://arxiv.org/abs/2010.13995

[7] Dimension Market Research, "Voice Recognition Security Market Size, CAGR, Trends and Forecast 2034," 2025, accessed: 2025-06-02. [Online]. Available: https://dimensionmarketresearch.com/report/voice-recognition-security-market/

[8] S. E. Griffin and C. C. Rackley, "Vishing," in *Proceedings of the 5th Annual Conference on Information Security Curriculum Development*, ser. InfoSecCD '08. New York, NY, USA: Association for Computing Machinery, 2008, pp. 33–35. [Online]. Available: https://doi.org/10.1145/1456625.1456635

[9] D. Forrest, "Challenges in Voice Biometrics: Vulnerabilities in the Age of Deepfakes," Feb. 2024, accessed: 2025-04-21. [Online]. Available: https://bankingjournal.aba.com/2024/02/challenges-in-voice-biometrics-vulnerabilities-in-the-age-of-deepfakes/

[10] A. Alali and G. Theodorakopoulos, "Partial Fake Speech Attacks in the Real World Using Deepfake Audio," *Journal of Cybersecurity and Privacy*, vol. 5, Mar. 2025.

[11] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio Deepfake Detection: A Survey," Aug. 2023. [Online]. Available: http://arxiv.org/abs/2308.14970

[12] PlayHT, "PlayHT: AI Voice Generator and Text-to-Speech Platform," accessed: 2025-05-06. [Online]. Available: https://play.ht/

[13] ResembleAI, "Resemble AI: AI Voice Generator and Deepfake Detection for Enterprise," accessed: 2025-05-06. [Online]. Available: https://www.resemble.ai/

[14] LovoAI, "LOVO AI: AI Voice Generator and Text-to-Speech Platform," accessed: 2025-05-06. [Online]. Available: https://lovo.ai/

[15] X. Li, P.-Y. Chen, and W. Wei, "Where Are We in Audio Deepfake Detection? A Systematic Analysis Over Generative and Detection Models," Oct. 2024. [Online]. Available: http://arxiv.org/abs/2410.04324

[16] S. Borzì, O. Giudice, F. Stanco, and D. Allegra, "Is Synthetic Voice Detection Research Going into the Right Direction?" Tech. Rep., 2022.

[17] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, "ASVspoof 2019: A Large-scale Public Database of Synthesized, Converted and Replayed Speech," Nov. 2019. [Online]. Available: http://arxiv.org/abs/1911.01601

[18] M. Neelima and I. S. Prabha, "Hybrid Feature Optimization for Voice Spoof Detection Using CNN-LSTM," *Traitement du Signal*, vol. 41, pp. 717–727, Apr. 2024.

[19] C. Zhang, Y. Chengzhu, and J. H. L. Hansen, "An Investigation of Deep Learning Frameworks for Speaker Verification Anti-spoofing," Tech. Rep., 2017.

[20] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-September. International Speech Communication Association, 2019, pp. 1068–1072.

[21] Q. Shen, M. Guo, Y. Huang, and J. Ma, "Attentional Multi-Feature Fusion for Spoofing-Aware Speaker Verification," *International Journal of Speech Technology*, vol. 27, 06 2024.

[22] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, "Channel-wise Gated Res2Net: Towards Robust Detection of Synthetic Speech Attacks," Jul. 2021. [Online]. Available: http://arxiv.org/abs/2107.08803

[23] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A Comparison of Features for Synthetic Speech Detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2015-January. International Speech and Communication Association, 2015, pp. 2087–2091.

[24] P. Abdzadeh Ziabari and H. Veisi, "A Comparison of CQT Spectrogram with STFT-based Acoustic Features in Deep Learning-based Synthetic Speech Detection," vol. 11, pp. 119–129, Jan. 2023.

[25] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)," 2019, accessed: 2025-03-11. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/3443

[26] R. Reimao and V. Tzerpos, "FoR: A Dataset for Synthetic Speech Detection," Tech. Rep., 2019. [Online]. Available: https://www.kaggle.com/percevalw/englishfrench-translations/kernels

[27] H. Ma, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, and R. Fu, "CFAD: A Chinese Dataset for Fake Audio Detection," Jul. 2022. [Online]. Available: http://arxiv.org/abs/2207.12308

[28] K. T. Mai, S. Bray, T. Davies, and L. D. Griffin, "Warning: Humans Cannot Reliably Detect Speech Deepfakes," *PLoS ONE*, vol. 18, Aug. 2023.

[29] C. Palacios-Castrillo, R. Palacios, R. Gesteira-Miñarro, A. Chávez-Macías, and G. López, "Analysis of the Security and Privacy of Smart Personal Assistants with Real and Synthetic Voices," *Journal of Information Security and Applications ()*, vol. Under Review, 2025.

[30] Z. S. Bojkovic, B. M. Bakmaz, and M. R. Bakmaz, "Hamming Window to the Digital World," *Proceedings of the IEEE*, vol. 105, no. 6, pp. 1185–1190, 2017.

[31] H. M. Fayek, "Speech Processing for Machine Learning: Filter Banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between," 2016. [Online]. Available: https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

[32] M. Todisco, H. Delgado, and N. Evans, "A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients," Tech. Rep., 2016.

[33] ——, "Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification," Tech. Rep., 2017.

[34] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification," Apr. 2018. [Online]. Available: http://arxiv.org/abs/1804.09618

[35] W.-C. Huang, Y.-C. Wu, K. Kobayashi, Y.-H. Peng, H.-T. Hwang, P. L. Tobing, Y. Tsao, H.-M. Wang, and T. Toda, "Generalization of Spectrum Differential based Direct Waveform Modification for Voice Conversion," Jul. 2019. [Online]. Available: http://arxiv.org/abs/1907.11898

[36] T. Kinnunen, J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, and Z. Ling, "A Spoofing Benchmark for the 2018 Voice Conversion Challenge: Leveraging from Spoofing Countermeasures for Speech Artifact Assessment," Apr. 2018. [Online]. Available: http://arxiv.org/abs/1804.08438

[37] T. Kinnunen, L. Juvela, P. Alku, and J. Yamagishi, "Non-parallel voice conversion using i-vector plda: Towards unifying speaker verification and transformation," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Institute of Electrical and Electronics Engineers Inc., Jun. 2017, pp. 5535–5539.

[38] Y. Wu, K. Chen, T. Zhang, Y. Hui, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," Nov. 2022. [Online]. Available: http://arxiv.org/abs/2211.06687

[39] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.