



PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE CIENCIAS EXACTAS NATURALES Y AMBIENTALES

CARRERA DE CIENCIA DE DATOS

Detección de Deepfakes de Audio mediante técnicas deep learning: revisión
sistemática de literatura (2020-2025)

Monografía previa a la obtención del título de Ingeniero en Ciencia de Datos

Sebastián David Reyes Santamaría

Quito, 2026

Certifico que la Monografía de Ciencia de Datos de **Sebastián David Reyes Santamaría** ha sido concluida de conformidad con las normas establecidas; por lo tanto, puede ser presentada para la calificación correspondiente.

Juan Camilo Anzieta Reyes

Tutor de la monografía

Quito, 25 de junio de 2026

DEDICATORIA

A mi padre, a mi madre y a mi hermano, por ser mi sostén y mi mayor motivación.

AGRADECIMIENTOS

A mi tutor, Juan Camilo Anzieta Reyes, por la guía brindada durante todo el desarrollo de este trabajo y por su revisión y sus aportes en la temática, que dieron rigor y dirección a esta monografía.

TABLA DE CONTENIDOS

1. Resumen	1
2. Abstract	2
3. Introducción	3
4. Objetivos	5
4.1. Objetivo general	5
4.2. Objetivos específicos	5
5. Marco Teórico y Antecedentes Históricos	6
5.1. Marco conceptual	6
5.2. Evolución de la generación de deepfakes	7
5.3. Evolución de la detección de deepfakes (enfoque deep learning)	18
5.4. Evolucion de los conjuntos de datos y desafíos	24
6. Metodología	27
6.1. Diseño del estudio	27
6.2. Estrategia de búsqueda	27
6.3. Proceso de selección	28
6.3.1. Primer screening	29
6.3.2. Segundo screening	29
6.3.3. Tercer screening	30
7. Estado del arte	32
7.1. Datasets y estrategias de entrenamiento frente al sesgo de conjunto cerrado	32
7.2. Explicabilidad en la detección de deepfakes de audio	46
7.3. Detección de deepfakes de audio en lenguas menos representadas	52
8. Conclusiones	59
9. Referencias	61

1. RESUMEN

La rápida evolución de la síntesis de voz permite generar audio prácticamente indistinguible del real, lo que habilita usos maliciosos como la suplantación de identidad, el fraude por voz y la desinformación. La detección basada en aprendizaje profundo es la principal línea de defensa, pero su acelerado avance dificulta mantener una visión clara del estado del arte. Esta monografía presenta una revisión sistemática de la literatura, conducida según la metodología PRISMA 2020, sobre la detección de deepfakes de audio mediante aprendizaje profundo entre 2020 y 2025. A partir de las bases IEEE Xplore y Scopus, un cribado en tres etapas redujo 1544 registros a 39 estudios, organizados en torno a tres ejes: los conjuntos de datos y las estrategias de entrenamiento, la explicabilidad de los modelos (XAI) y la detección en lenguas distintas del inglés. El análisis revela que el sesgo de conjunto cerrado es el problema estructural del campo: los detectores fallan ante sistemas de síntesis, corpus o lenguas no vistos, y los modelos de fundación autosupervisados ofrecen la mayor robustez. Se concluye que la generalización es el reto central que el área tiene todavía por delante.

Palabras clave: Detección de deepfakes de audio, Aprendizaje profundo, Revisión sistemática de literatura (PRISMA), Explicabilidad (XAI), Generalización, Lenguas de bajo recurso.

2. ABSTRACT

The rapid evolution of speech synthesis makes it possible to generate audio that is practically indistinguishable from real speech, enabling malicious uses such as identity impersonation, voice fraud, and disinformation. Deep learning-based detection is the main line of defense, but its fast pace makes it hard to keep a clear view of the state of the art. This work presents a systematic literature review, conducted following the PRISMA 2020 methodology, on deep learning-based audio deepfake detection between 2020 and 2025. Drawing on the IEEE Xplore and Scopus databases, a three-stage screening reduced 1544 records to 39 studies, organized around three axes: datasets and training strategies, model explainability (XAI), and detection in languages other than English. The analysis reveals that the closed-set bias is the structural problem of the field: detectors fail against unseen synthesis methods, corpora, or languages, and self-supervised foundation models offer the greatest robustness. We conclude that generalization is the central open challenge that the field still has ahead.

Keywords: Audio deepfake detection, Deep learning, Systematic literature review (PRISMA), Explainability (XAI), Generalization, Low-resource languages.

3. INTRODUCCIÓN

La rápida evolución tecnológica ha permitido la creación de modelos tanto de aprendizaje automático como de aprendizaje profundo (deep learning), capaces de generar audio sintético indistinguible del real, producido por un humano. En la actualidad diversas arquitecturas y modelos basados en difusión han impulsado aplicaciones beneficiosas en accesibilidad, doblaje automatizado y asistentes virtuales. Sin embargo, estas mismas tecnologías han facilitado la proliferación de audios sintéticos o manipulados para fines maliciosos “deepfake de audio”, lo que genera nuevos riesgos cada vez más presentes, tales como la divulgación de fake news, la suplantación de identidad, el acceso no autorizado a verificaciones de voz o el ciberacoso.

Ante esta problemática, se han desarrollado diversos modelos capaces de distinguir entre audios auténticos y sintéticos donde resaltan dos enfoques principales: Por un lado, los métodos tradicionales (clásicos) que incluyen el diseño manual de características acústicas e implementación de algoritmos; por otro, aquellos basados en aprendizaje profundo (deep learning) que, mediante redes neuronales o sus variantes, aprenden automáticamente los patrones acústicos en los datos, identificando los audios modificados de los reales.

Dentro del enfoque del deep learning, el campo evoluciona a un ritmo acelerado: constantemente surgen nuevos métodos y arquitecturas, o se modifican las existentes, lo que dificulta mantener una visión clara del panorama global y actual de la problemática. Debido a esta rápida evolución, se vuelve necesaria una revisión sistemática de la literatura sobre el estado del arte. Por ello, el primer objetivo de este trabajo es establecer una línea base para entender dicho estado del arte actual de la investigación en los últimos 5 años.

A su vez, la literatura actual muestra un desbalance importante, dado que la mayoría de los estudios, arquitecturas, conjuntos de datos y recursos están pensados para el idioma inglés, lo cual limita la comprensión y posible generalización del comportamiento de los modelos en contextos multilingües o en lenguas menos representadas como el español, lo que genera un sesgo lingüístico. Debido a esto resulta fundamental cuestionar si los modelos entrenados predominantemente en inglés pueden generalizarse de manera efectiva a otras lenguas. Esta

limitación no solo afecta la aplicabilidad práctica de los modelos de detección, sino que también plantea interrogantes sobre la validez de las conclusiones obtenidas cuando se extrapolan a contextos lingüísticos diferentes.

Por otro lado, a pesar de su efectividad, el enfoque basado en aprendizaje profundo presenta limitaciones importantes, especialmente en lo referente a la interpretabilidad de los modelos. La naturaleza de “caja negra” de estos sistemas dificulta comprender qué patrones acústicos son utilizados específicamente para distinguir entre audio sintético y real. Esta falta de transparencia ha impulsado nuevas líneas de investigación centradas en la explicabilidad e interpretabilidad de los modelos, conocidas como Explainable Artificial Intelligence (XAI), donde al mejor de nuestro conocimiento no existe ninguna revisión bibliográfica que analice estas técnicas de explicabilidad en la temática de detección de deepfakes de audio.

En consecuencia, se plantean las siguientes interrogantes en torno al problema expuesto: ¿Cuáles son los datasets y las arquitecturas deep learning más efectivos para detectar deepfakes de audio y cómo han evolucionado en los últimos 5 años? ¿Qué características acústicas son determinantes y cómo pueden interpretarse mediante técnicas de explicabilidad? ¿Qué desafíos específicos presenta la detección en idiomas menos explorados?

4. OBJETIVOS

Considerando la problemática expuesta y las limitaciones identificadas en el estado actual de la investigación, este trabajo busca proporcionar una visión integral y actualizada sobre la detección de deepfakes de audio mediante técnicas de deep learning.

4.1 OBJETIVO GENERAL

Realizar una sistematización clara y estructurada de la evolución de los enfoques basados en deep learning para la detección de deepfakes de audio entre 2020 y 2025. En particular, se prevé la identificación y clasificación de los principales conjuntos de datos, estrategias de entrenamiento y métricas de evaluación empleadas en la literatura.

Asimismo, se espera caracterizar las limitaciones y desafíos recurrentes reportados en escenarios multilingües, especialmente en lenguas de bajo recurso como el español, así como las oportunidades de mejora propuestas para fortalecer la generalización de los modelos. Finalmente, se anticipa la recopilación y descripción de las técnicas de explicabilidad (XAI) utilizadas en este dominio, aportando una visión integral sobre cómo se interpreta el comportamiento de los modelos de detección de deepfakes de audio.

4.2 OBJETIVOS ESPECÍFICOS

- Aplicar la metodología PRISMA para realizar una revisión sistemática de la literatura (2020–2025) y sintetizar el estado del arte en detección de deepfakes de audio basada en deep learning.
- Analizar los conjuntos de datos, estrategias de entrenamiento y métricas de evaluación utilizados en estudios de detección de deepfakes de audio.
- Identificar las posibles limitaciones y desafíos en contextos multilingües, con énfasis en lenguas de bajo recurso como el español, así como las oportunidades de mejora reportadas.
- Examinar la aplicación de técnicas de explicabilidad (XAI) en modelos de detección de deepfakes de audio.

5. MARCO TEÓRICO Y ANTECEDENTES HISTÓRICOS

Actualmente, ya existen trabajos de revisión bibliográfica centrados exclusivamente en la detección de deepfakes de audio, por ejemplo, el trabajo de Shaaban et al. (2023), que aborda una comparación en la evolución de arquitecturas de deepfake de audio, comparando desde el enfoque clásico hasta el enfoque deep learning.

Dichos trabajos presentan alcances consolidados y una cobertura exhaustiva de las arquitecturas y técnicas de detección actuales, por lo que abordar exclusivamente el estado del arte de los modelos de detección se limitaría a una repetición de lo ya trabajado por otros autores. Sin embargo, al mejor de nuestro conocimiento, persisten brechas en la literatura que justifican esta revisión, las cuales se harán evidentes a lo largo de los capítulos.

5.1 MARCO CONCEPTUAL

Antes de revisar la evolución histórica tanto de los modelos de generación como los de detección de audio sintético, resulta necesario entender que dentro del problema actual de detección de deepfakes, los atacantes con fines maliciosos “spoofers” cuentan con diversas formas de obtener audio sintético, denominadas “modalidades de síntesis”, además de diversos “escenarios de detección”, por ello es conveniente definirlos antes de continuar.

El “Text-to-Speech” (TTS) consiste en generar voz artificial directamente desde texto. El sistema recibe una secuencia textual y produce una señal de audio que intenta sonar natural o parecerse a una persona específica. Los sistemas modernos pueden clonar voces con pocos segundos de muestra. En el contexto de deepfakes, el TTS es importante porque permite crear frases completamente nuevas que la persona nunca dijo. Es decir, no modifica un audio existente, sino que sintetiza la voz desde cero; además, es el enfoque que mayor interés y evolución ha presentado.

“Voice Conversion” (VC) funciona diferente. Aquí sí existe un audio original pronunciado por un hablante real. El objetivo es transformar ese audio para que conserve el contenido lingüístico (palabras y ritmo) pero cambie la identidad vocal hacia otra persona. Por ejemplo,

el spoofer habla y el sistema convierte su voz para que suene como otra persona. Esta técnica es especialmente peligrosa porque preserva emociones, pausas y prosodia reales del hablante original, haciendo el deepfake más convincente.

Un ejemplo de escenario de detección es “Partial Spoof”, mencionada por autores como L. Zhang et al. (2023). Dentro de este escenario no todo el audio es falso, sino que solo una fracción del segmento ha sido manipulada o reemplazada, mientras que el resto es audio genuino. Esto lo hace mucho más difícil de detectar porque los sistemas de verificación pueden encontrar regiones auténticas y clasificar erróneamente el audio completo como legítimo. Es el equivalente a falsificar solo una firma en un documento real, mientras que TTS o VC serían el método por el que se generó la firma falsificada como tal.

5.2 EVOLUCIÓN DE LA GENERACIÓN DE DEEPPKAKES

A grandes rasgos el esfuerzo por sintetizar el habla comenzó con intentos mecánicos de replicar la anatomía humana para la producción de sonidos. En 1769, Wolfgang von Kempelen presentó su máquina de habla, mediante la cual, un operador podía producir sonidos vocálicos al manipular manualmente una boca de caucho mientras que palancas adicionales controlaban los sonidos nasales y las sibilantes. Aunque monótona, esta máquina demostró que el habla podía descomponerse en componentes físicos controlables.

En 1845, Joseph Faber presentó “Euphonia”, una máquina que utilizaba teclas de piano para pronunciar las letras del abecedario, controlando la articulación mediante réplicas mecánicas de la lengua, labios y mandíbula. Este tipo de dispositivos establecieron la idea de que para crear voz, se debe modelar físicamente la producción del sonido en el tracto vocal, predominante durante la época.

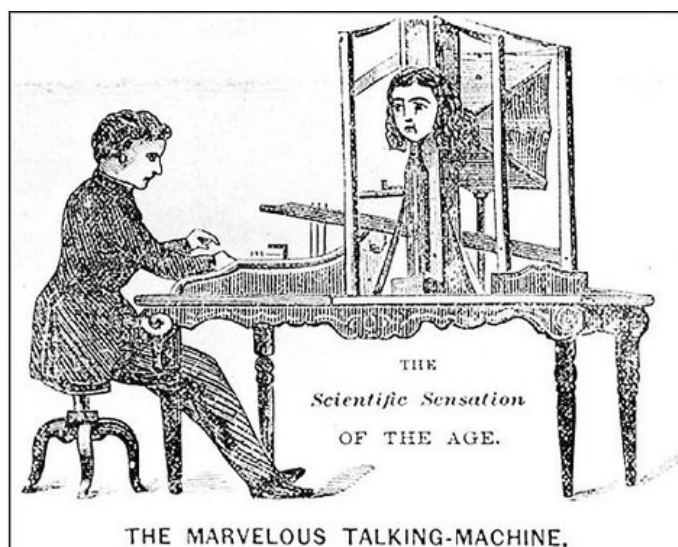


Figura 5.1

Representación de la máquina Euphonia de Joseph Faber (c. 1845), El sistema incluía generalmente un maniquí con apariencia femenina. A pesar de su grado de sofisticación, la invención no tuvo éxito comercial, y su creador terminó suicidándose años después. Imagen extraída de: IDIS (2026).

Posteriormente, la transición a la electrónica fue liderada por Homer Dudley con el VODER en 1939. Este fue el primer sintetizador de voz completamente electrónico. Utilizaba un oscilador para sonidos vocálicos y un generador de ruido para consonantes, pasando estas señales a través de una red de filtros de paso controlados por un operador.

Durante los años 50 y 60, surgieron dos enfoques dominantes en el campo. El primero fue la síntesis por Formantes, basada en el modelado de las resonancias del tracto vocal (formantes). Sistemas como el “Parametric Artificial Talker” (PAT) de Walter Lawrence y el “Orator Verbis Electris” (OVE I) de Gunnar Fant (ambos en 1953) fueron pioneros dentro de este ámbito. El segundo fue el “Linear Predictive Coding” (LPC), una técnica de procesamiento de voz que modela la señal como una combinación lineal de muestras pasadas para representar su envolvente espectral. Lemmetty (1999)

A medida que la capacidad de procesamiento aumentó, la industria se alejó del modelado matemático puro en favor de la reutilización de datos reales. Las primeras técnicas se basaban

en la síntesis concatenativa, cuyo principio consistía en unir fragmentos pregrabados de audio para construir nuevas palabras o frases. Dentro de este paradigma, se distinguen tres enfoques fundamentales.

El primero es la síntesis por dífonos, que utiliza como unidad básica el dífono, es decir, la transición acústica entre dos fonemas consecutivos (Olive, 1977). Su principal limitación radica en que requiere modificaciones de señal como el ajuste de tono lo que da lugar a una calidad percibida como artificial. El segundo enfoque fue, la selección de unidades, el cual recurre a bases de datos de entre 10 y 50 horas de audio, y selecciona el segmento más adecuado basándose en el contexto prosódico (Taylor, 2009). Sin embargo, su flexibilidad es reducida, ya que cambiar de voz o estilo requiere nuevas grabaciones extensas. Finalmente, la síntesis paramétrica estadística (SPSS) marcó una nueva dirección en la síntesis de audio (Yoshimura et al., 1999), al emplear vocoders paramétricos para modelar las características acústicas básicas del habla mediante parámetros estadísticos, basados en Hidden Markov Models (HMM) estos sistemas aún dependían fuertemente de feature engineering manual.

Hasta esta época los resultados mantenían una calidad notoriamente artificial, además de que requerían una especialización técnica para su uso, dos limitantes que no los hacían atractivos para que los spoofers adoptaran su uso para la generación de deepfakes. Sin embargo, esto cambia con la llegada de nuevas metodologías que incorporan redes neuronales dentro de sus arquitecturas.

Durante el periodo de 2016 a 2020 predomina lo que llamaremos el “paradigma clásico”, con lo que nos referiremos al periodo de tiempo donde ganan relevancia y se mejoran los modelos que tienen como objetivo generar espectrogramas o producir audio a partir de dichos espectrogramas. En términos simples, un espectrograma es una representación visual del sonido que muestra cómo cambian las frecuencias a lo largo del tiempo; esta representación se vuelve crítica durante este periodo al actuar como el puente entre el audio crudo y los modelos que necesitan una representación numérica para operar.

La mayoría de los modelos de síntesis de audio utilizan una variante diseñada para aproximar la percepción del sistema auditivo humano, conocida como espectrograma mel. Esta

es una representación bidimensional del audio en la que se cuenta con una representación del tiempo (eje x) una representación de las frecuencias (eje y) y un color o intensidad que muestra la magnitud del sonido en cada punto específico. Lo que cambia en relación a un espectrograma normal es que el eje y de frecuencias, se proyecta en la escala mel (logarítmica) que es la forma de medir frecuencias como las percibe el oído humano.

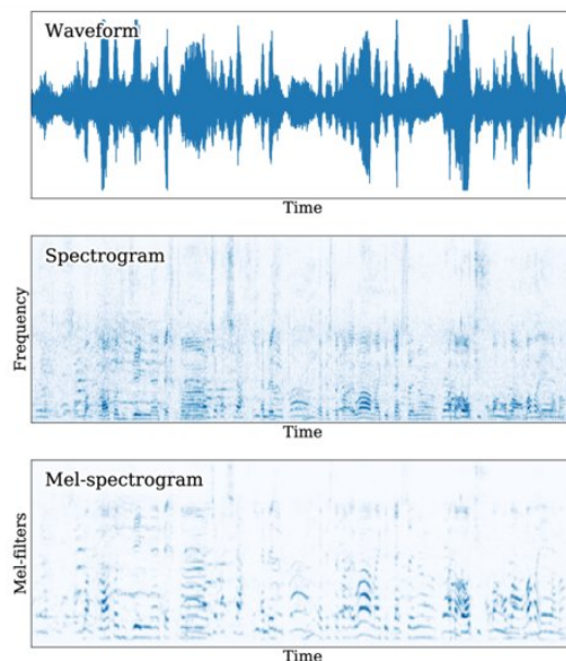


Figura 5.2

Comparación entre distintas representaciones de una señal de audio: (arriba) forma de onda en el dominio del tiempo, (centro) espectrograma en escala lineal de frecuencia, y (abajo) espectrograma mel, donde las frecuencias se transforman a una escala perceptual no lineal. Imagen obtenida de (de Benito et al., 2019).

En el paradigma clásico el flujo de trabajo estándar para la generación de voz sintética consistía en un sistema en cascada de dos etapas. En la primera etapa, un modelo acústico recibe una entrada (texto en TTS o audio en VC - Voice Conversion), utiliza un “encoder” para extraer representaciones lingüísticas profundas, las cuales posteriormente son procesadas por un “decoder” encargado de predecir secuencialmente un espectrograma mel. Finalmente, en la segunda etapa, un “vocoder” neuronal interpreta esta imagen espectral para reconstruir la onda de audio.

Texto o audio $\rightarrow$ encoder $\rightarrow$ decoder $\rightarrow$ espectrograma mel $\rightarrow$ vocoder $\rightarrow$ audio

El año 2016 marcó un cambio en la síntesis de audio, los autores van den Oord et al. (2016) introdujeron WaveNet, el modelo se basaba en redes convolucionales causales y fue pionero en generar ondas de audio directamente y de alta calidad, en lugar de predecir parámetros acústicos que luego debían ser procesados por un vocoder tradicional.

Este modelo construye el sonido desde su nivel más básico y granular, prediciendo un único valor numérico (una muestra) a la vez. La capacidad del modelo para generar estas secuencias se basa en la factorización de la probabilidad conjunta de una forma de onda

$$\mathbf{x} = \{x_1, \dots, x_T\}$$

Los autores van den Oord et al. (2016) definieron la probabilidad conjunta de la señal de audio mediante la siguiente formulación autorregresiva:

$$p(\mathbf{x}) = \prod_{t=1}^T p(x_t \mid x_1, \dots, x_{t-1}) \quad (5.1)$$

Lo relevante de esta ecuación es que indica que la probabilidad de una muestra de audio específica (x_t) depende de todas las muestras generadas en los pasos de tiempo anteriores ($x_1, \dots, x_{t-1}$) lo que marca una mejora al mantener una coherencia temporal en el resultado que se traduce en audio con menos discontinuidades artificiales.

Otro avance importante en este modelo es que este puede ser guiado para imitar a un hablante específico (presente en los datos de entrenamiento e identificado) mediante el uso de una identidad de hablante (condicionamiento global) sin necesidad de entrenar un sistema propio por cada individuo. En términos matemáticos, la función de activación de las capas de la red se

modifica para incluir información adicional

$$z = \tanh(W_{f,k} * x + V_{f,k}^T h) \odot \sigma(W_{g,k} * x + V_{g,k}^T h) \quad (5.2)$$

Donde h es el vector que contiene esta identidad del hablante y $V_{*,k}^T h$ es una proyección lineal que se difunde “broadcast” a lo largo de toda la dimensión temporal del audio. Esto asegura que el estilo o timbre del hablante se mantenga constante de inicio a final en el audio generado. Gracias al condicionamiento global, un mismo modelo empezó a imitar la identidad de múltiples hablantes (presentes en los datos de entrenamiento) simplemente cambiando un vector de entrada.

Paralelamente, sistemas autorregresivos de dos etapas (encoder-decoder) como Tacotron mejoraron el procesamiento de entrada, extrayendo características lingüísticas, mejorando los espectrogramas mel obtenidos. En 2018, Google presentó Tacotron 2, que simplificó drásticamente el flujo de trabajo al integrar un modelo de predicción de espectrogramas mel, por lo que el flujo de trabajo común paso a ser el siguiente:

Entrada $\rightarrow$ Tacotron 2 $\rightarrow$ Espectrograma mel $\rightarrow$ Wavenet (version modificada) $\rightarrow$ Audio

En los años siguientes, modelos como FastSpeech (Ren et al., 2019) y FastSpeech 2 (Ren et al., 2022) introdujeron arquitecturas no autorregresivas basadas en transformers, marcando la transición de la generación de audio paso a paso a una en paralelo, logrando aceleraciones considerables de generación al predecir todas las muestras a la par, manteniendo una calidad comparable a Tacotron 2, optimizando la primera parte del flujo de trabajo (encoder-decoder) y estableciendo un primer acercamiento al como los modelos actuales pueden ser producidos a escala masiva.

Ren et al. (2019) buscaban resolver el problema de lentitud inherente a los modelos autorregresivos, para ello utiliza una estructura llamada “Feed-Forward Transformer” (FFT) que

procesa toda la secuencia de texto al mismo tiempo, otro concepto central es el regulador de longitud, que actúa como un predictor de la duración, minimizando la diferencia entre la duración real del habla y la predicha. Matemáticamente dada una secuencia de fonemas $H_{pho} = [h_1, h_2, \dots, h_n]$ y una secuencia de duraciones predichas $D = [d_1, d_2, \dots, d_n]$, la secuencia expandida H_{mel} se define como:

$$H_{mel} = LR(H_{pho}, D, \alpha) \quad (5.3)$$

El “Regulador de Longitud” (LR) toma cada fonema y lo repite tantas veces como indique su duración predicha (d_i). generando así una representación temporal alineada con el espectrograma mel objetivo. El parámetro (α) introduce además un mecanismo de control sobre la velocidad del habla de forma manual: valores superiores a 1 producen una locución más lenta, mientras que valores inferiores la aceleran.

Un año mas tarde Ren et al. (2022) resolvieron el problema del mapeo conocido como “uno a muchos” marcando la base de la varianza en los deepfakes con el modelo FastSpeech 2. La idea es simple pues un mismo texto puede leerse con diferentes tonos, velocidades y emociones. Usar unicamente texto como entrada en modelos no tiene información suficiente para predecir todas estas variaciones. Para solucionar este problema, el modelo introduce el “adaptador de varianza“, que añade información explícita en tres dimensiones clave: duración, pitch (tono) y energía (volumen) a la secuencia de fonemas antes de generar el audio.

Dentro de la dimensión de duración, el regulador de longitud es idéntico a su predecesor, con ligeros cambios. El tono es el rasgo más complejo de modelar Ren et al. (2022) introducen el uso de la “Transformada de Wavelet Continua” (CWT). Dada una función de contorno de pitch F_0 , se convierte en un espectrograma de pitch $W(\tau, t)$ mediante :

$$W(\tau, t) = \tau^{-1/2} \int_{-\infty}^{+\infty} F_0(x) \psi\left(\frac{x-t}{\tau}\right) dx$$

El modelo descompone la curva del tono (F_0) en diferentes escalas o componentes usando una función matemática llamada wavelet (ψ). El contorno original se recupera mediante la transformada de wavelet continua inversa (iCWT), lo que permite al modelo aprender tanto variaciones rápidas como lentas del tono de forma mucho más precisa, lo que se traduce en una mejora en la emoción percibida en el audio y la entonación de las frases.

La energía en el procesamiento de audio es una medida de la magnitud de la señal en un momento dado, esta indica la magnitud a nivel de marco de los espectrogramas mel y afecta directamente el volumen y la prosodia. en esta dimensión los autores Ren et al. (2022), dividen la señal de audio cruda en pequeños fragmentos (frames). A cada uno se le aplica la “Transformada de Fourier de Tiempo Reducido” (STFT) para pasar del dominio del tiempo al dominio de la frecuencia.

Este valor representa la magnitud total del sonido en ese intervalo de tiempo específico, Los valores de energía se agrupan en 256 niveles posibles de forma uniforme representado mediante un vector de características “embedding”. Así, cada fragmento de audio generado lleva información sobre qué tan fuerte debe sonar.

En síntesis, el adaptador de varianza permite que FastSpeech 2 supere en eficiencia a los modelos autorregresivos manteniendo una calidad comparable al proporcionar información explícita. Además otorga un control total al usuario, permitiendo manipular el tono y el volumen del deepfake de forma independiente simplemente ajustando los parámetros del adaptador facilitando en gran medida la generación de deepfakes.

En paralelo al desarrollo de FastSpeech y FastSpeech 2 surgieron vocoders basados en redes generativas adversarias (GAN), estas Redes propuestas por Ian Goodfellow en 2014, fueron diseñadas para generar datos falsos pero realistas, la idea fundamental es entrenar dos redes neuronales para que compitan entre sí. La primera red corresponde a un generador que crea datos falsos y la segunda a un discriminador que busca detectar si los datos son reales o falsos, mejorando ambas redes en el proceso de competencia. En 2019, la arquitectura MelGAN introdujo un cambio de paradigma al demostrar que una GAN podía superar a los modelos de flujo en velocidad y eficiencia manteniendo una calidad aceptable en la generación de audio.

La innovación crítica de Kumar et al. (2019) no residía solo en su generador, sino en su esquema de discriminación. Introdujeron el Discriminador de Múltiples Escalas (MSD), compuesto por tres discriminadores (D_1, D_2, D_3) que operan a diferentes escalas. (D_1) analiza el audio real, mientras que (D_2) y (D_3) analizan el audio submuestreado por factores de 2 y 4 respectivamente. Matemáticamente esto obligaba a que cada discriminador aprendiera características de diferentes rangos de frecuencia. Además, MelGAN implementó la pérdida de ajuste de características (feature matching loss), que penaliza las diferencias entre las activaciones internas de los discriminadores para el audio real frente al sintético, estabilizando así el entrenamiento adversario.

Solo un año después Kong et al. (2020) presentan HiFi-GAN tomando la base tecnológica de MelGAN y añadiendo mecanismos específicos para resolver la naturaleza compleja y periódica del habla humana. Mientras que MelGAN se enfocaba en la escala del audio, HiFi-GAN introduce el discriminador multi-periodo (MPD), mediante el cual, el audio se divide en muestras igualmente espaciadas usando periodos basados en números primos (2, 3, 5, 7, 11), lo que permitía capturar estructuras periódicas implícitas que otros modelos ignoraban, resultando en una voz mucho más realista.

Una característica crítica para los deepfakes es la capacidad de clonar voces nuevas. HiFi-GAN demostró ya una excelente capacidad de generalización, logrando invertir espectrogramas mel de hablantes que no formaron parte de su entrenamiento permitiendo lo que se denomina clonación “zero-shot”. Sin embargo dada su naturaleza como vocoder, el modelo depende totalmente de la entrada, si el modelo que genera el espectrograma produce una imagen ruidosa, los resultados reflejaran esta falencia. Dado que el modelo solo convierte frecuencias en sonido, este no tiene la capacidad de deducir el estilo, la emoción o el entorno acústico de una persona a partir de una muestra breve; esa responsabilidad recaía en el modelo acústico previo.

Junto a este problema de la acumulación de errores, una de las innovaciones más importantes en el procesamiento directo del audio fue el reconocimiento de que las transformaciones al dominio de la frecuencia, como el espectrograma de magnitud, descartaban información de fase potencialmente diagnóstica. Esta observación impulsó en la comunidad el desarrollo de modelos “end-to-end”, que operan directamente sobre la forma de onda cruda en el dominio del tiempo,

marcando el inicio de un nuevo paradigma de modelos llamado “aprendizaje en contexto”, éstos representan un cambio importante pues ya no separaba el problema en múltiples módulos especializados (encoder - decoder - vocoder), los nuevos modelos aprenden la generación de voz de forma mucho más unificada.

La idea revolucionaria dentro de este paradigma para la síntesis de voz, fue tratar el audio como si fuera lenguaje, el audio ya no se trabajaba como una señal continua muy compleja, sino como una secuencia de tokens discretos, es decir como elementos seleccionados de un conjunto finito de símbolos, tal como funcionaría en el lenguaje natural. En vez de trabajar con el waveform continuo (sumamente largo y redundante en arquitecturas deep learning) o su representación en espectrogramas, el audio o texto se transforma en una secuencia compacta de índices enteros, mediante un codec neuronal que preserva la inteligibilidad, el timbre, la prosodia y otros detalles perceptualmente importantes. Este proceso empieza con un encoder que recibe la señal de audio y la convierte en una representación continua de menor dimensionalidad, posteriormente se necesita representar el audio mediante símbolos discretos. Aquí entra un proceso llamado cuantización vectorial.

La cuantización vectorial convierte los vectores producidos por el encoder en índices enteros pertenecientes a un codebook. El codebook funciona como un diccionario aprendido de patrones acústicos frecuentes. El sistema guarda el índice del vector más cercano dentro de ese diccionario. Así, el audio termina representado como secuencias de IDs discretos. El decoder del codec neuronal realiza el proceso inverso. Recibe la secuencia de tokens discretos y reconstruye el waveform original, lo que da como resultado un nuevo flujo de trabajo:

Waveform $\rightarrow$ Encoder $\rightarrow$ Quantization $\rightarrow$ Tokens discretos $\rightarrow$ Decoder $\rightarrow$ Audio reconstruido

Si el audio puede representarse como tokens discretos, entonces generar voz puede verse como un problema de modelado autorregresivo exactamente igual al lenguaje. En 2023 los autores C. Wang et al. (2023) presentaron VALL-E al trabajar para Microsoft, un modelo donde la arquitectura transformer predice el siguiente token acústico dado el texto y algunos tokens

previos de audio. El sistema aprende distribuciones estadísticas sobre secuencias acústicas. La principal fórmula de VALL-E busca maximizar la probabilidad :

$$\max p(C \mid x, \tilde{C})$$

Donde C son los códigos de audio a generar, x es la secuencia de fonemas del texto y $\tilde{C}$ es el “prompt acústico” (3 segundos del hablante). Resumiendo el modelo deduce cual seria la secuencia de sonidos mas probable (C) que seguiría esa voz ($\tilde{C}$) para ese mensaje (x). El modelo descompone la generación del audio en dos etapas por eficiencia, un primer modelo autorregresivo (AR) que genera la estructura básica de la voz predeciedo los tokens de la primera capa del cuantizador, junto a un segundo modelo no autorregresivo (NAR) que rellena la calidad acústica procesando el resto de los niveles de forma paralela reduciendo la complejidad computacional. Los autores C. Wang et al. (2023) advierten explícitamente sobre el riesgo de suplantación de identidad “spoofing” debido a la facilidad de clonar voces con muestras mínimas recomendando ya desde este punto la construcción de modelos de detección o “modelos anti-spoofing”.

Solo un año después Chen et al. (2024) presentan VALL-E 2 como una evolución a su predecesor, mostrando una mayor estabilidad y velocidad de generación ademas de ser el primer sistema que alcanza la paridad humana en los benchmarks LibriSpeech y VCTK que actuaban como pruebas para saber que tan naturales eran los resultados en comparación con los producidos por un locutor real.

Finalmente llegamos al estado del arte a partir de finales de 2024 y principios de 2025 con modelos como Maya1 desarrollado por Maya Research (2025), representando la culminación de la integración entre los Grandes Modelos de Lenguaje (LLM) y la síntesis de audio. Su arquitectura utiliza un transformador de solo decodificador de 3 mil millones de parametros, diseñado para predecir tokens del codec SNAC. La relevancia técnica de SNAC radica en su estructura jerárquica, que divide cada trama de audio en siete tokens operando a diferentes resoluciones como se explica en la tabla 5.1

Tabla 5.1

Niveles jerárquicos de SNAC y su función acústica en la representación de la señal de voz.
(Elaboración propia)

Nivel SNAC	Resolución poral	tem- Función acústica
Nivel 1 (Coarse)	~12 Hz	Ritmo, cadencia y estructura temporal básica.
Nivel 2 (Mid)	~23 Hz	Articulación fonética y claridad de palabras.
Nivel 3 (Fine)	~47 Hz	Detalles espectrales, textura de la voz y micro-matices.

El panorama de síntesis de audio que hemos recorrido refleja un campo en constante expansión, donde arquitecturas neuronales, flujos de difusión y modelos de lenguaje convergen hacia sistemas cada vez más naturales, eficientes y controlables. Este vertiginoso avance no ha pasado desapercibido: la misma capacidad que permite generar voz sintética indistinguible de la real plantea desafíos críticos de autenticidad y confianza, impulsando en paralelo el desarrollo de métodos capaces de detectar y distinguir el audio generado artificialmente del genuino.

5.3 EVOLUCIÓN DE LA DETECCIÓN DE DEEPFAKES (ENFOQUE DEEP LEARNING)

El contexto histórico de la detección ha mostrado una transformación durante la última década, en el ámbito de la acústica forense, el problema estaba ligado a áreas previas como los sistemas de reconocimiento de un hablante o la verificación del mismo, evolucionado desde el análisis rudimentario de artefactos espectrales hasta sistemas complejos basados en modelos de fundación y arquitecturas de grafos.

Para mejorar la profundidad de análisis, en esta sección nos concentraremos en la historia a partir de 2015, pues anterior a esta fecha, la investigación en spoofing estaba dividida. Se

desarrollaban contra-medidas para algoritmos específicos de síntesis TTS o VC sin protocolos de evaluación comunes, lo que resultaba en una falta de generalización. El punto de inflexión ocurrió con la creación de la iniciativa Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof). Esta iniciativa de Z. Wu et al. (2015) introdujo el corpus Spoofing Awareness for Speaker Verification (SAS), compuesto por diez algoritmos de síntesis de texto a voz (TTS) y conversión de voz (VC) bajo condiciones acústicas controladas de laboratorio, lo que proporcionó por primera vez una base de datos estandarizada, métricas rigurosas y un foro para la evaluación competitiva de sistemas.

Las técnicas de generación de la época, como los vocoders tradicionales y los modelos de parámetros ocultos de Markov producían audios con fragmentos digitales evidentes “firmas”, tales como artefactos en el espectro de frecuencias, los cuales un clasificador entrenado podía aprender a identificar con relativa consistencia. Los sistemas de detección en el periodo de 2015-2017, seguían un flujo de trabajo dual: un componente de extracción de características seguido de un clasificador estadístico. Este flujo de trabajo puede resumirse en:

Audio $\rightarrow$ pre-procesamiento $\rightarrow$ extracción de características $\rightarrow$ modelo de clasificación $\rightarrow$ score

Durante el pre-procesamiento se buscaba conservar solo la información útil de la voz. Para esto se usaba técnicas como la del “detector de actividad de voz” (VAD), que separa los segmentos con voz de silencios, pausas y ruido, técnica que mostró ser muy útil en reconocimiento automático del habla. Después se aplicaba alguna técnica de extracción de características, para esta etapa, los Mel-Frequency Cepstral Coefficients (MFCCs), dominaron en popularidad dado que históricamente representaban el estándar en los sistemas de “reconocimiento automático del habla” (ASR, por sus siglas en inglés), sin embargo resultaron insuficientes para la detección de deepfakes, porque su diseño, inspirado en la audición humana, tiende a ignorar las altas frecuencias donde los sintetizadores de la época también podían fallar. Esto motivó el desarrollo y la adopción de características diseñadas específicamente para este problema que se detallan mejor en la Tabla 5.2.

Tabla 5.2

Características acústicas desarrolladas para suplir las limitaciones de los MFCCs en detección de voz sintética. LFCC y CQCC amplían la cobertura espectral de alta frecuencia, mientras que MGD incorpora información de fase ignorada por los métodos clásicos.

Caract.	Mecanismo	Ventaja principal	Limitación
LFCC	Filtros distribuidos linealmente en frecuencia.	Conserva información de alta frecuencia que el escalado Mel suprime, detectando artefactos del vocoder.	Alta sensibilidad al ruido de fondo y variaciones de grabación.
CQCC	Transformada Q Constante (CQT) con resolución espectral variable.	Mayor resolución en frecuencias bajas; detecta irregularidades armónicas no visibles con FFT clásica.	Costo computacional considerablemente mayor.
MGD	Retardo de grupo modificado; extrae información de fase.	Evidencia inconsistencias de fase propias de sintetizadores que solo modelan magnitud espectral.	Difícil de aplicar fuera de entornos controlados.

En la penúltima etapa los clasificadores estadísticos eran predominantemente modelos de mezclas gaussianas (GMM). Estos modelos aprendían la distribución estadística de las características de la voz real frente a la sintética. Sin embargo, su principal limitación era la rigidez, un GMM entrenado para detectar un vocoder específico fallaba catastróficamente al encontrarse con una técnica de síntesis nueva, un fenómeno conocido como el problema de “ataques no vistos”. Otras metodologías comunes en esta fase incluían el uso de máquinas de soporte vectorial, el uso de vectores de identidad junto al análisis discriminante lineal probabilístico, entre otras. El último componente es el score: un valor numérico continuo que representa el nivel de confianza del modelo, donde valores altos indican mayor probabilidad de audio genuino y valores bajos de

audio sintético, en lugar de una etiqueta binaria directa.

La diversidad metodológica del paradigma clásico es sumamente modificable y explicar cada variación excede los alcances de esta revisión. Sin embargo, muchas de las limitaciones, se resuelven con el auge del deep learning. Como se menciono, la iniciativa ASVspoof fue clave para la mejora de los modelos, la edición de 2017 del desafío amplió el alcance del problema al incorporar ataques de repetición, en los que una grabación auténtica es capturada y reproducida mediante un altavoz físico frente a un sistema de verificación. Este tipo de amenaza difería sustancialmente de los ataques de síntesis, pues la señal no era generada algorítmicamente sino físicamente degradada. Para este sub-problema, la huella forense relevante no residía en los artefactos del vocoder sino en las propiedades acústicas del entorno. Esta edición impulsó las primeras aplicaciones de redes neuronales convolucionales (CNN) al problema, señalando una transición progresiva hacia el aprendizaje de representaciones automáticas.

La maduración del aprendizaje profundo trajo consigo un cambio de paradigma fundamental a partir de 2018. La detección de deepfakes de audio se centró en tratar los espectrogramas como imágenes bidimensionales y aplicarles arquitecturas diseñadas originalmente para visión por computadora, tales como ResNet. Este enfoque eliminaba en gran medida la necesidad de diseñar características a mano, permitiendo que la red neuronal aprendiera por sí misma cuáles regiones del espectro resultaban relevantes para distinguir la voz real de la sintética.

Dentro de esta corriente, uno de los aportes más influyentes fue la introducción de la Light Convolutional Neural Network (LCNN) por X. Wu et al. (2018), cuya innovación central residía en una función de activación denominada Max-Feature-Map (MFM). A diferencia de la función ReLU convencional, la MFM introduce un mecanismo de competencia entre pares de mapas de características: en lugar de simplemente suprimir valores negativos, selecciona el máximo entre dos canales correspondientes, forzando a la red a retener únicamente las activaciones más discriminativas. Esta operación, expresada formalmente como $y_{ij}^k = \max(x_{ij}^k, x_{ij}^{k+N/2})$, redujo significativamente la cantidad de parámetros del modelo sin sacrificar su capacidad de representación, facilitando el procesamiento eficiente de espectrogramas de alta dimensionalidad. Aunque la LCNN fue concebida originalmente para el reconocimiento facial con etiquetas ruidosas, su arquitectura ligera y su robustez frente al ruido motivaron una rápida adopción

en la comunidad antispoofing, donde demostró ser especialmente efectiva para detectar ataques de “acceso lógico” concepto adoptado en el desafío ASVspoof 2019 para formalizar los tipos de ataque como se explica a continuación.

El reto introdujo una distinción crítica entre dos categorías de ataque que hasta entonces se trataban de manera menos sistemática: los ataques de acceso lógico (LA), generados por TTS y VC, y los ataques de acceso físico (PA), correspondientes a los escenarios de repetición ya presentes en 2017. Esta bifurcación reconocía que las firmas forenses de ambos tipos de ataque eran fundamentalmente diferentes y requerían estrategias de detección distintas. Desde el punto de vista de los datos, el corpus de esta edición amplió considerablemente la variedad de métodos de síntesis incluidos, incorporando sistemas basados en el estado del arte de la época con modelos ya explicados como WaveNet y Tacotron. Este salto cualitativo en la sofisticación de los ataques tensionó los sistemas de detección existentes y aceleró la transición hacia modelos más expresivos.

En paralelo al trabajo sobre espectrogramas, surgió otra línea de investigación centrada en la integración de modelos de secuencia, motivada por la incapacidad de las CNN convencionales para capturar adecuadamente las propiedades temporales del habla sintética. Como se explicó, los vocoders neuronales de la época arrastraban errores acumulativos en la síntesis debido a su flujo de trabajo, estos errores se manifestaban como discontinuidades en la prosodia, irregularidades en el ritmo de las pausas y transiciones fonéticas con una regularidad mecánica poco natural. Para modelar estas dependencias de largo alcance se recurrió a las redes neuronales recurrentes (RNN) y, en particular, a las unidades “Long Short-Term Memory” (LSTM), capaces de retener información a lo largo de secuencias extensas de características acústicas. Sin embargo, las RNN no estaban exentas de limitaciones: el conocido problema del desvanecimiento del gradiente seguía afectando su desempeño en secuencias suficientemente largas, lo que terminaría motivando la exploración de arquitecturas alternativas basadas en mecanismos de atención.

Con la llegada de los modelos “end-to-end” surgió SincNet propuesto por los autores Ravanelli y Bengio (2019), una arquitectura que planteó una modificación simple pero efectiva en la primera capa convolucional: en lugar de dejar que los filtros tomaran formas arbitrarias durante el entrenamiento (como en una CNN estándar), los reemplazó por filtros de paso de banda paramétricos basados en la función sinc. Cada filtro solo necesitaba aprender dos valores,

la frecuencia de corte inferior (f_1) y la superior (f_2), con lo que el núcleo del filtro quedaba completamente determinado por:

$$g[n, f_1, f_2] = 2f_2 \text{sinc}(2\pi f_2 n) - 2f_1 \text{sinc}(2\pi f_1 n) \quad (5.4)$$

Al reducir tanto el espacio de parámetros que el modelo debía explorar, la convergencia se aceleró y cada filtro resultante tenía una interpretación física directa, lo que mejoraba notablemente la interpretabilidad de las características aprendidas. Sobre este principio se construyó RawNet2 Tak et al. (2021), que sumaba a los filtros sinc iniciales bloques residuales ResNet y un mecanismo de re-escalado de mapas de características “Feature Map Scaling” (FMS), el cual ajustaba adaptativamente la relevancia de cada filtro según el contexto global del audio.

Posteriormente, con el objetivo de tratar de forma más específica los ataques de conversión de voz, DeepRawNet (Alharbi et al., 2026) extendió este enfoque incorporando funciones de activación PReLU de pendiente negativa aprendible a diferencia de la ReLU estándar que descarta por completo los valores negativos, esta variante podía capturar artefactos de baja energía que los ataques de VC solían dejar como huella junto a convoluciones transpuestas que preservaban la resolución temporal fina necesaria para detectar micro-discontinuidades en los límites fonéticos.

En 2021, Jung et al. (2021) introdujeron AASIST, una arquitectura cuya innovación central fue replantear la detección de deepfakes en términos de teoría de grafos. La motivación era directa: los artefactos que dejan los sintetizadores de voz no se manifiestan de forma aislada solo en el espectro o solo en el tiempo, sino en la relación entre ambas dimensiones, es decir, en cómo ciertas frecuencias cambian de manera anómala a lo largo del audio. Las arquitecturas anteriores no podían capturar este tipo de patrones cruzados al analizar ambas dimensiones por separado.

Para resolver este problema, AASIST toma la salida de un codificador externo basado en una variante de RawNet2 y construye dos grafos paralelos: uno espectral, cuyos nodos repre-

sentan subbandas de frecuencia, y uno temporal, cuyos nodos representan segmentos del audio en el tiempo. Mediante una capa de atención heterogénea (HS-GAL), el modelo permite que la información navegue entre ambos grafos, conectando así ambas dimensiones en el análisis. Un nodo global de pila (Stack Node) consolida esta información para la decisión final, funcionando de forma similar al token [CLS] en los modelos Transformer. El resultado fue un salto de rendimiento notable con reducciones relativas del orden del 20 % en la tasa de error igual (EER) y una variante ligera, AASIST-L, con apenas 85.000 parámetros que mantenía resultados comparables en entornos con recursos computacionales limitados.

Desde aproximadamente 2022, el paradigma dominante en la detección de deepfakes de audio pasó al aprovechamiento de modelos pre-entrenados mediante aprendizaje auto supervisado (SSL). La idea detrás de este enfoque es bastante directa: modelos como Wav2Vec 2.0 Baevski et al. (2020) o HuBERT Hsu et al. (2021), entrenados sobre decenas de miles de horas de habla humana real sin ninguna etiqueta de clase, aprenden representaciones muy detalladas sobre los sonidos naturales de la voz humana. Al adaptarlos a la tarea de detección, actúan como detectores de anomalías: cualquier desviación del patrón de habla natural aprendido durante el entrenamiento inicial se interpreta como una posible falsificación. Esto los hace especialmente valiosos cuando se desconoce el método de síntesis del atacante, ya que el detector no busca la huella de un sistema en específico, sino cualquier desviación de la naturalidad fonética general.

5.4 EVOLUCION DE LOS CONJUNTOS DE DATOS Y DESAFÍOS

La evolución de los conjuntos de datos ha ido de la mano con el progreso técnico descrito en esta sección, y entender las limitaciones de cada corpus resulta tan importante como conocer las arquitecturas que los aprovecharon. A partir de ASVspoof 2021 (Delgado et al., 2021) las condiciones acústicas dejaron de ser relativamente limpias. Esta edición del reto introdujo efectos de codificación telefónica y transmisión de red como compresión MP3 y OPUS en los datos de evaluación, exponiendo la fragilidad de los modelos cuando se pierden componentes espectrales de alta frecuencia durante la transmisión.

El desafío ADD 2023 de Yi et al. (2023) amplió el problema hacia un frente distinto: la detección parcial. A diferencia de los corpus anteriores, que etiquetaban un audio completo

como real o falso, aquí los modelos debían localizar temporalmente los fragmentos manipulados dentro de audios que mezclaban segmentos auténticos con fragmentos generados por IA. Este escenario, conocido como “audio inpainting”, plantea un reto específico: identificar el punto de costura digital donde el ruido de fondo o la fase de la señal cambia abruptamente al insertar el fragmento sintético, algo que los clasificadores de audio completo no estaban diseñados para resolver.

ASVspoof 5 (X. Wang et al., 2024), incorporó datos recolectados mediante “crowdsourcing” de más de 4,000 hablantes en una variedad de entornos acústicos y tipos de micrófono sin precedentes, además de incluir por primera vez ataques adversariales deliberados como parte integral del corpus de evaluación. Otra de las innovaciones más importantes es la sustitución del tradicional EER por otras métricas basadas en funciones de costo de detección (minDCF y a-DCF), que permiten evaluar el impacto económico y de seguridad real de los fallos del sistema. Finalmente, los resultados del reto demuestran que, ante la perfección de los nuevos ataques, el uso de modelos SSL, como wav2vec 2.0 resulta ser la estrategia más efectiva frente a la complejidad del reto.

Pese a que ASVspoof se consolidó como la forma de evaluación estándar de los modelos de detección, el reto no se encuentra exento de limitaciones importantes. La primera y más crítica a nivel técnico es que el corpus representa condiciones de laboratorio, aunque un sistema presente buenas métricas durante el reto, estas decaen drásticamente ante condiciones reales sin detectar audios deepfake fuera del conjunto de evaluación, lo que definiremos como “sesgo de conjunto cerrado”. En este contexto podemos mencionar lo que dicen N. Müller et al. (2022) cuyos resultados fueron contundentes: sistemas que alcanzaban tasas de error cercanas al 1 % en ASVspoof 2019 presentaban EER de entre el 20 % y el 40 % al evaluarse sobre otro corpus más robusto propuesto por los autores, evidenciando que los modelos no habían aprendido los artefactos propios del audio sintético. A esto se le suma el desequilibrio lingüístico: hasta ASVspoof 5, el inglés dominaba sin contrapeso, un problema para sistemas multilingüaje, dado que los artefactos de síntesis dependen de las propiedades fonéticas de cada lengua.

Otra limitación a mencionar es que los sistemas entrenados con corpus anteriores a 2022 fallan de manera sistemática ante audio generado por modelos de flow-matching contemporáneos.

La idea detrás de este enfoque es bastante directa: si se añaden al audio sintético perturbaciones matemáticamente optimizadas, imperceptibles al oído humano, es posible modificar la decisión del clasificador de falso a real. este enfoque explota los gradientes del propio modelo detector para encontrar estas perturbaciones de manera eficiente.

En respuesta, se han propuesto arquitecturas de defensa que combinan módulos de reconstrucción de señal para exponer la firma del ataque con funciones de pérdida basadas en aprendizaje de tripletes, de modo que los audios reales se agrupan en el espacio de representación y tanto los sintéticos como los adversariales quedan apartados H. Wu et al. (2020), mejorando la robustez ante ataques no vistos durante el entrenamiento.

Dado todo el contexto histórico del problema, el capítulo 7 examina cómo la investigación reciente ha respondido a las siguientes grietas: qué datasets y estrategias de entrenamiento permiten superar el sesgo de conjunto cerrado, cómo hacer interpretable la decisión del detector, y qué ocurre cuando estos sistemas operan sobre lenguas distintas del inglés. En los tres casos, las respuestas son parciales.

6. METODOLOGÍA

6.1 DISEÑO DEL ESTUDIO

Con el fin de garantizar transparencia, reproducibilidad y trazabilidad en todas las etapas del proceso, el diseño metodológico considera las directrices establecidas por PRISMA 2020, lo que incluye el desarrollo de una estrategia de búsqueda exhaustiva en bases de datos, la construcción de criterios de inclusión (IC), criterios de exclusión (EC) y la aplicación de procedimientos estandarizados para la selección de los trabajos que marcan el estado del arte a analizar en el capítulo 7. Todo el proceso de filtrado se encuentra detallado en las secciones del presente capítulo, para un resumen más visual de lo trabajado en todo el proceso, dirigirse a la figura 6.1.

6.2 ESTRATEGIA DE BÚSQUEDA

Se empleó una búsqueda mediante *queries* (búsqueda avanzada). Inicialmente, se consideraron tres bases de datos debido a su relevancia en el campo del deep learning, así como a su amplia cobertura y rigor en publicación e indexación: [1] IEEE Xplore, [2] Scopus y [3] ACM Digital Library. Se excluyó la base de datos [3] debido a limitaciones técnicas en la exportación de metadatos en formato BIBTEX, requisito indispensable para el procesamiento sistemático en la herramienta parsifal.

En un inicio, se consideraron documentos publicados entre 2015 y 2025 (posteriormente se limitó a 2020–2025), en idioma español o inglés. Durante la búsqueda, se utilizaron diferentes queries, desde muy generales hasta otras más específicas, con el fin de validar la temática. La búsqueda más amplia arrojó un conjunto de 49.677 resultados. dicho conjunto contenía documentos duplicados y trabajos de temáticas no pertinentes; sin embargo, evidencia la amplitud de la literatura en torno a los deepfakes desde enfoques interdisciplinarios.

La query seleccionada para continuar con el proceso dado que generó el conjunto más adecuado, filtrando eficazmente un número considerable de documentos irrelevantes sin ser ex-

cesivamente restrictiva, fue:

```
("deepfake" OR "spoofing" OR "anti-spoofing") AND
("audio" OR "speech") AND
("deep learning" OR "neural network" OR "transformer")
```

6.3 PROCESO DE SELECCIÓN

Se recuperaron 2.999 documentos. Durante la importación a parsifal, se perdieron 678 registros de Scopus por errores de formato BIBTEX, de IEEE Xplore solo se tomaron 57 artículos que correspondían a aquellos de acceso libre, reduciendo el conjunto a 1.568 documentos. Tras la eliminación de 24 documentos duplicados entre ambas bases, el conjunto se redujo un total de 1.544 artículos que continuaron a una fase de selección, para esta fase todo el conjunto ya cumplía con los siguientes criterios de inclusión (IC) establecidos:

- **IC1:** Publicaciones revisadas por pares (artículos de revista, conferencias o workshops indexados).
- **IC2:** Estudios disponibles dentro de las bases de datos seleccionadas al momento de la revisión.
- **IC3:** Trabajos publicados en inglés o español.
- **IC4:** Artículos recuperados por el query de búsqueda, independientemente del enfoque metodológico.

Posteriormente, se aplicaron tres etapas de filtrado sucesivas; en cada una, un documento avanzaba a la siguiente únicamente si no satisfacía ningún criterio de exclusión (EC) definido en cada screening.

6.3.1 PRIMER SCREENING

El primer screening corresponde a una depuración inicial del conjunto, enfocada en verificar la alineación temática de los estudios con el objetivo propuesto. Para ello, se aplicaron ECs orientados principalmente a descartar trabajos que: (i) no se centraban en la modalidad auditiva, (ii) abordaban enfoques de spoofing no acústico (como diarización, deepfakes de texto, entre otros), y (iii) aunque trataban la temática de interés, no presentaban evaluación cuantitativa ni resultados experimentales que permitieran su comparación con otros estudios.

- **EC1:** Trata únicamente deepfakes en imagen, video u otras modalidades no auditivas sin un componente exclusivamente centrado en audio falsificado.
- **EC2:** No involucra manipulación o detección de deepfakes en audio-imagen (otros enfoques de spoofing distintos al audio, diarización, análisis de emociones, biometría no acústica, textos, etc.).
- **EC3:** No presenta evaluación cuantitativa ni experimentos, aunque trate la temática (opiniones, revisiones narrativas, ensayos, análisis legales o éticos).

El conjunto se redujo de 1.544 a 428 artículos, con un total de 1.116 documentos descartados: 922 por EC1, 78 por EC2 y 116 por EC3.

6.3.2 SEGUNDO SCREENING

En el segundo screening se aplicaron ECs más estrictos con el objetivo de refinar el conjunto de estudio y asegurar un mayor nivel de relevancia e impacto dentro del campo investigativo. En esta etapa, dado que aún se contaba con un número elevado de publicaciones, el período de análisis se restringió a trabajos publicados entre 2020 y 2025. Asimismo, se excluyeron documentos sin acceso a texto completo, ya sea por limitaciones institucionales o por no contar con acceso libre. Finalmente, se establecieron umbrales en indicadores de calidad de revistas y conferencias.

- **EC1:** Año de publicación fuera del periodo 2020–2025.
- **EC2:** No se cuenta con acceso al texto completo, por limitaciones de acceso institucional o falta de acceso libre a través de DOI
- **EC3:** Revista o journal no Q1 en Scimago Journal Rank y SJD 2024 <0.6 y H-index <100 . Para papers de conferencia sin revista H-index <120 .

En cuanto al criterio EC2, cabe mencionar que se identificaron cinco documentos de alta relevancia para los objetivos del estudio, para los cuales se solicitó acceso directamente a los autores. Sin embargo, no fue posible obtener dicho acceso, por lo que finalmente fueron excluidos del análisis.

Respecto al criterio EC3, este se flexibilizó de forma selectiva para incluir trabajos que, aunque no cumplieran estrictamente los umbrales establecidos, aportaban información relevante en las áreas específicas a analizar en el capítulo 7 (i) propuestas de datasets multilingües, (ii) técnicas de explicabilidad (XAI) y (iii) entrenamiento de modelos basados en idiomas distintos al inglés.

En resumen durante esta fase, se excluyeron 275 registros, distribuidos en 70 documentos por EC1, 123 por EC2 y 82 por EC3, reduciendo el conjunto a 153 estudios.

6.3.3 TERCER SCREENING

En esta fase final, se realizó una lectura rápida y comparativa de cada documento con ayuda de inteligencia artificial como herramienta de apoyo para la comparación de resúmenes y extracción de criterios técnicos, reservando el juicio de inclusión al criterio del autor. El enfoque no se centró exclusivamente en la calidad técnica, sino en el aporte diferencial de cada estudio para responder a las preguntas de investigación planteadas. Se asignó una menor prioridad a aquellos artículos cuyo enfoque se limitaba estrictamente al reporte de métricas de rendimiento, sin una discusión profunda sobre los mecanismos de detección o la naturaleza del audio.

- **EC1:** Estudios que, si bien presentan calidad técnica, utilizan datasets convencionales (ma-

yormente en inglés) sin profundizar en la discusión multilingüe, técnicas de explicabilidad o propuestas de nuevos datasets.

- **EC2:** Trabajos centrados predominantemente en técnicas de aprendizaje automático tradicional en lugar de arquitecturas de Deep Learning.

Como resultado final, se excluyeron 114 registros, distribuidos en 109 documentos por EC1 y 5 por EC2. El conjunto se redujo a 39 estudios relevantes, los cuales constituyen la base del análisis y la síntesis de resultados de esta revisión sistemática.

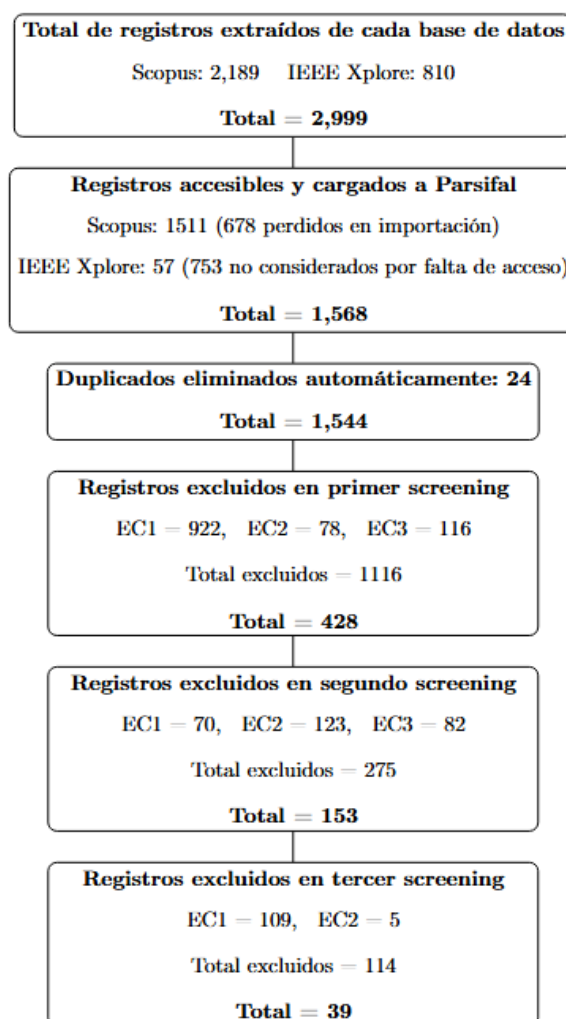


Figura 6.1

Diagrama PRISMA del proceso de identificación, depuración y selección inicial de los documentos incluidos en la revisión sistemática. Se muestra el total de registros obtenidos, pérdidas por importación, eliminación de duplicados y exclusiones aplicadas según criterios de exclusión en cada etapa de screening.

7. ESTADO DEL ARTE

7.1 DATASETS Y ESTRATEGIAS DE ENTRENAMIENTO FRENTE AL SESGO DE CONJUNTO CERRADO

El sesgo de conjunto cerrado es una limitación que estructura la investigación actual en detección de audios deepfake. Un detector entrenado sobre un corpus finito de sistemas de síntesis aprende a reconocer las firmas particulares de esos sistemas; sin embargo, cuando se enfrenta a un generador no visto durante el entrenamiento, su capacidad discriminativa colapsa. N. M. Müller et al. (2024) demostraron que esta caída no se debe a que los nuevos ataques sean intrínsecamente más difíciles de detectar, sino a que sus distribuciones estadísticas son simplemente distintas lo que denominan la “brecha de diferencia”, evidenciando que los modelos memorizan las firmas particulares del corpus de entrenamiento en lugar de aprender rasgos generales del audio sintético.

La literatura reciente responde a esta limitación de dos maneras complementarias. La primera, amplía el alcance del corpus de evaluación con nuevas formas de ataque o “vectores de ataque” que no habían sido incluidas en los benchmarks estándar como ASVspoof 2019: audio generados por codecs neuronales, modelos de difusión, voces cantada e idiomas distintos del inglés. El segundo explora estrategias de aprendizaje, intentando aprender representaciones invariantes con el método de síntesis, en lugar de adaptado a firmas específicas.

El primer vector de ataque del que podemos hablar fue el audio generado por modelos de lenguaje basados en codecs neuronales. Los modelos como VALL-E (descritos en la sección 5.2) generan la forma de onda directamente desde representaciones discretas aprendidas por codecs como EnCodec o SoundStream, sin pasar por la etapa de vocoder de los modelos clásicos.

En respuesta a esta necesidad, Lu et al. (2024) construyeron el corpus Codecfake para evidenciar esta brecha: el corpus cuenta con más de un millón de muestras en inglés y mandarín, generadas con siete métodos de codec distintos. Los resultados fueron contundentes: los modelos entrenados sobre ASVspoof 2019 que alcanzan tasas de error iguales (EER) menores al 1 % sobre

vocoders obtienen un EER promedio de entre el 36 % y el 44 % sobre audio basado en codecs. El entrenamiento sobre Codecfake redujo el EER promedio a 0,177 % ante codecs vistos y a un 0,884 % ante codecs no vistos, aunque con una pérdida parcial de capacidad para detectar vocoders convencionales, concluyendo que: “aún es necesario crear un conjunto de datos de codecs variado que abarque los métodos de generación de códecs neuronales y planteando la pregunta central de si es posible un detector generalizado”.

En paralelo, Bhagtani et al. (2024) evaluaron si los detectores del estado del arte son capaces de identificar audio producido por modelos de síntesis recientes basados en difusión: ProDiff, DiffGAN-TTS, ElevenLabs, UnitSpeech y XTTS, que permiten clonar una voz a partir de apenas tres segundos de referencia. El corpus propuesto abarca 52 horas de audio sintético distribuidas entre 11 hablantes y cinco métodos generativos. Los seis detectores evaluados, con EER menores al 6 % en ASVspoof 2019, se degradan de forma unánime: el mejor de ellos (PS3DT) alcanza un 11,45 % y Wav2Vec2-AASIST sube de un 0,30 % a un 30,18 %, lo que confirma que la detección correcta, se desplaza con cada generación de sintetizadores. Otro detalle a mencionar es que los autores muestran que los métodos basados en modelos SSL preentrenados mostraron una degradación relativamente menor a la de los sistemas supervisados clásicos, aunque la caída en rendimiento sigue siendo importante.

Otros autores como L. Zhang et al. (2022) extendieron el problema desde otra dirección: en lugar de clasificar falsificaciones completas, propusieron el corpus PartialSpoof para detectar segmentos sintéticos insertados dentro de un audio mayoritariamente genuino. El dataset ofrece etiquetas a múltiples resoluciones temporales, de 20 ms a 640 ms lo que lo convierte en el primer benchmark orientado específicamente a la localización del punto de costura digital. Un detector construido para falsificaciones completas no está preparado para esta tarea; Los autores muestran que esta exige estrategias de detección propias. Los resultados fueron que un detector entrenado sobre ASVspoof 2019 LA con EER de 0,83 % en su propio corpus sube a 14,19 % al evaluarse sobre PartialSpoof, evidenciando nuevamente que los clasificadores tampoco capturan los artefactos de inserción propios de las falsificaciones parciales. Su modelo propuesto Wav2Vec 2.0 Large logra 0,49 % en PartialSpoof y mantiene 0,90 % en ASVspoof 2019 LA, demostrando que ambos escenarios son compatibles cuando el entrenamiento cubre ambos vectores de ataque.

La distancia entre los resultados en laboratorio y el entorno real fue documentada sistemáticamente por N. Müller et al. (2022). Su dataset In-the-Wild recoge 20,8 horas de audio genuino y 17,2 horas de deepfakes públicamente disponibles de 54 figuras públicas, sin ningún control sobre el método de síntesis empleado. Los modelos con EER cercano al 1 % en ASVspoof 2019 presentan EER de entre el 20 % y el 40 % sobre este corpus, materializando la consecuencia práctica del sesgo de conjunto cerrado para audios en condiciones reales.

Otros ejemplos de dominio distintos en los que el sesgo de conjunto cerrado se hace presente es el caso de la voz cantada, que constituye un dominio acústicamente distinto del habla: sus propiedades prosódicas y la presencia de acompañamiento musical dificultan la generalización directa de los detectores sobre el problema. Zang, Zhang et al. (2024) documentan esta brecha con SingFake, el primer dataset para detección de deepfakes en voz cantada: 28,93 horas de audio genuino y 29,40 horas de deepfakes recolectados de plataformas de contenido generado por usuarios, en cinco lenguas (mandarín, cantonés, japonés, inglés y persa) con 40 cantantes. Los resultados sobre cantantes no vistos (T02) con modelos entrenados en ASVspoof 2019 LA son contundentes: AASIST obtiene un EER de 58,12 % sobre mezclas de canción y Wav2Vec2-AASIST un 56,75 %, equivalentes en ambos casos a clasificación aleatoria. Incluso tras reentrenar sobre SingFake, el conjunto T04 (cantantes persas no vistos en entrenamiento) representa el reto más difícil para los modelos al contener muestras de la lengua no vista junto a un género musical no visto, en este conjunto el EER no baja del 38,54 % evidenciando que la lengua y el género musical representan barreras de generalización independientes.

Tabla 7.1

Analogía entre las tareas de síntesis y detección en habla hablada y voz cantada. SVS y SVC son los vectores de ataque; SVDD es la tarea de detección que los enfrenta.

Tarea (canto)	Equivalente (habla)	Función
SVS (<i>Singing Voice Synthesis</i>)	TTS	Genera voz cantada desde partitura o letra con notas
SVC (<i>Singing Voice Conversion</i>)	VC	Transforma la voz de un cantante para que suene como otro
SVDD (<i>Singing Voice Deepfake Det.</i>)	Anti-spoofing	Detecta si una grabación cantada es genuina o sintética

Para estudiar esta brecha con control preciso sobre los métodos generativos y superar la falta de reproducibilidad de SingFake, cuyas fuentes de síntesis son en su mayoría desconocidas al ser recolectadas en condiciones In-the-Wild (redes sociales), Zang, Shi et al. (2024) construyeron CtrSVDD, el primer benchmark controlado para SVDD: este cuenta con 220,798 clips de voz cantada a cappella en 307,98 horas totales, generados con 14 métodos documentados (7 SVS y 7 SVC) sobre vocales de 164 cantantes en mandarín y japonés. La evaluación confirma dentro del dominio, la misma brecha de conjunto cerrado documentada en habla: el mejor modelo con un EER global de 13,75 %—desciende a 0,96 % ante A10 (Soft-VITS-SVC, visto en entrenamiento), pero sube a 26,83 % ante DiffSinger (A12, basado en difusión) y a 19,03 % ante ACESinger (A14, sistema comercial), ambos fuera del entrenamiento.

El primer reto dedicado a SVDD, Y. Zhang et al. (2024), recibió 47 equipos en la pista CtrSVDD, de los cuales 37 superaron los sistemas base (EER 12,03 % y 11,16 %); el equipo ganador obtuvo 1,65 %, con estrategias dominantes basadas en características SSL. Paralelamente, Nguyen-Duc et al. (2025) compararon modelos de espectrograma (AST, Whisper) frente a modelos de forma de onda (UniSpeech-SAT, HuBERT) sobre SingFake, encontrando que los primeros generalizan mejor ante cantantes no vistos (T02): AST alcanza un F1 de 0,8514 frente al 0,5908 de HuBERT, lo que los autores atribuyen a que los espectrogramas capturan mejor las variaciones de frecuencia continua características del canto.

En el dominio de idioma Purohit et al. (2024) propusieron GGMDDC, el primer corpus diseñado explícitamente para abarcar diversidad lingüística: 80,000 muestras de las cuales 40,000 eran reales provenientes de VoxLingua107 y 40,000 sintéticas generadas con HiFi-GAN en diez lenguas: ruso, francés, árabe, español, vietnamita, mandarín, inglés, hindi, portugués y sánscrito. El corpus obtiene una “puntuación de opinión media” (MOS) de 4,52, el cual representan la métrica estándar para evaluar la calidad percibida y la naturalidad del audio sintético generado, además de un MOS subjetivo de 4,39, los más altos entre los datasets comparados en el trabajo. Sin embargo, el propio trabajo identifica una limitación crítica: el generador HiFi-GAN fue entrenado exclusivamente sobre datos en inglés (LJ Speech y VCTK), lo que compromete la representación de los patrones fonéticos propios de cada lengua en las muestras sintéticas generadas.

Tabla 7.2

Corpus y benchmarks nuevos revisados en esta sección, ordenados por eje temático. Las lenguas abreviadas siguen el código ISO 639-1 cuando aplica. “Codec” indica síntesis directa desde codec neuronal; “Difusión” indica modelos generativos basados en procesos de difusión probabilística.

Corpus	Referencia	Lengua(s)	Real	Fake	Novedad principal
<i>Nuevos vectores de ataque</i>					
Codecfake	(Lu et al., 2024)	en, zh	132,277	925,939	7 codecs neuronales (EnCodec, DAC, etc.); primer corpus para audio LLM-based
Bhagtani et al.	(Bhagtani et al., 2024)	en	24,226	25,000	5 generadores de difusión (ProDiff, ElevenLabs, XTTS); prueba de obsolescencia de detectores
<i>Escenarios de detección complejos</i>					
PartialSpoof	(L. Zhang et al., 2022)	en	—	—	Etiquetas multi-resolución (20–640 ms); falsificaciones parciales en utterance real
In-the-Wild	(N. Müller et al., 2022)	en	≈20,8 h	≈17,2 h	Deepfakes reales de 54 figuras públicas; primera prueba de generalización en condiciones reales
<i>Voz cantada</i>					
SingFake	(Zang, Zhang et al., 2024)	en, zh, jap, persa	—	—	Primer dataset in-the-wild para SVDD; 5 lenguas, 40 cantantes
CtrSVDD	(Zang, Shi et al., 2024)	zh, jap	13,596 clips	79,173 clips	Benchmark controlado; 14 métodos SVS/SVC; EER 1,65 % (líder SVDD 2024)
<i>Cobertura multilingüe</i>					
GGMDDC	(Purohit et al., 2024)	10 lenguas	40,000	40,000	Mayor diversidad lingüística en un único corpus; MOS 4,52

La incorporación de modelos de fundación preentrenados mediante aprendizaje autosupervisado (SSL) como front-end de extracción de características es la respuesta más estudiada al sesgo de conjunto cerrado. La idea detrás de este enfoque es directa: un modelo como XLS-R, preentrenado sobre 436.000 horas de habla en 128 lenguas sin ninguna etiqueta de autenticidad, aprende representaciones generales del habla natural, actuando como detector de anomalías ante cualquier desviación de ese patrón aprendido, con independencia del sistema de síntesis concreto.

Pascu et al. (2024) demostraron que este principio, en su versión más simple representaciones XLS-R de 2B parámetros congeladas más un clasificador de regresión logística de apenas 2.000 parámetros entrenables, reduce el EER promedio de RawNet2 de un 30,9 % a un 8,8 % sobre un benchmark de ocho datasets fuera del dominio, con mejoras en todos ellos salvo uno. El análisis comparativo que utiliza once variantes del modelo, mostró que el tamaño del modelo preentrenado y la diversidad de lenguas en el preentrenamiento son los factores más determinantes: la variante de 2B parámetros entrenada en datos multilingües supera consistentemente a las variantes monolingües de 94M, evidenciando que la exposición a múltiples lenguas durante el preentrenamiento genera representaciones más generalizables. Finalmente, el trabajo introduce la calibración probabilística como dimensión de evaluación, demostrando que los detectores SSL también producen predicciones más fiables que los modelos supervisados clásicos.

Sobre este mismo principio, Q. Zhang et al. (2024) propusieron aprovechar de forma más granular la información disponible en las distintas capas del transformer de XLS-R, mediante un módulo de selección sensible de capas (SLS, *Sensitive Layer Selection*). Las capas iniciales e intermedias del transformer capturan características a distinto nivel de abstracción respecto a las finales, y la combinación ponderada de todas ellas en lugar de usar solo la última ofrece señales más discriminativas para la detección. La combinación XLS-R + SLS produjo el primer sistema que bajó del umbral del 2 % en el dataset ASVspoof 2021 DF (EER 1,92 %), al tiempo que lograba un 7,46 % en In-the-Wild; los experimentos de ablación confirmaron que ni el ajuste fino ni la aumentación de datos resultan prescindibles incluso cuando el front-end aporta representaciones ricas.

Xie et al. (2023) añadieron a esta línea un objetivo de aprendizaje explícito de invarianza al dominio. Su sistema W2V-ASDG combina el front-end XLS-R con un módulo ASDG (*Ag-*

gregation and Separation Domain Generalization) que aplica dos mecanismos simultáneamente: un discriminador adversarial de dominio de un solo lado, que fuerza a las representaciones del habla real a ser indistinguibles entre dominios mediante una capa de inversión de gradiente, y una función de pérdida de tripletes, que agrupa el habla real y dispersa la sintética en el espacio de características. Entrenado con tres corpus de dominios distintos (ASVspoof 2019, WaveFake, FakeAVCeleb), el modelo obtiene un EER promedio de 4,60 % sobre 13 condiciones de evaluación que incluyen chino y japonés, a pesar de que el conjunto de entrenamiento era exclusivamente en inglés, lo que lo convierte en uno de los primeros sistemas con transferencia cross-lingual documentada sin entrenamiento en las lenguas objetivo.

Una línea distinta no actúa sobre el front-end sino sobre la forma en que se organiza el propio entrenamiento. Si el problema es que el detector se ajusta a los ataques que vio, una respuesta natural es simular durante el entrenamiento la aparición de ataques nuevos. Con esta motivación, Z. Wang y Hansen (2024) parten de un codificador basado en RawNet2 y le añaden un esquema de meta-aprendizaje episódico: en cada iteración reservan un tipo de ataque como si fuera desconocido y obligan al modelo a clasificarlo a partir del resto, de modo que aprende un espacio de representación independiente del método de síntesis concreto. A ello suman ejemplos adversariales generados con “descenso de gradiente proyectado” (PGD), no como amenaza sino como aumentación que expone al modelo a perturbaciones difíciles, desacoplando sus estadísticas mediante una normalización por lotes auxiliar. La combinación alcanza un EER del 0,87 % sobre los ataques no vistos de ASVspoof 2019 LA, una reducción relativa cercana al 48 % respecto al modelo base, aunque la validación se mantiene dentro del mismo corpus.

No todos los trabajos seleccionados persiguen la generalización fuera del dominio; otra parte del conjunto busca exprimir al máximo la capacidad discriminativa dentro de un corpus combinando arquitecturas y representaciones de entrada. Zaman et al. (2024) exploran de forma sistemática esta dirección al acoplar un *Transformer* a distintas redes (ResNet34, VGG16, Bi-LSTM y una máquina de soporte vectorial) y alimentarlas con tres representaciones: espectrograma mel, MFCC y la “transformada de fourier de tiempo reducido” (STFT). El hallazgo es que la STFT separa mejor las clases que las otras dos representaciones, y que su unión con un ResNet34-Transformer logra el mejor resultado del trabajo, con un EER del 3,22 % en el conjunto de evaluación de ASVspoof 2019 LA. El aporte es mostrar qué combinación de característica y

arquitectura híbrida extrae señales más discriminativas, si bien toda la evaluación ocurre dentro del mismo corpus y no mide el comportamiento ante generadores externos.

El costo computacional del entrenamiento con grandes corpus fue abordado por Azeemi et al. (2023). Su método de poda autosupervisada parte de las representaciones latentes de wav2vec 2.0 para cada muestra de entrenamiento, agrupa esas representaciones mediante k -means y retiene las muestras más alejadas de los centroides es decir, las más informativas para el modelo descartando los ejemplos prototípicos redundantes. Aplicado a ASVspoof 2019, el método es un 91 % más eficiente en tiempo de cómputo que los métodos supervisados equivalentes: con el 90 % de los datos eliminados, AASIST-L alcanza un EER de 10,30 % frente al 13,74 % del descarte aleatorio, lo que traduce el potencial de los embeddings SSL hacia un uso más eficiente de los datos de entrenamiento.

Las estrategias anteriores siguen aprendiendo qué distingue lo real de lo falso a partir de ejemplos de ambas clases. Pianese et al. (2022) proponen un giro más radical: prescindir por completo del audio sintético en el entrenamiento. Su enfoque de una sola clase (*one-class*) reformula la detección como un problema de verificación de hablante, partiendo de la idea de que, si se conoce la identidad reclamada, basta con comprobar si la voz analizada coincide con un conjunto de referencia de grabaciones genuinas de esa persona; cualquier desviación se marca como falsa. Como el modelo solo aprende de voz real usando extractores de *embeddings* de hablante entrenados sobre VoxCeleb2, la generalización no depende de ningún método de síntesis. Los resultados fueron contundentes: mientras los detectores supervisados, casi perfectos en ASVspoof 2019, caen al nivel del azar sobre FakeAVCeleb e In-the-Wild, el enfoque one-class mantiene un área bajo la curva (AUC) promedio cercana al 91 % en los tres corpus, evidenciando una robustez ante generadores no vistos que los clasificadores binarios no logran.

Hay además una vulnerabilidad distinta del conjunto cerrado pero igual de relevante: los ataques adversariales, en los que perturbaciones diminutas añadidas a la entrada bastan para invertir la decisión del detector. Frente a ellos H. Wu et al. (2020) proponen una defensa pasiva, es decir, que no necesita anticipar el algoritmo de ataque ni reentrenar el clasificador. La idea consiste en anteponer al detector un modelo autosupervisado (Mockingjay), preentrenado para reconstruir habla a partir de fragmentos enmascarados, que actúa como una capa de purificación:

al proyectar la entrada contaminada sobre representaciones de alto nivel del habla natural, atenua el ruido adversarial capa por capa e impide su transferencia al clasificador. A diferencia del entrenamiento adversarial, que expone al modelo a perturbaciones conocidas, este enfoque neutraliza ataques de caja negra sin haberlos visto, y los autores muestran que el preentrenamiento es el factor que explica su eficacia.

El diagnóstico de por qué los detectores fallan en condiciones fuera del dominio fue formalizado por N. M. Müller et al. (2024) mediante la descomposición de la brecha de rendimiento en dos componentes: la “brecha de dificultad” (*hardness gap*) que mide si el nuevo dataset es intrínsecamente más difícil y la “brecha de diferencia” (*difference gap*) que mide si las distribuciones de los datos simplemente no coinciden con las del entrenamiento. Experimentos sobre cuatro modelos (LCNN, RawNet2, WhisperDF, SSL-W2V2) y tres conjuntos de evaluación (ASVspoof 2021 LA, DF e In-the-Wild) mostraron que la brecha de dificultad es prácticamente despreciable: la degradación se atribuye casi exclusivamente a la diferencia entre distribuciones. En el caso de In-the-Wild, el gap total de entre el 31 % y el 78 % de EER se debe en su totalidad a este componente de diferencia, lo que implica que aumentar la capacidad del modelo no resuelve el problema de generalización: los modelos aprenden atajos estadísticos específicos del corpus de entrenamiento.

Mientras el trabajo anterior formaliza por qué falla la generalización, Khan et al. (2023) la miden de forma empírica en la evaluación cruzada más amplia del período: entrenan las contramedidas sobre un corpus y las prueban sobre otro distinto (ASVspoof 2019, ASVspoof 2021 y VSDC), cubriendo todo el espectro metodológico características diseñadas a mano, aprendizaje profundo y sistemas *end-to-end* con cuatro clasificadores. La conclusión es que ninguna contramedida sostiene su rendimiento al cambiar de corpus: incluso los sistemas *end-to-end* más fuertes se desploman, como AASIST, que pasa de un EER del 0,83 % en ASVspoof 2019 a un 15,84 % en ASVspoof 2021. Entre las características clásicas, los coeficientes cepstrales de Q constante (CQCC) son los más estables, pero en evaluación cruzada su función de costo de detección se acerca al máximo, lo que confirma que el buen desempeño dentro de un corpus no anticipa el comportamiento en otro.

Chetia Phukan et al. (2024) realizaron el análisis más sistemático sobre qué tipo de

representación SSL generaliza mejor para la detección en condiciones multilingüe y cross-corpus. Evaluaron tres familias de modelos de fundación con distintas coberturas lingüísticas: XLS-R (128 lenguas), Whisper (96 lenguas), MMS (más de 1.400 lenguas), frente a alternativas monolingües (WavLM, Wav2Vec2), encontrando que la heterogeneidad del preentrenamiento supera consistentemente a la homogeneidad: los modelos multilingües superan a los monolingües en todos los benchmarks evaluados. Sobre este principio proponen “Multilingual Integration Output” (MIO), una fusión bilineal de representaciones de múltiples modelos que alcanza el estado del arte en ASVspoof 2019 e In-the-Wild, consolidando el argumento de que la diversidad lingüística en el preentrenamiento es un sustituto parcial del etiquetado específico para cada lengua.

Sobre este diagnóstico, los trabajos más recientes combinan varias técnicas para cerrar la brecha en lugar de confiar en una sola. Doan et al. (2024) articulan tres estrategias de entrenamiento alrededor de un front-end XLS-R: aprendizaje contrastivo supervisado, que agrupa la voz real y separa la sintética; minería de negativos difíciles por re-síntesis de audio, que genera pares real-falso con el mismo contenido lingüístico para forzar al modelo a ignorarlo y fijarse en los artefactos; y un muestreo proactivo de lotes que equilibra cada minibatch. Todo ello se condensa con una red de cuello de botella variacional que descarta información superflua. El sistema obtiene un EER del 2,07 % sobre la pista de deepfakes de ASVspoof 2021 y del 3,78 % sobre In-the-Wild, manteniendo el rendimiento donde otros caen al cuádruple; su análisis de ablación señala además que el muestreo de lotes es el componente más determinante.

Un segundo trabajo añade una dimensión hasta entonces ausente: la privacidad del contenido. X. Li et al. (2024) parten de que el deepfake imita el timbre y la prosodia del hablante, no lo que dice, por lo que la detección no debería exigir acceso al contenido hablado. Su sistema desacopla la señal mediante un codec neuronal con HuBERT como guía semántica en tokens semánticos y acústicos, y entrega al detector únicamente los acústicos, de modo que el contenido lingüístico nunca queda expuesto. A pesar de operar sin esa información, alcanza un EER del 2,02 % sobre su corpus multilingüe CVoiceFake, que abarca cinco lenguas (inglés, chino, alemán, francés e italiano), mientras que las tasas de error de palabra superiores al 93 % confirman que el contenido resulta irrecuperable. El trabajo muestra que generalización multilingüe y preservación de la privacidad no son objetivos incompatibles.

Una mirada de conjunto permite ordenar este panorama. Ranjan et al. (2023) realizan un meta-análisis de los sistemas publicados sobre ASVspoof 2019 y extraen un patrón claro: en los ataques de acceso lógico los tres mejores operan sobre forma de onda cruda combinada con redes de grafos, y entre ellos figura AASIST-L con apenas 85.000 parámetros, lo que nuevamente indica que la generalización no depende del tamaño del modelo; en los ataques de acceso físico, en cambio, siguen dominando las características diseñadas a mano basadas en la transformada Q constante. La observación más relevante para esta sección es que ninguno de los sistemas comparados fue evaluado fuera de su corpus de origen, por lo que el propio ranking de arquitecturas se apoya en una medición que no garantiza nada sobre su comportamiento real; los autores señalan esta ausencia de evaluación cruzada como una de las brechas críticas del campo.

Tabla 7.3

Estrategias de entrenamiento para la generalización fuera del dominio revisadas en esta sección. EER reportado sobre el dataset de referencia indicado entre paréntesis. “—” indica que el paper no reporta EER como métrica principal.

Propuesta	Referencia	Mecanismo central	EER clave
XLS-R + regresión logística	(Pascu et al., 2024)	Front-end SSL congelado (2B params.); sin ajuste fino; 8 datasets OOD	8,8 % (OOD prom.)
XLS-R + SLS Classifier	(Q. Zhang et al., 2024)	Selección adaptativa ponderada de 24 capas del transformer	1,92 % (ASV21 DF)
W2V-ASDG	(Xie et al., 2023)	Discriminador adversarial de dominio + triplet mining	4,60 % (13 cond. OOD)
MiO (fusión bilineal)	(Chetia Phukan et al., 2024)	Modelos multilingües (XLS-R, Whisper, MMS) > monolingües; fusión MiO	SOTA ASV19 + IWA
Meta-Learning + PGD	(Z. Wang y Hansen, 2024)	Meta-aprendizaje + ejemplos adversariales (PGD) sobre RawNet2	0,87 % (ASV19)
Trident of Poseidon	(Doan et al., 2024)	Contrastive learning + Hard Negative Mining + Batch Sampling + VIB	2,07 % (ASV21 DF)
ResNet34-Transformer	(Zaman et al., 2024)	Comparativa de arquitecturas híbridas; mejor: ResNet34 + STFT	3,22 % (eval)
SafeEar	(X. Li et al., 2024)	Tokens acústicos únicamente (sin contenido semántico); CVoiceFake 5 lenguas	2,02 % (CVoiceFake)
Poda autosupervisada	(Azeemi et al., 2023)	wav2vec2 + k -means; retiene muestras informativas (91 % más eficiente)	10,30 % (90 % poda-do)
One-class (verificación)	(Pianese et al., 2022)	Solo voz real del hablante reclamado; generaliza a síntesis no vista	AUC 91 % (prom.)
Defensa SSL (Moc-kingjay)	(H. Wu et al., 2020)	SSL como capa de purificación ante ataques adversariales de caja negra	—

Tabla 7.4

Estudios de diagnóstico, evaluación y análisis comparativo de la generalización fuera del dominio revisados en esta sección. A diferencia de los cuadros anteriores, recoge trabajos cuyo aporte no es un corpus ni una estrategia de entrenamiento, sino el análisis de la brecha de conjunto cerrado.

Estudio	Referencia	Tipo	Aporte principal
Harder or Different	(N. M. Müller et al., 2024)	Análisis de la brecha	Descompone la caída fuera de dominio en “dificultad” y “diferencia”; la diferencia entre distribuciones explica casi todo el gap
Battling Voice Spoofing	(Khan et al., 2023)	Evaluación cross-corpus	Ninguna contramedida sostiene su rendimiento entre corpus; AASIST pasa de 0,83 % a 15,84 % de EER
Uncovering the Deceptions	(Ranjan et al., 2023)	Meta-análisis	Raw audio + GNN domina en LA; AASIST-L (85K parámetros) en el top-2; ningún sistema evaluado cross-corpus
Comparative Sing-Fake	(Nguyen-Duc et al., 2025)	Comparativa de modelos	El espectrograma (AST) generaliza mejor que la forma de onda ante cantantes no vistos (F1 0,85 vs 0,59)
SVDD 2024	(Y. Zhang et al., 2024)	Challenge	Primer reto de detección en voz cantada; 47 equipos, EER líder 1,65 % con características SSL

Recorridos los dos caminos con que la literatura responde al sesgo de conjunto cerrado, emerge un cuadro coherente. Por el lado de los datos, el período 2020–2025 amplió el problema en frentes que los benchmarks estándar no contemplaban: nuevos paradigmas generativos como los codecs neuronales y la difusión, dominios acústicos distintos como la voz cantada y, sobre todo, lenguas más allá del inglés. Por el lado de las estrategias de entrenamiento, la incorporación de

modelos de fundación preentrenados mediante aprendizaje autosupervisado se consolida como la aproximación más robusta, complementada por el meta-aprendizaje, la detección de una sola clase y los objetivos explícitos de invarianza al dominio. Aun así, ninguna de estas respuestas cierra del todo la brecha entre el dominio conocido y el desconocido.

Dos hallazgos de esta sección convergen en el mismo diagnóstico. La “brecha de diferencia” de N. M. Müller et al. (2024) muestra que la degradación fuera del dominio no proviene de una mayor dificultad intrínseca sino del desajuste entre distribuciones, y la ventaja de la heterogeneidad de Chetia Phukan et al. (2024) muestra que los modelos preentrenados en muchas lenguas generalizan mejor que los monolingües. Ambos apuntan a que es la diversidad del preentrenamiento, y no la capacidad del modelo, el factor que gobierna la generalización. Queda, sin embargo, una pregunta que estas métricas no responden: qué aprende realmente el detector cuando decide que una voz es falsa. Esa necesidad de abrir la caja negra motiva el siguiente eje de esta revisión, dedicado a la explicabilidad.

7.2 EXPLICABILIDAD EN LA DETECCIÓN DE DEEPPAKES DE AUDIO

La pregunta central que guía esta sección es la siguiente: qué aprende un detector cuando decide que una voz es falsa. En los contextos forenses y periodísticos donde estas herramientas acabarán usándose, una decisión sin justificación es una decisión difícil de sostener; y, sobre todo, entender en qué se apoya el modelo es el camino más directo para diagnosticar por qué se derrumba fuera de su dominio de entrenamiento. La “inteligencia artificial explicable” (XAI, por sus siglas en inglés) atiende esa necesidad. Dentro del corpus revisado, el problema se aborda por dos caminos complementarios. El primero explica la caja negra a posteriori: toma un detector ya entrenado y le aplica herramientas de atribución para revelar en qué regiones de la señal se concentra su decisión. El segundo invierte el orden y parte de características acústicas que son interpretables por sí mismas, de modo que la explicación es intrínseca al rasgo y no un añadido posterior.

El primer camino no modifica el detector, lo interroga. Yu et al. (2024) aplican “Shapley Additive exPlanations” (SHAP) a un modelo de doble rama, una sobre la forma de onda y otra sobre el espectrograma, que alcanza un EER de 3,39 % en ASVspoof 2019 LA con apenas 0,52

millones de parámetros. La atribución revela que el soporte de la decisión no es difuso: en la rama de la forma de onda se concentra en la amplitud de las altas frecuencias, y en la del espectrograma en la frecuencia fundamental y sus armónicos por debajo de 4 kHz. Can y Soyhan (2025) llegan a una conclusión convergente por una vía distinta: combinan un vector acústico interpretable de 51 dimensiones con embeddings de redes convolucionales, lo examinan con importancia por permutación y mediante el uso de la técnica “Gradient-weighted Class Activation Mapping” (Grad-CAM). La permutación señala los primeros coeficientes cepstrales, sus derivadas y las bandas de contraste espectral como los rasgos decisivos, mientras que Grad-CAM confirma que la atención recae sobre la estructura de formantes y la textura de picos y valles del espectro, y no sobre artefactos accidentales de grabación; su sistema añade además la atribución del generador concreto, con un 94,59 % de exactitud binaria y un 87,09 % al identificar cuál de trece sistemas produjo el audio en ASVspoof2021-LA. Dos herramientas, dos arquitecturas y un mismo veredicto: lo que el modelo mira es la firma armónica y formántica de la voz.

La atribución no solo dice en qué frecuencias se fija el modelo, sino en qué momentos. Liu et al. (2024) estudian el escenario de la manipulación parcial, donde solo un fragmento breve de una elocución es sintético, y aplican Grad-CAM a una contramedida basada en aprendizaje autosupervisado. Para no quedarse en la inspección de ejemplos sueltos, introducen una métrica, la cuantificación de contribución relativa, que mide a qué tipo de segmento atiende el modelo a lo largo de todo el conjunto de evaluación. El hallazgo es que: un detector entrenado con audio parcialmente falso aprende a fijarse en las regiones de transición, las costuras que la concatenación deja entre la parte auténtica y la sintética, y no en las diferencias internas de cada fragmento como hace un modelo entrenado solo con audio íntegramente falso. Más revelador aún es el análisis de los errores: cuando el sistema se equivoca, deja precisamente de atender esas transiciones y se dispersa hacia el habla.

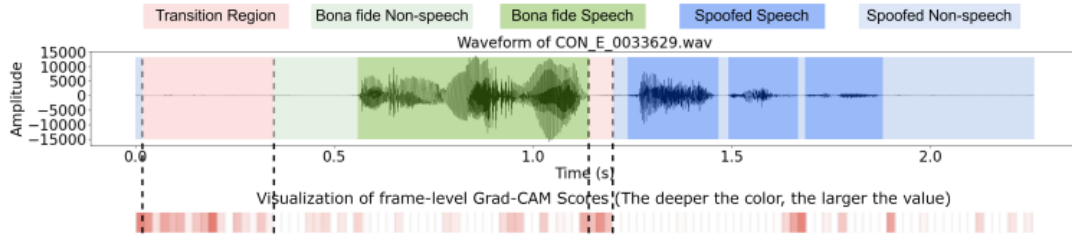


Figura 7.1

Forma de onda de un audio parcialmente falso con sus segmentos etiquetados (región de transición, habla y no-habla auténticas, habla y no-habla sintéticas) y, debajo, las puntuaciones Grad-CAM por marco: cuanto más intenso el color, mayor es la contribución del marco a la decisión del modelo. La atención se concentra en las regiones de transición.

Imagen obtenida de (Liu et al., 2024).

Estos tres trabajos explican modelos que no fueron diseñados para ser explicados. Una alternativa es construir la transparencia dentro de la propia arquitectura. M. Li y Zhang (2024) proponen una representación de activación de clase temporal que integra el mecanismo interpretativo en el modelo en lugar de aplicarlo después: sobre un front-end autosupervisado wav2vec 2.0 XLS-R, un módulo de atención por canales produce un mapa que localiza los marcos temporales decisivos al mismo tiempo que clasifica. Entrenar con etiquetas múltiples que distinguen el tipo de ataque, en lugar de una única etiqueta binaria de auténtico frente a falso, permite al modelo aprender las características propias de cada familia generativa y mejora de forma marcada el rendimiento; con etiqueta binaria el error se dispara hasta un 4,26 % de EER, frente al 0,51 % y un min t-DCF de 0,0165 que el sistema completo logra en ASVspoof2019 LA. La interpretabilidad por diseño muestra que transparencia y rendimiento no son objetivos enfrentados.

Tabla 7.5

Métodos de explicabilidad aplicados a modelos profundos de detección, revisados en esta sección. Recoge los trabajos que explican un detector ya entrenado mediante atribución a posteriori o que integran la interpretabilidad en la propia arquitectura.

Estudio	Referencia	Herramienta XAI	Hallazgo principal
Explainable Spec-trograms	(Yu et al., 2024)	SHAP (post-hoc)	El soporte se concentra en la F0 y los armónicos bajo 4 kHz y en la amplitud de las altas frecuencias; EER 3,39 %
Spectro-Temporal-CNN	(Can y Soyhan, 2025)	Grad-CAM + importancia por permutación	Atención a los formantes y a la textura de picos y valles; atribuye el generador (87,09 % entre 13 sistemas)
Partial Spoof CMs	(Liu et al., 2024)	Grad-CAM + métrica RCQ	El detector se fija en las regiones de transición; al equivocarse deja de atenderlas
Temporal Class Activation	(M. Li y Zhang, 2024)	Interpretabilidad por diseño (T-CAM)	Localiza los marcos discriminativos; el entrenamiento multietiqueta baja el EER a 0,51 %

El segundo camino renuncia a abrir la caja negra y elige, en su lugar, rasgos cuyo significado físico es conocido de antemano. M. Li et al. (2022) ofrecen la base de este enfoque al comparar dieciséis características, divididas entre físicas y perceptuales, sobre el corpus mandarín de ADD 2022. Su resultado desafía la preferencia habitual por los coeficientes cepstrales: la predicción lineal perceptual, una representación inspirada en la audición humana, es la más discriminativa en la pista de audio íntegramente falso, con un EER del 19 % que supera a las características físicas en los tres clasificadores probados; y en la pista de manipulación parcial, donde casi todas fracasan, la combinación de coeficientes cepstrales de Q constante con predicción lineal perceptual reduce el error hasta el 7 %. El trabajo es el primero en demostrar de

manera sistemática, mediante el análisis de las distribuciones de voz real y sintética, que los rasgos perceptuales como el tono, la relación armónico-ruido o el jitter y el shimmer encierran información que los físicos no capturan.

Esa pista perceptual conduce a la prosodia, el terreno donde la voz sintética muestra sus costuras. K. Li et al. (2023) se centran en el jitter y el shimmer, las microvariaciones de frecuencia y amplitud entre periodos glóticos consecutivos, y constatan que el habla generada exhibe diferencias estadísticas claras, sobre todo en el shimmer, cuya inestabilidad de amplitud es mucho mayor en lo falso. Al estimar la frecuencia fundamental con algoritmos más robustos, YIN y SWIPE, y fusionar el shimmer con el espectrograma, reducen el EER del 41,29 % al 35,77 % en ADD2023. Gao et al. (2020) elevan esta intuición a hipótesis: la voz de máquina es demasiado consistente y no reproduce las microperturbaciones que el aparato fonador humano genera de forma inevitable. Comparan ataques de imitación humana con voz sintética y establecen que los deepfakes son más peligrosos que un imitador profesional; luego muestran que rasgos ligados a la producción del habla, como la entropía espectral de la secuencia de frecuencia fundamental, la envolvente espectral y los parámetros aperiódicos, separan lo real de lo sintético de manera consistente entre corpus distintos, y que fusionarlos con representaciones convencionales reduce la brecha entre los conjuntos de desarrollo y evaluación. La naturalidad, entendida como inconsistencia, se vuelve así un rasgo discriminativo.

Los dos caminos confluyen cuando la atribución se aplica sobre rasgos ya interpretables. Bisogni et al. (2024) prescinden del aprendizaje profundo y entrenan clasificadores ligeros sobre un conjunto de características hechas a mano, coeficientes cepstrales, medidas espectrales, energía, tasa de cruces por cero y croma, evaluándolas en tres corpus con un protocolo independiente del hablante para evitar el sesgo biométrico: 89 % de exactitud en ASVspoof2019, 94,5 % en FakeAVCelebV2 y 94,67 % en el conjunto In-The-Wild. Lo relevante para este eje es que aplican SHAP sobre esos rasgos transparentes y obtienen un resultado estable a través de los tres corpus: el primer coeficiente cepstral y el ancho de banda espectral son los indicadores más fiables de lo sintético, con la tasa de cruces por cero como apoyo recurrente. La explicación señala así firmas acústicas detectables y comunes a tecnologías de generación distintas, y tiende un puente entre los dos caminos, porque usa la atribución del primero sobre la materia interpretable del segundo.

Tabla 7.6

Estudios basados en características acústicas interpretables revisados en esta sección. A diferencia del cuadro anterior, la explicación no proviene de una herramienta externa sino del significado físico o perceptual de los propios rasgos.

Estudio	Referencia	Características	Hallazgo principal
Physical vs Perceptual	(M. Li et al., 2022)	PLP, CQCC, jitter, shimmer, HNR, tono	Las perceptuales superan a las físicas: PLP con EER 19 %; CQCC+PLP 7 % en manipulación parcial
Jitter & Shimmer	(K. Li et al., 2023)	Jitter y shimmer (F0 con YIN/SWIPE)	El shimmer es el más discriminativo; EER de 41,29 % a 35,77 % en ADD2023
Human vs Machine	(Gao et al., 2020)	Entropía de F0, envolvente espectral, parámetros aperiódicos	La voz sintética es “demasiado consistente”; los rasgos reducen la brecha entre desarrollo y evaluación
Acoustic Features XAI	(Bisogni et al., 2024)	MFCC, medidas espectrales, ZCR, cromagrama + SHAP	El primer coeficiente cepstral y el ancho de banda espectral son firmas universales de lo sintético

Recorridos ambos caminos, el cuadro que emerge es notablemente coherente. Ya se interrogue a un modelo profundo con SHAP o Grad-CAM, ya se construya la explicación desde rasgos perceptuales y prosódicos, el sustrato que distingue la voz auténtica de la sintética es siempre el mismo: la frecuencia fundamental y sus armónicos, la estructura de formantes y las microperturbaciones de la producción vocal que la síntesis aún no replica del todo. Que rutas tan distintas converjan en el mismo lugar físico sugiere que la huella del deepfake de audio no es un artefacto arbitrario del clasificador, sino una propiedad del propio proceso generativo. La explicabilidad, además, deja de ser un fin en sí mismo y se revela como instrumento de diagnóstico, pues los mismos análisis que localizan las regiones de transición o las bandas decisivas son los que permiten entender cuándo y por qué un detector deja de generalizar. Queda, sin embargo, un límite que esta sección no resuelve: casi toda la evidencia procede de corpus en

inglés y mandarín, y nada garantiza que esas firmas acústicas se conserven en otras lenguas y estructuras fonéticas. Esa pregunta, la de si lo aprendido sobre unas pocas lenguas se sostiene en las demás, motiva el último eje de esta revisión.

7.3 DETECCIÓN DE DEEPPAKES DE AUDIO EN LENGUAS MENOS REPRESENTADAS

El eje anterior cerró con una advertencia: las firmas acústicas que distinguen lo auténtico de lo sintético se establecieron casi siempre sobre el inglés y el mandarín, las dos lenguas que acaparan los corpus de anti-spoofing. Nada garantiza que esa huella se conserve en otras estructuras fonéticas, y la cuestión no es menor. Muller et al. (2024) lo plantean sin rodeos al presentar su corpus multilingüe: el sesgo lingüístico de los datos, predominantemente en inglés, limita la aplicabilidad de los detectores y termina por discriminar a los hablantes de otras lenguas, justo cuando los deepfakes proliferan en medios no anglófonos. El problema tiene dos caras que conviene separar. La primera es de recursos: para la mayoría de las lenguas del mundo no existen datos con los que entrenar un detector. La segunda es de generalización: un modelo entrenado en una lengua no transfiere lo aprendido a otra con la fiabilidad que cabría esperar. Este último eje recorre ambas, primero a través de la evidencia que aportan los trabajos centrados en una lengua concreta y después a través de las estrategias que buscan cruzar la frontera lingüística.

Conviene empezar por los casos particulares, porque son los que ponen cifras a la brecha. El árabe ilustra el punto de partida. Almutairi y Elgibreen (2023) construyeron el primer conjunto sintético de voz árabe, generado con FastSpeech 2 y un vocoder HiFi-GAN sobre un único hablante de árabe estándar moderno, y propusieron Arabic-AD, un detector basado en el modelo autosupervisado HuBERT que alcanza un EER del 0,027 % y un 97 % de exactitud sobre habla con acentos. La cifra impresiona, pero su lectura honesta es la contraria a la que sugiere: se obtiene en un escenario de un solo hablante y dentro del mismo corpus, de modo que mide la facilidad de la tarea cuando los datos abundan en la lengua objetivo, no la capacidad de generalizar. Lo que el trabajo demuestra de verdad es que, en cuanto existe un dataset nativo, detectar deepfakes en árabe deja de ser un problema abierto.

La cara contraria aparece al cambiar de dominio. Awaad et al. (2024) entrenan un de-

tector ligero, un ensemble de XGBoost y LightGBM frente a una red neuronal de capas densas, sobre rasgos acústicos extraídos de la señal, para voz árabe generada por conversión de voz basada en recuperación “Retrieval-based Voice Conversion” (RVC). Dentro del árabe el sistema funciona, con exactitudes entre el 92 % y el 95 %, pero su rendimiento se desploma al 38 % cuando la voz de prueba está en inglés, y la red neuronal cae directamente al 0 %. El mismo modelo, la misma arquitectura, y un simple cambio de lengua basta para invalidarlo. La conclusión que los propios autores subrayan es que la dependencia de la lengua pesa tanto como la dependencia del método de síntesis.

El urdu, una lengua de muchos hablantes pero pocos recursos digitales, repite el patrón con un matiz instructivo. Ahmad et al. (2025) reunieron grabaciones reales de podcasts y generaron sus falsificaciones con RVC para construir un corpus de 125 hablantes, y reportan exactitudes muy altas, hasta un 98,89 % con una red recurrente con memoria a largo plazo. Frente a ese optimismo, Munir et al. (2024) ofrecen un retrato mucho más sombrío. Su conjunto, con 20 451 elocuciones auténticas y 16 830 deepfakes producidos por Tacotron y VITS sobre un corpus fonéticamente rico que cubre 62 de los 67 fonemas del urdu, somete a evaluación al modelo AASIST-L, una contramedida del estado del arte. El resultado es un EER de 0,495 frente a VITS y de 0,524 frente a Tacotron, es decir, en torno a 0,5, el valor del puro azar: el detector no distingue. La discrepancia entre ambos trabajos no es contradictoria sino reveladora, pues señala que el resultado depende por completo de la calidad del ataque y del protocolo, y que un sistema puntero falla cuando el sintetizador es competente y la evaluación, exigente. A ello se añade que ni siquiera los humanos discriminan bien: en su estudio con oyentes, el área bajo la curva fue de 0,63, con uno de cada tres audios falsos tomado por real.

El español aporta el experimento más limpio sobre la naturaleza de la brecha. Tamayo Flórez et al. (2023) presentan HABLA, el primer corpus de anti-spoofing en español, con cinco acentos latinoamericanos, más de 22 000 muestras reales y 58 000 falsificaciones generadas por seis estrategias de síntesis. Su diseño compara tres familias de modelos: los preentrenados en inglés, los entrenados desde cero en español y los preentrenados en inglés y reajustados después en español. Los primeros superan el 45 % de EER en todos los casos al evaluarse sobre el español y clasifican correctamente apenas el 62,6 % de las muestras reales, hasta el punto de que el acento colombiano les resulta indistinguible, con solo un 40 % de aciertos. Los modelos entrenados en

la lengua objetivo, en cambio, bajan el EER por debajo del 3 %. El reajuste a partir del modelo inglés ofrece el mejor equilibrio entre ambas lenguas y reduce el número de épocas necesarias, lo que sugiere que el conocimiento aprendido en inglés no es inútil, sino que necesita reorientarse hacia la fonética de destino.

Por otro lado, Hoang et al. (2024) construyen el primer dataset vietnamita para verificación de hablante robusta al spoofing, con 164 000 elocuciones auténticas y 174 000 falsas de 1 382 hablantes, y lo someten a tres familias de ataque: repetición, conversión de voz por RVC y perturbaciones adversariales. Los números cuantifican cada vulnerabilidad por separado: un modelo ECAPA-TDNN preentrenado en inglés se degrada un 158 % al pasar al vietnamita, los ataques de repetición elevan el EER en torno a un 1 100 % y las perturbaciones adversariales alrededor de un 780 %. La lectura conjunta confirma que los sistemas de verificación actuales, ajustados sobre habla inglesa, no llegan preparados a una lengua nueva.

Tabla 7.7

Trabajos de detección de deepfakes en lenguas menos representadas revisados en esta sección. Cada uno aporta un corpus nativo y, con él, una medida concreta de la brecha frente a los modelos entrenados en inglés.

Lengua	Referencia	Corpus / mé- todo	Evidencia sobre la brecha
Árabe	(Almutairi y El-gibreen, 2023)	Arabic-AD (HuBERT); primer sintético árabe	Intra-lengua casi resuelto: EER 0,027 % con datos nativos
Árabe	(Awaad et al., 2024)	RVC + en-semble XG-Boost/LightGBM	92–95 % en árabe; se desploma al 38 % en inglés
Urdu	(Ahmad et al., 2025)	URFV (RVC); 125 hablantes	Hasta 98,89 % intra-urdu con datos propios
Urdu	(Munir et al., 2024)	Tacotron + VITS; 62/67 fonemas	AASIST-L con EER 0,495–0,524: no discrimina
Español	(Tamayo Flórez et al., 2023)	HABLA; 5 acentos latino-americanos	Modelos en inglés: EER >45 %; nativos: <3 %
Vietnamita	(Hoang et al., 2024)	VSASV; verificación de hablante	ECAPA-TDNN en inglés se degrada un 158 % en vietnamita

Si la evidencia anterior diagnostica la brecha, una segunda línea de trabajo intenta cerrarla sin necesidad de un corpus nuevo para cada lengua. El estudio de referencia es el de Ba et al. (2023), el primero en abordar de forma sistemática la perspectiva cross-lingual. Para medirla construyeron DECRO, un conjunto de evaluación con subconjuntos en inglés y en chino generados con las mismas técnicas de síntesis, lo que aísla el efecto de la lengua de cualquier otra variable. El resultado cuantifica el desplome: tres detectores del estado del arte entrenados en inglés ven caer su exactitud media del 99,92 % al 70,15 % al evaluarse en chino, y su EER se multiplica por más de doscientos, del 0,09 % al 27,78 % en el caso de RawNet2.

La observación clave, sin embargo, es esperanzadora: el inglés, por su abundancia de algoritmos de síntesis, enseña al modelo a reconocer los artefactos de generación, y ese conocimiento es en parte transferible. Sobre esa idea aplican adaptación de dominio adversarial para despojar a la representación de su dependencia de la lengua, y consiguen rebajar el EER hasta el 10,67% aun cuando más del 80 % del habla de prueba procede de algoritmos no vistos.

La otra vía no transfiere desde una lengua, sino que parte de muchas a la vez. El preentrenamiento autosupervisado a gran escala sobre cientos de idiomas produce representaciones que ya no son monolingües, y varios de los trabajos revisados en la sección 7.1 apuntan en esa dirección desde el ángulo lingüístico. Q. Zhang et al. (2024) se apoyan en XLS-R, un modelo preentrenado en 128 lenguas, y sostienen que esa exposición multilingüe masiva es el factor más determinante de la generalización, por encima de la arquitectura concreta del clasificador. Chetia Phukan et al. (2024) lo confirman de manera directa al mostrar que los modelos preentrenados multilingües superan sistemáticamente a los monolingües en la detección de deepfakes. La heterogeneidad lingüística del preentrenamiento, lejos de ser ruido, es la fuente de la robustez.

Estas estrategias no eliminan la necesidad de datos diversos, y aquí aparecen sus límites y los recursos que tratan de paliarlos. Purohit et al. (2024) recopilan un corpus multilingüe de diez lenguas, pero advierten de un sesgo sutil: su generador, entrenado sobre inglés, captura de forma limitada los fonemas propios de las demás lenguas, de modo que un dataset nominalmente multilingüe puede arrastrar la huella de la lengua dominante en sus falsificaciones. Frente a esa estrechez, el corpus de Muller et al. (2024) ofrece la respuesta de mayor escala: 687,4 horas de voz sintética producidas por 140 sistemas de síntesis sobre 51 lenguas, concebido expresamente para combatir el sesgo anglófono y democratizar el entrenamiento de detectores. La disponibilidad de recursos así es la condición material para que las dos vías anteriores, la transferencia y el preentrenamiento multilingüe, puedan siquiera evaluarse con rigor.

Tabla 7.8

Estrategias de generalización cross-lingual y multilingüe revisadas en esta sección. A diferencia del cuadro anterior, estos trabajos no se atan a una lengua sino que buscan cruzar la frontera lingüística o abarcar muchas lenguas a la vez.

Estrategia	Referencia	Base		Hallazgo principal
Transferencia cross-lingual	(Ba et al., 2023)	DECRO (inglés/chino)	(in- + adaptación de dominio	Inglés→chino: exactitud del 99,92 % al 70,15 %; la adaptación adversarial baja el EER al 10,67 %
Preentrenamiento multilingüe	(Q. Zhang et al., 2024)	XLS-R (128 lenguas)		La exposición multilingüe masiva es el factor más determinante de la generalización
Preentrenamiento multilingüe	(Chetia Phukan et al., 2024)	PTMs tilingües vs.monolingües	mul-	Los modelos multilingües superan sistemáticamente a los monolingües
Corpus multilingüe	(Purohit et al., 2024)	GGMDDC (10 lenguas)	(10	Útil, pero el generador entrenado en inglés limita los fonemas de otras lenguas
Corpus multilingüe	(Muller et al., 2024)	MLAAD (51 lenguas, 140 TTS)		687,4 h de síntesis concebidas para combatir el sesgo anglófono

El panorama que dibuja este último eje es coherente con los dos anteriores y, a la vez, los matiza. Los avances en datasets, estrategias de entrenamiento y explicabilidad se construyeron sobre un puñado de lenguas, y la evidencia reunida aquí muestra que esa base estrecha no se sostiene por sí sola: un detector que roza la perfección en inglés puede quedar reducido al azar en urdu, en vietnamita o en español. La brecha, sin embargo, no es un destino. Donde existe un corpus nativo el problema se vuelve abordable, y donde no, la transferencia de dominio y el preentrenamiento sobre decenas de lenguas ofrecen caminos verificados para acortar la distancia. Lo que el conjunto del estado del arte deja claro es que la generalización, ya sea entre métodos de síntesis, entre corpus o entre lenguas, es el problema central que la detección de deepfakes

de audio tiene aún por delante, y que las firmas acústicas, la diversidad de los datos y la transparencia de los modelos son las tres palancas con las que la investigación reciente intenta moverlo.

8. CONCLUSIONES

Esta monografía sistematizó, mediante una revisión PRISMA de 39 estudios publicados entre 2020 y 2025, la investigación reciente sobre la detección de deepfakes de audio con aprendizaje profundo. Sus hallazgos responden a las tres preguntas planteadas en la introducción y convergen en un mismo diagnóstico de fondo, que se expone a continuación.

La trayectoria de las arquitecturas es nítida: de los clasificadores sobre características diseñadas a mano se pasó a las redes que operan sobre la forma de onda y, finalmente, a los modelos de fundación preentrenados mediante aprendizaje autosupervisado, como wav2vec 2.0 y XLS-R, que constituyen hoy la aproximación más efectiva. Lo determinante, sin embargo, no es qué arquitectura obtiene el menor error, sino que ese error solo es bajo dentro del corpus en que se entrena: el rendimiento se desploma ante sistemas de síntesis, corpus o lenguas no vistos. Ese sesgo de conjunto cerrado, y no una arquitectura concreta, es el verdadero objeto de estudio del periodo, y la evidencia señala de forma consistente a la diversidad del preentrenamiento, antes que al tamaño del modelo, como el factor que mejor lo mitiga.

Sobre las características acústicas determinantes y su interpretación, la explicabilidad ofrece una respuesta consistente: con independencia de la herramienta empleada, los trabajos apuntan al mismo sustrato acústico, ligado a la producción natural de la voz, que la síntesis todavía no replica del todo. Más allá de justificar decisiones, la explicabilidad se revela como un instrumento de diagnóstico, pues ayuda a entender cuándo y por qué un detector deja de generalizar.

En cuanto a las lenguas menos exploradas, la revisión arroja el resultado más contundente: un detector casi perfecto en inglés puede quedar reducido al azar en lenguas de bajo recurso como el español, y solo los corpus nativos y las estrategias de transferencia recuperan el rendimiento. Los tres ejes confluyen así en una única conclusión de fondo: la generalización, ya sea entre métodos de síntesis, entre corpus o entre lenguas, es el problema central que la detección de deepfakes de audio tiene todavía por delante.

Esta revisión tiene límites que conviene reconocer. El corpus final, acotado a 39 estudios de IEEE Xplore y Scopus entre 2020 y 2025, no agota una literatura que crece a gran velocidad, y la ausencia de una evaluación experimental propia impide contrastar de primera mano las cifras reportadas por cada trabajo. A ello se suma la heterogeneidad de corpus, métricas y protocolos entre estudios, que dificulta una comparación estrictamente homogénea, una limitación que el propio campo arrastra.

De cara al futuro, las líneas más prometedoras se desprenden de esas mismas carencias. La construcción de corpus nativos en lenguas de bajo recurso, con el español como caso de interés, es la condición para evaluar con rigor la generalización cross-lingual; la explicabilidad, usada como herramienta de diagnóstico y no solo de justificación, puede acelerar la comprensión de los fallos fuera de dominio; y la adopción de protocolos de evaluación cruzada estandarizados resulta indispensable para que el buen desempeño dentro de un corpus deje de confundirse con la robustez real. El campo dispone hoy de detectores precisos en condiciones conocidas; el reto pendiente es que esa precisión sobreviva al encuentro con lo desconocido.

9. REFERENCIAS

- Ahmad, O., Khan, M. S., Jan, S. y Khan, I. (2025). Deepfake Audio Detection for Urdu Language Using Deep Neural Networks. *IEEE Access*, 13, 97765-97778. <https://doi.org/10.1109/ACCESS.2025.3571293>
- Alharbi, L. A., Alnahas, J., Javed, A., Alhakamy, A., Nawaz, M., Malik, H., Ilyas, H. y Alfaifi, F. (2026). DeepRawNet: empowering deepfake audio detection through dynamic enhancements. *PeerJ Computer Science*, 12, e3670. <https://doi.org/10.7717/peerj-cs.3670>
- Almutairi, Z. M. y Elgibreen, H. (2023). Detecting Fake Audio of Arabic Speakers Using Self-Supervised Deep Learning. *IEEE Access*, 11, 72134-72147. <https://doi.org/10.1109/ACCESS.2023.3286864>
- Awaad, R., Hedy, A., Fadel, M. y Zaky, A. B. (2024). Real-Time Deepfake Detection for AI-Generated Arabic Speech. *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, 198-201. <https://doi.org/10.1109/NILES63360.2024.10753151>
- Azeemi, A. H., Qazi, I. A. y Raza, A. A. (2023). Self-Supervised Dataset Pruning for Efficient Training in Audio Anti-spoofing. *Interspeech 2023*, 2773-2777. <https://doi.org/10.21437/Interspeech.2023-481>
- Ba, Z., Wen, Q., Cheng, P., Wang, Y., Lin, F., Lu, L. y Liu, Z. (2023). Transferring Audio Deepfake Detection Capability across Languages. *Proceedings of the ACM Web Conference 2023*, 2033-2044. <https://doi.org/10.1145/3543507.3583222>
- Baevski, A., Zhou, H., Mohamed, A. y Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. <https://arxiv.org/abs/2006.11477>
- Bhagtani, K., Yadav, A. K. S., Bestagini, P. y Delp, E. J. (2024). Are Recent Deepfake Speech Generators Detectable? *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, 277-282. <https://doi.org/10.1145/3658664.3659658>
- Bisogni, C., Loia, V., Nappi, M. y Pero, C. (2024). Acoustic features analysis for explainable machine learning-based audio spoofing detection. *Computer Vision and Image Understanding*, 249, 104145. <https://doi.org/10.1016/j.cviu.2024.104145>
- Can, Z. y Soyhan, B. (2025). Spectro-Temporal-CNN Fusion for Deepfake Speech Detection and Spoof System Attribution. *IEEE Access*, 13, 185802-185817. <https://doi.org/10.1109/ACCESS.2025.3625746>

- Chen, S., Liu, S., Zhou, L., Liu, Y., Tan, X., Li, J., Zhao, S., Qian, Y. y Wei, F. (2024). VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers. <https://arxiv.org/abs/2406.05370>
- Chetia Phukan, O., Kashyap, G. S., Buduru, A. B. y Sharma, R. (2024). Heterogeneity over Homogeneity: Investigating Multilingual Speech Pre-Trained Models for Detecting Audio Deepfake. En K. Duh, H. Gomez y S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2496-2506). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-naacl.160>
- de Benito, D., Lozano-Diez, A., Toledano, D. y Gonzalez-Rodriguez, J. (2019). Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset. *EURASIP Journal on Audio, Speech, and Music Processing*, 2019. <https://doi.org/10.1186/s13636-019-0152-1>
- Delgado, H., Evans, N., Kinnunen, T., Lee, K. A., Liu, X., Nautsch, A., Patino, J., Sahidullah, M., Todisco, M., Wang, X. y Yamagishi, J. (2021). ASVspoof 2021: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan. <https://arxiv.org/abs/2109.00535>
- Doan, T.-P., Dinh-Xuan, H., Ryu, T., Kim, I., Lee, W., Hong, K. y Jung, S. (2024). Trident of Poseidon: A Generalized Approach for Detecting Deepfake Voices. *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2222-2235. <https://doi.org/10.1145/3658644.3690311>
- Gao, Y., Lian, J., Raj, B. y Singh, R. (2020). Detection and Evaluation of Human and Machine Generated Speech in Spoofing Attacks on Automatic Speaker Verification Systems. *2021 IEEE Spoken Language Technology Workshop (SLT)*, 544-551. <https://api.semanticscholar.org/CorpusID:226281976>
- Grinberg, P., Kumar, A., Koppiseti, S. y Bharaj, G. (2025). What Does an Audio Deepfake Detector Focus on? A Study in the Time Domain. *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1-5. <https://doi.org/10.1109/ICASSP49660.2025.10887568>
- Hoang, V., Pham, V. T., Xuan, H. N., Nhi, P., Dat, P. y Nguyen, T. T. T. (2024). VSASV: a Vietnamese Dataset for Spoofing-Aware Speaker Verification. *Interspeech 2024*, 4288-4292. <https://doi.org/10.21437/Interspeech.2024-1972>

- Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R. y Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. <https://arxiv.org/abs/2106.07447>
- IDIS. (2026). *Euphonia*. IDIS. Consultado el 10 de junio de 2026, desde <https://proyectoidis.org/euphonia/>
- Jung, J.-w., Heo, H.-S., Tak, H., Shim, H.-j., Chung, J. S., Lee, B.-J., Yu, H.-J. y Evans, N. (2021). AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks. <https://arxiv.org/abs/2110.01200>
- Khan, A., Malik, K. M., Ryan, J. y Saravanan, M. (2023). Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures. *Artif. Intell. Rev.*, 56(Suppl 1), 513-566. <https://doi.org/10.1007/s10462-023-10539-8>
- Kong, J., Kim, J. y Bae, J. (2020). HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. <https://arxiv.org/abs/2010.05646>
- Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brebisson, A., Bengio, Y. y Courville, A. (2019). MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. <https://arxiv.org/abs/1910.06711>
- Lemmetty, S. (1999). *Review of Speech Synthesis Technology* [Tesis de maestría, Helsinki University of Technology, Department of Electrical y Communications Engineering] [Master's Thesis. Supervisor: Matti Karjalainen]. http://research.spa.aalto.fi/publications/theses/lemmetty_mst/
- Li, K., Lu, X., Akagi, M. y Unoki, M. (2023). Contributions of Jitter and Shimmer in the Voice for Fake Audio Detection. *IEEE Access*, 11, 84689-84698. <https://doi.org/10.1109/ACCESS.2023.3301616>
- Li, M., Ahmadiadli, Y. y Zhang, X.-P. (2022). A Comparative Study on Physical and Perceptual Features for Deepfake Audio Detection. *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 35-41. <https://doi.org/10.1145/3552466.3556523>
- Li, M. y Zhang, X.-P. (2024). Interpretable Temporal Class Activation Representation for Audio Spoofing Detection. *Interspeech 2024*, 1120-1124. <https://doi.org/10.21437/Interspeech.2024-2156>

- Li, X., Li, K., Zheng, Y., Yan, C., Ji, X. y Xu, W. (2024). SafeEar: Content Privacy-Preserving Audio Deepfake Detection. *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 3585-3599. <https://doi.org/10.1145/3658644.3670285>
- Liu, T., Zhang, L., Das, R. K., Ma, Y., Tao, R. y Li, H. (2024). How Do Neural Spoofing Countermeasures Detect Partially Spoofed Audio? *Interspeech 2024*, 1105-1109. <https://doi.org/10.21437/Interspeech.2024-2009>
- Lu, Y., Xie, Y., Fu, R., Wen, Z., Tao, J., Wang, Z., Qi, X., Liu, X., Li, Y., Liu, Y., Wang, X. y Shi, S. (2024). Codefake: An Initial Dataset for Detecting LLM-based Deepfake Audio. *Interspeech 2024*, 1390-1394. <https://doi.org/10.21437/Interspeech.2024-2428>
- Maya Research. (2025). Maya1: Open Source Voice AI with Emotional Intelligence.
- Müller, N., Czempin, P., Diekmann, F., Froghyar, A. y Böttinger, K. (2022). Does Audio Deepfake Detection Generalize? *Interspeech 2022*, 2783-2787. <https://doi.org/10.21437/Interspeech.2022-108>
- Müller, N. M., Evans, N., Tak, H., Sperl, P. y Böttinger, K. (2024). Harder or Different? Understanding Generalization of Audio Deepfake Detection. *Interspeech 2024*, 2705-2709. <https://doi.org/10.21437/Interspeech.2024-247>
- Muller, N. M., Kawa, P., Choong, W. H., Casanova, E., Golge, E., Muller, T., Syga, P., Sperl, P. y Bottinger, K. (2024). MLAAD: The Multi-Language Audio Anti-Spoofing Dataset. *2024 International Joint Conference on Neural Networks (IJCNN)*, 1-7. <https://api.semanticscholar.org/CorpusID:267034946>
- Munir, S., Sajjad, W., Raza, M., Abbas, E., Azeemi, A. H., Qazi, I. A. y Raza, A. A. (2024). Deepfake Defense: Constructing and Evaluating a Specialized Urdu Deepfake Audio Dataset. En L.-W. Ku, A. Martins y V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 14470-14480). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-acl.861>
- Nguyen-Duc, M., Nguyen, L. V., Nguyen-Ho-Nhat, H., Nguyen, T.-H. y Lee, O.-J. (2025). A Comparative Study of Deep Audio Models for Spectrogram- and Waveform-Based SingFake Detection. *IEEE Access*, 13, 95739-95752. <https://api.semanticscholar.org/CorpusID:278798921>

- Olive, J. (1977). Rule synthesis of speech from dyadic units. *ICASSP '77. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2, 568-570. <https://doi.org/10.1109/ICASSP.1977.1170350>
- Pascu, O., Stan, A., Oneata, D., Oneata, E. y Cucu, H. (2024). Towards generalisable and calibrated audio deepfake detection with self-supervised representations. *Interspeech 2024*, 4828-4832. <https://doi.org/10.21437/Interspeech.2024-1302>
- Pianese, A., Cozzolino, D., Poggi, G. y Verdoliva, L. (2022). Deepfake audio detection by speaker verification. *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1-6. <https://api.semanticscholar.org/CorpusID:252568301>
- Purohit, R. M., Shah, A. J. y Patil, H. A. (2024). GGMDDC: An Audio Deepfake Detection Multilingual Dataset. *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1-6. <https://api.semanticscholar.org/CorpusID:275947747>
- Ranjan, R., Vatsa, M. y Singh, R. (2023). Uncovering the deceptions: an analysis on audio spoofing detection and future prospects. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. <https://doi.org/10.24963/ijcai.2023/756>
- Ravanelli, M. y Bengio, Y. (2019). Speaker Recognition from Raw Waveform with SincNet. <https://arxiv.org/abs/1808.00158>
- Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z. y Liu, T.-Y. (2022). FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. <https://arxiv.org/abs/2006.04558>
- Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z. y Liu, T.-Y. (2019). FastSpeech: Fast, Robust and Controllable Text to Speech. <https://arxiv.org/abs/1905.09263>
- Shaaban, O. A., Yildirim, R. y Alguttar, A. A. (2023). Audio Deepfake Approaches. *IEEE Access*, 11, 132652-132682. <https://doi.org/10.1109/ACCESS.2023.3333866>
- Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N. y Larcher, A. (2021). End-to-end anti-spoofing with RawNet2. <https://arxiv.org/abs/2011.01108>
- Tamayo Flórez, P. A., Manrique, R. y Pereira Nunes, B. (2023). HABLA: A Dataset of Latin American Spanish Accents for Voice Anti-spoofing. *Interspeech 2023*, 1963-1967. <https://doi.org/10.21437/Interspeech.2023-2272>
- Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge University Press.

- van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. y Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. <https://arxiv.org/abs/1609.03499>
- Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S. y Wei, F. (2023). Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. <https://arxiv.org/abs/2301.02111>
- Wang, X., Delgado, H., Tak, H., Jung, J.-w., Shim, H.-j., Todisco, M., Kukanov, I., Liu, X., Sahidullah, M., Kinnunen, T., Evans, N., Lee, K. A. y Yamagishi, J. (2024). ASVspooF 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale. <https://arxiv.org/abs/2408.08739>
- Wang, Z. y Hansen, J. H. L. (2024). Toward Improving Synthetic Audio Spoofing Detection Robustness via Meta-Learning and Disentangled Training With Adversarial Examples. *IEEE Access*, 12, 99894-99911. <https://doi.org/10.1109/access.2024.3421281>
- Wu, H., Liu, A. T. y Lee, H.-y. (2020). Defense for Black-Box Attacks on Anti-Spoofing Models by Self-Supervised Learning. *Interspeech 2020*, 3780-3784. <https://doi.org/10.21437/Interspeech.2020-2026>
- Wu, X., He, R., Sun, Z. y Tan, T. (2018). A Light CNN for Deep Face Representation with Noisy Labels. <https://arxiv.org/abs/1511.02683>
- Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., Sahidullah, M. y Sizov, A. (2015). ASVspooF 2015: the first automatic speaker verification spoofing and countermeasures challenge. *Interspeech 2015*, 2037-2041. <https://doi.org/10.21437/Interspeech.2015-462>
- Xie, Y., Cheng, H., Wang, Y. y Ye, L. (2023). Learning A Self-Supervised Domain-Invariant Feature Representation for Generalized Audio Deepfake Detection. *Interspeech 2023*, 2808-2812. <https://doi.org/10.21437/Interspeech.2023-1383>
- Yi, J., Tao, J., Fu, R., Yan, X., Wang, C., Wang, T., Zhang, C. Y., Zhang, X., Zhao, Y., Ren, Y., Xu, L., Zhou, J., Gu, H., Wen, Z., Liang, S., Lian, Z., Nie, S. y Li, H. (2023). ADD 2023: the Second Audio Deepfake Detection Challenge. <https://arxiv.org/abs/2305.13774>
- Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. y Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. *6th European Conference on Speech Communication and Technology*, 2347-2350. <https://doi.org/10.21437/Eurospeech.1999-596>

- Yu, N., Chen, L., Leng, T., Chen, Z. y Yi, X. (2024). An explainable deepfake of speech detection method with spectrograms and waveforms. *J. Inf. Secur. Appl.*, 81(C). <https://doi.org/10.1016/j.jisa.2024.103720>
- Zaman, K., Samiul, I. J. A. M., Sah, M., Direkoglu, C., Okada, S. y Unoki, M. (2024). Hybrid Transformer Architectures With Diverse Audio Features for Deepfake Speech Classification. *IEEE Access*, 12, 149221-149237. <https://doi.org/10.1109/ACCESS.2024.3478731>
- Zang, Y., Shi, J., Zhang, Y., Yamamoto, R., Han, J., Tang, Y., Xu, S., Zhao, W., Guo, J., Toda, T. y Duan, Z. (2024). CtrSVDD: A Benchmark Dataset and Baseline Analysis for Controlled Singing Voice Deepfake Detection. *Interspeech 2024*, 4783-4787. <https://doi.org/10.21437/Interspeech.2024-2242>
- Zang, Y., Zhang, Y., Heydari, M. y Duan, Z. (2024). SingFake: Singing Voice Deepfake Detection. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12156-12160. <https://doi.org/10.1109/ICASSP48485.2024.10448184>
- Zhang, L., Wang, X., Cooper, E., Evans, N. y Yamagishi, J. (2022). The PartialSpoof Database and Countermeasures for the Detection of Short Fake Speech Segments Embedded in an Utterance. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31, 813-825. <https://doi.org/10.1109/TASLP.2022.3233236>
- Zhang, L., Wang, X., Cooper, E., Evans, N. y Yamagishi, J. (2023). The PartialSpoof Database and Countermeasures for the Detection of Short Fake Speech Segments Embedded in an Utterance. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 813-825. <https://doi.org/10.1109/taslp.2022.3233236>
- Zhang, Q., Wen, S. y Hu, T. (2024). Audio Deepfake Detection with Self-Supervised XLS-R and SLS Classifier. *Proceedings of the 32nd ACM International Conference on Multimedia*, 6765-6773. <https://doi.org/10.1145/3664647.3681345>
- Zhang, Y., Zang, Y., Shi, J., Yamamoto, R., Toda, T. y Duan, Z. (2024). SVDD 2024: The Inaugural Singing Voice Deepfake Detection Challenge. *2024 IEEE Spoken Language Technology Workshop (SLT)*, 782-787. <https://doi.org/10.1109/slt61566.2024.10832284>