Detección automática de audio generado por Inteligencia Artificial

Victoria García Martínez-Echevarría
Universidad Pontificia Comillas
Escuela de Ingeniería ICAI
victoria.garcia@alu.comillas.edu

Rafael Palacios Hielscher
Massachusetts Institute of Technology
Cybersecurity at MIT Sloan
rafael.palacios@iit.comillas.edu

Gregorio López López
Universidad Pontificia Comillas
Instituto de Investigación Tecnológica
gllopez@icai.comillas.edu

***Resumen*—La accesibilidad y el creciente realismo de los sistemas modernos de síntesis y conversión de voz han intensificado la necesidad de mecanismos fiables para detectar audios generados por Inteligencia Artificial en contextos con requisitos críticos de seguridad. Este artículo analiza la detección automática de voz sintética en un entorno de duración controlada (2-4 segundos) mediante el reentrenamiento de dos sistemas de referencia del conjunto *Logical Access* del *ASVspoof 2019 Challenge*. El primero es una red convolucional residual entrenada como clasificador binario, mientras que el segundo adopta una estrategia de aprendizaje de una sola clase con arquitectura ResNet-18. Además, se comparan los resultados de un estudio de identificación realizado con usuarios a través de una plataforma web. Los resultados experimentales muestran que el enfoque de una sola clase generaliza mejor frente a ataques no vistos; en cambio, los humanos cometen más errores, especialmente ante voces sintéticas modernas de alta calidad.**

***Index Terms*—*One-class learning*, detección de *spoofing*, *deepfakes*, redes neuronales convolucionales (CNNs), síntesis de audio**

I. Introducción

Durante la última década, los avances en Inteligencia Artificial (IA) han dado lugar a sistemas que permiten crear contenido sintético cada vez más realista de múltiples tipos, (texto, imágenes, vídeo y audio). En particular, los sistemas actuales de síntesis de habla y conversión de voz pueden generar voces que reproducen con gran fidelidad las características y los matices de la voz humana [1]. Sin embargo, esto ha provocado que aumente la preocupación sobre la autenticidad y la confianza en las interacciones mediadas por la voz. Desde el punto de vista de la seguridad, las voces sintéticas pueden usarse para suplantar la identidad y atacar servicios basados en la identificación por voz, como la autenticación biométrica y los asistentes personales inteligentes.

Ya se han reportado casos reales que evidencian el potencial de los *deepfakes* de voz para eludir medidas de seguridad y comprometer la confianza en las transacciones digitales, e incluso los delicuentes pueden usar *deepfakes* parciales para estafar tanto a personas como a sistemas automatizados. Estas amenazas han motivado líneas de investigación específicas sobre audio sintético y medidas de detección, así como estándares de evaluación (como los retos ASVspoof) [2], [3] y datasets públicos.

En la literatura, los primeros sistemas de detección se basaban en modelos estadísticos como GMM (*Gaussian Mixture Models*) con características acústicas clásicas, incluyendo MFCC (Mel-Frequency Cepstral Coefficients), LFCC (Log-Frequency Cepstral Coefficients), CQCC (Constant Q Cepstral Coefficients), etc. [4].

Con la adopción del aprendizaje profundo, arquitecturas como CNN-RNN (*Convolutional Recurrent Neural Networks*) o ResNet (*Residual Networks*) con función de pérdida *One-Class Softmax* [5] mejoraron notablemente los resultados, siendo esta última capaz de detectar ataques no vistos durante el entrenamiento con un EER (Equal Error Rate) del 2,19 % en el conjunto LA del *ASVspoof 2019 Challenge*. Trabajos posteriores incorporaron mecanismos de atención y variantes avanzadas como Res2Net [6], consolidando la tendencia hacia arquitecturas profundas especializadas. A pesar de los avances, la detección robusta de voces sintéticas sigue siendo un desafío, especialmente cuando los modelos se enfrentan a métodos de síntesis no conocidos durante el entrenamiento.

II. Metodología

Para estudiar la detección automática de audio generado por IA, adaptamos y reentrenamos modelos de referencia del ASVspoof 2019 Challenge (partición *Logical Access*, LA), filtrando el conjunto de datos para incluir únicamente muestras de entre dos y cuatro segundos, con el fin de evitar sesgos asociados a la duración. Además, diseñamos un experimento complementario para evaluar la percepción humana en la misma tarea [7].

II-A. Características del Audio

Se consideran cuatro representaciones acústicas como entrada a los modelos. Los espectrogramas corresponden a la magnitud logarítmica de la STFT (*Short-Time Fourier Transform*), calculada con ventanas de Hamming de tamaño 2048 y solapamiento del 25 %; conservan información espectral detallada y permiten que las redes aprendan patrones discriminativos a partir de una entrada poco procesada [8]. Los MFCC se obtienen proyectando el espectro sobre un banco de filtros en escala Mel y aplicando una DCT, utilizándose los primeros 24 coeficientes por cuadro; ofrecen una representación compacta aunque pueden perder sensibilidad ante distorsiones sutiles [8]. Los CQCC se basan en la CQT, que emplea intervalos de frecuencia espaciados geométricamente, ofreciendo mayor resolución en bajas frecuencias; han demostrado superar a los coeficientes cepstrales tradicionales al resaltar artefactos en todo el espectro. Los LFCC siguen un procedimiento similar al de los MFCC pero con un banco de filtros en escala lineal, preservando mejor los detalles de alta frecuencia donde suelen aparecer anomalías del audio sintético [9].

II-B. Diseño de los Modelos

El primer modelo, basado en la propuesta de Alzantot et al. (2019) [8], emplea una arquitectura ResNet profunda con bloques residuales que incluyen capas convolucionales, normalizaciones, activaciones *leaky-ReLU* y *dropout*, seguidos de capas totalmente conectadas que producen una predicción binaria. Se entrenan tres variantes independientes con espectrogramas, MFCC y CQCC como entrada, y el entrenamiento minimiza una pérdida de entropía cruzada estándar.

El segundo modelo, propuesto por Zhang et al. (2020) [5], adopta un enfoque de aprendizaje de una sola clase (*one-class Learning*) mediante una arquitectura ResNet-18 sobre características LFCC. En lugar de una clasificación binaria, modela únicamente la distribución del habla real mediante la función de pérdida OC-Softmax, que proyecta los embeddings sobre una hiperesfera unitaria y penaliza desviaciones respecto al centro compacto de la clase auténtica; durante la inferencia, las muestras alejadas de dicha región se consideran potencialmente falsificadas.

II-C. Recopilación de Datos de Usuarios

Para complementar la evaluación automática, se utilizó un servidor interactivo [7] en el que los participantes clasificaban clips de voz como *Humano*, *Suena humano*, *Suena a IA* o *IA*. El servidor incluía 701 muestras del conjunto de evaluación filtrado de ASVspoof 2019 (LA) y ocho muestras externas generadas con herramientas comerciales (PlayHT [10], Resemble AI [11] y LOVO [12]), todas restringidas a audios de entre dos y cuatro segundos.

El experimento se estructuró en sesiones de 10 clips, completándose 108 sesiones y obteniéndose 1.080 decisiones en total. Para la comparación con los modelos, las cuatro opciones se fusionaron en dos categorías binarias: Humano (Humano + Suena humano) e IA (Suena a IA + IA).

III. Implementación y Experimentos

Los experimentos consisten en reentrenar modelos de detección de *spoofing* en el subconjunto filtrado de ASVspoof 2019 y evaluarlos tanto en las particiones de desarrollo y evaluación como en un conjunto externo de muestras generadas por herramientas de terceros, comparando su rendimiento con el de oyentes humanos.

III-A. Conjunto de Datos

Se utilizó la partición LA (*Logical Access*) de ASVspoof 2019, organizada en tres particiones (entrenamiento, desarrollo y evaluación) con distribución equilibrada de hablantes [13]. Como se menciona en la sección anterior, se filtraron las muestras para conservar únicamente audios de entre dos y cuatro segundos; y, aunque este filtrado reduce el número total de muestras, el equilibrio entre clases se mantiene similar al del conjunto original (véase Tabla I).

Adicionalmente, se incorporó un pequeño conjunto externo de ocho muestras (seis generadas con PlayHT [10], ResembleAI [11] y LOVO [12], y dos reales) para evaluar la generalización frente a técnicas modernas no vistas [7].

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 %

Tabla I: Conteos y proporciones (SP-BF) antes y después de filtrar los datos por duración.

III-B. Entrenamiento y Validación

El proceso de entrenamiento siguió estrictamente las configuraciones recomendadas en los trabajos originales [5], [8], preservando arquitecturas, optimizadores y demás hiperparámetros para garantizar que las diferencias de rendimiento se debieran al conjunto de datos filtrado. La principal variable experimental fue el número de épocas: para cada tipo de representación de entrada se entrenaron múltiples versiones del modelo variando la duración del entrenamiento, aplicando en algunos casos detención temprana cuando no se observaban mejoras adicionales en la función de pérdida.

El rendimiento se evaluó mediante la precisión (accuracy) y el Equal Error Rate (EER), métrica estándar en detección de spoofing que identifica el punto en que la tasa de aceptación falsa (FAR) y la tasa de rechazo falso (FRR) coinciden.

III-C. Análisis del Rendimiento

El rendimiento de los modelos se analiza desde tres perspectivas: (i) entrenamiento y desarrollo, (ii) conjunto de evaluación y (iii) conjunto externo. En los dos últimos casos, los modelos se enfrentan a datos no vistos y generados con técnicas distintas a las del entrenamiento, lo que permite evaluar su capacidad de generalización.

Para realizar una comparación clara entre representaciones de audio, se seleccionó una única versión representativa por tipo de característica, priorizando buen rendimiento con menor número de épocas y tomando como criterio la precisión en evaluación. Las configuraciones elegidas fueron: 100 épocas para el modelo basado en espectrogramas, 75 épocas para MFCC, 100 épocas para CQCC y 75 épocas para LFCC.

En los conjuntos de entrenamiento y desarrollo, todos los modelos alcanzan precisiones muy altas (Tabla II). La precisión en entrenamiento es prácticamente perfecta en todos los casos, mientras que en desarrollo apenas disminuye, salvo en los modelos basados en MFCC, donde se observa mayor sensibilidad al sobreajuste. Los valores de EER en desarrollo son superiores a los reportados en los trabajos originales, lo que no resulta sorprendente dado que el filtrado reduce el tamaño del dataset y restringe la variabilidad temporal.

Model	Train accuracy	Dev accuracy	Dev EER	Original Dev EER
spect_100	1.000000	0.995358	0.004263	0.0011
mfcc_75	0.999668	0.972577	0.043822	0.0334
cqcc_100	0.999934	0.995001	0.006728	0.0001
lfcc_75	1.000000	0.997143	0.007395	0.0020

Tabla II: Precisión y valores EER de cada modelo en los conjuntos de entrenamiento y desarrollo.

El comportamiento cambia notablemente en el conjunto de evaluación (Tabla III), donde todas las arquitecturas experimentan una caída de precisión, especialmente el modelo basado en espectrogramas. En contraste, el modelo basado en LFCC obtiene los mejores resultados, con una precisión del 93,46 % y los valores de EER más bajos (≈0,035), lo que confirma la mayor capacidad de generalización del enfoque de aprendizaje de una sola clase frente a ataques no vistos.

Model	Eval Accuracy	EER	Original Eval EER
spect_100	0.797348	0.125166	0.0968
mfcc_75	0.842393	0.147221	0.0933
cqcc_100	0.833338	0.104786	0.0769
lfcc_75	**0.934573**	0.036393	0.0219

Tabla III: Precisión y valores EER de cada modelo en el conjunto de evaluación.

En el conjunto externo, el rendimiento global disminuye respecto al conjunto de evaluación (Figura 1), especialmente en el modelo CQCC. El modelo LFCC vuelven a obtener los mejores resultados, con un 87,5 % de precisión, aunque el tamaño reducido del conjunto limita la robustez estadística de estas conclusiones.

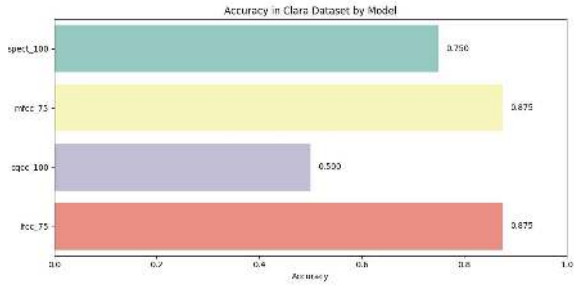

Figura 1: Precisión de cada modelo en el conjunto externo.

IV. Resultados

Dado que este trabajo analiza dos fuentes de rendimiento distintas, los resultados de modelos automáticos y usuarios se presentan por separado.

IV-A. Modelos Automáticos

Como se describió en la sección anterior, se seleccionó una configuración representativa por tipo de característica (espectrogramas a 100 épocas, MFCC a 75, CQCC a 100 y LFCC a 75). Los resultados detallados por partición se recogen en las tablas II y III: todos los modelos alcanzaron precisión casi perfecta en entrenamiento y desarrollo, pero el rendimiento cae ante técnicas no vistas, siendo LFCC el más robusto (93,46 % en evaluación y 87,5 % en el conjunto externo) y CQCC el que peor generaliza.

Para analizar qué técnicas resultaron más difíciles, se calculó la precisión media por método (Figura 2). Las técnicas A17 y A18 destacan como las más problemáticas, con precisiones medias del 19,7 % y 51,0 %, respectivamente. Ambas se basan en sistemas avanzados de conversión de voz no paralelos [13]: A17 emplea una arquitectura VAE con modificación directa de forma de onda [14], y A18 un marco de *i-vectors* y PLDA con aprendizaje por transferencia [15].

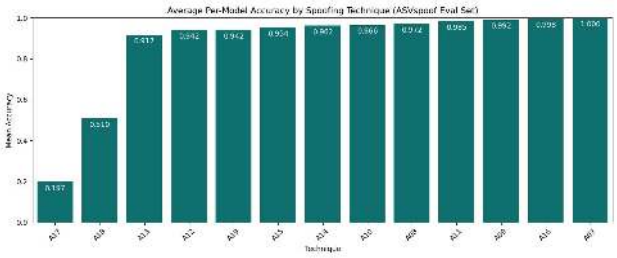

Figura 2: Precisión media por método de *spoofing* en el conjunto de evaluación.

En el conjunto externo *Clara* (Tabla IV), todos los modelos clasificaron correctamente las muestras de PlayHT y LOVO, mientras que ResembleAI resultó la técnica más engañosa: solo los modelos MFCC y LFCC acertaron la mitad de sus muestras, y CQCC no clasificó correctamente ninguna. Estas diferencias pueden explicarse por las distintas arquitecturas internas de cada herramienta, que generan artefactos acústicos diferentes y afectan de manera desigual a cada modelo.

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

Tabla IV: Precisión de cada modelo en el conjunto de datos *Clara* por técnica de *spoofing*.

En definitiva, los resultados automáticos muestran que el enfoque de una sola clase basado en LFCC ofrece la mejor generalización, aunque ningún modelo es completamente robusto frente a todos los métodos no vistos.

IV-B. Usuarios

En el experimento con el servidor web [7] participaron 80 personas en 108 sesiones, generando 1.080 respuestas: 972 sobre ASVspoof y 108 sobre el conjunto externo. La Tabla V resume la precisión global y por subconjunto.

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 %

Tabla V: Precisión global y por clase, junto con la proporción de muestras de IA y humanas en cada conjunto.

IV-B1. Rendimiento global

La precisión global fue del 66,85 %, con una tasa de error del 33,15 %. Los participantes identificaron mejor el audio humano (74,9 %) que el sintético (62,4 %), y mostraron una asimetría clara: tendieron más a aceptar como humano un audio falso (37,6 %) que a rechazar erróneamente uno auténtico (25,1 %).

IV-B2. Análisis por subconjunto

Al separar los datos por conjunto, se observa que el rendimiento humano varía significativamente. En ASVspoof,

la precisión global fue del 69,86 %, con tasas relativamente equilibradas (67,8 % en sintéticos y 73,4 % en auténticos). En el conjunto externo, el rendimiento cayó drásticamente hasta el 39,81 %: aunque la identificación de muestras humanas fue alta (93,1 %), el 79,7 % de las muestras falsificadas se clasificaron erróneamente como humanas, lo que indica que los oyentes aceptan con facilidad voces sintéticas modernas de alta calidad. Cabe resaltar que, dado el tamaño reducido de este conjunto, las conclusiones deben interpretarse con cautela, aunque la tendencia es consistente y significativa.

IV-B3. Dificultad por método

Las técnicas más engañosas para los participantes fueron las tres herramientas del conjunto externo (PlayHT, ResembleAI y LOVO, con precisiones del 6,7 %, 27,3 % y 29,3 % respectivamente) y la técnica A10 de ASVspoof (26,6 %), basada en Tacotron 2 con vocoder WaveRNN [16], lo que confirma que los sistemas TTS neuronales modernos superan fácilmente la percepción humana (Figura 3).

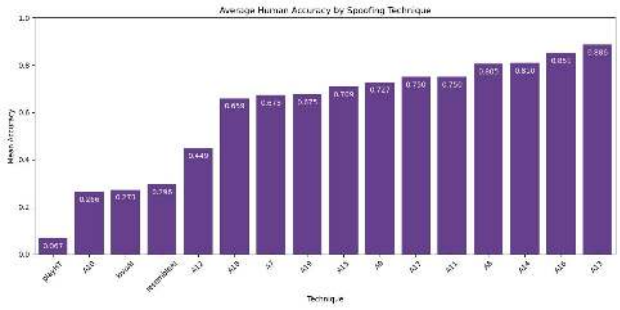

Figura 3: Precisión media por técnica de *spoofing* para los humanos.

IV-C. Comparación entre modelos y humanos

Comparando ambas fuentes de resultados, los modelos automáticos superan sistemáticamente a los humanos: el mejor modelo (LFCC) alcanza más del 93 % en evaluación, frente al 70 % humano en ese subconjunto, y ambos caen en el conjunto externo, aunque los humanos lo hacen de forma más pronunciada (por debajo del 40 %).

Modelos y humanos fallan por razones distintas: los humanos son especialmente susceptibles al realismo perceptual de voces modernas, mientras que los modelos sufren ante técnicas no vistas o altamente adaptativas. A pesar de estas limitaciones, los detectores automáticos ofrecen una ventaja clara y consistente, lo que refuerza la necesidad de desarrollar sistemas robustos frente a la evolución continua de la síntesis de voz.

V. Conclusión

En este trabajo se reentrenaron y analizaron distintos modelos de detección de voz sintética sobre un subconjunto filtrado del dataset ASVspoof 2019 (restringido a audios de entre dos y cuatro segundos), evaluándolos tanto en sus particiones estándar como en un conjunto externo, y comparando su rendimiento con el de usuarios humanos en la misma tarea.

El modelo basado en LFCC con *one-class learning* ofrece la mejor capacidad de generalización frente a técnicas no vistas, mientras que los modelos basados en espectrogramas, MFCC y CQCC, pese a alcanzar rendimientos casi perfectos en entrenamiento y desarrollo, presentan caídas significativas en escenarios más exigentes. Por su parte, los usuarios mostraron una dificultad considerable para identificar voces sintéticas modernas, con una tendencia marcada a clasificarlas como humanas. Cabe destacar que las técnicas más problemáticas para los modelos no siempre coincidieron con las que engañaron a los humanos, evidenciando diferencias en los criterios de detección. En conjunto, los modelos superaron consistentemente a los humanos, lo cual subraya la importancia de desarrollar sistemas automáticos robustos para contrarrestar las amenazas emergentes de los *deepfakes* de voz.

Como líneas de trabajo futuro, se plantea ampliar el rango de duraciones, incorporar grabaciones en distintos entornos y ampliar el conjunto externo con más herramientas comerciales de síntesis.

Referencias

[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] 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/

[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. 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.

[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] 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.

[7] 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.

[8] 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.

[9] M. Todisco, H. Delgado, and N. Evans, "Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification," Tech. Rep., 2017.

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

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

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

[13] X. W. et al., "ASVspoof 2019: A Large-scale Public Database of Synthesized, Converted and Replayed Speech," Nov. 2019. [Online]. Available: http://arxiv.org/abs/1911.01601

[14] W.-C. H. et al., "Generalization of Spectrum Differential based Direct Waveform Modification for Voice Conversion," Jul. 2019. [Online]. Available: http://arxiv.org/abs/1907.11898

[15] 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.

[16] J. S. et al., "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.